La construcción de la sociedad del conocimiento es uno de los principales retos en la actualidad. Los acelerados cambios que está sufriendo nuestra sociedad especialmente en términos de ampliación del saber humano, provocan la necesidad de nuevos escenarios a fin de satisfacer las demandas y desafíos que trae consigo el siglo XXI.

No estaría mal asumir desde el principio que, en los nuevos tiempos que corren, son necesarias nuevas formas de enfrentarnos a él, nuevas formas de abordar los problemas, y nuevas formas de comprenderlos, nuevas formas de plantear las relaciones con las personas, y también nuevas herramientas de comunicación, que van a requerir que las personas las dominemos, tanto desde un punto de vista instrumental, como sintáctico y semántico para la construcción con ellas de mensajes.

Un proceso de estas características sólo puede llevarse acabo con la ayuda de la Investigación Tecnológica, que en muchos casos son disruptivas y modifican profundamente los hábitos sociales y culturales. Por eso nos proponemos suscitar la reflexión en torno al papel que desempeñan las comunicaciones móviles y electrónicas en el marco de una sociedad globalizada.

Fernando Leal Ríos

ISBN: 978-607-410-105-8

innovación editorial lagares
M E X I C O

Comunicaciones Móviles y Electrónica: Investigación Tecnológica

Comunicaciones Móviles y Electrónica:
Investigación Tecnológica

Editores:
Mariby Lucio Castillo,
Carlos Portes Flores,
Ma. Magdalena Flores Morelos,
Everardo Huerta Sosa,
Armando Vega Pérez,
Miguel Walle Vázquez

Comunicaciones Móviles y Electrónica: Investigación Tecnológica

Comunicaciones Móviles y Electrónica: Investigación Tecnológica

Mariby Lucio
Carlos E. Portes
Ma. M. Flores
Everardo Huerta
Armando Vega
Miguel A. Walle

innovación editorial lagares
M E X I C O

Comunicaciones Móviles y Electrónica: Investigación Tecnológica

D.R. © 2009 por Innovación Editorial Lagares de México, S.A. de C.V.
Av. Álamo Plateado No. 1–402 Fracc. Los Álamos Naucalpan, Estado de México
C.P. 53230 Teléfono: (55) 5240-1295 al 98
email: editor@lagares.com.mx

ISBN: 978-607-410-105-8

Diseño de Portada: Enrique Ibarra Vicente

Cuidado Editorial: Rosaura Rodríguez Aguilera

Primera edición agosto, 2010

IMPRESO EN MÉXICO / PRINTED IN MEXICO

Mariby Lucio Castillo.
U.A.M. Agronomía y Ciencias
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México.

Luis Javier de la Cruz Llopis.
Departamento de Ingeniería Telemática
Universidad Politécnica de Cataluña
Campus Nord,
C. Jordi Girona 1-3, Barcelona, España.

Carlos Enrique Portes Flores.
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México

Carlos del Río Bocio.
Dpto. de Ingeniería Eléctrica y Electrónica
Universidad Pública de Navarra
Pamplona, España.

Ma. Magdalena Flores Morelos.
U.A.M. Agronomía y Ciencias
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México.

Aldo Luis Pérez Méndez.
U.A.M. Reynosa - Rodhe
Universidad Autónoma de Tamaulipas
Reynosa Tamaulipas, México.

Tania Guerrero Meléndez.
Departamento de Ingeniería Telemática
Universidad Politécnica de Cataluña
Campus Nord,
C. Jordi Girona 1-3, Barcelona, España.

Vicente P. Saldivar Alonso.
Departamento de Ingeniería Telemática
Universidad Politécnica de Cataluña
Campus Nord.
C. Jordi Girona 1-3 Barcelona, España.

Marco A. Panduro Mendoza.
U.A.M. Reynosa - Rodhe
Universidad Autónoma de Tamaulipas
Reynosa Tamaulipas, México.

Everardo Huerta Sosa.
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México.

Brenda Valdez Reyna.
U.A.M. Mante Centro
Universidad Autónoma de Tamaulipas
Cd. Mante Tamaulipas, México.

Miguel Walle Vázquez.
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México.

Armando Vega Pérez.
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México.

Ernesto Guillermo Amaya García.
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México.

Ignacio Matías Maestro
Dpto. de Ingeniería Eléctrica y Electrónica
Universidad Pública de Navarra
Pamplona, España.

René Dominguez Cruz
U.A.M. Reynosa - Rodhe
Universidad Autónoma de Tamaulipas
Reynosa Tamaulipas, México.

Arturo Medina Puente
UAM, Agronomia y Ciencias
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria, Tamaulipas

Ángel Dorantes Salazar.
U.A.M. Agronomía y Ciencias
Universidad Autónoma de Tamaulipas
Campus Victoria
Cd. Victoria Tamaulipas, México.

Carlos E. Martínez del Ángel.
Facultad de Comercio y Administración
de Tampico
Universidad Autónoma de Tamaulipas
Tampico Tamaulipas, México.

Julio Laria Menchaca.
Facultad de Ingeniería
"Arturo Narro Siller"
Universidad Autónoma de Tamaulipas
Tampico Tamaulipas, México.

ÍNDICE

1. Estudio de Esquemas de Control de Acceso al Medio para Redes de Comunicaciones Móviles Celulares.

Ma.Magdalena Flores Morelos, Aldo Luis Méndez

1. Introduccion

1.1 Control de Acceso al Medio (MAC)

En el entorno de las comunicaciones móviles, es importante destacar que la compartición de recursos es esencial y necesaria, puesto que el interfaz aire, es el único medio de transmisión.

Los protocolos MAC son el conjunto de reglas y acuerdos establecidos entre usuarios para lograr una correcta transmisión de información utilizando un medio común. A partir de 1970 con la creación del protocolo ALOHA una gran variedad de protocolos de acceso múltiple han sido desarrollados. De acuerdo a [16] estos se dividen como se muestra en la figura 1.

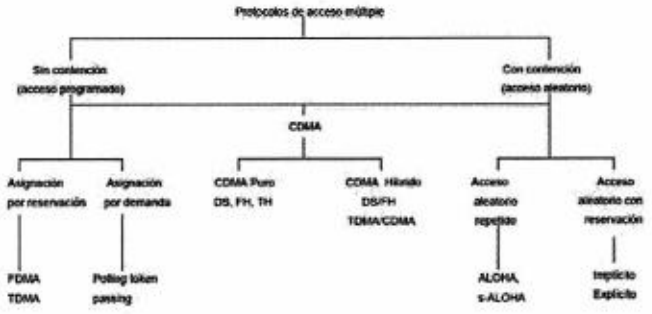


Figura 1. Clasificación de los protocolos de acceso múltiple

1.1.1 PROTOCOLOS SIN CONTENCIÓN

Estos protocolos evitan que dos o más usuarios accedan al canal al mismo tiempo programando la transmisión de los usuarios, es decir, los usuarios transmiten en un orden programado de manera que cada transmisión se realizará con éxito.

Algunos protocolos que pertenecen a esta clasificación son: FDMA (Acceso Múltiple por División De Frecuencia), TDMA (Acceso Múltiple por División de Tiempo), sondeo por paso de testigo, entre otros.

1.1.2 PROTOCOLOS CON CONTENCIÓN

En los protocolos con contención (acceso aleatorio) todo el ancho de banda es proveído al usuario como un solo canal para ser accesado aleatoriamente, por eso las colisiones entre paquetes pueden ocurrir y estos deben ser retransmitidos [7]. El problema más importante en este tipo de protocolos es cómo resolver los conflictos cuando dos o más usuarios transmiten al mismo tiempo.

PLANTEAMIENTO DEL PROBLEMA

Uno de los campos de las telecomunicaciones que ha merecido mayor atención en los últimos 20 años ha sido la resolución de colisiones en un canal multiacceso, es decir, resolver el acceso de un numeroso grupo de usuarios a un canal común para comunicarse [13].

Cuando un mismo recurso es susceptible de ser utilizado por varios usuarios independientes entre ellos aparece la necesidad de establecer un protocolo de acceso múltiple, a fin de gestionar y asignar el recurso escaso en cuestión. Si no considerase ningún tipo de protocolo, podrían ocurrir conflictos si más de un usuario quisiera acceder al recurso al mismo tiempo [5].

Los protocolos de acceso múltiple deberán evitar, o cuanto menos ser capaces de resolver, con alta probabilidad, los citados conflictos. Las razones que llevan a compartir los recursos son debido a que estos son escasos o cuanto menos caros, por lo que es necesario optimizar su uso. En el caso que nos ocupa estaremos considerando protocolos con contención en el cual las transmisiones de los usuarios no están planificadas, por lo que no se tiene un conocimiento exacto de los usuarios dispuestos a transmitir. Un protocolo de acceso se muestra en la figura 2, en el cual la parte correspondiente de petición de canal (RACH) (acceso aleatorio) es la que nos interesa analizar en su comportamiento.

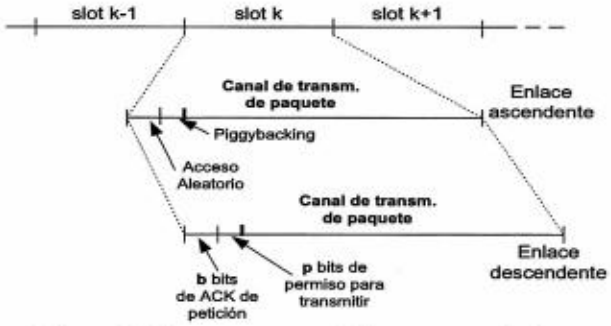


Figura 2. Diagrama temporal de un protocolo de acceso múltiple en comunicaciones móviles.

En el entorno de las comunicaciones móviles, es importante destacar que la mencionada compartición de recursos es más esencial y necesaria, puesto que la interfaz aire, que es el soporte de la transmisión, es único.

Por esta razón es deseable poder invocar herramientas matemáticas para evaluar e investigar las prestaciones de los protocolos complejos de acceso múltiple en comunicaciones móviles celulares.

Considerando el contexto anterior, se desprende entonces que resulta particularmente relevante el modelado matemático, como herramienta para la simulación de las prestaciones de la técnica de acceso en los futuros sistemas móviles celulares de tercera generación (3G), considerando los procesos aleatorios de llegada y de petición de recursos del terminal móvil dentro de un entorno celular en exteriores. Con ello se podrán determinar parámetros de operación asociados con el sistema de comunicaciones, tales como: paquetes generados versus paquetes recibidos (throughput), retardo medio de propagación (tiempo transcurrido desde que se genera un paquete hasta que se recibe correctamente por la estación base), throughput vs. retardo medio, throughput versus paquetes en modo de retransmisión, entre otros.

2. PROTOCOLO ALOHA RANURADO (S-ALOHA)

2.1 MODELADO DE S-ALOHA

Como en los procesos de llegada y de petición de recursos del terminal móvil (TM) son aleatorios, es necesario modelar matemáticamente el sistema para

poder determinar los parámetros de operación asociados con el sistema de comunicaciones. Para esto se analiza un sistema con un número finito de usuarios.

Consideramos el caso en que S-ALOHA (ALOHA ranurado) es usado por un grupo de M terminales móviles (TMs) y cada uno con un buffer unitario. Todos los paquetes son del mismo tamaño, requiriendo T segundos para transmitir, el cual es la duración del slot (ranura de tiempo).

A continuación es descrito el modelo de transmisión utilizado por S-Aloha. Cada TM está en unos de los estados: vacío y bloqueado (backlog). En el estado vacío el TM no tiene un paquete en su buffer y no participa en cualquier actividad de asignación de recurso. Cuando en este estado, el usuario genera un paquete en cada slot con probabilidad σ y no generación de un paquete en un slot con una probabilidad $1-\sigma$; la generación de un paquete es independiente de cualquier otra actividad. Lo anterior indica que la generación de paquete es un proceso independiente.

Una vez que un paquete es generado su transmisión se intenta transmitir inmediatamente en el siguiente slot. Si la transmisión tuvo éxito el TM retorna al estado vacío y el proceso de generación de paquete inicia de nuevo. Si la transmisión no tiene éxito el TM cambia a un estado de backlog (bloqueo) y programa la retransmisión del paquete de acuerdo a una distribución geométrica independiente con parámetro v. En otras palabras, en cada slot el TM retransmitirá el paquete con probabilidad v y se abstendrá de hacerlo con una probabilidad $1-v$ [11]. Mientras en el estado backlog el TM no genera ningún nuevo paquete. Cuando el paquete finalmente es transmitido con éxito el TM regresa al estado vacío.

Los slots del sistema son numerados secuencialmente k=0, 1, ... y $\tilde{N}(k)$ denota el número de TMs en backlog en el inicio del k-ésimo slot. La variable aleatoria $\tilde{N}(k)$ es referida como el estado del sistema. El número de TMs en backlog en el inicio del $(k+1)$-ésimo slot depende del número de TMs en backlog en el inicio del k-ésimo slot y el número de TMs que pasan de un estado a otro dentro del slot. Ya que el estado de transición de los TMs es independiente de las actividades en cualquier slot previo, el proceso $\{\tilde{N}(k), k=0, 1, ...\}$ es una cadena de Markov. Porque el número de TMs en backlog no puede exceder M, por lo que la cadena es finita; así, si todos los estados se comunican, esta cadena de Markov es ergódica, lo que indica que existe una distribución de estado estable.

El diagrama de transición para el sistema es mostrado en la figura 3. Las transiciones superiores son posibles entre cada estado y todos los estados de número más altos, por tanto una colisión entre cualquier número de paquetes es posible. Las transiciones inferiores son solamente posibles hacia el estado adyacente debido a que solamente un paquete puede ser transmitido con éxito en un slot, en el cual el tiempo en backlog es reducido en una unidad. Note además la transición

perdida del estado 0 a 1 la cual está limpia debido a que si todos los TMs estuvieron en estado vacío y un solo TM generó y transmitió un paquete, éste no causará una colisión y por lo tanto no llega a estar en el estado backlog.

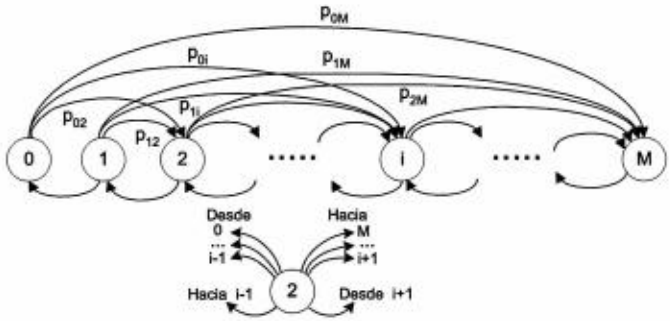


Figura 3. Diagrama de transición de estados para una población finita.

2.1.1 PROBABILIDADES DE ESTADOS ESTABLES

Para el propósito de este análisis introducimos la siguiente notación. Sea p_i la probabilidad en estado estable del sistema estando en el estado i, esto es $\pi_i = \lim_{k \to \infty} \Pr ob[\tilde{N}(k) = i]$. Además, p_{ij} es la probabilidad de transición de estado estable, es decir, $p_{ij} = \lim_{k \to \infty} \Pr ob[\tilde{N}(k) = j \mid \tilde{N}(k-1) = i]$ [2]. Finalmente, denotamos por P la matriz cuyos elementos son p_{ij} y por p el vector renglón cuyos elementos son p_i. De lo mencionado anteriormente, el vector de probabilidad en estado estable es la solución al conjunto finito de ecuaciones lineales.

$$\pi = \pi P, \qquad \sum_i \pi_i = 1 \tag{1}$$

el cual garantiza la existencia de una única solución [11] y [8]. Por eso debemos construir la matriz $\boldsymbol{P}$ y derivar la solución deseada.

Debido a que el proceso de retransmisión de cada TM es un proceso geométrico independiente, la probabilidad que *i* TMs en backlog programara una retransmisión en un slot dado es una distribución binomial, expresada como:

$$\Pr ob[i \text{ usuarios en backlog transmitan en un slot / j en backlog}] = \binom{j}{i} v^i (1-v)^{j-i} \tag{2}$$

De una manera similar, obtenemos para los usuarios en estado vacío:

$$Prob[i \text{ usuarios en estado vacío transmitan en un slot / j en backlog}] = \binom{m-j}{i}\sigma^{i}(1-\sigma)^{M-j-i} \qquad (3)$$

debido a que cuando j usuarios están en estado backlog, M-j TMs están en estado vacío.

La matriz $\boldsymbol{P}$ puede ser construida aplicando las ecuaciones (2) y (3) como sigue:

1. Una transición del estado i al estado j<i-1 es imposible e implica que $\boldsymbol{P}_{ij} = 0$ para estos casos.
2. Considere la transición del estado i al estado i-1. Esto indica una reducción en el backlog, el cual es posible solamente si un solo paquete en backlog fue transmitido y ningún paquete nuevo fue generado.
3. La transición del estado i al mismo estado puede venir de dos razones distintas. La primera resulta de la circunstancia en la cual ningún nuevo paquete fue generado (y transmitido) mientras varios usuarios en backlog intentaron retransmitir. Los usuarios transmisores colisionan y quedan en backlog; porque ninguna transmisión de nuevos paquetes fue realizada y los TMs en backlog no cambia. La segunda razón para esta situación resulta de una situación en la cual ninguno de los TMs intenta la retransmisión y un solo TM vacío transmite. En este caso el TM transmite con éxito y por lo tanto el usuario permanece en el estado vacío. Los dos casos anteriores se pueden resumir por la unión de dos eventos independientes: "Ningún TM en backlog tiene éxito y ningún usuario intenta transmitir" y "Ningún TM en backlog intenta transmitir y solo un TM vacío trasmite".
4. La siguiente transición a considerar es del estado i al estado i+1. Debido a que el backlog incrementa, una colisión debió haber tenido lugar. Además, si el backlog se incrementa en una unidad, exactamente un usuario vacío ha intentado transmitir junto con al menos un TM en backlog.
5. El último caso es la transición del estado i al estado j>i+1. Aquí el backlog se incrementa por dos o más, indicando que j-i TMs vacíos generaron paquetes y colisionaron. La actividad de los TMs en backlog no es im-

portante en este caso debido a que la colisión es generada solamente por TMs vacíos.

Los casos anteriores los podemos resumir en la siguiente expresión [11] y [2]:

$$p_{ij} = \begin{cases} 0 & j<i-1 \\ \left[iv(1-v)^{i-1}\right](1-\sigma)^{M-i} & j=i-1 \\ \left[1-iv(1-v)^{i-1}\right](1-\sigma)^{M-i}+\left[(M-i)\sigma(1-\sigma)^{M-i-1}\right](1-v)^{i} & j=i \\ \left[(M-i)\sigma(1-\sigma)^{M-i-1}\right]\left[1-(1-v)^{i}\right] & j=i+1 \\ \binom{M-i}{j-i}\sigma^{j-i}(1-\sigma)^{M-j} & j>i+1 \end{cases} \quad (4)$$

Es fácilmente verificar que $\sum_j P_{ij} = 1$. Igualmente, se nota que es cero; este resultado es correcto y esperado debido a que toma por lo menos dos paquetes que colisionaron y porque ninguno de los TMs estaba en backlog antes, por lo que es imposible tener un solo TM en backlog al final del slot.

2.1.2 ANÁLISIS DE CAUDAL EFICAZ (THROUGHPUT)

Para evaluar el throughput del sistema considere el instante en el inicio de cada slot. Debido a que la actividad dentro de un slot dado es independiente de la actividad en cualquier slot previo estos instantes son puntos de renovación. Por esto, la fracción de tiempo que transporta información útil —el throughput— es igual a la fracción promedio de slots en una transmisión con éxito. Si denotamos por $\mathbf{P}_{succ}$ como la probabilidad de un slot con éxito entonces:

$$S = P_{succ} \quad (5)$$

Para que un slot llegue a tener éxito solamente una transmisión puede tomar lugar dentro de si mismo. Esto indica que todos los TMs en backlog quedan en silencio y solo un nuevo TM transmite, o un solo TM en backlog transmite mientras ningún nuevo paquete es generado. Dado que hay *i* TMs en backlog esto puede ser expresado como:

$$P_{succ}(i) = \Pr ob[slot \text{ con éxito} \mid \text{i usuarios en backlog}]$$
$$= (1-v)^i (M-i)\sigma(1-\sigma)^{M-i-1} + iv(1-v)^{i-1}(1-\sigma)^{M-i} \quad (6)$$

El throughput total es denotado como:

$$S = P_{succ} = E[P_{succ}(i)] = \sum_{i=0}^{M} P_{succ}(i)\pi_i \quad (7)$$

Note que todos los usuarios son estadísticamente idénticos, el throughput individual es dado por el valor de S de la ec.(7) dividida entre M.

Como un caso especial, consideramos una situación en el cual no necesitamos distinguir entre paquetes en backlog y nuevos paquetes, es decir, $v = \sigma$. Sustituyendo esto dentro de la ec.(6) resulta:

$$P_{succ}(i) = M\sigma(1-\sigma)^{M-1} \quad (8)$$

indicando que $P_{succ}(i)$ es independiente de i. Este resultado, por supuesto, no sorprende debido a que si suspendemos la distinción entre TMs en backlog y vacíos no podemos esperar la probabilidad de éxito que dependa del número de TMs en backlog. Además, porque $P_{succ}(i)$ es independiente de i obtenemos de la ec.(7) una expresión cerrada para el throughput:

$$S = E[P_{succ}(i)] = M\sigma(1-\sigma)^{M-1} \quad (9)$$

Continuando con la consideración de no distinguir entre TMs en backlog y de vacíos, denotamos por G el tráfico ofrecido, que es el número de intentos de transmisión por slot; en nuestro caso es igual a $M\sigma$. Sustituyendo este valor dentro de la ecuación del throughput dada en ec.(9), obtenemos:

$$S = G\left[1 - \frac{G}{M}\right]^{M-1} \quad (10)$$

Bajo estas circunstancias y considerando que M incrementa a infinito se encuentra que en límite $S = Ge^{-g}$, un resultado idéntico como para el esquema de S-Aloha con población infinita. Se puede concluir que el modelo de población infinita es el límite del modelo de población finita; esto se presenta cuando no hay distinción entre TMs en backlog y TMs vacíos, además si el número de usuarios se incrementa forzando a que la razón de arribo promedio sea finita.

2.1.3 RETARDO PROMEDIO

En la sección anterior se mencionó que el throughput es la razón relación media de salida de paquetes del sistema. Si el sistema es estable entonces esta razón puede ser igual a la razón media de generación de nuevos paquetes. Ahora, cuando el sistema está en estado i hay **M - i** TMs vacíos cada uno generando paquetes en cada slot con probabilidad σ. Así, la razón media de generación de nuevos paquetes en el estado i es **(M - i)**σ.

Tomando el promedio se obtiene:

$$S = E[(M-i)\sigma] = \sum (M-i)\sigma\,\pi_i = (M-\tilde{N})\sigma \tag{11}$$

donde $\tilde{\mathbf{N}}$ es el número promedio de TMs en backlog.

Denotamos por b la razón media en el cual los paquetes (TM*s* con paquetes) pasan a backlog; entonces de acuerdo al teorema de Little [3], la cantidad media de tiempos perdidos en backlog, es el número medio de TMs bloqueados respecto a los que pasan al estado backlogged, o sea $\tilde{\mathbf{N}}/\mathbf{b}$. Definiendo también $(S - b)/S$ como una fracción de los paquetes que no están más en estado backlogged. Se puede definir el retardo medio como:

$$\hat{D} = \frac{S-b}{S} \bullet 1 + \frac{b}{S}\left(\frac{\tilde{N}}{b} + 1\right) = 1 + \frac{\tilde{N}}{S} \tag{12}$$

Usando el valor de $\tilde{N}$ de la ec. (12) y el valor de S tomado de la ec. (7), se obtiene el retardo expresado como:

$$\hat{D} = 1 - \frac{1}{\sigma} + \frac{M}{S} \tag{13}$$

Esta última ecuación expresa la relación throughput-retardo. Se debe hacer notar que esta representación es paramétrica debido a que σ influye en el valor de *S*. Así, el throughput primero incrementa con σ hasta conseguir la máxima capacidad; después de esto el throughput decrece con el aumento de la carga. El retardo incrementa monótonamente con σ.

Considerando otra el caso especial en el que σ = *v*, además el throughput dado en la ec. (9), entonces se obtiene

$$\hat{D} = 1 + \frac{1-(1-\sigma)^{M-1}}{\sigma(1-\sigma)^{M-1}} \tag{14}$$

Dos interesantes observaciones puede hacerse con respecto al último resultado. Primero, manteniendo constante el producto **M**σ e incrementando *M* muestra cada vez un incremento en el retardo. La segunda observación interesante relaciona al retardo esperado cuando σ tiende a cero. Tomando el límite se encuentra que $\hat{D}(\sigma \to 0) \to M$, un resultado que parecería sorprendente al inicio. Cuando es muy pequeña, difícilmente tendrá lugar una colisión, y en la mayoría de los casos el retardo será de un solo slot. Sin embargo, en el raro caso de una colisión los TMs que colisionaron pasarán al estado backlog, y se mantendrá en este estado por un largo tiempo debido a que el tiempo de espera para un paquete en backlog es inversamente proporcional a . Juntando todo lo mencionado se encuentra que más paquetes teniendo un retraso de una unidad, pero pocos paquetes tienen retardos extremadamente largos, produciendo un retardo medio combinado de M slots.

2.2 SIMULACIÓN

Después de presentar la técnica para modelar las prestaciones de S-Aloha, esto es llevado a un escenario de simulación, en el cual los parámetros son los siguientes:

- Número de terminales móviles: 80.
- La medida fundamental de análisis estará basada en slots.
- Se asume que cada paquete puede estar contenido en el slot.
- Una terminal no puede generar un nuevo paquete hasta que el actual paquete ha sido transmitido.
- La probabilidad de transmitir un paquete nuevo generado es 1, independientemente del valor actual de p.

Las respuestas que se presentarán en las figuras siguientes consideran el comportamiento de S-Aloha en cuanto al throughput, retardo promedio y número de estaciones en estado bloqueo. La figura 4 muestra el comportamiento del throughput en función de la carga ofrecida al canal.

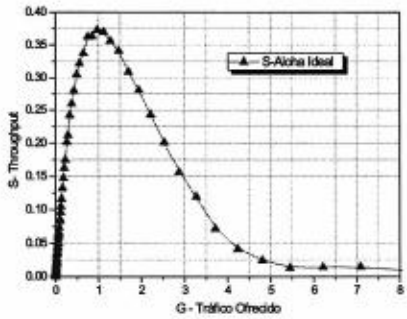


Figura 4. Comportamiento del throughput en S-Aloha

De la figura 4 se puede visualizar que el valor del throughput máximo es obtenido cuando el tráfico ofrecido normalizado es 1 y cuyo valor de throughput es 0.368. Además, se observa que existen tres regiones: lineal, saturación y congestión. En la región lineal ($0 \leq G < 0.8$), la relación de paquetes que entran y los paquetes que salen es 1. La región de saturación ($0.8 \leq G < 1.2$) es cuando algunos paquetes que se generan llegan a salir. La región de congestión ($G \geq 1.2$) se presentan cuando al aumentar el tráfico ofrecido al sistema se vuelve inestable debido a que se presentan demasiadas colisiones.

Otro parámetro importante a considerar es el retardo promedio, este comportamiento se observa en la figura 5.

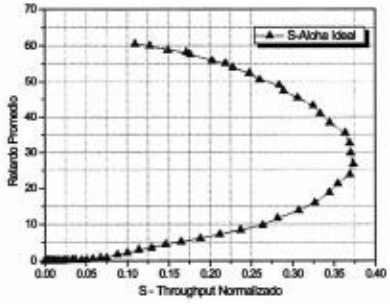


Figura 5. Comportamiento del retardo promedio en S-Aloha.

De la figura 5 se observa que el retardo promedio es elevado en la región de bajo tráfico y alcanza un máximo throughput en un valor de 27. Esto es debido a que cuando se presentan colisiones, el TM retransmite hasta que tenga éxito su transmisión, provocando un retardo desde el momento que genera la información hasta que la transmite con éxito. Además, cuando se llega a un valor máximo de throughput y sigue aumentando el tráfico al canal el retardo también aumenta.

Hasta este momento se han obtenido resultados con respecto al throughput y retardo, en el cual se presenta inestabilidad y valores elevados de retardo. En la figura 6 se muestra el comportamiento de un tercer parámetro que es estaciones en modo bloqueo.

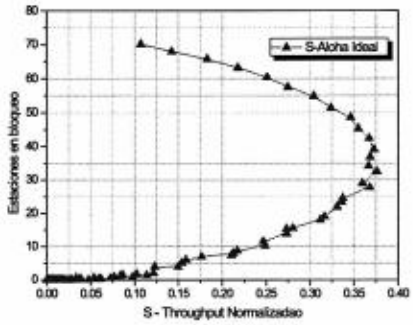


Figura 6. Comportamiento del número de terminales en backlog para S-Aloha

De la figura anterior, se observa que en la región de bajo tráfico el número de estaciones en estado bloqueo es elevado, en el cual a un máximo throughput el número de estaciones en estado bloqueo es 33. Y conforme aumenta el tráfico, después de alcanzar el máximo throughput, el número de estaciones en estado bloqueo incrementa también. Esto es debido porque al presentarse colisiones las estaciones quedan bloqueadas, por lo que es necesario un algoritmo que sea capaz de resolver de una manera óptima y rápida las colisiones y así disminuir el número de estaciones en estado bloqueo.

3. ACCESO MULTIPLE POR SENSADO DE PORTADORA

El protocolo CSMA es una evolución del protocolo de acceso al medio Aloha; el cual fue creado con el fin de incrementar la utilización de los canales de comunicación [9]. Las siglas CSMA significan: Carrier Sense Multiple Access. El nombre carrier proviene del hecho que existe una onda portadora sobre el canal de comunicación que es sensada por terminales de acceso.

3.1 FUNCIONAMIENTO

En este trabajo se hará uso del protocolo de acceso múltiple CSMA no persistente. En este protocolo cuando el canal está libre transmite y en caso de que el canal esté ocupado, espera un tiempo aleatorio e intenta de nuevo [14].

Usando el sensado de portadora, es posible determinar si otras terminales de acceso están transmitiendo [9]. En la figura 7 se muestra el proceso de transmisión y colisión entre terminales.

De la figura 7 podemos observar que el usuario de la terminal 1 realiza el sensado de portadora en el canal de comunicación, al detectar el canal libre transmite su paquete. Por otra parte, el usuario de la terminal 2 realiza el sensado de portadora, al detectar que el terminal 1 se encuentra transmitiendo detiene el proceso de transmisión y espera un tiempo aleatorio. Cuando el usuario del terminal n intenta transmitir y realiza el sensado de portadora, al detectar que el terminal 1 aún se encuentra transmitiendo interrumpe el proceso de transmisión y espera un tiempo aleatorio.

Es importante mencionar tanto el terminal 2 como el terminal n se encuentran realizando su periodo de espera aleatorio. Al terminar el periodo de espera, el terminal 2 realiza el sensado de portadora detectando el canal libre; simultáneamente la terminal n finaliza su periodo de espera y realiza el sensado de portadora detectando el canal libre también. Debido a que ambas terminales realizaron el sensado de portadora casi simultáneamente, las dos terminales detectaron el canal libre. Por este motivo el terminal 2 y el terminal n retransmiten, y es aquí donde ocurre la colisión [10].

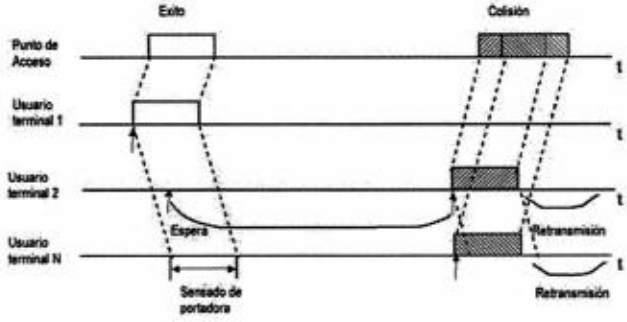


Figura 7. Proceso de transmisión y retransmisión np-CSMA.

3.2 MODELADO DE CSMA

En el protocolo CSMA no persistente cuando los paquetes son generados en una terminal de acceso, dicha terminal inicia el sensado de portadora. Si el resultado del sensado es "libre", el paquete es transmitido inmediatamente al punto de acceso. Sin embargo, si el resultado del sensado es "ocupado", la terminal de acceso detiene el sensado de portadora, espera un tiempo aleatorio, y entonces inicia el sensado de portadora de nuevo. El tiempo de espera es un punto clave para realizar sistemas con una salida alta [14].

El caudal eficaz (o salida del sistema) es expresado en ec.(15), dicha expresión es obtenida de acuerdo a [1].

$$S = \frac{U}{B+I}, \tag{15}$$

donde:

- B es el tiempo de espera en un periodo "ocupado",
- I es el tiempo en un periodo "libre", y
- U es el tiempo en el que no ocurren colisiones y la transmisión de paquetes se realiza exitosamente.
- S es el caudal eficaz

Este lo podemos observar en la figura 8, donde el tiempo en el que cada paquete es normalizado y los retardos de transmisión se definen como 1 y a, respectivamente. De tal manera que, transmitir exitosamente un paquete generado en un terminal móvil (TM) en un tiempo t_1 al punto de acceso es igual a la probabilidad de no generar paquetes en el periodo de $t1$ a $t1 + a$.

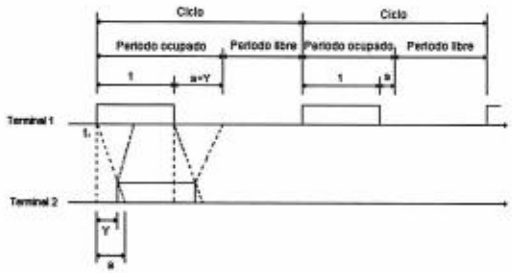


Figura 8. Periodicidad de paquetes del protocolo np-CSMA.

Así, el tiempo esperado en el cual no ocurren colisiones está dado por :

$$U = Ge^{-aG}. \tag{16}$$

El tiempo esperado de un periodo "libre" obedece a la siguiente distribución exponencial:

$$I = \frac{1}{G}. \tag{17}$$

El último paquete transmitido en el periodo de t_1 a $t_1 + a$ está definido como t_1 + Y , y el valor esperado de Y está definido como $\overline{Y}$, de tal manera que, el tiempo de un periodo ocupado está definido por:

$$B = 1 + a + \overline{Y}. \tag{18}$$

Realizando la función de distribución para Y, se obtiene la ec. (19) que representa la salida del sistema.

$$S = \frac{Ge^{-aG}}{G(1+2a)+e^{-aG}}. \tag{19}$$

donde:

a representa retardo de propagación
G representa el tráfico de la red

3.3 SIMULACIÓN

Al igual como en el proceso llevado a cabo en S-ALOHA, ahora también se va a simular como primer parámetro el caudal eficaz (throughput), S, que representa el porcentaje de paquetes transmitidos con éxito a través del canal.

Los parámetros considerados son:

- El radio de servicio medido en metros = 100
- Número de TMs de acceso =100
- Número de slots = 10000

- Buffer unitario
- No se puede generar otro paquete hasta que se transmita el que se encuentra en su buffer.

Es importante mencionar que cada TM de acceso genera paquetes aleatoria e independientemente. Tal generación de paquetes está dada por una distribución de Poisson.

El resultado del caudal eficaz es mostrado en la figura 9. Podemos decir que en un sistema ideal si no son generados paquetes de transmisión y todos los paquetes transmitidos son destruidos por colisiones, la salida *S* alcanza un valor mínimo de 0. De otra manera si todos los paquetes son transmitidos exitosamente, la salida alcanza un valor máximo de 1.

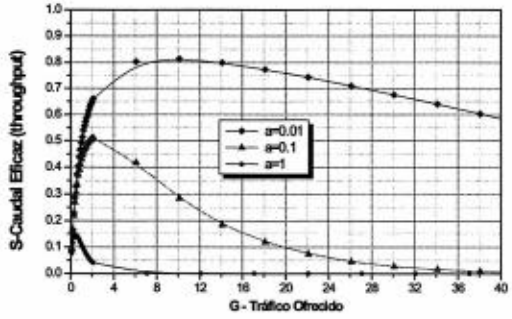


Figura 9. Caudal eficaz (throughput) para CSMA.

De la figura 9 podemos mencionar que a un mayor retardo de propagación, el sensado de portadora toma un periodo mayor de tiempo, lapso por el cual otro TM que intenta transmitir puede interpretar como libre el canal de comunicación. Al suceder esto aunado al tráfico de la red se generan las colisiones, disminuyendo considerablemente la eficiencia del sistema. Por otra parte, hay una zona de inestabilidad y ésta es causada por el incremento de tráfico en la red ya que al aumentar el tráfico a través del canal de comunicación aumenta también la probabilidad de colisiones en el mismo.

Con respecto al retardo de transmisión promedio (ver figura 10) podemos decir que es el periodo desde que el paquete es generado en un TM de acceso, transmitido hacia el punto de acceso y recibido en el punto de acceso destino. El retardo de transmisión promedio depende de la longitud del paquete.

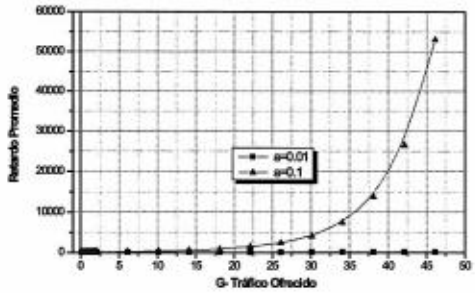


Figura 10. Tráfico y retardo promedio de transmisión de np-CSMA

En la figura 10 se muestran dos resultados. Cuando a=0.1 el retardo promedio de transmisión se incrementa considerablemente en forma exponencial, mientras que cuando a=0.01 el incremento del retardo promedio de transmisión es casi despreciable. Esto se debe a que al incrementar el valor de a el periodo que le toma a un paquete desde que es generado en un TM, pasar por un punto de acceso y ser recibido en un TM destino es mucho mayor; por lo cual se reduce considerablemente cuando aplicamos valores pequeños de a.

4. ACCESO MULTIPLE POR SEÑAL DE INHIBICION (ISMA)

Uno de los mejores protocolos de comunicación para un paquete de acceso aleatorio es S-ALOHA, pero tiene una deficiencia este sufre de manera brusca las colisiones entre los paquetes. La segunda categoría es el protocolo CSMA, el cual ofrece una alta capacidad, pero su rendimiento es afectado por el terminal oculto. El protocolo ISMA es el tercer tipo de protocolo de acceso aleatorio en donde cada estación base (EB) controla el flujo de paquetes en los terminales móviles (TMs). Este protocolo también reduce los dos problemas que presentan los protocolos de acceso múltiple, que son las colisiones de transmisión de paquetes, y el problema del terminal oculto, este protocolo también es definido como la emisión de un canal de comunicación de salida (base-terminal), y este evento el canal de acceso múltiple de entrada (terminal- base) está ocupado. En la figura 11 se puede mostrar el funcionamiento del protocolo ISMA.

Otro de los aspectos de ISMA es que éste es una porción del retardo de inhibición, el cual es un parámetro decisivo en la designación de paquetes de un sistema de comunicación.

En la misma figura 11 el punto de acceso manda una señal ocupada a todas las terminales de acceso, cuándo esta recibiendo paquetes de las terminales, y el punto de acceso manda una señal inactiva esto es debido a que las terminales no están enviando paquetes al punto de acceso. Por otra parte cuando las terminales reciben una señal inactiva, éstas deben decidir si van o no transmitir paquetes al punto de acceso, y cuando las terminales están recibiendo una señal ocupada la transmisión de paquetes para cada terminal es inhabilitada, por eso es que al protocolo se le llama ISMA [15].

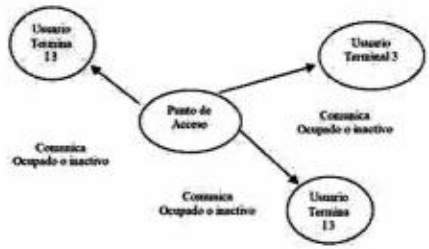


Figura 11. Protocolo ISMA.

Existen dos subclases en los que se pueden dividir el protocolo ISMA: ISMA no ranurado y ranurado. El protocolo que se va a analizar en esta investigación va ser el protocolo ISMA ranurado.

4.1 FUNCIONAMIENTO DEL PROTOCOLO ISMA RANURADO

El protocolo ISMA se beneficia de contar con un canal que está dividido en ranuras. Ya que en éste existe una fuerza que obliga a los paquetes a ser transmitidos y recibidos por la estación base en unos instantes de tiempo al inicio de cada ranura, por lo que únicamente podrá producirse una colisión en dichos instantes. Si suponemos que el canal de bajada está ranurado por lo que la señal de subida solo será efectiva a partir de la primera ranura posterior de la transmisión, por lo que existe un retardo de una ranura para esta señal con respecto del inicio de la transmisión del paquete. En la figura 12 se muestra el funcionamiento.

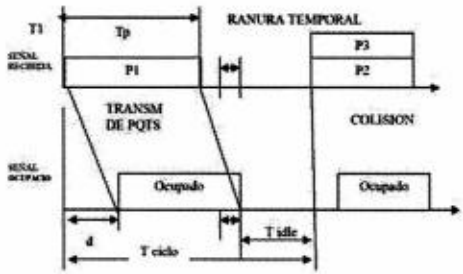


Figura 12. Esquema del protocolo ISMA ranurado.

4.2 MODELO MATEMÁTICO PARA ISMA RANURADO

Entrando al análisis matemático, es preciso definir el tiempo de ciclo, en la cual contribuirán los siguientes factores,

$$T_c = d + T_P + T_{idle} = d + 1 + T_{idle} \tag{20}$$

Donde nuevamente se ha considerado la duración de un paquete, $T_p = 1$, y d es la duración de tiempo temporal de una ranura según esta unidad de tiempo.

Para el calculo del tiempo promedio T_{idle} basta con recurrir a una estadística de llegadas, considerando que todos los paquetes lleguen dentro de una ranura, cuando el canal no este ocupado, ya que este supondrá que habrán transmisiones de paquetes al inicio de cada ranura, por lo que cuando el canal se encuentre ocupado el tiempo T_{idle} tomará un valor mínimo d esto es una ranura si no se produce ninguna llegada en la primera ranura, y al final del periodo BUSY, una o mas si no se producen llegadas en la segunda ranura, lo cual tomará el valor de $2d$, y así sucesivamente, llegando a la siguiente ecuación.

$$E[T_{IDLE}] = d * P_a(0,d)(1 - P_a(0,d)) + 2d * P_a(0,2d)(1 - P_a(0,d)) + \ldots\ldots \tag{21}$$

$$= \sum_{K=0}^{\infty} kde^{-Gkd}\left(1 - e^{-Gd}\right) = \frac{de^{-Gd}}{1 - e^{-Gd}}$$

$$E[T_c] = \frac{1 + d - e^{-Gd}}{1 - e^{-Gd}} \tag{22}$$

Para el cálculo del caudal eficaz en el caso ranurado, bastará con considerar el número medio de paquetes transmitidos correctamente en un tiempo de ciclo N_p, que será de 1 siempre que se produzca alguna llegada y esta sea única en una ranura temporal.

Para el cálculo de N_p bastara solo con aplicar la probabilidad de que exista una única llegada condicionada en el momento que se haya producido una llegada, pues el tiempo medio de ciclo queda definido como.

$$N_p = 1 * P[1\_llegada / hay\_llegadas] = \frac{P[1\_llegada]}{P[hay\_llegadas]} = \frac{P_a(1,d)}{1 - P_a(0,d)} = \frac{Gde^{-Gd}}{1 - e^{-Gd}} \tag{23}$$

De esta manera se obtiene el caudal eficaz como el cociente entre N_p y $E[T]_c$, esto es

$$S_{SLOTTED} = \frac{Ge^{-Gd}}{1 + d - e^{-Gd}} \tag{24}$$

4.3 SIMULACIÓN

El protocolo será llevado a un ambiente de simulación con los siguientes parámetros:

- El número de terminales 20.
- El número de paquetes 10000
- Seguir un proceso de Poisson en la cual la probabilidad de transmisión será de 0.01 a 1, para cada probabilidad de 10000 ranuras.
- Buffer unitario.
- Cuando un usuario genera un paquete éste no puede generar otro hasta que lo transmita.

El resultado obtenido del caudal eficaz en función de tráfico ofrecido es mostrado en la figura 13.

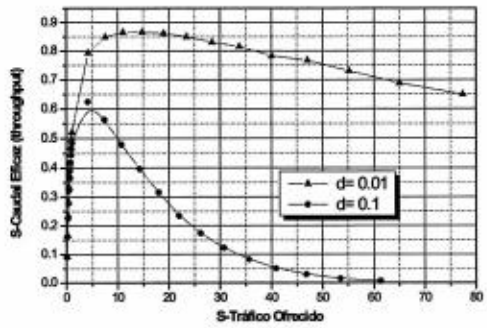


Figura 13. Comportamiento del caudal eficaz en ISMA ranurado.

En la figura 13 se observa el caudal eficaz contra el tráfico en donde se hace una comparación con dos diferentes valores que toma el retardo de propagación normalizado, en el cual se refleja claramente que la eficiencia del sistema es muy bajo, cuando toma valores mayores a 0.01, esto es debido a que el caudal eficaz empeora al incrementar el retardo de propagación *d*, pues para valores mas elevados suponen una mayor probabilidad de colisiones, por lo que la eficiencia del sistema es bajo cuando toma el valor de 0.1. Comparando con el retardo de 0.1, se ve claramente, que al aumentar el retardo de transmisión a 1 el caudal eficaz decrece, por lo que ISMA ofrecería un mal comportamiento, no solo se pueden producir colisiones a lo largo de toda la transmisión sino que se estará desaprovechando el canal, ya que la información indicada por la base no correspondería al estado real del sistema, ya que esta estaría ocupando la región de incertidumbre de todo el paquete.

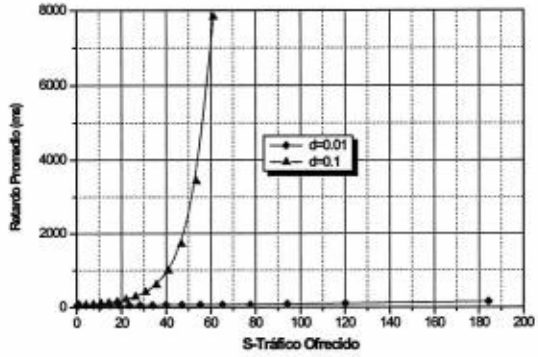


Figura 14.Retardo promedio de ISMA rasurado.

En la figura 14 se observa el retardo contra el caudal eficaz, podemos observar que cuando se tiene que a= 0.01, podemos ver que el caudal alcanza una eficiencia de 0.89, cuándo alcanza este valor el retardo de promedio de tiempo se puede ver que va ser de 25.38 milisegundos, se podría decir la transmisión del paquete al TM de acceso es más rápida, cuando a=1 se puede ver que el sistema tiene una eficiencia de 0.38, con un retardo de 11.2 milisegundos, por lo que la transmisión del paquete será lenta.

5. ACCESO MULTIPLE POR RESERVACION DE PAQUETE (PRMA)

5.1 INTRODUCCIÓN

En la actualidad se ha extendido el interés en redes inalámbricas y móviles para comunicaciones en voz y datos principalmente, por lo que es necesario desarrollar protocolos que controlen el acceso de usuarios al canal radio común. Estos protocolos pueden ser diseñados, así que el espectro de frecuencia asignado para las comunicaciones, es usado de la manera más eficiente posible con objeto de maximizar la capacidad y eficiencia de los sistemas.

En la investigación se ha considerado un protocolo de transferencia de información paquetizada que es PRMA — Acceso Múltiple por Reservación de Paquete — que fue propuesto por [6]. En la figura 15 se ilustra la relación entre PRMA y las otras técnicas de acceso [12].

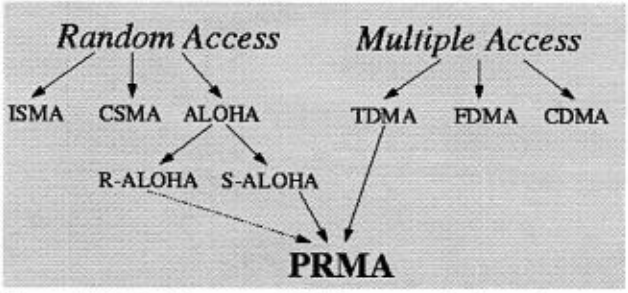


Figura 15. Relación entre PRMA y otras técnicas de acceso.

5.2 FUNCIONAMIENTO

5.2.1 UN PROTOCOLO PARA PAQUETE DE VOZ

Con el protocolo diseñado para acceso múltiple en una sola celda, la red tiene una topología de estrella con la estación base como el nodo central. En [6] PRMA es propuesto como un multiplexor para paquetes terminales de voz y datos aleatorios. Las terminales de voz dispersas en el espacio transmiten paquetes de longitud fija en ranuras de tiempo, hacia la estación base (EB). La distribución del tiempo del slot se obtiene de la distribución del tiempo que retroalimentó la EB.

Después de cada slot de tiempo, la EB difunde a todas los terminales móviles (TMs) un paquete de retroalimentación basado en la información que recibió en ese slot. Si la EB es capaz de decodificar el encabezado de un paquete que llegó, el paquete de retroalimentación identifica el TM que envió el paquete a la EB.

Si la EB no es capaz de decodificar el encabezado de un paquete que llega, la EB difunde un paquete de retralimentación "nulo" para indicar este resultado. La EB no necesita indicar por qué no es posible decodificar un encabezado que llega. Las razones posibles son: no se transmitió el paquete (estado vacío-idle), se transmitió más de un paquete (colisión); se transmitió un paquete pero fue perjudicado por condiciones del canal adversas (errores del paquete). Para simplificar el problema se considerarán canales libres de errores.

5.2.2 ACCESO AL CANAL Y PERMISO

El canal PRMA es ranurado, y las ranuras son agrupadas en tramas. La tasa de la trama es idéntica a la tasa que llega de los paquetes de voz. El periodo del slot es el tiempo de transmisión de un paquete de voz. Los TMs clasifican a cada slot como "reservado" o "disponible" de acuerdo al mensaje de retroalimentación que recibió de la EB al final del slot. En la siguiente trama, un slot reservado puede ser usado solamente por el TM que lo reservó. Un slot disponible puede ser usado por cualquier TM, que tiene información para transmitir a la EB. Cuando inicia una descarga de voz de multipaquete, los TMs contienden por el siguiente slot de tiempo disponible. Sobre la recepción satisfactoria del primer paquete de la descarga, la EB concede al TM una reservación para uso exclusivo del mismo slot de tiempo en las tramas subsecuentes. Al final de la descarga, el TM detiene la transmisión. Su slot de tiempo reservado vacío causa que la EB difunda un mensaje de retroalimentación nulo para indicar a todos los TMs que el slot está otra vez disponible.

Si dos TMs simultáneamente transmiten un paquete en un slot disponible, ocurre una colisión. La EB falla en detectar uno u otro paquete y ambos TMs tienen que retransmitir los paquetes. En la práctica, cuando llegan paquetes que colisionan a la EB con niveles de señal diferentes substancialmente, la EB puede ser capaz de detectar al paquete con la señal más fuerte. Esto se menciona como la captura de paquete. Aunque la captura mejorará el funcionamiento de PRMA, ignoramos sus efectos en el análisis y asumimos que todos los paquetes que colisionan requieren ser retransmitidos.

Como en R-ALOHA, un TM que contiende, transmite un paquete en un slot disponible si tiene permiso para transmitir. El permiso ocurre en cada una de los TMs con una probabilidad fija, como se determinó por un generador de números pseudo-aleatorios. El TM intenta transmitir el paquete inicial de una descarga hasta que la EB reconoce la recepción satisfactoria del paquete, o hasta que el TM descarta al paquete porque ha sido retrasado demasiado tiempo. El tiempo máximo de espera del paquete, D_{max} s, es determinado por el límite de retardo que soporta la transmisión de voz. D_{max} es un parámetro del diseño del sistema de PRMA. Si un TM pierde el primer paquete de una descarga, continúa contendiendo por una reservación para enviar los paquetes subsecuentes. Ésta pierde paquetes adicionales cuando sus tiempos de espera exceden el límite del retardo.

Puesto que PRMA es un multiplexor estadístico, cuando el tráfico se acumula, las colisiones de paquetes se incrementan y los TMs encuentran retardos en el acceso al canal. Las fuentes de datos absorben estos retrasos como penalidades del funcionamiento. Las conversaciones requieren liberar pronto la información y los TMs de voz descartan los paquetes retrasados. Este paquete perdido perjudica la calidad de la voz. Una medida clave del funcionamiento de PRMA es el número de terminales de voz que pueden compartir un canal con un nivel de tolerancia dada de la probabilidad de descarte del paquete. Este proceso de acceso al canal y probabilidad de descarte de paquete se muestra en la figura 16.

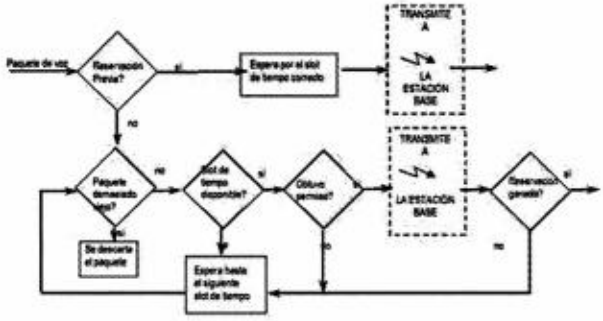


Figura 16. Proceso de transmisión de PRMA.

5.2.4 EJEMPLO DE OPERACIÓN DE PRMA

En la figura 17 existen 8 slots por trama y la EB estableció en la trama *k*-1, que en la trama *k*, 6 slots están reservadas por los TMs 11, 5, 3, 1, 8 y 2, 2 slots están disponibles, la 3 y 7. Al inicio de la trama los TMs 6 y 4 están contendiendo para accesar al canal, los dos obtienen permiso de transmitir en la ranura 3 y como colisionan nunca obtienen la reservación. En el slot 7 ambos TMs fallan en obtener permiso para transmitir y permanecen en estado de contención al inicio de la trama *k*+1, mientras tanto, en la trama *k*-1, el TM 3 transmitió su paquete final, por lo tanto en la trama *k* (slot 4) no usa esa reservación. La EB transmite un mensaje de retroalimentación para el slot 4 en trama *k* que indica que estará disponible en la trama *k*+1.

En la trama *k*+1 ni el TM 6 ni el 4 tienen permiso de transmitir en el slot 3, en slot 4, el TM tiene permiso pero el TM 6 no, el TM 4 gana acceso para el slot 4, TM 6 obtiene permiso para transmitir en slot 7 y reserva esa ranura en la trama *k*+2. En *k*+1 el TM 8 deja el slot 6, una ráfaga empieza en TM 12 el cual entra en estado de contención. En la trama *k*+2, el TM 12 gana una reservación (slot 3) y el TM 1 libera su reservación (slot 5).

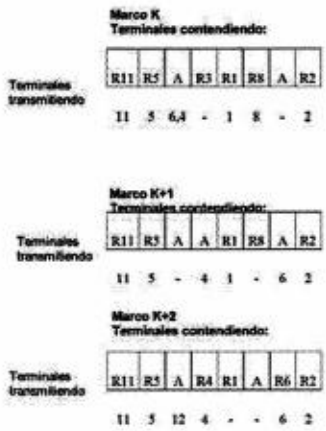


Figura 17. Ejemplo de operación de PRMA.

5.3 VARIABLES DEL SISTEMA

5.3.1 DETECTOR DE VOZ

Una fuente de voz crea un patrón de ráfagas y silencios, y esto puede ser descubierto por un detector de actividad de voz. Para el análisis presentado en este trabajo se presenta solamente el detector de actividad de voz lento.

El detector de voz lento es modelado como un proceso de Markov de dos estados (ver figura 18). Donde la probabilidad que una ráfaga con duración media de t1 finalice en un slot de tiempo de duración τ segundos es

$$\gamma = 1 - \exp\left(-\tau / t_1\right) \tag{25}$$

Esta es la probabilidad de pasar de una transición de un estado activo (hablando), TLK, a un estado de silencio, SIL. De la misma manera, la probabilidad de que un silencio, de duración media t2 segundos, finalice durante un slot de tiempo de duración τ segundos es dada por

$$\sigma = 1 - \exp\left(-\tau / t_2\right) \tag{26}$$

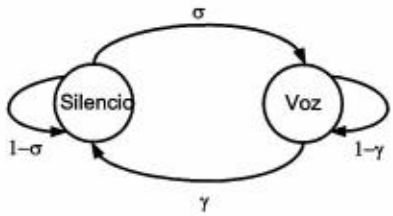


Figura 18. Detector de actividad de voz lento.

5.3.2 ORGANIZACIÓN DE TRAMAS Y RANURAS

El canal del enlace ascendente es primero organizado en ranuras (slots), de tal manera que cada ranura puede llevar un paquete de un TM a la EB, los tiempos de la ranura, son agrupados en tramas, dentro de una trama, las terminales reconocen cada ranura como reservada o disponible con base a la información de retroalimentación de tipo difusión en la trama anterior, como parte del tráfico de enlace descendente (ver figura 19).

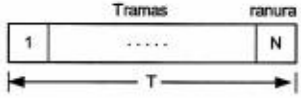


Figura 19. Trama para PRMA

El número de ranuras en cada trama esta dado por

$$N = \text{int}\left[\frac{R_c T}{R_s T + H}\right] \tag{27}$$

donde N es el número de ranuras por trama para todos los TMs, R_c es la tasa del canal, R_s es la tasa de la fuente, T es el tiempo de duración de trama y H es el número de bits del encabezado.

5.3.3 LÍMITE DEL RETARDO Y TAMAÑO DEL BUFFER

Un TM de voz contiene un 'buffer' del tipo FIFO para almacenar los paquetes que están en espera de transmitir. Si el 'buffer' está lleno cuando arriba un nuevo

paquete, el TM desecha el paquete que más tiempo ha estado almacenado y almacena el nuevo paquete.

Con este mecanismo de desechar paquetes, el tamaño del 'buffer' requerido es

$$B = \left\lceil \frac{D_{\max}}{T} \right\rceil \tag{28}$$

donde D_{max} es el máximo retardo de transmisión para la voz y T es la duración de la trama.

En el análisis de PRMA, la variable D es definida como el máximo tiempo de espera medido en ranuras de tiempo,

$$B = \left\lceil \frac{D_{\max}}{\tau} \right\rceil \tag{29}$$

donde τ es la duración de la ranura de tiempo. Note que $T = \tau N$ no implica automáticamente $BN = D$. En general $BN \geq D$, con igualdad solamente si D es un número entero múltiplo de N.

5.4 MODELO DEL SISTEMA PRMA

Usamos el detector lento para el modelado de la voz. Según este modelo, una fuente de voz que crea unas series de ráfagas y vacíos es clasificado por el detector de actividad de voz a estar en el modo de ráfaga y modo de silencio respectivamente. Los paquetes de voz son generados por el TM cuando la fuente de voz está en modo de ráfaga.

5.4.1 MODELO DEL TERMINAL

Una ranura de tiempo es la unidad de tiempo en un sistema PRMA. Las transiciones en un modo terminal ocurren al final de cada ranura de tiempo. El modelo de Markov para un terminal es mostrado en la figura 19. Las etiquetas en las ramas de cadena son las probabilidades de transición. Consideramos primero el caso donde un terminal permanece en la misma célula durante una ráfaga. Cuando un terminal está inactivo, está en modo silencio (S). En el inicio de una ráfaga, éste entra en el modo contención (C), este evento ocurre con una probabilidad σ, en cada ranura de tiempo. El 'slot' en el que el TM maneja con éxito su primer paquete es reservado por la EB en tramas subsecuentes. El terminal que ha obte-

nido una reservación entra al modo reservación (R_n-*1*), donde *n-1* es el número de 'slots' que permanecen antes de que el terminal empiece otra vez a transmitir.al final de un 'slot' de tiempo, el terminal en modo R_n-*1*, con probabilidad 1, irá al modo R_n-*1* y así sucesivamente. De R_0, el terminal retorna a R_n-*1* si tiene más paquetes para transmitir. De otra manera, retorna al modo silencio. La probabilidad de una transición del estado R_0 al modo silencio es la probabilidad que la ráfaga en la más reciente trama. Esta probabilidad de transición es dada como

$$\gamma_f = 1-(1-\gamma)^N \approx N\gamma \tag{30}$$

donde γ es la probabilidad que la ráfaga finalice en uno de los N ranuras de la trama.

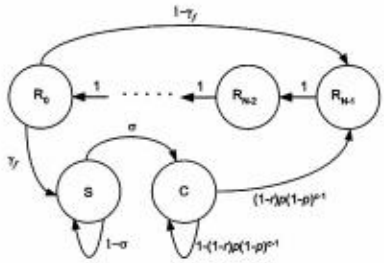


Figura 19. Modelo de Markov para PRMA.

Para el análisis de prestaciones del sistema PRMA se utiliza la técnica EPA (Equilibrium Point Analysis), que es una técnica analítica de aproximación en que se asume que el sistema está en un estado de equilibrio. El número de ecuaciones a resolver en un proceso de Markov incrementa tanto como el número de terminales incrementa. Sin embargo, utilizando EPA el número de ecuaciones a resolver crece linealmente con el número de 'slots' en una trama. Un estado de equilibrio es definido como el estado en que el incremento esperado en el número de TMs en cada modo es cero.

Definimos ahora Ω como el conjunto de modos en el cual el terminal puede estar. Los modos son: $\Omega=\{S, C, R_0, ..., R_{n-1}\}$. Entonces para una célula dada el estado del sistema es dado por el vector ($S, C, R_0, ..., R_{n-1}$) donde S es el número de TMs en una célula en modo S, C es el número de terminales en modo C y así sucesivamente. Además, s, c, ri ($0 \le i \le n-1$) denotan los valores de equilibrio de las variables de estado mencionadas anteriormente. Para que un TM en estado de contención transmita un paquete con éxito, debe cumplirse las condiciones

siguientes: el TM puede transmitir solamente en un 'slot' de tiempo que no esté reservado, éste puede tener permiso para transmitir, y no retornar al modo de silencio. También, ninguno de los otros TMs los cuales están contendiendo por un 'slot' puede tener permiso para transmitir en ese 'slot'.

Para este análisis, se observa que en equilibrio todos los 'slots' tienen una misma probabilidad, r, de reservación:

$$r_0 = r_1 = r_2 = \ldots = r_{n-1} = r \tag{31}$$

- En equilibrio, las probabilidades de transición del estado de contención (C) al estado
- R_{N-1}, es el producto de las probabilidades
- $1\text{-}r$ probabilidad que haya un spot disponible
- p probabilidad de permiso, es un parámetro del sistema; y
- $u(c)$ es la probabilidad que ningún otro TM contendiente tiene permiso para transmitir en el spot actual.

El valor de $u(c)$ es dado por,

$$u(c) = \begin{cases} (1-p)^{(c-1)}, \ldots\ldots\ldots c \geq 1 \\ 1, \ldots\ldots\ldots\ldots\ldots\ldots c < 1. \end{cases} \tag{32}$$

La ecuación de equilibrio en R_{N-1} puede ser escrita de acuerdo a los flujos de salida y entrada,

$$r(1-\gamma_f) + cpu(1-r) = r, \tag{33}$$

con $u=u(c)$ dada en la ec. (33). Similarmente en modo silencio (S) se obtiene,

$$r\gamma_f = s\sigma \tag{34}$$

Por otra parte, siendo M el número total de TMs en todos los estados $N+2$, tenemos

$$s + c + Nr = M \tag{35}$$

Sustituyendo los valores de ecs. (33) y (34) en ec. (35) se consigue

$$c+\left(N+\frac{\gamma_f}{\sigma}\right)\frac{cpu}{cpu+\gamma_f}=M \tag{36}$$

Para obtener el desempeño de PRMA se calcula el caudal eficaz que es definido como la fracción promedio de 'slots' de tiempo que con éxito transportan paquetes de las terminales a la EB y es dado por

$$\eta = r(1-\gamma_f)+cpu(1-r) \tag{37}$$

5.5 SIMULACIÓN

Para la simulación de las prestaciones de PRMA se consideran los parámetros que se muestran en la tabla I.

Parámetro	*Valor*
Tasa del canal, R_c	**720 kbps**
Duración de la trama, T	**16 ms**
Duración de la ranura, τ	**0.8 ms**
Ranuras por trama, N	**20 ranuras**
Duración promedio del periodo de voz, t_1	**1.00 s**
Duración promedio del periodo de silencio, t_2	**1.35 s**
Tasa de voz, Rs	**32000 bps**
Tamaño del paquete de voz	**576 bits**
Encabezado por paquete	**64 bits**
Número máximo de paquetes en el buffer	**2 paquetes**
Retardo máximo de voz	**32 ms**
Probabilidad de permiso para voz	**Variable**
TMs de voz	**Variable**

Los resultados obtenidos se muestran en la tabla II y III. Podemos decir, que de acuerdo a la tabla II, se obtiene un máximo de 36 conversaciones simultáneas con las probabilidades de permiso de 0.3 y 0.3. Se pensaría que al aumentar la probabilidad de permiso aumentaría el número de conversaciones, pero no sucede eso debido a que al aumentar la probabilidad de permiso produce mayor número de TMs que contiende y eso deriva en un aumento de colisiones y por consiguiente la ranura de tiempo queda disponible.

Tabla II. Número máximo de conversaciones simultáneas.

Probabilidad de permiso	*Número máximo de conversaciones simultáneas*
p=0.1	**26**
p=0.2	**33**
p=0.3	**36**
p =0.4	**36**
p =0.5	**34**
p =0.6	**34**

Se analizó el número máximo de conversaciones, ahora toca determinar cuál es el caudal máximo para PRMA. El resultado se presenta en la tabla III.

De la tabla III se concluye que al igual que en el resultado anterior el máximo caudal se obtiene con una probabilidad de permiso de 0.3 y 0.4. Además al aumentar la probabilidad de permiso da como resultado que los TMs colisionan y la ranura quede disponible, eso hace que la eficiencia del sistema disminuya.

Tabla III. Caudal eficaz para PRMA.

Probabilidad de permiso	*Número máximo de conversaciones simultáneas*	*Caudal eficaz (throughput)*
p=0.1	**26**	**0.55**
p=0.2	**33**	**0.68**
p=0.3	**36**	**0.77**
p =0.4	**36**	**0.78**
p =0.5	34	**0.73**
p =0.6	**34**	**0.70**

6. CONCLUSIONES

En este trabajo de investigación se ha llevado a cabo el análisis del modelado y simulación de las prestaciones de los protocolos S-ALOHA, CSMA, ISMA y PRMA utilizados en la fase de acceso al canal en sistemas de comunicaciones móviles. De acuerdo a los resultados del desempeño de cada uno de los protocolos estudiados se concluye:

- El bajo desempeño y el alto retardo puede ser mejorado utilizando la técnica de efecto captura. El efecto captura indica que aún cuando se haya presentado una colisión la estación podrá recibir la información de aquel terminal móvil que haya transmitido con la mayor potencia y esté arriba de un umbral.

- La inestabilidad en la región de alto tráfico puede ser mejorada usando un algoritmo de retransmisión considerando el canal de retorno.
- Así mismo, los esquemas estudiados pueden agregarse al del efecto del canal radio y determinar si afecta el desempeño del sistema.
- En PRMA, éste es un protocolo orientado a voz. Al aumentar la probabilidad de permiso de transmisión se creería que aumentaría el desempeño del sistema pero no es así, porque sucede que habrá más usuarios intentando transmitir y da como resultado un aumento de colisiones.
- Estos protocolos de control de acceso al medio son utilizados en la petición del canal radio, y para la transmisión puede combinarse con otro protocolo: ALOHA-CDMA, ISMA-CDMA, PRMA-CDMA, CSMA-CDMA.

REFERENCIAS

[1] Ahmad Bahai. "CSMA evaluation and enhancement with physical layer imperfections". EE 360 Projects.

[2] A.B. Carleial and M. E. Hellman, "Bistable behavior of ALOHA-type systems", IEEE Trans. Commun., vol. COM-23, pp.401-410, April 1975.

[3] D. Bertsekas and R. Gallager, Data Networks, 2nd ed. NJ: Prentice-Hall, 1992.

[4] D. Covarrubias, Notas de la materia "Comunicaciones Móviles Celulares de Tercera Generación", Abril-Julio del 2000, CICESE, Ensenada, B.C.

[5] D. Covarrubias, Notas de la materia "Protocolos de Acceso Múltiple en Comunicaciones Móviles Celulares", Agosto-Noviembre del 2000, CICESE, Ensenada, B.C.

[6] D. J. Goodman, R. A. Valenzuela, K. T. Gayliard and B. Ramamurth, "Packet Reservation Multiple Access for Local Wireless Communications", IEEE Transaction on Communications, Vol. 37, No. 8, pp. 885-890, August 1989.

[7] Eric Lawrey. Chapter 1. [En línea]. Disponible: http://www.skydsp.com/publications/4thyrthesis/chapter1.htm

[8] G. Bolch, S. Greiner, H. de Meer, and K. S. Tivedi, Queueing Networks and Markov Chains: Modeling and Performance Evaluation with Computer Science Applications, John Wiley & Sons, first edition, Canada 1998, pp. 726

[9] ITLP, Instituto Tecnológico de la Paz. "Protocolos de Comunicación".[En línea]. Disponible: http://www.itlp.edu.mx/publica/tutoriales/redes/tema14.htm

[10] Iván Briseño, Análisis de Prestaciones del Protocolo CSMA Aplicado a Comunicaciones Móviles, Tesis de Ingeniería, UAT, México, Julio de 2006.L.

Kleinrock, Queueing Systems Volum I: Theory, John Wiley & Sons, first edition, U.S.A. 1975, pp. 417.

[11] L. Kleinrock and S. S. Lam, "Packet Switch in a Multiaccess Broadcast Channel: Performance Evaluation", IEEE Trans. on Communications, Vol. COM-23, No. 4, pp.410-423, April 1975.

[12] M. D. Orange, "Performance of Microcellular PRMA in a Fading Channel Environment", in School of Engineering Report No. 557, The University of Auckland, New Zealand, Dec. 1995, [online: http://www.ele.auckland.ac.nz/students/orange/papers/ele_let2.pdf]

[13] N. Abramson, "Multiple Access in Wireless Digital Networks", Proc. Of the IEEE, Vol. 82, No. 9, pp.1360-1370, September 1994.

[14] O.K. Li, Victor. "Multiple Access Communication Networks". IEEE Communications Magazine, vol. 25, No. 6, June 1987.

[15] Oliver Rivera, Evaluación de Prestaciones del Protocolo ISMA para Sistemas de Comunicaciones Móviles, Tesis de Ingeniería, UAT, México, Julio de 2006.

[16] Shuji Tasaka, "Perfomance Analysis of Multiple Access Protocols - Research, reports and notes". 1986, 1a ed., Ed. The MIT Press. USA.

2. Análisis de Elementos Básicos para la Integración de una Estrategia de Seguridad en Redes Inalámbricas de Área Local.

Armando Vega Pérez, Marco Antonio Panduro Mendoza,
Carlos del Rio Bocio

1. Introducción

La libertad de tener la posibilidad de acceder a una red corporativa sin ser limitado por una infraestructura cableada ha hecho de las redes inalámbricas toda una institución en la industria de las telecomunicaciones. Esto ha desencadenado una gran demanda por parte de los usuarios de tener la posibilidad de acceder recursos de red en cualquier lugar.

Cada ambiente de red es susceptible a riesgos y las redes inalámbricas no son la excepción. De acuerdo a una encuesta de la división de Crímenes y Seguridad Computacionales del Buró Federal de Investigaciones (FBI), la única categoría que indica un incremento en tipos de ataques o posibilidades a mal uso es "abuso de redes inalámbricas". La naturaleza de broadcast de las redes inalámbricas las ha convertido en blancos perfectos para recibir ataques por parte de usuarios no autorizados.

Este problema es aún más exacerbado por el gran cúmulo de herramientas gratuitas de intrusión de seguridad que están disponibles en Internet, al igual que por las vulnerabilidades inherentes de las redes inalámbricas mismas. Una de las vulnerabilidades más explotadas es el protocolo WEP, el cual es un problema tan grave que muchas compañías han decidido abandonar el negocio de las redes inalámbricas.

Por otro lado, gran parte de las estrategias de despliegue de redes inalámbricas carecen de una integración efectiva con la infraestructura de servicios de autentificación de la organización en la cual se implementan. Este error común es fácil de mitigar y su corrección rinde frutos casi de manera inmediata, al cerrar

la brecha en cuanto al número de individuos que pueden hacer uso de la misma red debido a que esos usuarios son extraídos de una base de datos cuyo acceso se lleva a cabo de manera segura por parte de la infraestructura misma.

En otros casos, los problemas de seguridad van más allá del elemento meramente tecnológico. Muchas veces, la falta de planeación de la red inalámbrica es un factor determinante en relación a la cobertura y ubicación de la misma. Otros elementos como políticas de seguridad, procedimientos de acceso, políticas internas de uso de recursos de red y lineamientos de confidencialidad y protección de la información sirven como una estructura regulatoria complementaria que provee de apoyo a la infraestructura tecnológica, estableciendo limitaciones relacionadas a la manera en la cual se usa o debe usarse la red misma.

El objetivo de esta investigación es el de proveer un marco de trabajo para el despliegue de redes inalámbricas seguras. Se hace hincapié en adaptar los lineamientos y soluciones recomendados para satisfacer las necesidades específicas tanto de seguridad como de negocio.

De igual forma, la investigación hará un análisis exhaustivo de los esquemas de redes locales inalámbricos existentes (infraestructura), su uso en diversas aplicaciones situacionales, las debilidades inherentes de las tecnologías debido a su diseño o a prácticas comunes inadecuadas y los grupos de mejores prácticas orientadas a mitigar los problemas de seguridad más comunes mediante la aplicación de tecnologías, protocolos y esquemas de red existentes en el mercado.

Este trabajo de investigación se da ala tarea de abordar el problema que presentan las redes inalámbricas Wlan en cuanto a seguridad y control de acceso se refiere, buscando la estrategia de proveer a las mismas de una estrategia de seguridad que comprende dos aspectos mutuamente relacionados:

- El análisis e implementación de protocolos y tecnologías abiertas y propietarias que provean de seguridad a las comunicaciones que transitan por una red inalámbrica y que proporcionen un esquema de acceso seguro a la red.

- El establecimiento de un marco de referencia de recomendaciones y mejores prácticas que complemente y amplíe la seguridad física proveída por la infraestructura de tecnología existente.

Esta investigación no cubre el análisis de tecnologías celulares como 3G y 2.5G, las cuales tienen una orientación preponderantemente hacia las redes de telefonía celular.

2. Conceptos Básicos

El auge de las redes inalámbricas de datos ha vivido una trayectoria ascendente en los últimos 8 años. Básicamente, el acceso a la tecnología inalámbrica (access points, tarjetas de red inalámbricas) se ha hecho más fácil debido a los relativamente bajos precios por los cuales se pueden obtener los componentes necesarios para implementar una red inalámbrica. Gran parte de esos equipos son comercializados bajo el titulo de equipos SOHO (Oficina Pequeña, Oficina de Hogar por sus siglas en inglés) cuya instalación es inherentemente sencilla de llevar a cabo debido a que los usuarios de esos equipos son, por lo general, usuarios neófitos o personas con conocimientos básicos de instalación de equipos de cómputo.

Las ventajas actuales de la implementación de una red inalámbrica son:

- Disponibilidad: Los miembros de una organización pueden tener acceso a los recursos de información en cualquier lugar si necesidad de depender de una infraestructura de cables.

- Movilidad: Un usuario puede trasladarse de un lugar a otro dentro de un mismo edificio, de edificio a edificio e incluso a otra ciudad y las redes inalámbricas le seguirán proveyendo de una conexión a los recursos de información que desee acceder. Este punto presupone uno de los principales retos o amenazas en materia de seguridad de redes inalámbricas.

- Productividad: Debido a que las redes inalámbricas pueden proveer de conexiones en virtualmente cualquier lugar, esto permite a los usuarios poder seguir trabajando sin importar donde se encuentren. Este punto cobra mayor importancia en el sector de usuarios empresariales debido a factores como acceso a información de negocios y sistemas de información administrativa.

- Facilidad de instalación: Una red inalámbrica sencilla puede ser desplegada en cuestión de minutos y puede ser trasladada a otro lugar con la misma rapidez con la que fue instalada. Básicamente, esta ventaja tiende a ser mas notoria cuando se trata de redes cuya permanencia y complejidad no es demasiado alta. Sin embargo, la facilidad de instalación no significa que importantes factores de planeación y seguridad como la realización de un "site survey" y la configuración de protocolos y métodos de autentificación deban de ser omitidos.

- Escalabilidad: Las redes inalámbricas pueden adaptarse rápidamente para dar servicio a una población creciente de usuarios. Para lograr esto, se requiere de una menor cantidad de equipos inalámbricos que de equipos alámbricos. La escalabilidad también debe de ser determinada previamente, ya que se necesita saber el número exacto de usuarios que se agregan a la red y el número de equipos que han de soportar a dichos usuarios así como las aplicaciones que se trabajen sobre esta plataforma.

- Costo: Aunque los equipos inalámbricos son un poco más costosos que sus contrapartes alámbricos, sus precios no dejan de estar al alcance del bolsillo del usuario hogareño. Los equipos de escala empresarial tienden a ser muy costosos, pero cuentan con funciones de seguridad y administración más robustas, están construidos para usarse a la intemperie y pueden soportar una mayor cantidad de usuarios.

Por otro lado, las redes inalámbricas también presentan claras desventajas propias de su misma naturaleza. Esas desventajas consisten en:

- Seguridad: Con la ayuda del equipo adecuado y conocimientos pertinentes al funcionamiento de una red inalámbrica, cualquier persona puede capturar información que viaja por el aire y que es producto de una transmisión inalámbrica. Solo es necesario posicionarse a una distancia relativamente cercana a la red, desde donde se pueda tener una recepción aceptable y contar con las herramientas necesarias para iniciar la decodificación de la información.

- Distancia: El área de cobertura de las redes inalámbricas de datos que se utilizan en la actualidad se considera en decenas de metros (tomando en cuenta que se habla de redes parte de la familia del estándar 802.11 y derivados). Es necesario adquirir e instalar accesorios y equipos como antenas y repetidores que amplíen el área de cobertura de las redes.

- Confiabilidad: La tecnología inalámbrica está sujeta a los efectos de la interferencia, el medio ambiente y los obstáculos del terreno. La gran mayoría de las veces, estos fenómenos son corregibles pero es un hecho innegable que su efecto produce resultados no deseados y que el rendimiento de la red se ve diezmado (en diferentes magnitudes).

- Velocidad: La velocidad de las redes inalámbricas es baja por naturaleza y no es comparable a la de las redes cableadas, las cuales pueden ofrecer velocidades y tasas de transferencia mucho mas elevadas. Aún cuando se pueden usar componentes para dar mayor potencia a la señal, esto no se traduce en alzas en las tasas de transferencia y velocidad, lo cual sigue siendo una desventaja mayor.

Al hacer un análisis de las desventajas, se puede concluir que la seguridad es sin duda la más importante de todas y la que origina la mayor cantidad de retos. Aunque la confiabilidad es un área donde se explota el diseño de antenas y otras áreas que son de extrema importancia, la seguridad toma una postura más dinámica debido a que no bien se ha solucionado un problema cuando ya existe una nueva manera de romper la seguridad propuesta por la solución emergente. Herramientas como AirCrack permiten a un usuario poder descifrar llaves de protocolos como WPA, los cuales eran considerados como seguros en extremo por parte de usuarios y fabricantes de equipos de redes inalámbricas.

Generalmente, la seguridad en redes inalámbricas es vista como un problema modular que es o puede ser resuelto mediante la integración de varias tecnologías. Esta "solución integral" debe de ser diseñada de acuerdo a las necesidades de seguridad de la organización en cuestión. Esto se debe a que diversas organizaciones requieren de diferentes niveles de seguridad o de privacidad dependiendo de la información que manejen. Es evidente que una universidad no tiene los mismos niveles de seguridad o la necesidad de confidencialidad que un banco o el ejército. Esta determinación de las necesidades de seguridad no solo dictará la pauta en cuanto a lo referente a los niveles y medidas de seguridad, sino al grado de complejidad y costo de la solución a implementar (tipo de soluciones de software y hardware).

La búsqueda de mecanismos de seguridad sólidos que permitan desplegar redes de seguridad es y será siempre una prioridad tanto para la industria como para los investigadores. Actualmente, el panorama de la seguridad inalámbrica ha enfocado sus esfuerzos hacia un marco de trabajo que incluye las siguientes áreas clave:

1. Seguridad General: El mundo inalámbrico difiere mucho de su contraparte alámbrico. Debido a que los dispositivos inalámbricos tienden a ser compartidos e interactúan con redes en el exterior, las contraseñas han dejado de ser un método efectivo de seguridad debido a que éstas se encuentran constantemente en contacto con medios externos peligrosos e inseguros. En este sentido, las redes inalámbricas se pueden beneficiar

de un método de autentificación bipartito, en el cual el usuario hace uso de un elemento que posee y otro que conoce. En este caso, un dispositivo USB, un token de autentificación o su propia huella digital (leída mediante un dispositivo biométrico) son elementos que posee, mientras que una contraseña es el elemento que conoce.

2. Encriptación: Las redes inalámbricas deben de poder soportar esquemas de encriptación que provean de ese servicio de manera sólida y a lo largo de todo el trayecto que el mensaje debe de recorrer para llegar al destinatario. Anteriormente, soluciones tales como el estándar WEP (Wired Equivalent Privacy ó Privacidad Equivalente a Alámbrico) y WPA (Wi-Fi Protected Access o Acceso Wi-Fi Protegido) fueron propuestos para remediar los problemas de seguridad que plagaban a las redes inalámbricas, pero eventualmente se tornaron vulnerables cuando los atacantes estudiaron su funcionamiento y publicaron herramientas que los inutilizaban. Actualmente, WPA2 se perfila como una solución viable de encriptación pero es evidente que con el tiempo, una nueva solución (o soluciones) será necesaria.

3. Firmas Digitales: Desde sus inicios, las redes inalámbricas han sido (y siguen siendo, hasta cierto punto) vulnerables a ataques que replican e insertan paquetes falsificados o alterados en las sesiones de transmisión con la finalidad de engañar a los usuarios y/o obtener información de ellos. Es evidente que esquemas como PKI, (Public Key Infrastructure o Infraestructura de Llave Pública) donde se puede asegurar la integridad y confiabilidad de los datos, se han hecho necesarios para evitar ataques tipo "man-in-the-middle" (hombre en medio), donde un tercero se encarga de "secuestrar" paquetes de una sesión inalámbrica para después alterar la información contenida en ellos para engañar al usuario que recibe los paquetes.

4. Interoperabilidad: Debido a que opciones como el uso exclusivo de WEP o WPA dejan huecos importantes de seguridad, es evidente que las soluciones de seguridad se constituyen de diferentes productos (los cuales utilizan, frecuentemente, tecnologías propietarias) de manera independiente que deben de interactuar uno con otro para remediar las diversas vulnerabilidades de una red. Frecuentemente existe la posibilidad de que, debido a la naturaleza propietaria de esas tecnologías, éstas no puedan interactuar una con la otra de manera adecuada. Este hecho

ha enfocado a los órganos rectores de estándares a guiar a la industria hacia el desarrollo de aplicaciones y soluciones que puedan operar entre si de manera nativa y sobre cualquier plataforma de redes IP.

3. Soluciones de Seguridad Nativas de 802.11

Uno de los factores mas importantes a tener en consideración al momento de implementar o trabajar con redes inalámbricas Wireless Lan, es sin duda el tema de la seguridad, concepto que se puede visualizar y abordar desde un amplio rango de puntos de vista. Podríamos tomar como base dos o tres practicas viables para el desempeño con seguridad básica y o incrementar esto a razón de la prevención y mitigación de puntos de falla y vulnerabilidades propios de este tipo de tecnologías wireless Lan.

Por inicio podemos trabajar con de la elaboración de las políticas de uso y procedimientos que se deben seguir al pie de la letra y que su implementación de pie a el desarrollo funcional de la tecnología así como para dejar la preparación de nuestra plataforma para recibir con facilidad posibles cambios, expansión y escalabilidad de la misma todo bajo un esquema de crecimiento programado.

Este tipo de practicas en muchas de las ocasiones son dejadas fuera al momento de la necesidad de implementar estas redes, o simplemente son consideradas para tiempos posteriores, cuando los sistemas comienzan a ser puntos de falla o puntos potenciales de ataques en ambos sentidos de nuestra red, que pueden en un momento hacer que toda una infraestructura de red, llámese guiada o inalámbrica colapse.

También es sumamente importante la revisión minuciosa y la emisión e implementación de estas políticas, que pueden proporcionar un esquema conocido, saludable y funcional que puede ser utilizado o debe ser utilizado en conjunto con una suite que integra una serie de protocolos y estándares, que incrementan la seguridad o la vuelven mas robusta en el mismo criterio además de darle capacidad de escalabilidad y por ende mas funcional a una plataforma con servicios de red con usuarios y aplicaciones inalámbricas.

3.1 Wired Equivalent Privacy (WEP)

Uno de los mecanismos de seguridad básicos al momento de manejar redes Wlan es el protocolo de encriptación WEP, acrónimo de Privacidad equivalente a cableado (***Wired Equivalent Privacy***). Es el mecanismo original de autenticación y encriptación especificado por el estándar 802.11. WEP se basa en un algoritmo

de encriptación simple, que utiliza un generador de números pseudoaleatorio (PRNG) y cifrado de flujo RC4. Este tipo de cifrado es rápido y eficiente al momento de encriptar y desencriptar lo cual minimiza el impacto en el rendimiento de la red.

Su implementación en equipos de punto de acceso es simple, lo cual lo coloco como un buen candidato para establecer seguridad cuando el estándar 802.11 estaba en sus inicios. En ese momento con la implementación de 802.11 el mecanismo de RC4 no expresa que sea criptográficamente fuerte como algunos otros algoritmos como AES, por mencionar alguno, pero fue asimilado como lo suficiente fuerte para 802.11 en sus inicios.

El proceso realizado para la encriptación o desencriptacion de los datos con el uso de WEP, implica que el algoritmo RC4 cree una secuencia de bits pseudoaleatoria conocida como *keystream*. La llave WEP es utilizada como una semilla para el algoritmo pseudoaleatorio, de esta manera alguna de dos estaciones que tienen la misma llave de WEP pueden generar el mismo keystream.

El proceso de trabajo de este esquema implica que la estación transmisora hace uso de la operación XOR con la información no encriptada conocido como *texto plano* y con el keystream para crear los datos encriptado conociendo esto como cifrado de texto (*ciphertext*). En el proceso correspondiente a la recepción del bloque o frame continua, el receptor (que tiene la misma llave de WEP que el transmisor) genera un "keystream", invierte el proceso XOR y recupera nuevamente los datos.

La función XOR es un algoritmo matemático muy simple, en el cual los datos pueden ser considerados encriptados desde el punto de vista que nadie puede generar el keystream correcto inicialmente, invertir el XOR y recuperar el texto plano todo ello sin conocimiento de la llave WEP, traduciéndose esto en un método básico de seguridad.

Dentro de las características operacionales generales de WEP, este nos proporciona un panorama claro para asumir criterio de su aplicación, que se torna alrededor de RC4, así tenemos que:

- WEP implementa el algoritmo de encriptación RC4, en el sentido que permite a los atacantes determinar cuales bloques o frames están encriptados, utilizando llaves matemáticamente débiles. Un análisis cuidadoso de estos frames puede permitir a un atacante aprender la llave de encriptación para el acceso a una red. Algunas aplicaciones como el Aircrack toman ventaja de de estas debilidades

- La aplicación del algoritmo RC4 requiere que un atacante no capture dos textos cifrados o "ciphertext" (bloques de datos encriptados) que fueron encriptados con la misma llave o "keystream", ya que cada bloque o frame es encriptado con un diferente keystream, y es posible que el numero de posibles keystream sea agotado en un corto periodo de tiempo, permitiendo que este evento ocurra. El tiempo específico necesario para que esto ocurra depende de que tan ocupada se encuentre la red en cuanto a tráfico se refiere (mas frames o bloques por segundo agotan los keystream rápidamente) y el numero de posibles llaves que estén disponibles.
  El numero o cantidad de estas llaves o keystream esta basado en la longitud de un campo denominado Vector de Inicialización (IV). En WEP el Vector de Inicialización es de 24 bits (3 bytes) de longitud resultando que hay 2 ala 24 posibles keystreams.

- WEP no protege contra continuos ataques. Un atacante puede capturar bloques o frames y entonces retransmitir el frame sobre la red un tiempo mas adelante. El frame será indistinguible del original.

- WEP no protege contra falsificaciones. Igual y sin saber la llave WEP, es posible para un atacante modificar arbitrariamente bits en un bloque o frame encriptado de manera que no sea detectado por las estaciones receptoras.

- 802.11 no proporciona de ninguna manera centralizada y automática de manejo de llaves WEP, estas deben ser configuradas manualmente en todas las estaciones, y si una llave es comprometida o descubierta, el cambiar a una nueva llave se torna ciertamente difícil mas si la red contiene una cantidad considerable de estaciones a las cuales tiene que afectar.

Por los aspectos descritos y las características de operación con las que esta integrada esta estrategia de encriptación, se puede lograr un desempeño funcional, apoyado en RC4 dado su increíble velocidad y simplicidad, en tanto la otra cara de la moneda es de este aspecto es la implicación de que se puede lograr la captura y modificación de la información originada por una estación origen y ser interpretada de manera distinta por la entidad receptora y en el mas estricto y critico de los casos que al capturar una cantidad suficiente de paquetes de información, cifrada bajo un esquema básico de encriptación y que es estático, no

solo se puede modificar la informaron si no que con un poco mas de tiempo y mas captura de secuencias de información adicional, se puede traducir la llave que es el elemento final para el acceso total a la red por parte de un intruso. WEP es por tanto un algoritmo confidencial criptográfico opcional que lo especifica IEEE 802.11 y es utilizado para proveer confidencialidad de datos que es subjetivamente equivalente a la confiabilidad de una red local cableada medio que no emplea técnicas criptográficas para realizar privacidad.

3.2 Marco de trabajo del estándar 802.1x

Como una respuesta a las debilidades encontradas en WEP, la IEEE y la Alianza Wi-Fi han trabajado en desarrollar mecanismos de seguridad inalámbrica. Como resultado de esto, un grupo de nuevos protocolos de seguridad han emergido, incluyendo Acceso Protegido Wi-Fi (WPA), WPA2, variantes operativos de 802.1x (PEAP, EAP-FAST, EAP-TLS, etc.) hasta el estándar 802.11i reciente.

La encriptación es una contramedida muy necesaria, pero no suficiente para la protección y el acceso a las redes inalámbricas, ya que no pueden impedir accesos no deseados a la red. Poco a poco se han intentado diversas soluciones en este ámbito pero luego se iría demostrando su vulnerabilidad, hasta que en 2001 el IEEE fija el estándar IEEE802.1X que se aplica a todas las redes con o sin cables y en 2004 es ratificado para empleo en redes inalámbricas WiFi lo cual establece un marco para la seguridad para estas redes.

El estándar IEEE 802.1X define el control de acceso a redes basadas en puertos. Gracias a él se exige autenticación antes de dar acceso a las redes Ethernet. En el control de acceso a redes basadas en puertos se utilizan los elementos físicos que componen una infraestructura de conmutación de la red LAN para autenticar los dispositivos agregados al puerto de conmutación. No se pueden enviar ni recibir tramas en un Puerto de conmutación Ethernet si el proceso de autenticación ha fallado. A pesar de que se diseñó para redes Ethernet fijas, este estándar se ha adaptado para su uso en redes LAN inalámbricas con IEEE 802.11. Windows XP soporta la autentificación IEEE 802.1X para todos los adaptadores de red basados en redes LAN, incluyendo las Ethernet y las inalámbricas.

El estándar IEEE 802.1X define los términos siguientes:

PAE

El Puerto PAE (Port access entity), también denominado Puerto LAN, es una entidad lógica que soporta el protocolo IEEE 802.1X asociado con un puerto. Un Puerto LAN puede hacer las veces de autenticador, el solicitante o ambos.

Autenticador

Es un Puerto LAN que exige autenticación antes de permitir el acceso a los Servicios que se suministran a través de él. Para las conexiones inalámbricas, el autenticador es el Puerto lógico de LAN en un punto de acceso (AP) inalámbrico a través del cual los clientes que trabajan con conexiones inalámbricas que operan con infraestructuras acceden a la red fija.

Puerto solicitante

El puerto solicitante es un Puerto de la LAN que solicita acceso a los servicios disponibles a través del autenticador. En las conexiones inalámbricas, el demandante es el Puerto lógico de la LAN alojado en el adaptador de red LAN inalámbrica que solicita acceso a una red fija. Para ello se asocia y después se autentifica con un autenticador.

Independientemente de que se utilicen para conexiones inalámbricas o en redes Ethernet fijas, los puertos solicitante y de autenticación están conectados a través de un segmento LAN punto a punto lógico y físico.

El servidor de autenticación

Para corroborar los credenciales del Puerto solicitante, el de autenticación utiliza el servidor de autenticación. Este servidor comprueba los credenciales del solicitante en nombre del Autenticador y después le responde a éste indicándole si el solicitante tiene o no permiso para acceder a los Servicios que proporciona el Autenticador. Hay dos tipos de servidor de autenticación:

Punto de acceso como un componente de autentificación

El dispositivo punto de acceso debe configurarse utilizando los credenciales de los clientes que intentan conectarse. Normalmente no se implementan utilizando puntos de acceso inalámbricos por diversos motivos de rendimiento y desempeño y se otorgan a equipos definidos para un optimo y especifico desempeño.

Una entidad distinta

El punto de acceso reenvía los credenciales de la conexión que ha intentado establecerse a un servidor de autenticación distinto. Por lo general un punto de acceso inalámbrico utiliza el protocolo de autenticación remota RADIUS (Remote Authentication Dial-In User Service) para enviar los parámetros de las conexiones que han intentado conectarse al servidor RADIUS.

Puertos de acceso sin y con autenticación

El control de acceso basado en el autenticador define los siguientes tipos de puertos lógicos que acceden a la LAN conectada físicamente a través de un solo puerto LAN fijo:

Puerto de acceso sin autenticación

El Puerto de acceso sin autenticación hace posible el intercambio de datos entre el autenticador (AP inalámbrico) y otros dispositivos dentro de la red fija, independientemente de que se haya autorizado o no al cliente la utilización de la conexión inalámbrica. Un ejemplo ilustrativo es el intercambio de mensajes RADIUS entre un punto de acceso inalámbrico y un servidor RADIUS alojado en una red fija que ofrece autenticación y autorización a las conexiones inalámbricas. Cuando un usuario de una conexión envía una trama, el punto de acceso inalámbrico nunca la reenvía a través del puerto de acceso sin autenticación.

Puerto de acceso con autenticación

Gracias al Puerto de acceso con autenticación se pueden intercambiar datos entre un usuario de una red inalámbrica y la red física pero sólo si el usuario de la red inalámbrica ha sido autenticado. Antes de la autenticación, el conmutador se abre y no se produce el reenvío entre el usuario de la conexión inalámbrica y el de la red física. Una vez que la identidad del usuario remoto se ha comprobado a través de IEEE 802.1X, se cierra el conmutador y las tramas son reenviadas entre el usuario de la red inalámbrica y los nodos de la red con conexión física.

Protocolo de autenticación extensible (EAP)

Para poder ofrecer un mecanismo de autenticación estándar para IEEE 802.1X, IEEE escogió el protocolo de autenticación extensible (EAP). EAP es un protocolo basado en la tecnología de autenticación del protocolo punto a punto (PPP)- que previamente se había adaptado para su uso en segmentos de redes LAN punto a punto. Para la autenticación de conexiones inalámbricas; Windows XP utiliza el protocolo EAP Seguridad del nivel de transporte (EAP-TLS). El protocolo EAP-TLS se define en las peticiones de comentario RFC 2716 y se utiliza en entornos seguros y certificados. El intercambio de mensajes EAP-TLS ofrece una autenticación mutua, unas transferencias cifradas totalmente protegidas y una determinación conjunta para las claves de cifrado y firma entre el cliente remoto y el servidor de autenticación (el servidor RADIUS). Una vez que se ha realizado la autenticación y autorización correspondiente, el servidor RADIUS envía las claves de cifrado y firma al punto de acceso inalámbrico a través de un mensaje de Acceso-aceptado de RADIUS.

Windows XP ha elegido EAP-TLS- que trabaja con usuarios registrados y certificados digitales en el equipo del usuario- como método de autenticación para sus conexiones inalámbricas por las siguientes razones:

EAP-TLS no se necesita la clave de la cuenta del usuario.
EAP-TLS la autenticación es automática, sin intervención del usuario.
EAP-TLS utiliza certificados por lo que el esquema es más consistente

Aunque para ello con una simple aplicación se tenga que requerir de conectarse a la red cableada para obtener el certificado como primera instancia, siendo una de las características mas inconvenientes y poco funcionales recalcando que esto trabajando bajo un esquema relativamente básico que tal vez pueda ser mejorado con la incorporación de aplicaciones adicionales simultaneas que en un momento impliquen un gasto adicional.

Básicamente la filosofía consiste en el control de los puertos de acceso, dado que no se abrirá el puerto ni se permitirá la conexión, hasta que el usuario este autentificado y autorizado contra una base de datos alojada en el servidor radius o interactuando con una base de datos adicional como un servicio de directorio

El estándar 802.1x constituye la columna vertebral de la seguridad WiFi y es imprescindible y muy recomendable su utilización en toda red empresarial que pretenda lograr una seguridad robusta, debido a esto el estándar introduce importantes cambios en el esquema de seguridad de WiFi.

3.3 WI-FI PROTECTED ACCESS (WPA)

WPA (Acceso Protegido Wi-Fi), es un mecanismo para protección de las redes inalámbricas (Wi-Fi); creado para corregir las deficiencias del sistema previo WEP (Wired Equivalent Privacy - Privacidad Equivalente a Cableado). Las investigaciones han evaluado y encontrado varias debilidades en el algoritmo WEP (tales como la reutilización del vector de inicialización (IV), del cual se derivan ataques estadísticos que permiten recuperar la clave WEP, entre otros). WPA implementa gran parte del estándar IEEE 802.11i, y fue creado como una medida intermedia o emergente para ocupar el lugar de WEP mientras 802.11i llegaba a su madurez finalmente. WPA fue creado por "The Wi-Fi Alliance" (La Alianza Wi-Fi),

WPA fue diseñado para operar en conjunto con un servidor de autenticación (generalmente un servidor RADIUS), que distribuye claves diferentes a cada usuario (a través del protocolo 802.1x); sin embargo, también se puede utilizar

en un modo menos seguro de clave precompartida ([PSK] - Pre-Shared Key) para usuarios de casa o pequeña oficina. La información es cifrada utilizando el algoritmo RC4 (debido a que WPA no elimina el proceso de cifrado WEP, sólo lo fortalece), con una clave de 128 bits y un vector de inicialización de 48 bits.

Una de las mejoras de WPA sobre WEP, es la implementación del Protocolo de Integridad de Clave Temporal (TKIP - *Temporal Key Integrity Protocol*), que cambia claves dinámicamente a medida que el sistema es utilizado. Cuando esto se combina con un vector de inicialización (IV) mucho más grande, evita los ataques de recuperación de clave (ataques estadísticos) a los que es susceptible WEP.

Adicionalmente a la autenticación y cifrado, WPA también mejora la integridad de la información cifrada. El chequeo de redundancia cíclica (CRC - *Cyclic Redundancy Check*) utilizado en WEP es inseguro, ya que es posible alterar la información y actualizar el CRC del mensaje sin conocer la clave WEP. WPA implementa un código de integridad del mensaje (MIC - *Message Integrity Code*). Además, WPA incluye protección contra ataques de "repetición" (replay attacks), ya que incluye un contador de tramas.

Al incrementar el tamaño de las claves, el número de llaves en uso, y al agregar un sistema de verificación de mensajes, WPA hace que la entrada no autorizada a redes inalámbricas sea mucho más difícil. El algoritmo MIC fue el más fuerte que los diseñadores de WPA pudieron crear, bajo la premisa de que debía funcionar en las tarjetas de red inalámbricas más viejas; sin embargo es susceptible a ataques. Para limitar este riesgo, las redes WPA se desconectan durante 60 segundos al detectar dos intentos de ataque durante 1 minuto.

3.3.1 Temporal Key Integrity Protocol (TKIP) and (AES)

Fue el primer protocolo que se utilizo para reparar los huecos de seguridad suscitados con el uso de WEP, desde su diseño se contemplo que TKIP no era la solución perfecta para la seguridad de 802.11 pero que proporcionaría un mejor desempeño que lo que ocurría con WEP, aunque TKIP a diferencia de WEP solo utiliza llaves de 128 bits no considera mucha diferencia de los 64 bits que utiliza WEP, pero esto no es todo dado que la mayoría de los ataques de 802.11 con encriptación WEP eran independientes de la longitud de la llave.

TKIP trata esta debilidad por medio del mezclado de llaves por paquete y la reintroducción automática. Con el mezclado de llaves por paquete, cada estación es asignada con una llave estática de WEP que es la misma para todas las estaciones, esta llave es conocida como una llave temporal. Cada estación entonces combina esta llave con sus seis bytes de dirección MAC para crear una llave de

encriptación que es única para cada estación. Esto incrementa esto crea un numero muy largo de keystreams disponible para un BSS, desde que cada estación efectivamente esta usando un a llave diferente de WEP. Para posteriormente incrementar el número de keystreams disponibles, TKIP usa un Vector de Inicialización IV de seis bytes en lugar de uno de tres bytes como es el caso de WEP.

La combinación de una llave temporal con la dirección MAC de la estación se conoce como la fase numero uno "la llave intermedia" esta llave se sigue procesando por medio d un algoritmo mezclador para producir la llave de encriptación para el bloque o frame. El propósito del algoritmo mezclador es para hacer mas difícil para un atacante el determinar si un frame o bloque fue encriptado con una llave matemáticamente débil de WEP. Estas llaves temporales pueden estar cambiando periódicamente cada 10,000 frames, pero el intervalo de regeneración de las llaves puede ser configurado por el administrador. Todo este esquema se apoya con la utilización de un código de integridad del mensaje (MIC - Message Integrity Code). Que garantiza aun ma la integridad del mensaje o descarta ENCASO de alguna variación derivada de un potencial ataque según sea el caso.

TKIP es usado en WPA y 802.11i así como también AES como un mecanismo de llaves temporales que apoya e incrementa la integridad de la información pero TKIP esta considerado para poder ínter operar con los sistemas de la generación anterior dejando espacio para que AES se perfile como una opción mas robusta y pero que requiere de mas poder de procesamiento. El hardware diseñado para RC4 simplemente no tiene el poder de procesamiento necesario para manejar la encriptación de AES. Por tanto a partir del 2002 se ha puesto especial atención en hardware para el soporte de esta aplicación así como para soportar 802.11i.

3.4 Wi-Fi Protected Access 2 (WPA2)

WPA2 está basada en el nuevo estándar 802.11i. WPA, por ser una versión previa, que se podría considerar de "migración", no incluye todas las características del IEEE 802.11i, mientras que WPA2 se puede inferir que es la versión certificada del estándar 802.11i. El estándar 802.11i fue ratificado en Junio de 2004. La alianza Wi-Fi llama a la versión de clave precompartida WPA-Personal y WPA2-Personal y a la versión con autenticación 802.1x/EAP como WPA-Enterprise y WPA2-Enterprise.

Los fabricantes comenzaron a producir la nueva generación de dispositivos y puntos de accesos apoyados en el protocolo WPA2, que utiliza el algoritmo de cifrado AES (Advanced Encryption Standard). Con este algoritmo será posible cumplir con mejores requerimientos de seguridad. "WPA2 está idealmente pen-

sado para empresas tanto del sector privado cómo del público. Si bien parte de las organizaciones estaban aguardando esta nueva generación de productos basados en AES es importante resaltar que los productos certificados para WPA siguen siendo seguros de acuerdo a lo establecido en el estándar 802.11i

WPA es actualmente lo más simple para implementar. Dentro de las más básicas implementaciones todos los usuarios tienen que configurar la llave precompartida para alcanzar más de los beneficios de seguridad del 802.1x/EAP y TKIP. WPA también es probablemente el más extenso, si solamente ha reemplazado a WEP como el mecanismo estándar de seguridad para consumidores y dispositivos de punto de acceso. 802.1x/EAP es mas común en instalaciones en empresas, desde que tienen los recursos de las IT (tecnologías de la información) necesarias para configurar y manejar la infraestructura de seguridad (instalación de certificados en las maquinas por ejemplo).

Instalaciones empresariales son mas probables de tener una infraestructura de seguridad existente en el cual 802.1X/EAP pueda integrarse. De hecho, variantes de EAP seguras son específicamente diseñadas con esta integración en mente. 802.11i se ratifico en 2004 pero no es aun extensamente adoptado en el mercado, parte de la razón de una lenta adopción puede deberse a que los consumidores perciben que la solución existente (WPA u 802.1x/EAP) s buena y suficiente y una ultima razón podría deberse a al tiempo que demoro en el proceso de ratificación y el ver productos en el mercado.

Hay que tener presente que 802.11i no es un punto y aparte. Se trata básicamente de una evolución de tecnologías anteriores, fundamentalmente de WPA (Wi-Fi Protected Access), implementado ya hace tiempo por la industria, hasta el punto de que el nuevo estándar todavía se conoce también como WPA2.

En el mundo de la seguridad, es una verdad innegable que el peor tipo de característica de seguridad es aquella que infunde una falsa sensación de seguridad. Al fin y al cabo, si se sabe que algo no es seguro, se pueden tomar medidas de protección adecuadas; si se cree que algo es seguro, no es probable que se le preste mucha atención. El cifrado en las redes inalámbricas entra lamentablemente en la categoría de medidas de seguridad que se dan por sentado: ofrecen una sensación de satisfacción, pero con resultados dudosos.

La norma 802.11i incluye la Norma de codificación avanzada (AES), que soporta claves de 128 bits, 192 bits y 256 bits. La AES es una forma de codificación más fuerte que se encuentra en la actual especificación de Acceso protegido de Wi-Fi (WPA) y es la norma de seguridad de las redes inalámbricas que tienen información del gobierno de los Estados Unidos. La especificación 802.11i garantiza que la información enviada por estas redes esté encriptada y no pueda ser dañada por alguien que la intercepte.

La diferencia entre WPA y WPA2 es la forma, WPA fue definido por la Wi-Fi alliance y WPA2 estandarizado por la IEEE en forma de la norma 802.11i. En el fondo se reemplazó el código de autenticación de mensajes que usaba WPA (llamado Algoritmo de Michael) por CCMP (Counter Mode with Cipher Block Chaining Message Authentication Code Protocol) y RC4 es sustituido por AES como algoritmo de cifrado.

4. Integración Tecnología y Mejores Prácticas

Como se mencionó en el Capítulo 1, la implementación de una red inalámbrica segura presupone la integración elementos que pueden ser clasificados bajos dos grupos:

- Soluciones de Seguridad Nativas de 802.11: Son aquellas soluciones que forman parte, de manera natural, de los estándares establecidos por el grupo de trabajo 802.11 de la IEEE: WEP, TKIP, 802.1x/EAP, WPA, WPA2

- Soluciones Independientes de 802.11: Comprenden todas aquellas tecnologías que son implementadas "encima" de las soluciones naturales de 802.11. Son, por lo general, tecnologías propietarias de terceros con diferentes grados de interoperabilidad: VPNs, servicios de directorio basados en LDAP, portales cautivos, configuraciones físicas y consideraciones de diseño lógico de red (estudios "site survey")

Una vez más, las soluciones finales dependerán de los requisitos propios de privacidad y confidencialidad producto del análisis previo de la red. La posibilidad de mitigar múltiples problemas con una sola herramienta es distante y muy probablemente inexistente. Debido a esto, se analizarán algunas de las soluciones independientes que hoy en día se combinan con mayor frecuencia con las soluciones nativas de 802.11.

4.1 VPNs

Hoy en día, las VPNs (Virtual Private Networks o Redes Privadas Virtuales) son una de las soluciones de seguridad más populares tanto en redes alámbricas como inalámbricas. Su alto nivel de confiabilidad y el manejo que hacen de la encriptación de la comunicación de emisor a receptor son solo algunas de las razones

por las cuales son escogidas por las organizaciones para proteger las comunicaciones que entran a ellas desde el exterior de sus redes.

Una VPN trabaja mediante un mecanismo sencillo. Básicamente, simula el hecho de estar trabajando en una red privada, de la misma manera como si la computadora en cuestión se encontrara dentro de la red de la organización. Esto se lleva a cabo mediante circuitos virtuales (llamados "túneles") los cuales proveen de una conexión segura hacia adentro de la red de la organización, siendo que la comunicación se origina desde afuera. Para establecer el túnel en una VPN inalámbrica, el cliente tiene 3 opciones:

1. Puede contactar a un servidor de VPNs montado sobre cualquier sistema operativo de red (Windows 2003 Server, Linux, UNIX) que esté ejecutando el rol o servicio de Servidor VPN. Esta opción tiene la ventaja de un bajo costo y además, utiliza la infraestructura de servidores existente de la organización donde se implemente.

2. Puede contactar a un concentrador de VPNs. Esta pieza de hardware es similar a un switch de red que se encarga exclusivamente de validar e iniciar túneles entrantes de VPN. La ventaja de un concentrador de VPN es que puede soportar grandes cantidades de conexiones, ya que dedica todo su poder de procesamiento a una sola tarea.

3. Puede contactar al mismo access point inalámbrico, si es que el access point cuenta con el servicio de Servidor VPN interconstruido. Esta opción es particularmente útil porque se resume el rol de un servidor de VPN en una misma pieza de hardware.

Las VPNs hacen modificaciones a los paquetes de información en la red para establecer que fueron originados por un cliente que soporta VPNs y que está involucrado en ellas. Después de iniciado el túnel de VPN, los paquetes se forman normalmente según el modelo OSI, donde cada capa agrega un encabezado con información de control propio. La diferencia radica en que una vez que la capa 2 (Enlace de Datos) termina de añadir su información, el paquete regresa a la capa 3 (Red) donde se le añade una extensión de información que contiene la información de encriptación (algoritmo de encriptación y llaves) y la distinción que se trata de un paquete de VPN. Al llegar al destino, el paquete es recibido por el componente que controla la VPN (servidor, concentrador o access point) y es analizado buscando el encabezado de capa 3 donde se indica que los contenidos

del paquete están encriptados y que provienen de una computadora que usa un túnel de VPN para comunicarse.

Gran parte de la seguridad que las VPNs proveen a las redes inalámbricas radica básicamente en las capacidades de autentificación y encriptación que se introducen a priori y a posteriori de una conexión inalámbrica.

VPN puede implementar un numero finito de algoritmos de encriptación, de entre los cuales, IPSec el más comúnmente utilizado. Esto se debe a que, por si solo, el protocolo PPTP (Point-to-Point Tunneling Protocol o Protocolo de Túnel Punto a Punto y protocolo de comunicación nativo de las VPN) no provee un método seguro de encriptación debido a que utiliza un algoritmo compartido secreto (o "shared secret") donde ambas partes de la comunicación conocen de antemano el algoritmo y la llave de encriptación, haciéndolo vulnerable en caso de que un atacante obtenga la llave de encriptación. En contraste a esto, IPSec puede asegurar el flujo de paquetes, proveer de autentificación mutua y establecer parámetros criptográficos mediante el uso de tres elementos básicos:

- Encabezados de Autentificación (Authentication Headers o AH): AH asegura la integridad de la información durante la transmisión y provee servicios de autentificación. Recientemente, AH ha dejado de ser soportado por varios fabricantes, pero sigue siendo usado debido a que puede encriptar partes de la información que ESP no puede.

- Carga Útil de Seguridad Encapsulada (Encapsulating Security Payload o ESP): ESP es el elemento clave para encriptación. Su función es la de proveer mecanismos que encripten la parte que representa la carga útil del paquete de datos mediante el uso de estándares de encriptación como AES y 3DES. Recientemente, las funciones de ESP se han extendido un poco hacia los servicios de autentificación, pero esa sigue siendo un área asignada a AH.

- Intercambio de Llaves de Internet (Internet Key Exchange o IKE): Este protocolo tiene la finalidad de negociar, crear y administrar elementos llamados Asociaciones de Seguridad (Security Associations o SAs). Las SAs definen las propiedades y niveles de seguridad que han de ser implementados en una conexión IPSec. Debido al manejo de las SAs, IKE puede proveer niveles variables de seguridad definidos por el usuario.

4.2 Servicios de Directorio

Los sistemas de red tienden a ser distribuidos debido a su naturaleza descentralizada y debido a que satisfacen la necesidad del acceso a la información desde diversos puntos. De igual forma, los elementos que conforman la red (usuarios, equipos, recursos) tienden a distribuirse de forma dinámica y pueden ser agrupados según diferentes criterios: de acuerdo a su rol, ubicación física, departamento, función, naturaleza o privilegios. Es por esto que, actualmente, los servicios de red tienden a agruparse bajo un servicio o aplicación que cuente con las siguientes características:

- Cuente con un repositorio de datos que permita contener a todos los elementos de la red
- Permita hacer distinciones según criterios establecidos entre los elementos que contiene
- Permita buscar rápidamente uno o varios elementos dentro del directorio
- Utilice un método universal y estandarizado para comunicarse con aplicaciones que deseen realizar acciones de búsqueda y/o administración del directorio

Los Servicios de Directorio conforman una aplicación o grupo de aplicaciones que proveen todos los servicios anteriormente mencionados. En los últimos 10 años, estos servicios han dictado la pauta respecto a la manera en que la industria diseña los sistemas operativos de red y las soluciones de administración de red, básicamente por la lógica con la que manejan y administran grupos de objetos (principalmente usuarios). Debido a esta lógica, los Servicios de Directorio son especialmente útiles cuando una organización debe de designar permisos de acceso de red en base a una jerarquía establecida.

Las redes inalámbricas se apoyan en los Servicios de Directorio para diseñar una estrategia de acceso "AAC" (llamada "AAA" en ingles debido a las palabras que representa). Esta estrategia de seguridad tiene como objetivo, proveer de tres funciones básicas que aumentan los niveles de seguridad:

- Autentificación (Authentication): En coordinación con la infraestructura inalámbrica, los Servicios de Directorio proveen una base de datos de usuarios de la red y sus contraseñas. Al momento que un usuario intenta validarse para obtener acceso a los servicios, la red inalámbrica coteja la información del usuario con la información almacenada en el repositorio. Si concuerda con lo almacenado en el directorio, el usuario obtiene acceso a los recursos. Para esto, se puede utilizar un servidor RADIUS (Remote Access Dial-In User Service o Servicio de Acceso Remoto de Usuarios Mediante Marcado) que puede almacenar la información de usuarios o hacer que ese servidor RADIUS se comunique directamente con el Sevicio de Directorio usando LDAP (Lightweight Diretory Access Protocol o Protocolo de Acceso Ligero a Directorio), el cual es un protocolo nativo de los Servicios de Directorio que puede ser usado por cualquier aplicación que lo soporte.

- Autorización (Autorization): Esta fase define los privilegios que el usuario autentificado tiene para usar la red después de haber entregado y validado sus credenciales. Los privilegios de usuario son otorgados en base a los que el administrador de red designe o en base a los privilegios otorgados según la lógica de agrupación (grupo, departamento, ubicación, etc.). Esta función es extremadamente útil debido a que limita las capacidades del usuario, previniendo que este lleve a cabo ataques, tenga acceso a información confidencial o que lleve a cabo actividades que degraden el rendimiento de la red.

- Contabilización (Accounting): Después de que el usuario se autentificó y se le fueron concedidos privilegios de acceso a la red, el último paso consiste en registrar que el usuario ha entrado a la red en bitácoras. Estos recursos no solo registran la entrada del usuario, sino que también pueden contabilizar y registrar el acceso y uso de recursos, archivos y otros rubros según se crea pertinente. De igual forma, la función de contabilización debe de poder apoyarse en aplicaciones que provean reportes detallados que puedan ser generados de manera automática y que contengan datos de utilidad para los administradores de red.

Los Servicios de Directorio pueden ser implementados en cualquier tipo de solución, sea ésta de software propietario o libre. Entre los servicios mas populares se incluyen Active Directory de Microsoft, Tivoli de IBM, Novell Directory Services de Novell (hoy eDirectory), OpenLDAP y Fedora Directory Services.

4.3 Portal Cautivo

Un Portal Cautivo es una aplicación particularmente útil cuando se pretende controlar la autentificación de los usuarios de una red inalámbrica, forzándolos a proporcionar credenciales mediante una pagina web (la cual hace las veces de cliente de autentificación, generalmente) desde donde se puede partir a otros pasos en el proceso de autentificación o desde donde se puede redireccionar al usuario para que comience su experiencia de trabajo en la red inalámbrica.

En esta aplicación, el access point al cual se intenta conectar una computadora detecta mediante un intento de conexión a la red inalámbrica que se está tratando de iniciar una sesión de red. El access point envía entonces una respuesta HTTPS (Página Web Segura) hacia esa computadora, desplegando una página web donde se le especifica al usuario que debe de introducir sus credenciales, aceptar las condiciones de uso de la red o ambas. De esta forma, el access point recibe la respuesta por parte del usuario y coteja las credenciales proporcionadas contra la base de datos de un servidor de autenticación RADIUS o bien, hace que el servidor RADIUS coteje la información ante un Servicio de Directorio.

Una vez que las credenciales han sido cotejadas y se ha reconocido al usuario, el Portal Cautivo puede tomar diferentes formas en cuestión de lo que ha de requerir que el usuario haga. Algunos de los ejemplos más comunes son:

1. Puede proporcionar al usuario información relevante mediante una página de inicio como noticias, avisos, notificaciones, ofertas u otra información relevante. Esta información puede contener las limitaciones de las credenciales de usuario y alertar sobre las consecuencias del mal uso de la red.

2. Requerir al usuario instalar software necesario para navegar en la red o para interactuar con otros servicios. Opción especialmente útil cuando los usuarios deben instalar certificados de seguridad para utilizar servicios de red o cuando se debe descargar e instalar clientes de software para establecer conexiones de VPN seguras.

3. En redes comerciales que requieren de un pago para permitir el uso de la red, el Portal Cautivo evita el acceso no autorizado y redirecciona al usuario a una página donde éste tenga la posibilidad de hacer su pago para después otorgarle el acceso.

Es conveniente que las aplicaciones de Portal Cautivo sean montadas en servidores dedicados a proveer dicho servicio. Aún cuando ciertos portales cautivos pueden ser montados sobre el access point mismo, es muy probable que la ejecución de dicho portal degrade el rendimiento del access point debido a que éste no fue diseñado para manejar altas cargas de procesamiento. De igual forma, siempre se debe de permitir el acceso a las páginas del Portal Cautivo en caso de que se cuente con software de filtrado de contenido, ya que de lo contrario, la solución se vuelve completamente inútil.

4.4 Site Survey y Consideraciones Básicas de Seguridad en Dispositivos Inalámbricos

Después de haber analizado los diferentes mecanismos de seguridad que se pueden implementar en una red inalámbrica en forma de software y hardware, muchos usuarios podrían considerar que el trabajo de planeación está completamente terminado. Esta suposición dista en demasía de la realidad, ya que el encargado de la implementación de una red inalámbrica puede estar dejando pasar algunas de las medidas más básicas de seguridad que pueden representar una amenaza cuyo impacto tiende a crecer una vez que son detectadas. Esas vulnerabilidades son aquellas relacionadas con el comportamiento de las ondas de radiofrecuencia, con la planeación física de la red y la configuración de los equipos que forman parte de la red.

Antes de implementar una red inalámbrica es necesario conocer el ambiente físico en el cual será implementada. Al conjunto de tareas de planeación y determinación por el cual un implementador conoce el comportamiento y cobertura de las ondas de radiofrecuencia, al igual que los fenómenos que las afectan directamente se le conoce como Site Survey (o Estudio de Sitio). Su objetivo primario es determinar patrones de radiación y áreas de cobertura que puedan considerarse óptimas para proveer a los usuarios con una red inalámbrica confiable.

Cuando se relaciona el estudio de Site Survey con la seguridad de las redes inalámbricas, las áreas prioritarias se concentran en la determinación de patrones de radiación y detección/prevención de interferencia:

- Determinación de Patrones de Radiación: Dependiendo de las antenas usadas por los access points, las potencias de transmisión y la polarización de dichas antenas, el área de cobertura de una red puede variar drásticamente. Si el patrón de radiación no es determinado con precisión, éste puede extenderse más allá del área de cobertura permitida de

manera que usuarios no autorizados estén en posibilidad de interceptar señales inalámbricas fuera de un edificio, desde los pisos superiores e inferiores o habitaciones contiguas. Las consecuencias de este fenómeno son equivalentes a instalar un puerto Ethernet alámbrico carente de vigilancia y restricciones, fuera del edificio.

- Detección de Fuentes de Interferencia: Esta solución intenta determinar las fuentes actuales y probables de interferencia de radiofrecuencia. Motores eléctricos, aires acondicionados no blindados, hornos de microondas y teléfonos inalámbricos representan un acérrimo enemigo en contra del buen desempeño de una red inalámbrica. De igual forma, es necesario detectar los posibles puntos desde donde un intruso pueda lanzar un ataque de interferencia de manera que los servicios sean deshabilitados por el menor tiempo posible y que el problema pueda ser resuelto cuanto antes.

Paralelamente, los instaladores de redes inalámbricas deben tomar en cuenta ciertas configuraciones físicas y lógicas que deben de ser modificadas en los equipos antes de la implementación de red:

- Cambiar o eliminar las configuraciones por defecto de fábrica: Los atacantes conocen de antemano los parámetros con los cuales los fabricantes configuran sus equipos. Si no se han hecho modificaciones a estos parámetros, un atacante puede adivinar contraseñas de administración de access points u otros equipos inalámbricos y reconfigurarlos sin autorización.

- Ocultar los identificadores de red SSID: Las redes inalámbricas utilizan un identificador llamado SSID (Service Set Identifier o Identificador de Set de Servicio) para distinguir una red de otra. Este identificador es transmitido de manera que todo mundo pueda verlo y así poder tener acceso a la red. Resulta muy útil deshabilitar la transmisión de este elemento en redes cuya presencia no se desea anunciar de manera que parezca que no existe del todo.

- Asegurar la instalación física de los equipos de red: Los equipos inalámbricos pueden ser montados bajo un "camuflaje urbano" que permita utilizar elementos como plafones, lámparas, cajas y difusores de luz para ocultar su presencia. Existe la opción de construir estos elementos a la

medida y especificaciones deseadas, aunque ciertos fabricantes ofrecen soluciones de este tipo. Igualmente, es importante montar los equipos de manera segura, de forma que ninguna persona pueda removerlos de su lugar sin autorización. Para esto, es necesario consultar las guías de instalación del producto para conocer las opciones de montaje y aseguramiento.

5. Conclusiones

Con los esquemas de seguridad mencionados en este trabajo de investigación, mismos que parten desde características de encriptación por medio de llaves estáticas, pasando por mejoras en llaves dinámicas de encriptación y soluciones de autentificación por medio de la interacción de plataformas de software abierto (o según sean las posibilidades) hasta soluciones mas robustas como VPN o portales cautivos, las preocupaciones que en materia de seguridad han perseguido a las redes LAN inalámbricas tanto en empresas como en redes domesticas en los últimos años, finalmente se están disipando. Podríamos concluir que se rompe con el paradigma de seguridad de estas redes inalámbricas en comparación con las redes cableadas debido al uso de los estándares actuales disponibles, organismos certificadores como la IEEE y la Alianza Wi-Fi y soluciones integradas, abiertas o propietarias relacionadas con el control de acceso y encriptación de la información que viaja por el aire.

Aun cuando las primeras versiones de las redes inalámbricas no fueron diseñadas con seguridad, la cantidad de métodos y soluciones de seguridad esta creciendo a pasos agigantados. Con el advenimiento de 802.1x y 802.11i, existen ya opciones seguras en cuanto a estándares de encriptación y autentificación. Estas características de seguridad emergentes deben de ser implementadas para poder asegurar la información en una red inalámbrica. Con una plantación cuidadosa y su debida diligencia, una red inalámbrica puede ser tan segura como una de sus contrapartes alámbricas. El factor humano es tan importante como el factor tecnológico.

Aunque el progreso en el desarrollo de la próxima generación de normas de seguridad WLAN es alentador, las empresas deben continuar tomando medidas para garantizar la eficacia de sus programas de seguridad WLAN. El trabajo futuro deberá encaminarse hacia la estandarización de la soluciones de seguridad, de manera que cualquier solución de seguridad pueda operar de manera integral con otras soluciones sin alterar la esencia de los estándares o protocolos.

BIBLIOGRAFIA

[1] Arbaugh, W. *"An Inductive Chosen Plaintext Attack Against WEP/WEP2"*. IEEE Document 802.11-02/230, May 2001; grouper.ieee.org/groups/802/11.

[2] Arbaugh, W., Housley, R. *"Security problems in 802.11-based networks"*. Commun. ACM 46, 5 (May 2003)

[3] Arbaugh, W.A., Shankar, N., Wan, Y.C., *"Your 80211 wireless network has no clothes"*. IEEE Wireless Communications, vol. 9, Dec. 2002, pp 44 - 51.

[4] Bhagyavati, Summers, W. C., and DeJoie, A. *"Wireless security techniques: an overview"*. Proceedings of InfoSecCD '04: 1st Annual Conference on information Security Curriculum Development. 2004. ACM Press, New York, NY, 82-87.

[5] Borisov, N., Goldberg, I., Wagner, D. *"Intercepting mobile communications: The insecurity of 802.11"*. Proceedings of the International Conference on Mobile Computing and Networking, (July 2001), 180–189.

[6] *"Choosing a Strategy for Wireless LAN Security"*. Microsoft Corporation. Consultado el 12 Junio 2007 en http://www.microsoft.com/technet/security/guidance/cryptographyetc/peap_int.mspx

[7] *"Cisco 4400 Series Wireless LAN Controllers"*. Cisco Systems, Inc. Consultado el 30 Julio 2007 en http://www.cisco.com/en/US/customer/products/ps6366/index.html. (Sitio protegido. Requiere contraseña de acceso).

[8] Dworkin, M. *"Recommendation for Block Cipher Modes of Operation: Methods and Techniques"*. NIST Special Publication 800-38A, Dec. 2001.

[9] LaRocca, J., LaRocca, R., *"802.11: Demystified"*. New York: McGraw-Hill, 2002.

[10] Mehta, P.C. *"Wired Equivalent Privacy Vulnerability"*, LevelOne Security Essentials Track, April 2001

[11] Federal Trade Commission; 16 CFR Part 314; *"Standards for Safeguarding Customer Information"*; Final Rule, 2002, May 23. Federal Register, 67(100), 36484-36494. Consultado el 25 Mayo 2007 en http://www.ftc.gov/os/2002/05/67fr36585.pdf

[12] Ferguson, M.N. *"An improved MIC for 802.11 WEP"*. IEEE 802.11 doc 02-020r0, Jan. 17, 2002; en http://grouper.ieee.org/groups/802/11.

[13] Fluhrer, S., Mantin, I., Shamir, A. *"Attacks on RC4 and WEP"*. RSA Laboratories, Cryptobytes, vol. 5, no. 2, Summer/Fall 2002

[14] Fluhrer, S., Mantin, A., Shamir,A. *"Weaknesses in the key scheduling algorithm of RC4"*. Proceedings of the 8th Workshop on Selected Areas in Cryptography, LNCS 2259. Springer-Verlag, 2001.

[15] Giller, R., Bulliard, A. *"Security Protocols and Applications 2004: Wired Equivalent Privacy"*, Swiss Institute of Technology, Lausanne, Mar. 3, 2004

[16] IEEE Std 802.11, *"Standards for Local and Metropolitan Area Networks: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications"*, 1999.

[17] National Institute of Standards and Technology. *"Wireless Network Security. 802.11, Bluetooth and Handheld Devices"*. Consultado en http://csrc.nist.gov/publications/nistpubs/800-48/NIST_SP_800-48.pdf el 12 de Agosto de 2007.

[18] Peikari, C. Forgie, S. *"Cracking WEP"* (2003) Consultado el 26 Julio 2007 en http://www.airscanner.com/publications.html#articles.

[19] Planet3 Wireless. *"CWNA Certified Wireless Network Administrator Official Study Guide"*. 3ª Edicion. McGraw-Hill/Osborne. 2005.

[20] *"Seguridad en LAN inalámbricas con PEAP y contraseñas"*. Microsoft Corporation. Consultado el 5 Julio 2007 en http://www.microsoft.com/latam/technet/seguridad/guidance/lan/peap_1.mspx#ERE.

[21] *"Secure Wireless Access Point Configuration"*. Microsoft Corporation. Consultado el 5 Julio 2007 en http://www.microsoft.com/technet/security/midsizebusiness/default.mspx.

[22] Wesiman, C. *"The Essential Guide to RF and Wireless"*. 2da. Edición. Prentice Hall PTR. 2002.

3. Evaluación de Rendimiento de Protocolo MAC Basado en CDMA para Redes Ad-Hoc.

Carlos Enrique Portes Flores

1. Introduccion

1.1 Protocolo de Control de Acceso al Medio

Una red ad-hoc consiste en un grupo de nodos inalámbricos que están equipados con transmisores y receptores (transceptores). No existe una infraestructura cableada para soportar la comunicación entre estos nodos móviles, y cada nodo en la red es, tanto un ruteador como un punto de comunicaciones. Usualmente, cada nodo tiene la posibilidad de comunicarse con los otros nodos cuando todos ellos están esparcidos a lo largo de un área geográfica. Sin embargo, los nodos pueden esparcirse a través de rangos geográficos que son mayores que lo que pueden alcanzar las señales de comunicaciones. En este caso, los nodos pueden comunicarse a través de varios saltos, pero existe solo un medio que es compartido por todos los nodos que están en el mismo rango de comunicaciones y el ancho de banda de radio frecuencia es limitado. Más aún, las colisiones de paquetes son inevitables debido al hecho que el tráfico entrante es aleatorio y existe un tiempo de propagación "non-zero" entre transmisores y receptores. Por ende, se utilizan esquemas de Control de Acceso al Medio (MAC) para coordinar el acceso al canal en la red [1].

Debido a que no existe una autoridad centralizada que asigne radiofrecuencias específicas, time-slots o códigos a diferentes nodos que están totalmente distribuídos, las terminales tienen que contender ellas mismas por el acceso al medio. Por ende, los protocolos de Control de Acceso al Medio juegan un papel importante en el desempeño de la red móvil ad-hoc. Un protocolo MAC define como cada unidad móvil puede compartir el recurso limitado de ancho de banda inalámbrico de una manera eficiente.

La tendencia de las redes ad-hoc inalámbricas es usar sistemas adaptivos para ajustar los parámetros de transmisión. El objetivo es maximizar el rendimiento del caudal eficaz (throughput) en uso del canal. Adicionalmente, el bajo rendimiento en la región de tráfico bajo se da debido a que no hay más información que mandar y no debido a errores en la interferencia multiusuario. Podemos decir que el rendimiento del sistema está limitado por la técnica de acceso usada en redes ad-hoc.

La investigación de tasas de transmisión variable para redes ad-hoc es limitada. En [3], se presenta un protocolo MAC de tasa adaptiva que emplea la modificación de Acceso Múltiple por Detección de Portadora Evitando Colisión (CSMA/CA). Recientemente, las técnicas de transmisión adaptivas basadas en CDMA (Acceso Múltiple con División de Código) han sido investigadas para el mejoramiento del rendimiento en redes inalámbricas ad-hoc [4-5] y [6], pero estos esquemas solo manejan una y dos tasas de transmisión, respectivamente. Más aún, algunos trabajos tales como [4] usan ciertos umbrales para obtener la adaptabilidad del tráfico de red.

Por lo tanto, para poder mejorar el rendimiento (throughput) de una red inalámbrica ad-hoc, se expone un protocolo MAC basado en CDMA con un esquema de tasa adaptiva. En este esquema se controla la tasa de transmisión de acuerdo al tráfico ofrecido, y la interferencia de acceso múltiple es minimizada por medio del manejo de la ganancia de procesado.

En este trabajo se hace un estudio y evaluación del desempeño del protocolo MAC basado en CSMA-CDMA para redes inalámbricas ad-hoc. Logrando la maximización en el caudal eficaz con diferente tasas de transmisión.

2. TIPOS DE PROTOCOLOS DE CONTROL DE ACCESO AL MEDIO

Existe una amplia gama de protocolos inalámbricos de Control de Acceso al Medio. Jurdak et al. [12] proponen criterios de clasificación para protocolos de Control de Acceso al medio basados en características operativas consideradas como esenciales. Este criterio de clasificación toma en cuenta:

- Separación de canales y métodos de acceso: Se le considera de gran importancia debido a que este factor determina el número de usuarios que el protocolo puede soportar de manera efectiva. Se dividen en:
  - Protocolos de Canal Sencillo (Single Channel)
  - Protocolos de Canales Múltiples (Multiple Channel)

- Topología: Este factor toma en cuenta la visión general que el protocolo hace de la red en sí. Su importancia radica en la separación jerárquica de la red y la efectividad de dicha separación en relación a la administración de las funciones de red. Se divide en:

    - Topología Plana de Salto Sencillo (Single-hop Flat Topology): No se encargan de relevar información. Presuponen la existencia de redes con hosts de capacidades similares
    - Topología Plana de Saltos Múltiples (Multiple Hop Flat Topology): Presuponen la existencia de nodos de capacidades similares, pero consideran la red como una entidad mucho más escalable y expandible.
    - Topología de Cluster (Cluster Topology): Topología que agrupa estaciones de trabajo en unidades lógicas y designa una de ellas como cabeza de cluster. La cabeza de cluster se encarga de llevar a cabo funciones específicas de control y administración.
    - Topología Centralizada (Centralizad Topology): Se introduce el concepto de una terminal base que pueda encargarse de proveer información de transimisión y asignación de canales. El concepto en sí contradice las especificaciones de las redes ad-hoc, las cuales no necesitan de una terminal base para considerarse como infraestructura.

- Método de inicio de transmisión: El método utilizado por la entidad que da inicio a la transmisión en la red depende en gran parte de las aplicaciones usadas en la red misma. El enfoque de inicio puede ser:

    - De iniciación por emisor
    - De iniciación por receptor

- Eficiencia de consumo de batería: Este criterio es de gran importancia debido a la cantidad de tiempo que un host debe de trabajar en base a los requerimientos de protocolo, aplicación o carga de red. Los criterios a considerar para la eficiencia del manejo de la carga de baterías incluyen:

    - Control de Potencia del Transmisor: Se puede utilizar ya sea solo la potencia necesaria para llegar a los receptores en la red o niveles ilimitados de potencia.

- Modalidad de Hibernación: Métodos utilizados para la administración de la actividad de un host cuando éste entra en periodos prolongados de inactividad.
- Consideración del Nivel de Batería: Cambios en los roles de los hosts (administración y control) de una red basados en los niveles de potencia restantes en las baterías y el consumo de las mismas.
- Sobreflujo de Control Reducido: Fluctuaciones en los niveles de envío y recepción de información de control en base a los niveles de energía de los hosts.

- Carga de tráfico y escalabilidad: Consideraciones relativas a la capacidad de un protocolo para soportar diversos aspectos del tráfico y crecimiento de una red:

  - Carga Elevada: Redes que manejan altos volúmenes de tráfico sin importar el número de hosts.
  - Alta densidad: Enfoque orientado a redes donde el crecimiento en el número de hosts es la preocupación central.
  - Trafico de tiempo real: Tratamiento específico del tráfico de aplicaciones que utilizan esquemas de tiempo real y sus características inherentes, tales como voz y video.
  - Escenarios Selectivos: Se encarga de aspectos de red específicos tales como: redes con paquetes de tamaño anormal, aplicaciones no mencionadas anteriormente, etc.

- Alcance: Dependiendo del alcance, un protocolo se puede considerar como:

  - Muy Corto Alcance
  - Corto Alcance
  - Mediano Alcance
  - Largo Alcance

Protocol	Published	Channel	Topology	Trans. Initiation	Power efficient	Traffic load and scalability	Range
1. CSMA [4]	1975	Single	Single/Flat	Sender	No	Wired networks	Medium
2. BTMA [5]	1975	1 control/1 data	Centralized	Sender	No	Hidden terminal	Long
3. PRMA [6]	1988	Hybrid	Centralized	Sender	No	Voice	V. short
4. MACA [7]	1990	Single	Single/Flat	Sender	No	Hidden terminal	N/A
5. MACAW [8]	1994	Single	Centralized	Sender	No	Delivery guarantee	Medium
6. FAMA [9]	1995	Single	Single/Flat	Sender	No	Delivery guarantee	Medium
7. IEEE 802.11 [1]	1996	Multiple (FHSS/DSSS)	Single/Flat	Sender	No	Access point	Medium
8. HIPERLAN [2]	1996	Multiple (hybrid)	Clustered	Sender	Yes	Data relay	Short
9. MACA-BI [10]	1997	Single	Multiple/Flat	Receiver	No	Predictable traffic	Long
10. FPRP [11]	1998	Multiple (time)	Multiple/Flat	Sender	No	Voice	N/A
11. PAMAS [12]	1998	1 control/1 data	Multiple/Flat	Sender	Yes	Dense low load	Medium
12. Bluetooth [3]	1999	Multiple (FHSS)	Clustered	Master	Yes	Low rate PAN	V. short
13. Markowski [13]	1999	Multiple (time)	Single/Flat	N/A	Yes	Voice	N/A
14. HRMA [14]	1999	Hybrid	Multiple/Flat	Sender	No	Large packets	N/A
15. MCSMA [15]	1999	Multiple (frequency)	Single/Flat	Sender	No	High density	Medium
16. PS-DCC [16]	1999	Single	Single/Flat	Sender	Yes	High load	Medium
17. RIMA-SP [17]	1999	Single	Single/Flat	Receiver	No	Predictable traffic	N/A
18. ADAPT [18]	1999	Multiple (time)	Multiple/Flat	Sender	No	High load	Medium
19. CATA [19]	1999	Multiple (time)	Multiple/Flat	Sender	No	Low load	Medium
20. Jin [20]	2000	Hybrid	Clustered	Sender	Yes	Heterogeneous	N/A
21. MARCH [21]	2000	Single	Multiple/Flat	Sender	Implicit	Homogeneous	V. short
22. RICH-DP [22]	2000	Multiple (FHSS)	Multiple/Flat	Receiver	No	High load	Long
23. SRMA/PA [23]	2000	Multiple (time)	Multiple/Flat	Sender	Yes	Voice	N/A
24. DCA-PC [24]	2001	1 control/N data	Multiple/Flat	Sender	Yes	High density	Short
25. GPC [25]	2001	Single	Clustered	N/A	Yes	High density	N/A
26. VBS [26]	2001	N/A	Clustered	N/A	No	Voice	N/A
27. DPC/ALP [27]	2002	Single	Multiple/Flat	Sender	Yes	Heterogenous	Long
28. Lal [28]	2002	Multiple (space)	Multiple/Flat	Receiver	Implicit	High load/Density	Medium
29. GRID-B [29]	2002	1 control/N data	Multiple/Flat	Sender	No	High load/Density	Medium
30. MC MAC [30]	2002	Multiple (CDMA)	Multiple/Flat	Sender	No	High rate PAN	V. short
31. WCA [31]	2002	N/A	Clustered	N/A	Yes	Heterogeneous	N/A
32. DBTMA [32]	2002	2 control/1 data	Multiple/Flat	Sender	No	Hidden terminal	Short
33. MMAC [33]	2002	Multiple (space)	Multiple/Flat	Sender	Yes	High load	Medium
34. D-PRMA [34]	2002	Multiple (time)	Single/Flat	Sender	No	Voice	Medium

Tabla 1. Protocolos inalámbricos de Control de Acceso al Medio. **Fuente**: [12]

3. PROTOCOLO MAC BASADO EN CDMA

En este capítulo, se analiza el protocolo MAC usando CSMA/CA y un esquema de tasa de transmisión adaptivo basado en CDMA.

El mecanismo básico de IEEE 802.11 falla al resolver el bien conocido problema del terminal oculto [7], donde dos nodos que no se escuchan entre ellos mismos transmiten paquetes a un receptor común. La figura 1a ilustra este problema donde el nodo A sensa el canal buscando que esté desocupado mientras transmite un paquete a B. Pero el canal está ocupado en B debido a una comunicación con C. Debido a esto, el paquete de A no es recibido en B (sufre una colisión). Otro problema es la terminal expuesta. Como podemos ver en la figura 1b, el nodo A está transmitiendo un paquete de datos al nodo B. Debido a que C está fuera del rango del nodo B y el nodo D está fuera del rango del nodo A, el nodo C puede ahora transmitir paquetes de datos a D. Sin embargo, como el nodo C está en el rango del nodo A, el nodo C no puede transmitir paquetes de datos hasta que hasta después de que escuche la transmisión del nodo A.

La solución a estos problemas es usar el proceso de "*handshaking*". En este proceso, siempre que un paquete vaya a ser transmitido, el nodo transmisor envía primero un paquete de petición para mandar (request-to-send o RTS). Si el nodo receptor escucha el RTS, responde con un paquete corto de autorizado para mandar (clear-to-send o CTS). Después de este intercambio, el nodo transmisor transmite un paquete ACK. Los paquetes RTS/CTS incluyen la información acerca de cuánto tardará transmitir el siguiente paquete. Así, otros nodos cerca del nodo transmisor y el nodo oculto cercano al nodo receptor no iniciarán una transmisión durante este periodo. El proceso de "handshaking" presupone un reto más que debe ser resuelto por este mismo protocolo: el congestionamiento relacionado con el envío de múltiples mensajes de control. La adopción de un mecanismo de control de envío similar a la utilizada dentro del Protocolo de Acceso Múltiple con Handshake Reducido (MARCH) se encarga de la reducción de sobreflujo y procesamiento relacionados con dichos mensajes. MARCH emplea un mecanismo de envío reducido de mensajes RTS y CTS mediante un esquema de árbol con una propagación hacía los vecinos más cercanos. Una vez que un nodo emite un mensaje de RTS al momento de habilitarse como nodo de envío, el nodo receptor siguiente manda un mensaje de CTS al nodo que originó el mensaje de CTS (autorizándole el envío de información) y a su nodo más cercano. Subsecuentemente, éste último nodo enviará mensajes de CTS a sus nodos receptores más cercanos, indicándoles que el primer nodo utilizará el medio de transmisión. De ésta manera, se reduce el intercambio de mensajes RTS y CTS entre todos los nodos participantes hasta en un 50%, ya que solo se envía un solo mensaje RTS por parte del nodo que enviará información.

En el protocolo descrito previamente, CSMA/CA, los nodos transmiten a la misma tasa y se presenta el problema de congestión en la red. Por ende, se puede utilizar dos alternativas para poder transmitir con diferentes tasas y obtener adaptabilidad en el tráfico de la red. Estas mejoras en el rendimiento (throughput) de la red ad-hoc se obtienen mediante el uso de CDMA.

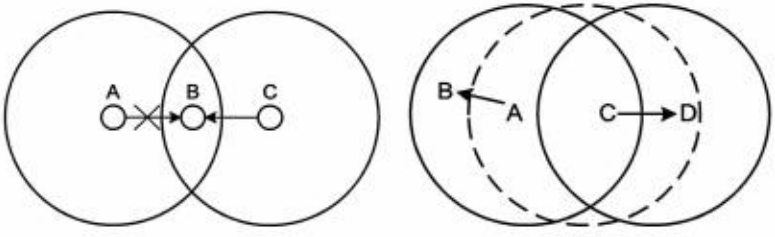


Fig. 1. Problemas de protocolo CSMA/CA, a) terminal oculta, b) terminal expuesta.

En el esquema basado en CDMA, cuando un nodo transmite un paquete, manda un RTS a todos los nodos adyacentes en un código común. El nodo receptor manda un CTS y comienza la transmisión en un código basado en el receptor. En este proceso de "handshaking" existen fallas debido a la congestión de red. A medida que aparece la falla en el proceso de "handshaking", el nodo pasa a un estado de *backoff* y debe entonces retransmitir la información. Esta falla parece por dos razones: 1) Colisiones en el proceso de "handshaking", 2) La interferencia (MAI) causada debido a los nodos activos que están transmitiendo a otros nodos. Entonces, usando CDMA y manejando la ganancia de procesado dinámicamente, podemos transmitir a diferentes tasas y obtener la adaptabilidad del tráfico de la red.

En CDMA, el ancho de banda es constante, es dado por

$$B=1/T_c, \tag{1}$$

donde T_c es el chip time.

Si (1) es multiplicado y dividido entre la duración del bit (T_b), obtenemos

$$B = \frac{1}{T_C} = \frac{T_b}{T_C}\frac{1}{T_b} \rightarrow \text{constant.} \tag{2}$$

Observamos que el ancho de banda se constituye de dos términos: T_b/T_c es la ganancia de procesamiento (P_G) y $1/T_b$ es la tasa de transmisión en bps. Por lo tanto, el ancho de banda puede ser expresado en términos de la ganancia de procesado y la tasa de transmisión

$$B = \frac{1}{T_c} = \frac{T_b}{T_c}\frac{1}{T_b} = P_G R_b \text{ es constante.} \tag{3}$$

Si la tasa de transmisión es fija, tendrá un cierto nivel de protección contra la interferencia. En caso de carga baja en el canal, esto se debe a que hay menos nodos adyacentes transmitiendo simultáneamente y el nivel de interferencia será bajo. Por ende, una ganancia de procesamiento alta no será necesaria. En tales circunstancias, se puede reducir la ganancia de procesado e incrementar la tasa de transmisión en la misma proporción, de manera que la información que fluye por unidad de tiempo en el canal es mayor y consecuentemente el retardo puede ser reducido. Por otro lado, sería un sistema con diversas tasas de transmisión

y diferentes niveles de protección a interferencia y se pueden desarrollar esquemas para seleccionar tasas de transmisión óptimas como una función de las condiciones de la carga del canal. Previamente, mencionamos que el problema de "handshaking" es debido a la colisión o interferencia MAI (Interferencia de acceso múltiple), así los dos esquemas en el manejo de tasas son los siguientes:

I) El protocolo MAC utiliza un grupo de ganancias de procesado durante el proceso de "handshaking". En este paso, el nodo monitorea el estado del canal. Si hay retransmisiones, esto se debe a una falla en el "handshaking" y debido a la aparición de paquetes de colisión RTS/CTS. Adicionalmente, existe un nivel alto de tráfico y debido a eso se debe de disminuir la tasa de transmisión a la mitad. Mientras el nodo transmite exitosamente varias veces, se deduce entonces que el tráfico en el canal es bajo y que la tasa se incrementa al doble.

II) Este algoritmo se maneja para cada nodo, el cual decide la tasa de transmisión a usar. Cuando un nodo necesita transmitir, calcula el número de nodos adyacentes (n) que intentan transmitir en un time spot dado. Podemos alcanzar el rendimiento de red ad-hoc óptimo por medio de variación de ganancia de procesado dinámica. La implementación exitosa de este procedimiento asegura que el sistema CDMA funcionará apropiadamente dentro de las variaciones de tráfico del canal. Sin embargo, es necesario lograr una combinación optima de bit rate para obtener, bajo la presencia simultánea de n nodos adyacentes, el rendimiento de procesamiento del sistema CDMA. Entonces, el SNR del nodo m con tasa r_m se calcula como[8]

$$SNR_m = Q\left(\left(\frac{n_0}{2E_b} + \frac{1}{3G_{p_j}}\left(\sum_{i=1}^{n}\frac{r_i}{r_j}n_i - 1\right)\right)^{-1/2}\right), \tag{4}$$

donde ni es el número de nodos en el sistema, r_0 es el bit rate básico, $G_{pi}=B/r_i$ es la ganancia de procesado de cada nodo, B es el ancho de banda. Adicionalmente, los bit rates son ordenados como $r_1>r_2>\ldots>r_n$ con las ganancias de procesado correspondientes $G_{pi}=B/r_i$. Todos los bit rates son múltiplos del más bajo bit rate r_n, por consiguiente $r_n=r_0$, $r_{n-1}=(n\text{-}1)r_0$, $r_{n-2}=(n\text{-}2)r_0,\ldots, r_1=zr_0$.

Con este resultado, la mejor combinación de tasas de transmisión se obtiene después de una búsqueda exhaustiva [9],

$$\max_{(n_{r_1}, n_{r_2}, \ldots, n_{r_m})} S(n_{r_1}, n_{r_2}, \ldots, n_{r_m}) \quad \text{subject to } n_{r_1}, n_{r_2}, \ldots, n_{r_m} = n, \tag{5}$$

donde n_{r1} es el número de nodos transmitiendo a zr bps, n_{r2} es el número usando $(z\text{-}1)r$ bps, y n_{rm} es el número a r bps (donde $r=r_0$ rate básico).

Los resultados de la evaluación de rendimiento del protocolo MAC basado son presentados en el siguiente capítulo.

4. SIMULACIÓN

El proceso de simulación fue originado en lenguaje C, en una computadora personal, que considera una red inalámbrica ad-hoc. En la simulación se condujeron 80 ciclos de simulación. Adicionalmente, los parámetros para las simulaciones están basados en propiedades de CDMA. Dado esto, la información recolectada de estas simulaciones constituye el rendimiento (throughput).

Se consideraron 250 nodos que se localizan de manera aleatoria en un área cuadrada que mide 1500m de lado. En lo respectivo al tráfico de datos, se usó el modelo presentado en [10], donde la aplicación de servicio de datos sigue una distribución de Pareto con valores fijos de α=1.1 y k=81.5. Adicionalmente, todos los nodos transmiten con la misma potencia. Otros parámetros usados en las simulaciones se muestran en la Tabla I.

Tabla I. Parámetros de Simulación.

Parámetro	*Valor*
Chip rate	4.096 Mcps
Tamaño Paquete de Datos	506 Bytes
Máxima Ganancia de Procesado	256
Mínima Ganancia de Procesado	8

Se presenta el rendimiento del sistema entre los esquemas con tasas adaptivas (esquema I y II) y por diferentes tasas de transmisión (tasas fijas). Se consideraron seis tasas (r_0, $2r_0$, $4r_0$, $8r_0$, $16r_0$ y $32r_0$, donde υ=1, 2, 4,..., 32) y transmisión a una razón de υr_0 permite una ganancia de procesamiento de P_G/υ.

En la Figura 2, se observa que con una tasa básica (r_0) el rendimiento de procesamiento es bajo en la región de tráfico bajo, lo cual se debe a que hay pocos nodos transmitiendo. Si incrementamos al doble la tasa de transmisión entonces la tasa de transmisión entonces la ganancia de procesamiento es reducida a la mitad. Esto es posible debido a que una alta ganancia de procesado y no es conveniente transmitir con una tasa baja cuando hay pocos nodos. Adicionalmente, mientras las tasas de transmisión se incrementan, se ilustra un decremento en el rendimiento. Esto es evidente porque la probabilidad correcta de paquetes disminuye [11]. Es por eso que es necesario manejar las tasas de transmisión dinámicamente para obtener adaptabilidad en el tráfico de red. El comportamiento ilustrado en la figura 2 muestra un mejor rendimiento en el esquema I con respecto a las tasas fijas. Esta mejora se debe a que el esquema I decide cuando incrementar o disminuir la tasa de transferencia en función del éxito o falla del proceso de "handshaking".

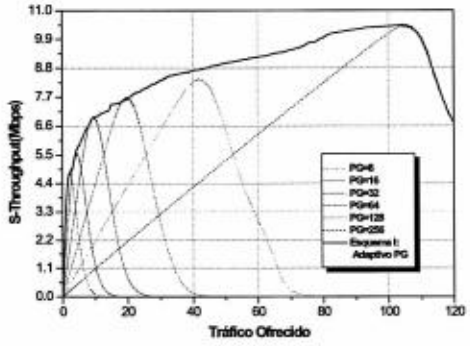


Fig. 2. Comportamiento de rendimiento de procesamiento (throughput) considerando el esquema adaptivo de tasa – I.

La Figura 3 ilustra la respuesta del esquema II. Se observa que el rendimiento de procesamiento mejora substancialmente con respecto al esquema-I y a las tasas fijas. Esta mejora se debe a la combinación óptima de tasas de transmisión. Este esquema es diferente a [4] debido a que utiliza umbrales y no es posible maximizar el rendimiento de procesamiento. Se concluye que el protocolo MAC propuesto (esquema I y II) es una alternativa para mejorar el rendimiento de redes inalámbricas ad-hoc basadas en CDMA.

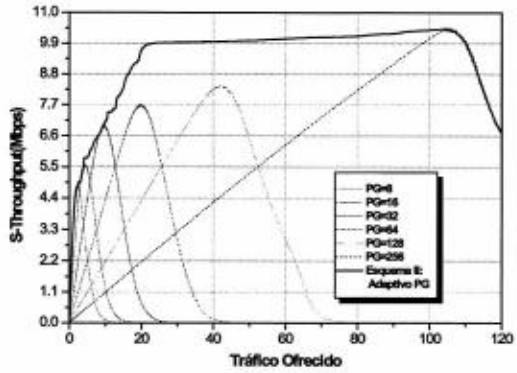


Fig. 3. Comportamiento de rendimiento de procesamiento (throughput) considerando el esquema adaptivo de tasa – II.

5. CONCLUSIÓN

Se ha analizado un protocolo de control de acceso al medio basado en CDMA. Los resultados muestran que nuestro esquema basado en CDMA mejora el rendimiento (throughput) comparado con tasas fijas. Los dos esquemas manejan tasas dinámicas como una función de las colisiones en el proceso de "handshaking" y las interferencias que producen los nodos adyacentes activos. Adicionalmente, al usar estos dos esquemas se obtiene adaptabilidad del tráfico de red.

En este trabajo los nodos usan un control perfecto de potencia, que se puede ampliar considerando que los nodos usan distintas potencias para la transmisión. Por consiguiente es necesario usar un control de potencia y esto nos llevaría a un ahorro de energía.

REFERENCIAS

[1] S. Kumar, S., Raghavan, V.S., Deng, J.: Medium Access Control Protocols for Ad-Hoc Wireless Networks: A Survey. Elsevier Ad-Hoc Networks Journal 4(2006) 326-358.

[2] Bao, J.Q., Tong, L.: A performance comparison between ad hoc and centrally controlled CDMA wireless LANs. IEEE Transactions on Wireless Communications 1(2002) 829-841.

[3] Djonin, D.N., Karmokar, K., Djonin, D.V., Bhargava, V.K.: Adaptive Rate Transmission in Ad-Hoc Wireless Networks. IEICE Transactions on Communications 85(2004) 1385-1392.

[4] Fantacci, R., Ferri, A., Tarchi, D.: Medium Access Control for CDMA Ad-Hoc Networks. Electronics Letters 40(2004) 1131-1132.

[5] Muqattash, A., Krunkz, M., Ryan, W.E.: Solving the Near-Far Problem in CDMA-Based Ad-Hoc Network. Ad Hoc Networks Journal 1(2003) 435-453.

[6] Su, T-S., Jeng, W-L., Hseih, W-S.: Enhancing the Performance of Ad-Hoc Wireless Using Code Division. Journal of Internet Technology 7(2006) 285-292.

[7] Kleinrock, L., Tobagi, F. A.: Packet Switching in Radio Channels: Part II - the Hidden Terminal Problem in Carrier Sense Multiple Access and the Busy Tone Solution. IEEE Transactions on Communications 23(1975) 1417–1433.

[8] Ottosson, T., Svensson, A.: On schemes for multirate support in DS/CDMA. Journal on Wireless Personal Communications 6(1998) 265–287.

[9] Méndez, A.: Contributions to the Medium Access Techniques in Third Generation Mobile Communication Systems in a DS-CDMA Environment (in Spanish), PhD Thesis, CICESE, Mexico (2003).

[10] ETSI TR 101 112, UMTS Selection Procedure for the Choice of Radio Transmission Technologies of the UMTS (UMTS 30.03 version 3.1.0), Technical Report, European Telecommunications Standard Institute (1997).

[11] Pursley, M.B.: Performance Evaluation for Phase-Coded Spread-Spectrum Multiple-Access Communications – Part I: System Analysis. IEEE Transactions on Communications 25(1977) pp.795-799.

[12] Jurdak, R. Videira Lopes, C. Balde, P. "A Survey, Classification and Comparative Análisis of Medium Access Control Protocols for Ad-hoc Networks". IEEE Communications Surverys and Tutorials Magazine, Vol. 6, #1, 2004.

4. Gestión de Recursos para Redes 3G Considerando Preferencia.

Mariby Lucio Castillo, Aldo Luis Mendez Perez

1. Introduccion

1.1 Gestión de recursos (*scheduling*)

Las políticas para programar el orden de transmisión para tráfico multimedia y la gestión de recursos, tiene un gran impacto en la eficiencia y prestaciones de protocolos de control de acceso al medio (MAC) en las redes 3G. Varios criterios pueden ser usados para el diseño de un gestor eficiente, por ejemplo, maximizar el caudal eficaz, minimizar los paquetes perdidos, mantener la calidad de servicio (QoS) y asignar recursos de acuerdo a una estructura de prioridad pre-definida. Varias disciplinas de gestión de recursos han sido propuestas para garantizar las prestaciones de servicio en redes alambicas, pero proveer QoS en redes móviles es más complejo que en redes fijas, debido a la movilidad y deterioros en el canal inalámbrico, que es altamente dependiente del interfaz aire.

Con el algoritmo de gestión de recursos se puede controlar, de manera efectiva, el retardo experimentado por un paquete desde que éste se genera, hasta que se recibe con éxito. Esto es debido a que maneja cierta prioridad en la transmisión de un paquete, estableciendo de esta manera cierto retardo de un paquete a otro. Además, con este algoritmo se puede establecer, el orden en el cual se llevará a cabo la transmisión de los paquetes que se encuentran en su memoria temporal de los terminales móviles (TMs).

Por lo tanto, es necesario nuevos algoritmos de gestión de recursos para redes móviles de tercera generación. Estos deben establecer equidad en la asignación de los recursos, ser flexible en el manejo del ancho de banda, adaptable a las condiciones de tráfico multimedia, y garantizar la calidad de servicio de acuerdo al estándar UMTS (Sistema Universal de Comunicaciones Móviles).

2. COMUNICACIONES MÓVILES DE TERCERA GENERACIÓN

La introducción del sistema GSM en Europa desde principios de la década de los noventa, con la incorporación de la tecnología digital, la posibilidad de utilizar un sistema normalizado en todos los países y la consiguiente reducción de precios que posibilitan un mercado más amplio y un marco más competitivo, ha puesto una auténtica revolución en la sociedad. La Tabla I muestra las diferentes fases en la implantación del sistema GSM, donde la primera de ellas corresponde básicamente a un entorno poco competitivo y con baja penetración de usuarios, en la que el operador realiza únicamente servicio de voz [Sallent, 1999]. A medida que el sistema gana aceptación y aumenta la competencia, las estrategias de los operadores se dirigen a un mayor uso de los recursos disponibles mediante la prestación de nuevos servicios. Llegados a este punto quedan todavía algunas potencialidades del sistema GSM por explotar como la transmisión de datos de manera más intensiva que permita descargar y ejecutar aplicaciones tipo telebancaria, consulta de cotizaciones mediante la tarjeta inteligente (*smartcard*) llamada SIM, siguiendo una arquitectura cliente-servidor.

Tabla I. Evolución típica del mercado y la competencia en GSM hacia 3G.

Medio	Impulsor de Mercado	Fase	Acción	Penetración
GSM Fase 1	Voz	Inicial	Atraer Usuarios	Poca
GSM Fase 2	Servicio (buzón de voz, prepago)	Sofisticación	- Diferenciación - Aumentar tráfico	10-30%
GSM Fase 2+	- Datos - Internet	Maduración y límites tecnológicos	- Aumentar tráfico - Desarrollo y aplicaciones de datos - Preparar siguiente paso	30- 40%
3G	Contenidos multimedia	Nueva era	Desarrollar mercado	40-50%

No obstante, la provisión de servicios más sofisticados topa definitivamente con las limitaciones tecnológicas del propio sistema GSM, un sistema que empezó a tomar forma en 1992 y que si bien se diseñó con perspectivas de futuro permitiendo un entorno multi-operador, no deja de estar básicamente optimizado para transmisión de voz. Por lo tanto, parece claro que es necesario ampliar las capacidades de los sistemas de comunicaciones móviles para satisfacer, a mediano plazo las nuevas necesidades, dando lugar a los que se conoce como sis-

temas de comunicaciones móviles de tercera generación o sistemas 3G. Algunos aspectos que se consideran importantes para la creación de una nueva generación de sistemas de comunicaciones móviles se plantean a continuación.

1) El intenso crecimiento de la penetración en el mercado de los sistemas de segunda generación, ha provocado que en algunos países el número de terminales móviles haya superado el número de líneas fijas de telefonía. A manera de ejemplo la Figura 1 muestra el incremento de la telefonía móvil en los últimos años en México, donde el crecimiento hace que se prevea a mediano plazo una saturación en las capacidades de los sistemas de segunda generación. Esto hace patente la búsqueda de nuevas bandas de frecuencia y de nuevos sistemas que permitan hacer frente a la demanda con un uso más eficiente de los recursos [COFETEL, 2003].

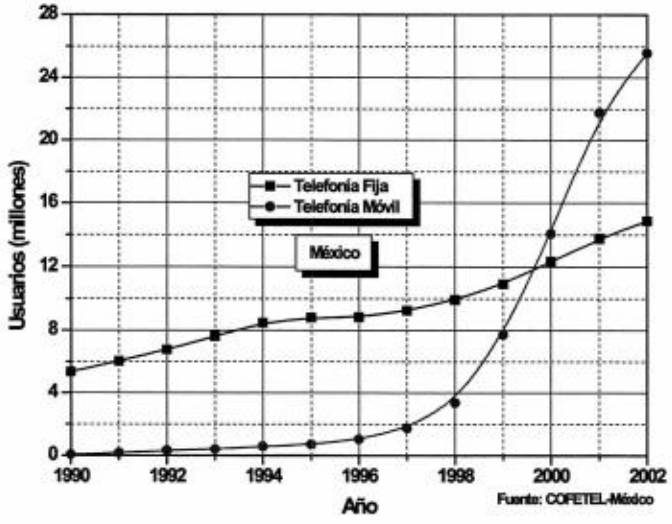


Figura 1. Crecimiento de la telefonía móvil en México.

En la Figura 2 se muestra, a modo de ilustración, las previsiones de crecimiento en los usuarios de telefonía móvil para la próxima década [UMTS, 1999], donde se observa como prácticamente se prevé doblar el número de usuarios cada 5 años, especialmente en lo que a la región Asia-Pacífico se refiere, motivo por el que, de hecho, es Japón uno de los países que lideren el camino hacia los sistemas de tercera generación.

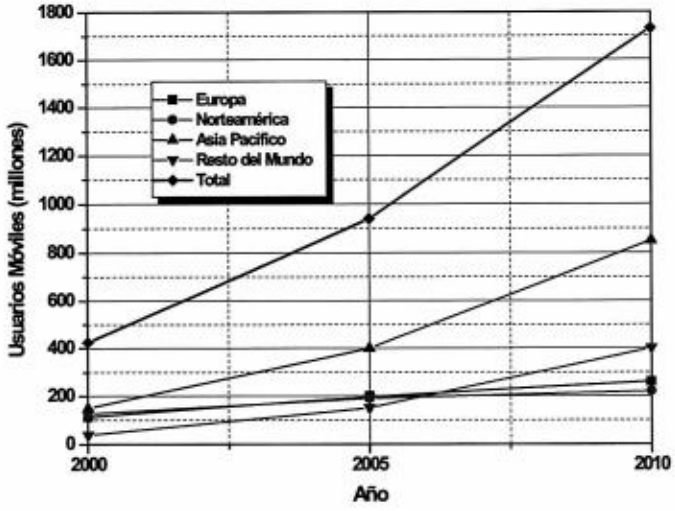


Figura 2. Predicción de crecimiento de la telefonía móvil.

2) Otro aspecto importante, es el crecimiento del acceso a Internet y de la evolución del mercado de las comunicaciones móviles. Es de esperar que la combinación de ambos en un acceso a Internet desde redes móviles, pueda suponer un enorme mercado potencial de cara a los próximos años [UMTS, 2002].

Si bien los sistemas celulares de segunda generación como GSM son capaces de ofrecer acceso a Internet, lo hacen en modo circuito, lo que presenta enormes limitaciones no sólo en términos de la velocidad de transmisión empleada sino también de la eficiencia en el uso de los recursos, pues este tipo de aplicaciones se caracterizan por generar la información a ráfagas con lo que durante buena parte del tiempo el circuito no es utilizado. Estas limitaciones redundan por un lado en una reducida capacidad para ofrecer este tipo de servicios, y por el otro en un precio de conexión para los usuarios mucho más elevado del que se puede llegar a ofrecer en una red fija.

En consecuencia, es deseable el diseño de nuevos sistemas que sean capaces de hacer frente a estas limitaciones con un uso más efectivo de los recursos, capaces de adaptarse a las nuevas características del tráfico mediante técnicas de transmisión orientadas a paquetes, constituyendo éste uno de los retos al que los sistemas de tercera generación deberán hacer frente.

3) La llamada sociedad de la información demandará cada vez más disponer de servicios como correo electrónico, acceso a redes corporativas, acceso a Internet, videoconferencia, comercio electrónico, multimedia y muchos otros.

Además, el usuario deseará mantenerse informado cuando se desplaza de un sitio a otro, disfrutando de los citados servicios en cualquier lugar y en cualquier momento. Para poder proporcionar la suficiente calidad de servicio en comunicaciones multimedia y acceso a Internet, claramente se necesitan velocidades de transmisión elevadas y sistemas de radio frecuencia (RF) que presenten una elevada eficiencia espectral.

La Figura 3 pone de manifiesto el mercado potencial que este tipo de servicios supone en el ámbito de las comunicaciones móviles, presentando las perspectivas de crecimiento del número de abonados totales a sistemas móviles en la Unión Europea y diferenciando aquéllos que únicamente emplearán servicios de voz o de datos de baja velocidad [UMTS, 1998]. En el 2005, habrá 200 millones de usuarios, de los cuales 32 millones usando servicios de multimedia. Para el año 2010, habrá 260 millones de usuarios y 90 millones utilizarán servicio de multimedia.

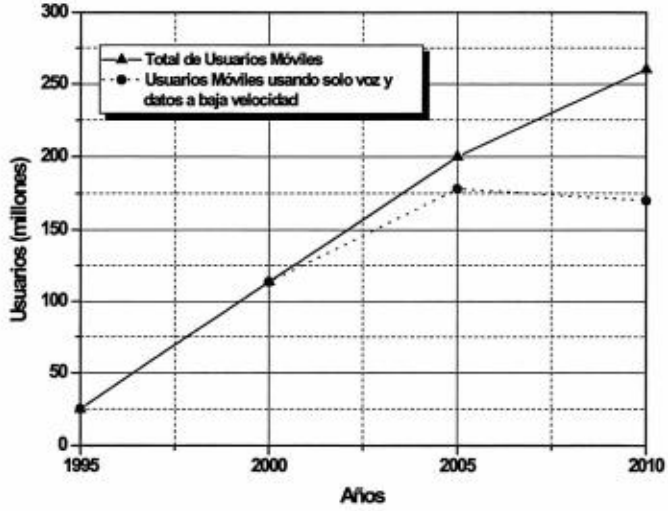


Figura 3. Previsiones de uso de servicios multimedia de alta capacidad para Europa.

Como se observa en la Figura 3, es a partir del año 2000 cuando empieza a existir un cierto porcentaje de usuarios (en Europa) con servicios de alta velocidad, con previsiones del 16 % para el año 2005 y del 30 % para el año 2010. Se prevé incluso que el número de usuarios móviles con requerimientos únicamente de voz o baja velocidad llegue a decrecer en tanto que los usuarios tenderán a aprovechar al completo las facilidades que puedan ofrecer los nuevos sistemas.

Por otra parte, y puesto que muchas aplicaciones multimedia están orientadas a trabajar en modo paquete, es esencial optimizar las técnicas de la tercera generación para soportar transmisión por paquetes así como velocidades de transmisión variable. De esta forma, los mismos recursos pueden ser compartidos por mucho usuarios, aprovechando la naturaleza de este tipo de comunicaciones para mejorar la eficiencia en su utilización.

4) Por último, reseñar también el interés en lograr un sistema de comunicaciones móviles verdaderamente global que permita una movilidad universal con operación entre redes pertenecientes a países diferentes, llegando incluso a reunir bajo un sistema común las tres zonas geográficas de mayor influencia que son Europa, Estados Unidos y Japón.

El ámbito de actuación de los sistemas de tercera generación se pretende englobar los diferentes entornos existentes bajo un único sistema, en función de la cobertura ofrecida, desde los sistemas vía satélite hasta los más reducidos entornos de interiores, con objeto de permitir una movilidad universal de terminales capaces de soportar aplicaciones personalizadas de muy variada naturaleza.

Para asegurar el éxito de los servicios 3G, se ha de proporcionar a los usuarios con comunicaciones muy eficientes, con una alta velocidad y calidad y, además, fáciles de utilizar. Por lo tanto, los sistemas de 3G deben ofrecer [Castro, 2001]:

- Transmisión simétrica/asimétrica de alta fiabilidad.
- Uso de ancho de banda dinámico, en función de la aplicación.
- Velocidades binarias mucho más altas: 144 kbps en alta movilidad, 384 kbps en espacios abiertos y 2 Mbps en baja movilidad.
- Soporte tanto de conmutación de paquetes (IP) como de circuitos.
- Soporte IP para acceso a Internet (navegación WWW), videojuegos, comercio electrónico, y vídeo y audio en tiempo real.
- Diferentes servicios simultáneos en una sola conexión.
- Calidad de voz igual a la ofrecida en la red fija.
- Soporte radioeléctrico flexible, con utilización más eficaz del espectro, con bandas de frecuencia comunes en todo el mundo.
- Personalización de los servicios, según perfil de usuario.
- Servicios dependientes de la posición (localización) del usuario.
- Incorporación gradual en coexistencia con los sistemas actuales de 2G.
- Itinerancia (*roaming*), incluido el internacional, entre diferentes operadores y tipos de redes.
- Ambientes de funcionamiento marítimo, terrestre y aeronáutico.
- Capacidad de terminales multibanda y multientorno.

- Economías de escala y un estándar global y abierto que cubra las necesidades de un mercado de masas.
- Provisión de un ambiente local virtual (VHE): el usuario podrá recibir el mismo servicio independiente de su ubicación geográfica.

Por lo anterior, ante la entrada de los sistemas móviles de tercera generación, el control de acceso al medio así como la gestión de recursos son dos aspectos importantes en el diseño de estos sistemas. Por lo cual, es necesario proponer estrategias para obtener:

- alta eficiencia,
- manejo de tráfico multimedia,
- flexibilidad en el ancho de banda,
- mayor capacidad en función del número de usuarios,
- velocidades de transmisión variable,
- tasas de error aceptable,
- control de admisión y
- asignación equitativa de recursos.

Estas estrategias nos deben llevar a proponer algoritmos con el fin de que el sistema sea tanto adaptable a las condiciones del tráfico como equitativo en la gestión de recursos (ancho de banda y códigos) para servicios multimedia, y además garantizando la calidad de servicio.

La investigación a nivel mundial sobre estos algoritmos, ha adquirido una relevancia muy significativa en los últimos años, debido fundamentalmente a la puesta en marcha experimental de los primeros sistemas comerciales de 3G.

Dentro de la parte que nos interesa investigar está la gestión de recursos, donde un gran número de estos han sido propuestos para redes alambicas [Zhang, 1995]. Sin embargo, estos algoritmos de gestión de recursos no pueden ser aplicados directamente a redes inalámbricas debido a las restricciones técnicas del interfaz aire. Por otra parte, los algoritmos actuales que poseen una adaptabilidad a las condiciones del tráfico como los propuestos en [Kim D. et al., 2001], [Vannithamby, 2000] solo manejan tráfico de datos, no garantizan calidad de servicio, no aplican algún control de admisión y tampoco hacen gestión de recursos. En el caso de [Sallent y Agustí, 2000] maneja solamente tráfico integrado de voz y datos, en cambio [Sandouk *et al.*, 1999] garantiza calidad de servicio. Otros trabajos de investigación que manejan tráfico integrado de voz/datos y además llevan a cabo una gestión de recursos son propuestos en [Kim J. *et al.*, 2001] y [Kang *et al.*, 2000].

Por lo anterior, es necesario el estudio de nuevas propuestas de sistemas adaptativos capaces de ajustar los parámetros de transmisión a las necesidades concretas de la información a transmitir. Además deben llevar a cabo una asignación equitativa de recursos, siempre con el objetivo de maximizar la eficiencia en el uso del canal y garantizando la calidad de servicio en un tráfico multimedia, ya que esto es de gran importancia por su aplicabilidad inmediata en el diseño de nuevos sistemas de comunicaciones móviles.

A continuación se analizará algunos algoritmos que se usan para la gestión de recursos para de redes móviles celulares.

3. ANÁLISIS DE ALGORITMOS DE GESTIÓN DE RECURSOS

En años recientes ha habido un gran crecimiento en la gestión de redes móviles. Con el uso creciente de redes móviles e inalámbricas en ambientes, ya sea para interiores, como para exteriores, ha aparecido el problema de proporcionar equidad en el acceso entre múltiples TMs, que contienden sobre un canal inalámbrico escaso y compartido. En redes alambicas, la asignación equitativa de recursos (*fair scheduling*) ha sido durante mucho tiempo un paradigma para proporcionar equidad en el enlace de acceso. Sin embargo, la adaptación de asignación equitativa de recursos a redes inalámbricas no es trivial debido a los problemas únicos en canales inalámbricos, tales como dependencia de localización, aleatoriedad de los errores y la contención por el canal. Por consiguiente, los algoritmos de gestión de recursos propuestos para redes alambicas no se aplican directamente a las redes inalámbricas. A continuación se mencionan algunos algoritmos de gestión de recursos para redes TDMA y CDMA [Fattah y Leung, 2002].

3.1 ALGORITMOS PARA REDES TDMA

En una red del tipo TDMA, solo una sesión puede transmitir en cualquier tiempo dado, se define el modelo de red como un sistema celular que consta de una estación base (EB) y un número de terminales móviles (TMs), la gestión del recurso es implementada en la EB, el intercambio de paquetes entre la EB y un TM esta caracterizado por la presencia de errores debido al canal. Se sabe que el estado del canal y el estado de los paquetes en la cola para todas las sesiones están disponibles en la EB, algunos algoritmos de gestión de recursos que funcionan de acuerdo a las condiciones anteriormente planteadas son las que a continuación se mencionan:

- Gestión de recursos de paquetes dependiente del estado del canal (Channel State Dependent Packet Scheduling (CSDPS)).
- Gestión de recursos de cola imparcial idealizada para sistemas inalámbricos (Idealized Gíreles Fair Queuing (IWFQ)).
- Gestión de recursos de cola imparcial independiente de la condición del canal (Channel-Condition-Independent Fair Queuing (CIF-Q))
- Gestión de recursos de aproximación imparcial basada en el servidor (Server-Based Fairness Approach (SBFA)).
- *Scheduling* de servicio imparcial para sistemas inalambricos (Wireless Fair Service (WFS)).

3.1.1 GESTIÓN DE RECURSOS DE PAQUETES DEPENDIENTE DEL ESTADO DEL CANAL

La idea principal de este tipo de gestión de recursos es impedir que se presenten errores en ráfaga en la capa de enlace en lugar de confiar en que la capa de transporte o la de aplicación se encarguen de la recuperación de los errores, otra característica importante es que el estado del canal es monitoreado para cada sesión, y si este se encuentra en un mal estado, entonces la transmisión del paquete es aplazada.

3.1.2 GESTIÓN DE RECURSOS DE COLA IMPARCIAL IDEALIZADA

Es implementado con un mecanismo de compensación para errores propensos en una sesión, cada sesión tiene una etiqueta de servicio que es mantenida el tiempo virtual final de su paquete Head-Of-Line (HOL). Si la sesión no tiene mucho trabajo atrasado, la etiqueta de servicio es puesta en ∞ .

3.1.3 GESTIÓN DE RECURSOS DE COLA IMPARCIAL INDEPENDIENTE DE LA CONDICIÓN DEL CANAL

Utiliza el algoritmo Start Time Fair Queuing (STFQ) como su modelo de servicio para la liberación de errores, debido a que establece que es más fácil basar la gestión de recurso en el tiempo de inicio que en el tiempo de término.

Start Time Fair Queuing: El algoritmo Start Time Fair Queuing reduce en gran manera el cálculo complejo que conlleva el algoritmo WFQ, evitando la necesidad de simular el flujo del servidor en tiempo real. El tiempo virtual en SFTQ se deriva de la etiqueta de inicio del paquete en servicio, otra ventaja de este método es que el STFQ es aplicable a tasas variables de servidores ya que no es necesario tomar en cuenta la tasa del servidor en el tiempo virtual del cálculo.

El precio a pagar por esta simplicidad radica en que las garantías de retardo se incrementan con el número de flujos.

3.1.4 GESTIÓN DE RECURSOS DE APROXIMACIÓN IMPARCIAL BASADA EN EL SERVIDOR

Es un sistema que puede acomodar cualquier scheduler de línea cableada como su modelo de servicio de liberación de errores, en SBFA, una porción del ancho de banda es reservada para una sesión hipotética llamada Long-Term Fairness Server (LTFS), la cual es utilizada para compensar las sesiones retrasadas. Cuando una sesión es seleccionada para transmitir, se le permite hacerlo siempre y cuando el canal se encuentre en buen estado, de otra manera, se crea una ranura (slot) para esta sesión y es colocada en la cola dentro de la sesión LTFS, siendo seleccionada en su lugar una sesión con un buen comportamiento del canal para ser transmitida.

3.2 GESTIÓN DE RECURSOS EN REDES CDMA

CDMA provee ciertas ventajas sobre TDMA y FDMA, como mayor capacidad en el sistema, handoff suaves, una planeación de frecuencia sencilla, e inherente diversidad de frecuencia contra desvanecimientos multitrayectoria, entre otras ventajas.
En los sistemas CDMA, los paquetes enviados por un número de TMs pueden ser recibidos simultáneamente por la EB (siempre y cuando tenga un adecuado control de potencia). Algunos gestores de recursos (schedulers) para el uso en redes CDMA son los que a continuación se mencionan:

- Gestión de recursos de paquete por paquete GPS (Packet by Packet GPS (PGPS)).
- Gestión de recursos de CDMA calendarizado (Scheduled CDMA (SCDMA)).
- Gestión de recurso dinámico (Dynamic Resource Scheduling (DRS)).
- Protocolo de control de acceso multimedia para sistemas inalámbricos con gestión de recursos basado en BER (Wireless Multimedia Access Control Protocol with BER Scheduling (WISPER)).
- Gestión de recursos en redes multisaltos (Scheduling in Multihop Networks).

3.2.1 GESTIÓN DE RECURSOS DE PAQUETE POR PAQUETE GPS

El servidor PGPS en CDMA es del tipo conservador del trabajo y opera entre diferentes sesiones para garantizar las tasas de transmisión y los índices de potencia. PGPS tiene un problema en cuanto al índice de potencia residual porque se asume que cada sesión debe tener un índice de potencia con valor fijo. Cuando un número de paquetes de diferentes sesiones son seleccionados para ser transmitidos, la suma de sus índices de potencia no debe ser cercano a uno, resultando en un índice de potencia residual que no es utilizado, para compensar lo anterior, PGPS trata de encontrar uno o más paquetes adicionales, los cuales puedan ser transmitidos, cumpliendo con la misma condición de que la suma de sus índices de potencia no debe ser cercano a uno, aunque cabe mencionar que esto no resuelve por completo el problema del índice de potencia residual.

3.2.2 SCHEDULING DE CDMA CALENDARIZADO

Es un híbrido del gestor de recursos CDMA y del gestor de recursos TDMA, en el cual la EB calendariza las transmisiones de los TMs como se muestra en la siguiente figura 4.

Los datos son intercambiados entre la EB y los TMs en una unidad de tamaño fijo llamada *capsule* la cual puede acomodar a uno o mas paquetes. SCDMA asume una operación en ranuras de tiempo (time-slots) en las cuales a los TMs se les permite transmitir simultáneamente en cada ranura de tiempo.

3.2.3 GESTIÓN DE RECURSO DINÁMICO

Es un sistema centralizado y adaptivo para proveer gestión de recursos a través de una óptima asignación de potencia y salto de códigos en un sistema W-CDMA.

En DRS, los TMs envían sus peticiones a la EB, la cual se encarga de clasificarlas de acuerdo a las características del tráfico de los servicios pedidos, y los coloca en dos colas separadas denominadas: cola garantizada y cola de mejor esfuerzo, las cuales se muestran en la Figura 5.

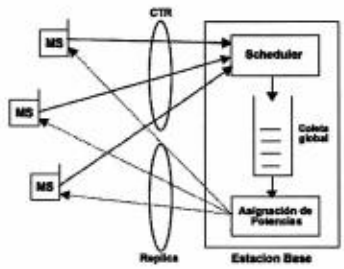


Figura 4. Gestor de recursos SCDMA

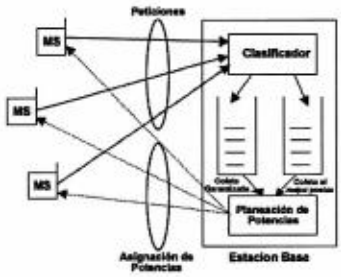


Figura 5. Gestión de recurso dinámico.

3.2.4 CONTROL DE ACCESO MULTIMEDIA PARA SISTEMAS INALÁMBRICOS CON GESTIÓN DE RECURSOS BASADO EN BER

Los paquetes son transmitidos dentro de tramas de longitud T en ambos enlaces de subida y de bajada. Existe un número de clases de tráfico, cada clase con un diferente requerimiento de BER, se asume que los paquetes llegan en grupo, y todos los paquetes en un grupo dado tienen el mismo tiempo de expiración. Para cada grupo, WISPER primero asigna un valor de prioridad que es directamente proporcional al número restante de paquetes en el grupo, e inversamente proporcional al tiempo restante antes de que el paquete expire y a la máxima tasa de transmisión para los TMs. El algoritmo selecciona los paquetes a ser transmitidos en orden de prioridad y trata de maximizar el número de paquetes que pueden ser transmitidos en una trama.

3.2.5 GESTIÓN DE RECURSOS EN REDES MULTISALTOS

Las redes de un solo salto y multisaltos son redes ad-hoc con una pequeña infraestructura de soporte, en particular, las estaciones base no están disponibles. En una red de un solo salto, cada TM se puede comunicar directamente con todas las otros TMs, al ir aumentando el número de TMs, una red multisaltos debe formarse, en la cual, no todas los TMs se puedan comunicar directamente con cualquier otro.

Los TMs, actuando como relevadores, son necesarios para guiar a los paquetes a su destino final. La principal dificultad en el diseño de gestores de redes multisaltos radica en el hecho de que no todos los TMs se pueden comunicar

directamente con cualquier otro y de que la topología de la red cambia rápidamente.

Después de dar una breve descripción de algunos algoritmos de gestión de recursos para redes móviles, en el siguiente capítulo se presenta un algoritmo de gestión de recurso basado en CDMA considerando preferencia.

4. Algoritmo de Gestión de Recurso: Cdma con Procesador Centralizado Considerando Preferencia (Preemption)

4.1 Modelo del Sistema

Para el algoritmo de gestión de recursos basado en CDMA se considera una red celular y solamente el enlace ascendente (del terminal móvil a la estación base). Además es empleado un control perfecto de potencia que es capaz de atenuar los desvanecimientos del canal, es decir, el terminal móvil (TM) transmite el nivel de potencia necesario para mantener la misma potencia recibida siempre. Además, el canal se puede considerar como ruido aditivo blanco Gaussiano (AWGN), por lo cual podemos utilizar la hipótesis Gaussiana para modelar la interferencia originada por otros terminales móviles (TMs) [Pursley,1977]. Slotted-ALOHA, como una parte del protocolo de acceso CDMA, se utiliza en el acceso de la petición, donde un TM elige una secuencia del código del PN (pseudo ruido) para el espectro ensanchado [Kueh, 2002].

Cada estación base EB controla todo el tráfico (tráfico de datos alta prioridad como transacciones financieras, vídeo, voz y datos del tipo WWW) generado en la célula. En nuestro estudio, nos centramos en una sola célula, con algunos TMs que se comuniquen con la EB usando un paquete común de canales en el enlace ascendente.

Para poder transmitir, el TM necesita generar un paquete de datos. Dependiendo de la naturaleza de los datos un modelo de tráfico es utilizado para generar tal paquete. Por ejemplo, para las llamadas de transacciones financieras se utiliza una distribución de Poisson, mientras que para las llamadas de los datos-WWW se utiliza una distribución de Pareto. En el caso de llamadas de voz se utiliza un proceso ENCENDIDO-APAGADO (ON-OFF), y para las llamadas de vídeo un proceso de MMPP. Una vez que el paquete es creado, el TM hace la petición de transmitir en la ranura de acceso usando para esto ALOHA. Después de que la EB reciba con éxito una petición para transmitir de un TM, obtiene información de esa MT, que incluye el tipo del servicio, cantidad de información generada, tiempo de generación, la ranura de tiempo donde la información fue generada, y el

estado del buffer. Con esta información la EB actualiza su tabla de las peticiones (TP) y asignan los recursos iniciando con la asignación de códigos a los TMs (Ver figura 6).

Si hay insuficientes códigos para todos los TMs presentes en la célula, la EB asigna los códigos basándose en la prioridad del servicio, donde las llamadas de transacciones financieras tienen la prioridad más alta, seguida por el vídeo, la voz y los datos-WWW, respectivamente. Según la información almacenada en su TP, la EB asigna códigos a los TM iniciando con el tráfico de las transacciones financieras que tengan sus buffers llenos. Si todavía hay códigos disponibles, éstos se asignan, primero, al vídeo después a la voz, y finalmente a los datos-WWW. De una manera similar si podemos todavía encontrar códigos disponibles, estos se asignan con la prioridad anterior y tomando en cuenta aquellos que están próximos a expirar su tiempo de vida.

En el proceso anterior de asignación de códigos, siempre que no haya suficientes códigos para las peticiones de las transacciones financieras, se propone que los códigos usados para datos-WWW se les retire momentáneamente y se asignen a las transacciones financieras. A los TMs de datos-WWW que se les haya quitado el código se almacenan en un buffer y se le asigna otra vez código tan pronto como uno llegue a estar disponible. Una punto importante por mencionar es que la EB toma la decisión de que acepta o no una nueva petición de un TM, basado en el razón de bit, número de códigos disponibles, y siempre y cuando se garantice la relación señal a ruido (SNR) para cada uno de los diferentes tipos de datos.

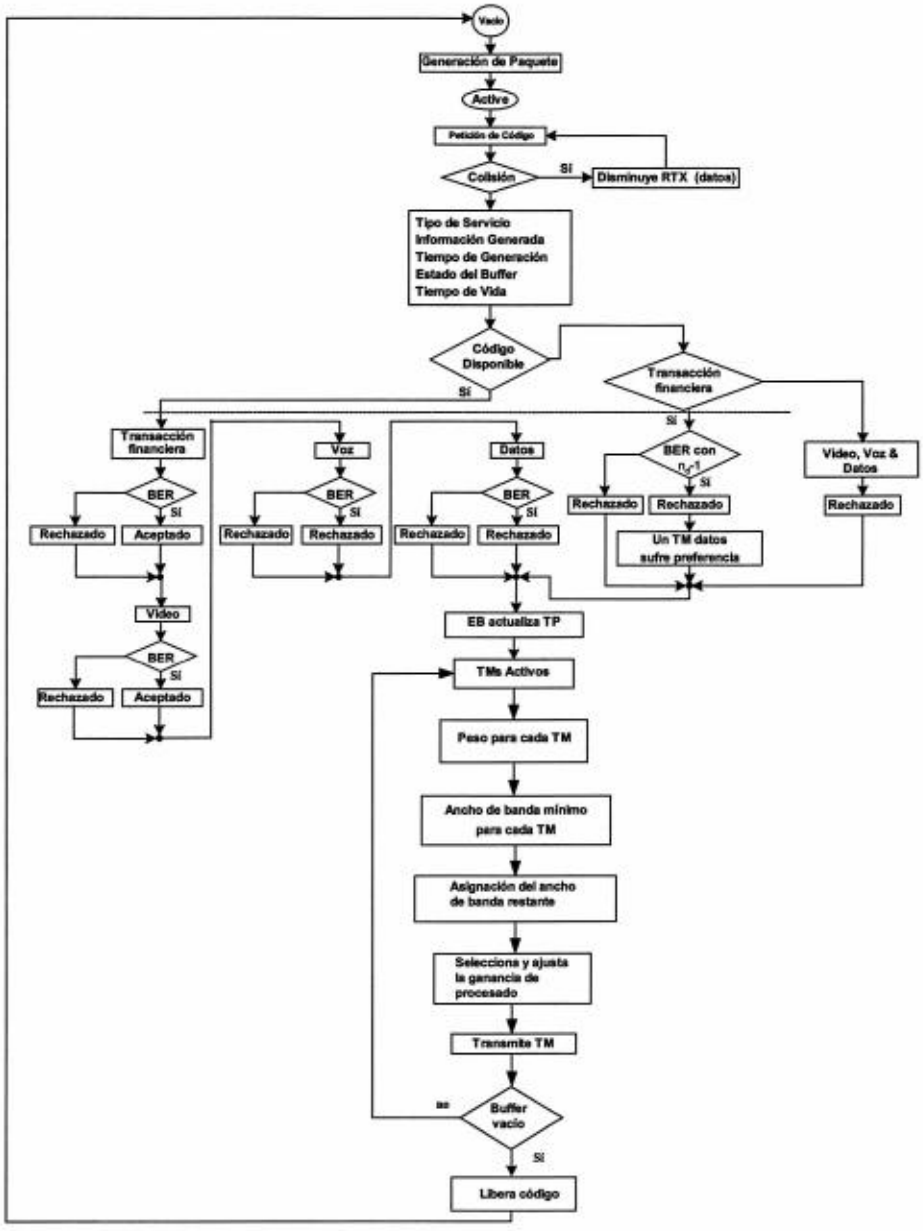


Figura 6. Diagrama de flujo para el algoritmo de gestión de recursos.

Después de asignar códigos a los TMs, la EB sabe la cantidad de TMs asignados a las transacciones financieras, al vídeo, a la voz, y a los datos-WWW, así como la tasa que cada TM está solicitando y a través del controlador centralizado (GPS-Generalized Processor Sharing) se calcula la tasa mínima requerida para cada TM. La tasa excedente es asignada de acuerdo a la carga de cada TM. Basándose en esta información, la EB puede seleccionar y ajustar la ganancia de procesado de cada TM listo para transmitir en la siguiente ranura de tiempo disponible. Si después de transmitir, un TM no tiene información en su buffer, libera sus códigos e pasa al estado vacío (IDLE). En caso de que el TM tenga información adicional a transmitir, informa a la EB para evitar una interrupción durante la próxima ranura de tiempo.

4.2 CONTROL DE ADMISIÓN

La asignación de códigos esta obligada a cumplir con los requisitos SNR del tráfico multimedia. Por lo tanto, una nueva llamada de petición se admite basándose en el SNR requerido

$$SNR\quad i\quad >\quad SNR_{hreshold} \tag{1}$$

donde $SNR_{threshold}$ es el umbral de SNR, el SNR_i es el sistema SNR desde que la petición i-th es aceptada.

Si el tráfico transmitido incluye voz, datos, transacciones financieras, y vídeo, entonces el SNR del tráfico de voz, SNR_v, se describe como ,

$$SNR_v = \frac{E_b^v}{n_o + \frac{n_v - 1}{\frac{3}{2}P_{G_v}}\alpha_v E_b^v + \frac{n_d}{\frac{3}{2}P_{G_d}}E_b^d + \frac{n_q}{\frac{3}{2}P_{G_q}}E_b^q + \frac{n_t}{\frac{3}{2}P_{G_t}}E_b^t} \tag{2}$$

donde E_b^v es la energía de bit del TM para voz, E_b^d es el la energía de bit del TM para datos, E_b^q es la energía de bit del TM para vídeo, E_b^t es la energía de bit para transacciones financieras, n_v es el número de TMs activos de voz, n_d es el número de TMs activos para datos, n_q es el número de TMs activos de video, n_t es el número de TMs activos para transacciones financieras, P_{G_v} es la ganancia de procesado para TMs de voz, P_{G_d} es la ganancia de procesado para TMs de datos, P_{G_q} es la ganancia de procesado de TMs de vídeo, P_{G_t} es la ganancia de proce-

sado para TMs de transacciones financieras, α_v es el factor de actividad de la voz donde $\alpha_v = t_1/(t_1+t_2)$ siendo t_1 el promedio de la duración de la voz activa y t_2 el promedio de la duración de los bloques silenciosos de la voz. Además, el término 3/2 se consigue según [Wu y Kohno, 1996].

Semejantemente, el SNR_d de los TMs de datos, SNR_q para vídeo y SNR_t para las transacciones son dadas por

$$SNR_d = \frac{E_b^d}{n_o + \frac{n_d - 1}{\frac{3}{2}P_{G_d}} E_b^d + \frac{n_v}{\frac{3}{2}P_{G_v}} \alpha_v E_b^v + \frac{n_p}{\frac{3}{2}P_{G_q}} E_b^q + \frac{n_t}{\frac{3}{2}P_{Gt}} E_b^t} \tag{3}$$

$$SNR_q = \frac{E_b^q}{n_o + \frac{n_q - 1}{\frac{3}{2}P_{G_q}} E_b^q + \frac{n_v}{\frac{3}{2}P_{G_v}} \alpha_v E_b^v + \frac{n_d}{\frac{3}{2}P_{G_d}} E_b^d + \frac{n_t}{\frac{3}{2}P_{G_t}} E_b^t} \tag{4}$$

$$SNR_t = \frac{E_b^t}{n_o + \frac{n_t - 1}{\frac{3}{2}P_{Gt}} E_b^t + \frac{n_d}{\frac{3}{2}P_{G_d}} E_b^d + \frac{n_v}{\frac{3}{2}P_{G_v}} \alpha_v E_b^v + \frac{n_p}{\frac{3}{2}P_{G_q}} E_b^q} \tag{5}$$

Si la EB recibe la misma densidad de energía de cada TM, entonces $SNR_v = SNR_d = SNR_q = SNR_t = SNR_{system}$ y $E_b^d = E_b^v = E_b^q = E_b^t = E_b$.

En nuestro sistema, las transacciones financieras y el flujo de información de voz son constantes, las tasas de transmisión son fijas, y la ganancia de procesado es constante.

También, el flujo de información de los datos-WWW y del vídeo no es continuo, las tasas de transmisión no son fijas, la ganancia de procesado no es constante, y se requiere un control dinámico de la ganancia de procesado.

Por lo tanto, podemos alcanzar el funcionamiento óptimo del sistema de CDMA utilizando la variación dinámica de la ganancia de procesado de acuerdo a [Oh y Wasserman, 1999], mientras que dinámicamente asignan un ancho de banda para cada MT al mismo tiempo. Así mismo, con esto es posible calcular el BER para cada tipo de tráfico en un ambiente de multimedia (voz, dato, vídeo).

4.3 ASIGNACIÓN DE LA TASA DE TRANSMISIÓN CON Y SIN PREFERENCIA

Para la asignación de la tasa mínima a cada sesión es usado el esquema de procesador centralizado (GPS), también conocido como Fluid-Flow Fair Queuing (FFQ) [Parekh y Gallager, 1993].

En el esquema de gestión de recursos basado en CDMA junto con procesador centralizado (CDMA/GPS), del tipo rate-scheduling, es posible transmitir diferentes tipos de datos a diferente tasa de transmisión. Este esquema es diferente a los que se usan en las redes TDMA ya que éstas están basadas en time-scheduling [Lu *et al.*, 1999] [Jeong *et al.*, 2001].

La capacidad total del enlace, C, en el esquema CDMA/GPS es compartida por N sesiones. Cada sesión *i* mantiene una conexión con la tasa del enlace $C_i(k)$ durante *k*-ésima ranura de tiempo, tal que,

$$\sum_{i=1}^{N} C_i(k) \leq C \tag{6}$$

Cualquier sesión *i* entra a un solo servidor a un sistema de colas con una tasa de servicio $C_i(k)$. A diferencia del servidor de colas convencional, la tasa de servicio $C_i(k)$ puede variar con el tiempo. Permitiendo que ϕ_i, sea el peso para la sesión *i*, donde $i = 1, 2, ..., N$, $N = n_d + n_v + n_q + n_t$, y $W_i(k)$, la cantidad de tráfico servido durante la ranura *k*. Entonces, según la disciplina de asignación del recurso del GPS, la ecuación 7 se debe mantener para cualquier sesión que sea continuamente reservada en la ranura *k*, es decir,

$$\frac{W_i(k)}{W_j(k)} \geq \frac{\Phi_i}{\Phi_j} \qquad j = 1, 2, \ldots, N \tag{7}$$

La cantidad de tráfico reservado de la sesión *i* durante la ranura *k* es el tráfico reservado en la ranura anterior más el tráfico estimado de la sesión *i* durante la ranura *k*. Si el tráfico reservado es cero entonces el tráfico servido es cero, y en caso de que el tráfico reservado no sea cero el tráfico servido en la *i* sesión es

$$W_i(k) = g_i T \tag{8}$$

donde $g_i = \frac{\phi_i}{\sum_{j=1}^{N} \phi_j} \cdot C$ es la tasa mínima garantizada para la sesión *i*, T es la longitud de la ranura de tiempo, y C es la cantidad máxima de tasa de servicio que se puede proporcionar por la red.

La tasa asignada al TM *i* puede ser determinada por $C_i(k) = \frac{W_i(k)}{T}$, así que la tasa mínima garantizada para el TM *i* es dada por

$$C_i(k) = \frac{\Phi_i}{\sum_{j=1}^{N} \Phi_j} \cdot C \tag{9}$$

y el recurso restante de la red será distribuido proporcionalmente en cargas individuales de ϕi.

Por lo tanto, es usada una asignación de pesos, donde una tasa mínima está garantizada para cada TM [Mendez, 2003],

$$\phi_1 \cdot P_{G_1} = \phi_2 \cdot P_{G_2} = \ldots = \phi_N \cdot P_{G_N} \tag{10}$$

donde $N = n_t + n_q + n_v + n_d$

Ahora, necesitamos entender cómo incluir la preferencia para satisfacer las peticiones de la transacción financiera. Siempre que una petición financiera que llega no encuentre ningún código fijo disponible éste tiene preferencia a un servicio de dato-WWW. Para seleccionar que usuario de dato-WWW se le va a quitar el código, se tomará en cuenta su historial y se empezará con aquellos que se les ha dado más tiempo el servicio.

La hipótesis con este esquema será que el rendimiento para transacción financiera será incrementado, es decir que se podrá manejar un mayor número de usuario para transacciones financieras. Esto debe ser obtenido garantizando la calidad de servicio. La siguiente sección está dirigida para probar experimentalmente esta hipótesis.

5. SIMULACIÓN

Para la simulación se asume enlace libre de error considerando una célula. Los parámetros para la simulación se basan en el estándar UMTS y consideran la siguiente caracterización del tráfico de servicios:

Tráfico de voz: Este modelo está basado en el modelo de Markov para un detector lento y se utiliza para generar patrones de dialogo de la conversación.

Tráfico de video: Es generado usando un modelo MMPP (Markov-Modulated Poisson Processor) basado en [Frost y Relamed, 1994].

Transacción financiera: Este modelo de tráfico se genera de acuerdo a una distribución de Poisson.

Para tráfico de datos-WWW se considera en el modelo presentado en [UMTS, 1997] referido al estándar UMTS, donde la aplicación del uso de los datos-WWW sigue una distribución de Pareto con unos valores fijos de α=1.1 y k=81.5 y una máxima medida de 66.666 bytes.

Puesto que varios tipos de tráfico coexisten en un sistema, el sistema debe observar el requerimiento SNR. Es decir, el es determinado por el requerimiento del tráfico de WWW-datos.

Además, el funcionamiento del sistema es limitado solamente por la interferencia, por lo tanto el ruido térmico es despreciable.

La tabla II muestra los valores de los parámetro (estándar UMTS) usados en la simulación. Para datos-WWW y vídeo la ganancia de procesado es variable, y puesto que estamos considerando los flujos de información continuos para la transacción financiera y la voz entonces la ganancia de procesado para esto es fijo.

Tabla II. Parámetros de simulación

Parámetro	*Valor*
Chip rate	4.096 Mcps
Tasa del canal WCDMA	2.0 Mbps
Modulació	BPSK (uplink)
Ganancia de procesado para vídeo	Variable
Ganancia de procesado para datos-WWW.	Variable
Tasa de la fuente de voz	16 kbps
Tasa de la fuente de datos-WWW	16 to 384 kbps
Tasa de la fuente para transacciones financieras	16 kbps
Tasa de la fuente para vídeo	16 to 384 kbps
Duración promedio de la ráfaga de voz	1.41 s
Duración promedio de los silencio para voz	1.74 s
BER para tráfico de vídeo	$\leq 10^{-5}$
BER para tráfico de voz	$\leq 10^{-3}$
BER para tráfico de datos-WWW	$\leq 10^{-9}$
BER para tráfico de transacciones financieras	$\leq 10^{-9}$

El primer resultado que presentaremos considera el tráfico multimedia compuesto por transacciones financieras, vídeo, y voz. Asignamos a la voz la prioridad más baja; por lo tanto, el servicio de la voz será *preempted* siempre que la transacción financiera necesite códigos adicionales. Los resultados bajo estas asunciones se demuestran en las figuras 7 y 8.

La Figura 7 muestra el comportamiento del rendimiento para el esquema CDMA/GPS como función del tráfico ofrecido. Para el tráfico lento, el incremento del rendimiento es aproximadamente lineal con respecto al tráfico ofrecido hasta que el rendimiento alcanza su valor máximo. En esta región, esencialmente no hay colisiones. Cuando el tráfico aumenta, causa una disminución del rendimiento. Esta disminución es resultado del incremento de las colisiones que son directamente proporcionales al número de TMs para un número fijo de códigos, y no se relaciona con el método usado para la gestión de recursos.

En la misma figura 7 muestra el rendimiento para el tráfico de voz cuando consideramos un tráfico multimedia, considerando los casos: con y sin preferencia. Observamos que si aplicamos preferencia a la voz esto da como consecuencia un aumento del rendimiento de las transacciones financieras, aproximadamente un 17% con respecto si sin preferencia. Esto es obvio porque, siempre que no haya códigos disponibles y la transacción financiera necesita más, tomará algunos códigos del servicio de voz. Este resultado sugeriría que se puede aplicar prefe-

rencia a la voz, sin embargo, todavía necesitamos verificar si los requisitos de QoS estén satisfechos o no.

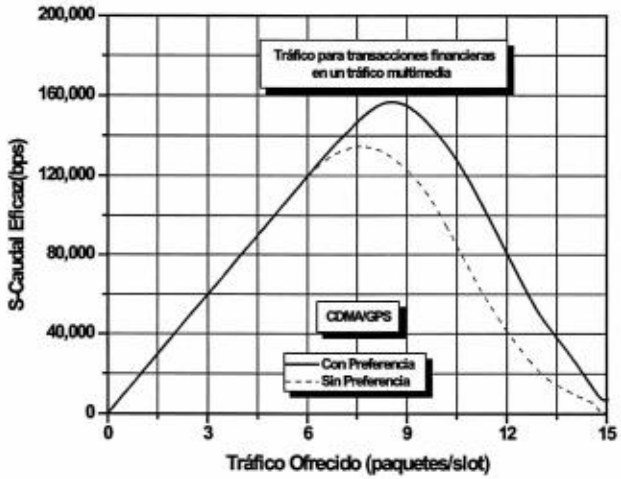


Figura 7. Desempeño para transacciones financieras.

De acuerdo con estándares de UMTS debemos tener como máximo el 1% de paquetes perdidos. Un paquete será forzado a perderse una vez que expire su ciclo de vida, o cuando el paquete generado en el buffer es desbordado. Este resultado se observa en la figura 8.

En la figura 8 observamos que si CDMA/GPS no considera la preferencia entonces podemos tener 27 usuarios de voz con un máximo del 1% de paquetes perdidos, sin embargo, cuando se aplica el preferencia el sistema puede manejar solamente 12 usuarios de voz. Por lo tanto, concluimos que no está recomendado aplicar preferencia a usuarios de voz.

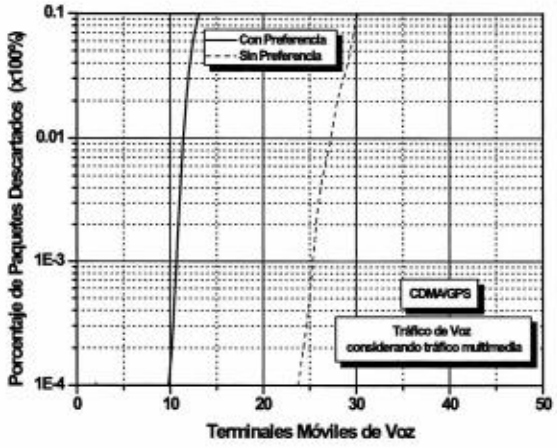


Figura 8. Porcentaje de paquetes descartados para tráfico de voz.

Consideremos ahora el tráfico multimedia formado por transacciones financieras, vídeo, voz, y datos (tipo WWW), la prioridad más baja será dada al tráfico de WWW-datos, puesto que no es un servicio en tiempo real. Bajo esta consideración el rendimiento para la transacción financiera observa en la figura 9.

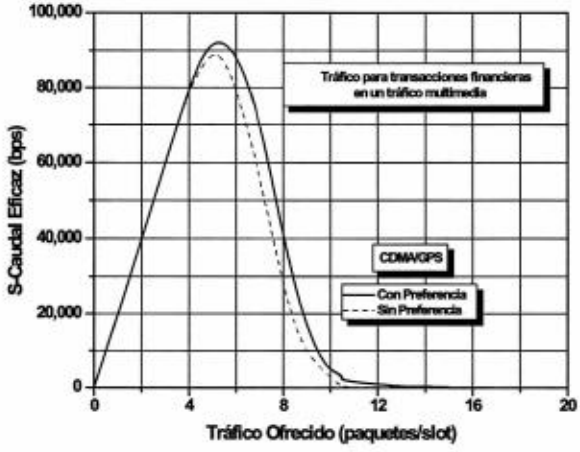
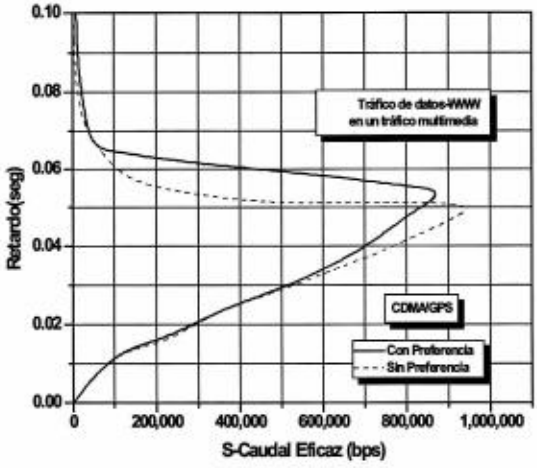


Figura 9. Desempeño para transacciones financieras en un tráfico multimedia

La figura 9 muestra que cuando aplicamos preferencia al servicio de datos-WWW se obtiene un aumento del rendimiento de las transacciones financieras.

El aumento en el rendimiento de las transacciones financieras es de 3.5%. Esta diferencia entre el aumento del rendimiento y la disminución es por las tasas variables de los servicios (datos-WWW) tienen una fuerte influencia en el rendimiento que los servicios de tasa constantes (transacciones financieras). Es necesario determinar si aplicando preferencia a los datos-WWW sigue garantizando el máximo retraso para el tipo de datos WWW, que es de cuatro segundos. Esto se muestra en la Figura 10.

En la Figura 10 podemos ver que, sin preferencia en el esquema CDMA/GPS, el tráfico de datos-WWW garantiza la QoS con un retraso de 50 mseg a lo máximo, sin embargo, cuando se aplica preferencia a los datos-WWW el retraso aumenta a 54 mseg. Este aumento en el retraso no es suficientemente grande para violar la QoS requerida.



6. CONCLUSIONES.

Con la aparición de nuevas tecnologías multimedia, del Internet y el uso de las redes inalámbricas, han estimulado el estudio de los algoritmos programables para proporcionar Garantías de QoS. Estas garantías están generalmente dentro de la forma del límite de retraso, garantizando tarifa e imparcialidad entre sesiones. Sigue habiendo mucho trabajo por hacer en la evaluación del diseño y el de funcionamiento de la radio planificación, especialmente para CDMA .

La evaluación del funcionamiento, para el esquema CDMA/GPS-DW (con y sin preemption) con la asignación dinámica de tasas, para el tráfico multimedia (transacciones financieras, vídeo, voz, y WWW-datos) se ha presentado. Además,

el control de la admisión fue propuesto para permitir al BS manejar las peticiones de las MTs basadas en el SNR requerido. Nuestros resultados de la simulación demuestran que el uso del preemption está recomendado siempre que no haya usuarios con estrictos requerimientos de retrasos (por ejemplo, WWW-datos). Esto es principalmente porque si aplicamos preemption a los usuarios de voz, no habrá garantías en el QoS relacionado con retraso y porcentaje de paquetes perdidos.

El análisis presentado se centra en los aspectos cualitativos que se deben considerar para el uso del preemption dentro de la estrategia de la asignación del recurso. Muchas preguntas interesantes siguen abiertas, siendo relacionadas a cómo introducir eficientemente el preemption en los panoramas de multimedia para el uso óptimo de recursos. La investigación presentada aquí es justo un primer paso en esta dirección.

REFERENCIAS.

Castro, J. P. 2001. The UMTS Network and Radio Access Technology: Air Interface Techniques for Future Mobile Systems. John Wiley & Sons. Primera Edición. Chichester, pp. 354

COFETEL-Comisión Federal de Telecomunicaciones, 2003. Estadísticas de Telecomunicaciones. México. [online] http://www.cft.gob.mx/.

Fantacci, R. and Naldi, Alessandro, 2000. Performance of a CDMA Protocol for Voice and Data Integration in *Personal Communication Networks, IEEE Transactions on Vehicular Technology*, 49(2): 307-320.

Fattah, H. y Leung, C. 2002. An Overview Scheduling Algorithms in Wireless Multimedia Networks". *IEEE Wireless Communications*. 9(5): 76-83.

Frost, V. S. and Melamed, B, 1994. Traffic Modeling for Telecommunications Networks, *IEEE Communications Magazine*, 32(3): 70-81.

Jeong, M., Morikawa, H. and Aoyama, T., 2001. A fair scheduling algorithm for wireless packet networks, *IEICE Transactions on Fundamentals*, 84-A(7): 1624-1635.

Kang, H., Kim, D., Lee, C. y Kim, K. 2000. A Throughput-Efficient Code Assignment Scheme for an Integrated Voice/Data Multi-Code CDMA System. Proc. *51st IEEE VTC'00*. p.1494-1497.

Kim, D. I, Hossain, E. y Bhargava, V. K. 2001. Integrated Error Control in Variable Spreading Gain WCDMA Systems, *Proc. IEEE ICC'2001*. p. 1362-1366.

Kim, J. B., Honig, M. L. y Jordan, S. 2001. Dynamic Resource Allocation for Integrated Voice and Data Traffic in DS-CDMA. *Proc. 54th IEEE VTC'01*. p.42-46.

Kueh, V., Tafazolli, R., and Evans, B. G. 2002.Performance of Prioritized W-CDMA RACH Transmission Over Satellite-UMTS, *in Proc. European Wireless 2002*, Florency-Italy, February 25-28, [online] www.ing.unipi.it/ew2002/proceedings/.

Lu, S., Bhargavan, V. and Srikant, R.,1999. Fair Scheduling in Wireless Packet Networks, *IEEE/ACM Transactions on Networking*, 7(4): 473-489.

Mendez, A., 2003. Contributions to the Medium Access Techniques in Third Generation Mobile Communication Systems in a DS-CDMA Environment, PhD Thesis, CICESE, Mexico. (in Spanish).

Mendez, A., Covarrubias, D. and Brizuela, C. 2004. Fair Scheduling with Dynamic Resource Allocation in CDMA/GPS System for IP-Multimedia Wireless Networks, *Journal of Circuits, Systems, and Computers*, 13(2): 253-269.

Oh, S.-J. and Wasserman, K. M.,1999. Dynamic Spreading Gain control in Multiservice CDMA Networks, *IEEE Journal on Selected Areas in Communications*, 17(5): 918-927.

Parekh, A. K. and Gallager, R. G., 1993. A Generalized Processor Sharing Approach to Flow Control in Integrated Services Networks: the Single-Node Case, *IEEE/ACM Transaction on Networking*, 1(3): 344-357.

Pursley, M. B. 1977. Performance Evaluation for Phase-Coded Spread-Spectrum Multiple-Access Communications - Part I: System Analysis. *IEEE Transactions on Communications*. 25(8): 795-799.

Rappaport,T.S.,2001.Wireless Communications, Prentice Hall PTR, 2nd Edition, pp. 736.

Sallent, O. 1999. Comunicaciones Móviles: Sistemas 3G. *Mundo Electrónico*. 304(11): 46-50 pp.

Sallent, O. y Agustí, R. 2000. Adaptive S-ALOHA CDMA as an Alternative Way of Integrating Services in Mobile Environments. *IEEE Transactions on Vehicular Technology*. 49(3): 936-947.

Sandouk, A., Yamazato, T., Katayama, M. y Ogawa, A. 1999. An Integrated Voice/Data CDMA Packet Communications with Multi-Code CDMA Scheme. *IEICE Transactions on Fundamentals*. 82(10): 2105-2113.

5. Evaluación de Prestaciones del Esquema ALOHA-CDMA Adaptable a las Condiciones del Tráfico.

Ángel Dorantes Salazar

I. Introduccion

Sabemos que los protocolos Aloha es probablemente la familia más rica de protocolos de acceso múltiple. Su popularidad es debido a su madurez ya que fue la primera técnica de acceso aleatorio, esto aunado a que son protocolos tan simples que su implementación es muy sencilla. Todas estas ventajas se podrían ver incrementadas con el soporte de CDMA, dada la robustez intrínseca de CDMA a las interferencias, a su flexibilidad en la banda disponible del canal cuando hay pocos usuarios entre otras.

Por todo lo anterior a la fase de petición de canal (S-Aloha), se agregará el entorno de CDMA (fase de transmisión), es decir, se trata de sistemas de comunicaciones móviles de tercera generación del tipo ALOHA-CDMA. Con el sistema ALOHA-CDMA lo que se pretende es tomar las ventajas que tiene cada sistema y obtener uno que sea adaptable a las condiciones del tráfico en el canal (figura 1).

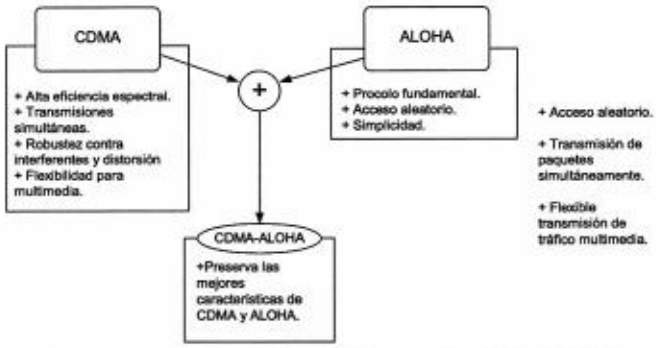


Figura 1. Características del sistema ALOHA-CDMA.

S-Aloha es usado para hacer la petición del canal debido a que su acceso al medio es en forma aleatoria, al adaptarlo con DS-CDMA se adicionan las características de éste, como ganancia de procesado, múltiples terminales móviles transmitiendo al mismo tiempo, códigos de ensanchado, entre otros. Consiguiendo con esto un manejo de tráfico, para lograr después un esquema adaptativo a las condiciones del tráfico en el canal.

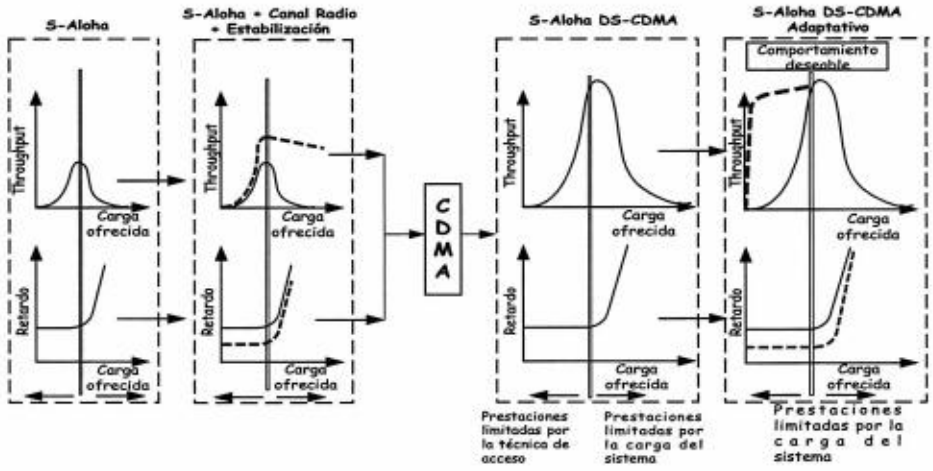


Figura 2. Comportamiento de un sistema S-Aloha DS-CDMA.
a) Comportamiento S-Aloha teórico. b) Comportamiento de S-Aloha mejorado.
c) Comportamiento típico S-Aloha/DS-CDMA. d) Comportamiento de S-Aloha/DS-CDMA deseado.

En la figura 2a observamos el comportamiento de S-Aloha teórico el cual presenta limitaciones tales como inestabilidad y baja eficiencia, esto por el bajo throughput y el alto retardo.

Para elevar la eficiencia se empleó el efecto captura para una distribución espacial y el efecto canal radio (Rayleigh y Shadow), de esta manera se elevó casi al doble el valor del throughput de S-Aloha teórico, además se disminuyó el retardo y el número de terminales móviles bloqueados en la región de bajo tráfico como lo visualizamos en la figura 2b.

El comportamiento que presenta S-Aloha/DS-CDMA es mostrado en la figura 2c, en donde observamos que hay un bajo throuhput y además un retardo elevado en la región de bajo tráfico. Por lo que es necesario un esquema que sea adaptable a las condiciones del tráfico, para tener como resultado un throughput alto y retardo bajo en la región de bajo tráfico, como se muestra en la figura 2d.

En este trabajo presentamos un análisis del esquema S-Aloha/DS-CDMA, en el cual se plantean las alternativas de esquemas de multi-velocidades en un entorno DS-CDMA (multi-ganancia de procesado, multi-códigos y multi-modulación) escogiendo el esquema óptimo, y así obtener un sistema S-AlohaDS-CDMA adaptable a las condiciones del tráfico en el canal.

II. MODELO DE S-ALOHA/DS-CDMA CON MAYOR NÚMERO DE CÓDIGOS QUE TERMINALES MÓVILES

El modelado del sistema S-Aloha/DS-CDMA considera N terminales móviles y K códigos disponibles donde $K \geq N$, estos terminales móviles pueden estar en uno de los dos modos en el inicio de un slot: modo vacío (I) o modo bloqueo (B). Bajo este esquema, el modelo para el sistema S-Aloha/DS-CDMA es mostrado en la figura 3, mientras en la Tabla I son definidos los parámetros utilizados en dicho sistema.

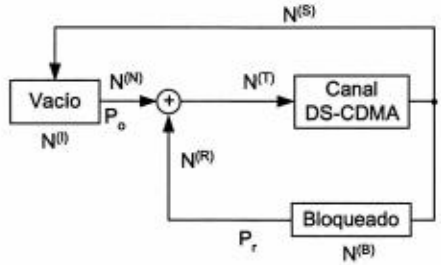


Figura 3. Modelo para el sistema S-Aloha/DS-CDMA con $K \geq N$.

Tabla I. Parámetros para el sistema S-Aloha/DS-CDMA con $K \geq N$.

N	Número de terminales móviles en el sistema.
$N_k^{(B)}$	Número de terminales móviles en modo bloqueo (B) en el inicio del slot k.
$N_k^{(I)}$	Número de terminales móviles en modo vacío (I) en el inicio del slot k.
$N_k^{(T)}$	Número total de terminales móviles transmitiendo paquetes en el slot k.
$N_k^{(R)}$	Número de terminales móviles en modo bloqueo transmitiendo paquetes en el slot k.
$N_k^{(N)}$	Número de terminales móviles en modo vacío transmitiendo paquetes en el slot k.
$N_k^{(S)}$	Número de paquetes recibidos correctamente en el slot k.
p_o	Probabilidad que un terminal móvil en modo I genere un paquete en un slot.
p_R	Probabilidad que un terminal móvil en modo B retransmita un paquete en un slot.
$P_c(n)$	Probabilidad de recibir correctamente un paquete cuando se presentan n transmisiones simultáneas.

En este sistema la distribución en equilibrio es caracterizada mediante la matriz de transición de estado, cuyos elementos p_{ij} son dados a partir de:

$$p_{ij} = \Pr\left(N_{k+1}^{(B)} = j \mid N_k^{(B)} = i\right) \quad 0 \le i \quad j \le N \tag{1}$$

y después de un análisis exhaustivo se obtiene[1] [2]:

$$p_{ij} = \sum_{n=0}^{N} \sum_{s=0}^{n} \binom{N-i}{j-i+s} p_o^{j-i+s} (1-p_o)^{N-j-s} \times \binom{i}{n-s+i-j} p_R^{n-s+i-j} (1-p_R)^{j+s-n} \times \Pr\left(N^{(S)} = s \mid N^{(T)} = n\right)$$

(2)

La expresión para la distribución en equilibrio del paquete, que llega en un slot, en términos de la distribución en equilibrio de la cadena de Markov es:

$$\Pr\left(N_k^{(T)} = n\right) = \sum_{i=0}^{N} \Pr\left(N_k^{(T)} = n \mid N_k^{(B)} = i\right) \pi_i , \tag{3}$$

$$\Pr\left(N_k^{(T)} = n\right) = \sum_{i=0}^{N} \left[\sum_{m=\max(0,n-i)}^{\min(n,N-i)} \binom{i}{n-m} \times p_R^{n-m} (1-p_R)^{i-n+m} \binom{N-i}{m} \times p_o^{m} (1-p_o)^{N-i-m} \right] \cdot \pi_i .$$

Entonces, el throughput lo podemos expresar como [1] [2]:

$$S = \sum_{n=0}^{N} \left[\sum_{s=0}^{n} s \times \Pr\left(N_k^{(S)} = s \mid N_k^{(T)} = n\right) \right] \times \Pr\left(N_k^{(T)} = n\right) \text{ paquetes/slot} , \tag{4}$$

Donde:

$$\Pr\left(N_k^{(S)} = s \mid N_k^{(T)} = n\right) = \binom{n}{s} \left[P_c(n)\right]^s \left[1 - P_c(n)\right]^{n-s} , \tag{5}$$

Siendo $P_c(n)$ la probabilidad de recibir correctamente un paquete cuando n terminales móviles han intentado transmitir en un slot:

$$P_c(n) = \left[1 - P_b(n)\right]^L \tag{6}$$

Donde L es el número de bits en el paquete y P_b es la probabilidad de error en el bit.

$$P_b(n)=Q\left(\sqrt{2\frac{E_b}{N_o}}\right) \quad \text{con} \quad \frac{E_b}{N_o}=\frac{1}{\frac{2(n-1)}{3G_p}}$$

III. MODELO DE S-ALOHA/DS-CDMA CON MAYOR NÚMERO DE TERMINALES MÓVILES QUE CÓDIGOS

El modelo para el sistema S-Aloha/DS-CDMA para $K<N$ es mostrado en la figura 4, mientras los elementos del sistema para esta sección son definidos tanto en la Tabla I como en la Tabla II.

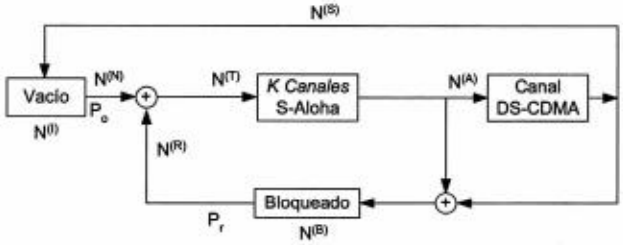


Figura 4. Modelo S-Aloha/DS-CDMA con $K<N$.

Tabla II. Lista de parámetros adicionales para el sistema S-Aloha/DS-CDMA con $K<N$.

K	Número de códigos-receptor en la estación base.
$N_k^{(A)}$	Número de terminales móviles que son recibidos en el slot k

Para el análisis realizado se tomó el procedimiento realizado en la Sección II, siempre y cuando se introduzcan las correcciones necesarias a la cadena de Markov. La primera corrección es la probabilidad de transición de una etapa del estado i al estado j calculada en la ecuación 2, donde la modificación es que el número de paquetes que se pueden transmitir correctamente no debe ser mayor al número de códigos disponibles (K). Haciendo esta modificación se obtiene [1] [2]:

$$p_{ij}=\sum_{n=0}^{N}\sum_{s=0}^{\min(n,K)}\binom{N-i}{j-i+s}p_o^{j-i+s}(1-p_o)^{N-j-s}\binom{i}{n-s+i-j}p_R^{n-s+i-j}(1-p_R)^{j+s-n}\times \Pr\left(N^{(S)}=s \mid N^{(T)}=n\right)$$

(17)

Otra modificación es, en la evaluación de la función de probabilidad del número de paquetes recibidos correctamente, condicionada al número de paquetes transmitidos. Esto separado en dos etapas: la transmisión y la recepción correcta de los paquetes. Con estas dos etapas se obtiene que:

$$\Pr\left(N_k^{(S)} = s \mid N_k^{(T)} = n\right) = \sum_{a=0}^{\min(n,K)} \Pr\left(N_k^{(S)} = s \mid N_k^{(A)} = a, N_k^{(T)} = n\right) \text{x } \Pr\left(N_k^{(A)} = a \mid N_k^{(T)} = n\right) \tag{18}$$

donde la probabilidad condicional del número de paquetes recibidos correctamente dado que hay n paquetes transmitidos y a paquetes recibidos es:

$$\Pr\left(N_k^{(S)} = s \mid N_k^{(A)} = a, N^{(T)}_k = n\right) = \begin{cases} \binom{a}{s} [P_c(n)]^s [1 - P_c(n)]^{a-s} & 0 \le s \le a, \\ 0 & \text{para otro caso.} \end{cases} \tag{19}$$

La expresión anterior es considerando que la estación base no envía reconocimiento de recepción a los terminales móviles, por este motivo los terminales móviles continúan transmitiendo sin importar que haya colisión, por lo que la probabilidad de recibir correctamente un paquete P_c depende de n. Para corregir esto, la estación base determina que terminales móviles han sido recibidos correctamente, además, los terminales móviles deben esperar el reconocimiento para que sigan transmitiendo. Ahora la probabilidad condicional se modifica a:

$$\Pr\left(N_k^{(S)} = s \mid N_k^{(A)} = a, N^{(T)}_k = n\right) = \begin{cases} \binom{a}{s} [P_c(a)]^s [1 - P_c(a)]^{a-s} & 0 \le s \le a, \\ 0 & \text{para otro caso,} \end{cases} \tag{20}$$

donde la probabilidad condicional queda en términos de $P_c(a)$, ya que, a pesar de que inicialmente transmiten n terminales móviles, al recibirse únicamente a códigos serán estos los que proseguirán la transmisión. Quedando por determinar la probabilidad de que se reciban a códigos cuando n terminales móviles inician la transmisión, y esto es:

$$Pr\left(N_k^{(A)} = a \mid N_k^{(T)} = n\right) = \begin{cases} \dfrac{\binom{K}{a} \sum_{i=0}^{\min\left(\left[\frac{n-a}{2}\right], K-a\right)} \binom{K-a}{i} \binom{n-a-i-1}{n-a-2i}}{\binom{n+K-1}{n}} & \text{si } 0 \le a \le \min(n, K), \\ 0 & \text{en otro caso.} \end{cases}$$

(21)

Con los nuevos cálculos se puede determina el throughput en el sistema para un ambiente real [1] [2]:

$$S = \sum_{n=0}^{N} \left[\sum_{s=0}^{\min(n,K)} s \times \Pr\left(N_k^{(S)} = s \mid N_k^{(T)} = n\right) \right] \times \Pr\left(N_k^{(T)} = n\right) \text{ paquetes/slot} \tag{22}$$

El modelado realizado en los apartados anteriores hay que añadirle algunos algoritmos para manejar dinámicamente las velocidades de transmisión y de esta manera el sistema S-Aloha/DS-CDMA sea adaptable a las condiciones del tráfico en el canal, por lo que proponen tres algoritmos y estos son tratados en el siguiente apartado.

Como se mencionó, al usar el esquema S-Aloha/DS-CDMA la respuesta típica que se tiene es la mostrada en la figura 1c. Se observa en esa figura que en la región de bajo tráfico el throughput es bajo, no porque existan demasiados interferentes (terminales móviles) sino porque hay pocos terminales móviles transmitiendo. Además del problema en el throughput, existe un retardo alto en la región de bajo tráfico. Por lo que se requiere que el sistema S-Aloha/DS-CDMA sea adaptable a las condiciones del tráfico, para tener así un throughput alto y retardo bajo en la región de bajo tráfico, tal y como se muestra en la figura 1d.Por lo tanto, es necesario analizar esquemas que nos permitan modificar esta situación y así tener un sistema adaptable a las condiciones del tráfico.

Hay que mencionar que el ancho de banda para CDMA (1.25 MHz) o W-CDMA (5 MHz) es fijo. Este ancho de banda es dado como el inverso de la duración de chip $1/T_c$:

$$\Delta B = \frac{1}{T_c} = \text{constante}$$

Por lo que si a este valor lo multiplicamos tanto en el numerador y denominador por la duración de bit T_b, el ancho de banda se mantiene fijo:

$$\Delta B = \frac{1}{T_c}\cdot\frac{T_b}{T_b} = \frac{T_b}{T_c}\cdot\frac{1}{\mathrm{T_b}} = \text{constante}$$

Donde T_b/T_c es la ganancia de procesado (G_p)y $1/T_b$ es la velocidad de transmisión en bits por segundo (R_b):

$$\Delta B = \frac{1}{T_c} = \frac{T_b}{T_c}\frac{1}{T_b} = G_p R_b = cons\,tan\,te \qquad (8)$$

Entonces, tomando en cuenta que la respuesta del throughput es baja en la región de bajo tráfico (figura 1c), esto indica que hay pocos terminales móviles transmitiendo, por lo que es necesario aumentar la velocidad de transmisión, pero manteniendo el ancho de banda constante. Al aumentar la velocidad de transmisión hay mayor flujo de bits de datos, y cuyo resultado será un aumento en el throughput. Así que se necesitan esquemas que nos permitan transmitir con diferentes velocidades, tales como multi-ganancia de procesado, multi-códigos y multi-modulación. Por lo que fue necesario analizar cual de estos esquemas es el óptimo para adaptarlo al sistema S-Aloha/DS-CDMA.

IV. ESQUEMAS DE MULTI-VELOCIDADES EN SISTEMAS DS-CDMA.

4.1 MULTI-MODULACIÓN

Con el esquema de multi-modulación la ganancia de procesado es la misma para todas las modulaciones, en donde la energía de bit es unitaria $E_b=PT_b=1$ y la relación señal a ruido debe ser la misma, así que al introducir modulaciones multinivel aparece la componente en cuadratura, que se manifestará como interferencia al terminal móvil de referencia, tanto en su rama en fase como en su rama en cuadratura, por lo que a pesar de que se puede aumentar el flujo de bits al canal radio aumenta el nivel de interferencia [3] 4].

4.2 MULTI-GANANCIA DE PROCESADO

En un esquema de multi-ganancia de procesado, se puede tener un sistema con distintas velocidades de transmisión y diferentes grados de protección a interferentes. Esto es, para una velocidad de transmisión ν hay una ganancia de procesado G_p, manteniendo constante el ancho de banda. Si se quiere aumentar la velocidad de transmisión, ya sea 2ν, 4ν u otra velocidad mayor, se debe disminuir la ganancia de procesado en ½, ¼, etc., por lo que $\alpha\nu \rightarrow G_p/\alpha$ donde α=1, 2, ...
Si los terminales móviles transmiten a una velocidad $\alpha\nu$ se debe conservar la misma relación señal a ruido, además el valor unitario en el nivel de energía por bit $E_b=PT_b=1$, por consiguiente se recibe un nivel α veces superior de potencia y para mantener el nivel unitario en la energía de bit es necesario que la duración de bit se reduzca α veces. Por lo que la relación señal a interferente resulta [5] 6][7].:

$$\frac{E_b}{N_o}=\frac{3G_p/\alpha}{2(n-1)}=\frac{1}{\alpha\dfrac{2(n-1)}{3G_p}} \tag{9}$$

Esta expresión nos indica que al transmitir con una velocidad de $\alpha\nu$ la ganancia de procesado disminuye en un factor de α, o que el nivel de interferencia aumenta en una proporción de α porque se transmite α veces más potencia. (hace falta algunas referencia para este esquema)

4.3 MULTI-CÓDIGOS

En este esquema, cuando un usuario necesita transmitir (y es permitido por la estación base) *m* veces la velocidad básica ν, convierte su flujo de bits de serie a paralelo, y con esto podrá transmitir sus paquetes con un múltiplo entero de la velocidad básica hasta un máximo de mν, con diferentes secuencias código y con la misma ganancia de procesado [8] 9][10].

Con la transmisión en paralelo se tendrá transmisiones a la velocidad básica ν y cada una de ellas aportará un nivel de interferencia de $\frac{2}{3G_p}$ para BPSK ($2\frac{2}{3G_p}$ para QPSK). Además por la subconcatenación de códigos, existe ortogo-

nalidad entre los códigos del mismo terminal móvil, por lo que α de los códigos no tendrán interferencias, esto es

$$\frac{E_b}{N_o} = \frac{1}{\frac{2(m\alpha - \alpha)}{3G_p}} = \frac{1}{\alpha \frac{2(m-1)}{3G_p}} \tag{10}$$

esta expresión es idéntica a la obtenida para multi-ganancia de procesado.

De acuerdo a lo anterior, se puede concluir que:

Los tres esquemas tratados provocan un aumento en el nivel de interferencia. Pero hay que mencionar que con el esquema de multi-modulación la interferencia aumenta debido a que en el receptor se presenta la interferencia tanto en la componente de fase, como en cuadratura. En cambio, en multi-ganancia de procesado el nivel de interferencia se incrementa porque se reduce la ganancia de procesado. Cuando usamos multi-códigos, aumenta el nivel de interferencia debido a que aumenta la potencia transmitida.

Por lo anterior, dada la mala relación señal a interferente el esquema de multi-modulación es descartado. Falta por determinar cual de los dos esquemas restantes es el óptimo para adaptarlo a DS-CDMA.

De acuerdo a los trabajos reportados por las prestaciones de los esquemas de multi-ganancia de procesado y multi-códigos son casi las mismas. No obstante:

- Una desventaja del esquema multi-código es que los terminales móviles transmitiendo a una alta velocidad necesitan de amplificadores muy lineales.
- En multi-ganancia de procesado cuando se transmite un paquete puede suceder que sea recibido todo el paquete o que se deseche todo el paquete. En cambio, en multi-códigos la información se divide en α paquetes distintos, por lo que se pensaría que al usar multi-códigos se tendría un throughput mayor que con multi-ganancia de procesado pero no es así, porque en multi-códigos cuando uno de los paquetes es decorrelado por un receptor y se presenta un error, a todos los demás paquetes les sucederá lo mismo.
- La velocidad de transmisión en multi-códigos se ve limitada a transmitir en múltiplos de la velocidad básica, en cambio con multi-ganancia de procesado esto no sucede porque hay mayor flexibilidad en el manejo de α.

Como se quiere un sistema MAC que sea adaptable a las condiciones del tráfico y dadas las características de transmisión se opta por trabajar con multi-ganancia de procesado. Por lo tanto, con el esquema de multi-ganancia de procesado, se tendrá un sistema con diversas velocidades de transmisión y distintos grados de protección a interferencias, dando pie a desarrollar algoritmos que sean capaces de seleccionar la velocidad de transmisión más adecuada las condiciones de carga del canal. A continuación se presentan las prestaciones de S-Aloha/DS-CDMA adaptable a las condiciones del tráfico en el canal radio, considerando un esquema de multi-ganancia de procesado.

V. MODELO DE S-ALOHA/DS-CDMA ADAPTATIVO A LAS CONDICIONES DEL TRÁFICO EN EL CANAL

Como primer paso para lograr que S-Aloha/DS-CDMA sea adaptativo a las condiciones del tráfico en el canal, es necesario obtener la mejor combinación de velocidades cuando estén presentes n terminales móviles al mismo tiempo. Esto se lo logra después de una búsqueda exhaustiva:

$$\max_{(n_v,n_{2v},n_{4v})} \quad S(n_v,n_{2v},n_{4v})$$
$$\text{sujeto a } n_v+n_{2v}+n_{4v}=n \tag{11}$$

donde n_v indica el número de terminales móviles transmitiendo a una velocidad de v b/s, n_{2v} el número terminales móviles usando $2v$ b/s, n_{4v} el número de terminales móviles con velocidad de $4v$ b/s, n el número total de terminales móviles simultáneos, y S es el throughput usando la combinación de las velocidades de transmisión. Entonces:

$$S(n_v,n_{2v},n_{4v})=n_v \times \left[P_{c,(v)}(n_v,n_{2v},n_{4v})\right] \times L + n_{2v} \times \left[P_{c,(2v)}(n_v,n_{2v},n_{4v})\right] \times 2L + \left[P_{c,(4v)}(n_v,n_{2v},n_{4v})\right] \times 4L \tag{12}$$

Tomando en cuenta el análisis hecho de cadenas de Markov en la Sección II, podemos evaluar las prestaciones óptimas del sistema, pero se debe hacer algunas modificaciones. La expresión usada para evaluar la BER debe ser modificada, con objeto de tomar en cuenta de que el canal radio será compartido por terminales móviles con diferentes velocidades, además la probabilidad de éxito del paquete bajo un control de potencia perfecta puede ser calculada como:

$$\left(\frac{E_b}{N_o}\right)_{\alpha v} = \frac{1}{\frac{2}{3G_p}\left[\sum_{\substack{i=1 \\ i\neq\alpha}}^{K} i \text{ x } n_{iv} + \alpha \text{ x } (n_{\alpha v} - 1)\right]} \tag{13}$$

por lo que la probabilidad de error en el bit es [11][12]:

$$P_{b,(\alpha v)}(n_v, n_{2v}, \ldots, n_{4v}) = Q\left(\sqrt{2\left(\frac{E_b}{N_o}\right)_{\alpha v}}\right), \tag{14}$$

y la probabilidad de recibir correctamente un paquete es [11][12]:

$$P_{c,(\alpha v)}(n_v, n_{2v}, \ldots, n_{Kv}) = \left[1 - P_{b,(\alpha v)}(n_v, n_{2v}, \ldots, n_{Kv})\right]^{\alpha L} \tag{15}$$

VI. ALGORITMOS PARA EL CAMBIO EN LA VELOCIDAD DE TRANSMISIÓN

En esta sección se presentan tres algoritmos para el cambio en la velocidad de transmisión. De los tres algoritmos uno es controlado por el terminal móvil y dos por la estación base. La idea de estos algoritmos es sensar el tráfico del canal con objeto de acomodar adecuadamente las velocidades de transmisión.

6.1 ALGORITMO CONTROLADO POR EL TERMINAL MÓVIL –MS

Este algoritmo es llevado a cabo por el terminal móvil y trabaja de la siguiente manera: cada terminal traza su propia evolución durante el tiempo de transmisión, esto es, los terminales móviles cuentan sus paquetes con éxito o erróneos. En la ausencia de errores, el terminal móvil asumirá una carga de tráfico baja y prueba una velocidad de transmisión más alta. Si ocurren errores, el terminal móvil decide que el canal está cargado demasiado y prueba una velocidad de transmisión más baja.

Específicamente, el terminal móvil necesita establecer solamente dos parámetros: el número de paquetes consecutivos con fallas antes de cambiar a una velocidad más baja (max_tr), y el número consecutivo de paquetes con éxito

antes de probar una velocidad más alta (min_suc). Estos parámetros pueden ser adaptables de acuerdo al tráfico del canal.

6.2 ALGORITMO CONTROLADO POR LA ESTACIÓN BASE –BS

- ***Algoritmo BS-I***: En este algoritmo la estación base decide que velocidades se pueden usar por medio de valores umbrales. Esto es, el umbral L_1 indica que para $n<L_1$ todos los terminales móviles usarán $4v$ b/s en el siguiente slot, L_2 indica que para $L_1<n<L_2$ todos los terminales móviles usarán $2v$ durante el siguiente slot, y para $n>L_2$ todos los terminales móviles usarán v b/s.

- ***Algoritmo BS-II***: Una mejora en las prestaciones obtenidas con el algoritmo BS-I, puede obtenerse si la estación base conoce con anticipo, el número exacto de terminales móviles listos para transmitir en un slot dado. Para este propósito, el slot es dividido en dos partes: la primera parte del slot es usado para indicar que un paquete es programado para transmitir, y en la segunda parte la información es transmitida eventualmente. Con esto, la estación base cuenta cuántos terminales móviles intentarán transmitir y selecciona la combinación óptima de las velocidades a ser empleadas.

VII. SIMULACIÓN Y ANÁLISIS NUMÉRICO.

La simulación es hecha para un entorno celular en exteriores, en donde se asume que cada paquete está contenido en el slot. Los paquetes de datos para cada terminal móvil son generados de acuerdo a un proceso de Bernoulli con probabilidades de generación de 10^{-3} a 1. Para cada probabilidad de generación de 10000 slots han sido transmitidos con objeto de obtener los valores promedios, donde la probabilidad de transmitir un paquete nuevo es 1, independientemente del valor actual de retransmisión. Adicionalmente, en la tabla III se consideran los parámetros de simulación.

Tabla III. Parámetros de simulación

PARÁMETRO	CANTIDAD
Población de terminales móviles - N	80
Número de códigos - K	25
Longitud del paquete - L	200 bits
Ganancia de procesado- G_p	128

Se presentan dos casos en nuestra simulación:

1. Velocidad de transmisión constante.
2. Algoritmo controlado por el terminal móvil.

7.1 VELOCIDAD DE TRANSMISIÓN CONSTANTE

Antes de simular el algoritmo de adaptación a las condiciones del tráfico controlado por la estación móvil, es necesario evaluar las prestaciones para el throughput para velocidades de transmisión constante. Para este caso, consideramos tres velocidades 1ν, 2ν y 4ν ($\alpha\nu$, donde α=1,2,4), ganancia de procesado de 128, y una longitud de paquete de 200 bits. Para obtener los resultados es necesario variar la ganancia de procesado dependiendo la velocidad a la que se esté transmitiendo ($\alpha\nu \rightarrow Gp/\alpha$). Los resultados obtenidos para el throughput con distintas velocidades se muestran en la figura 5.

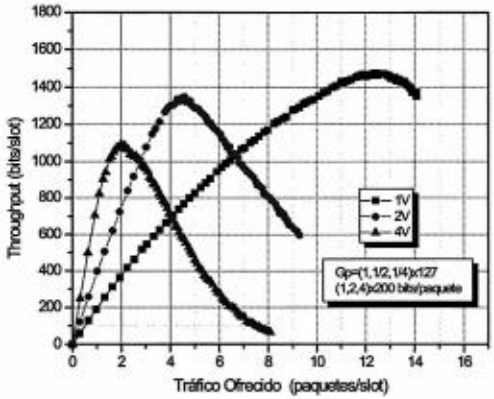


Figura 5. Comportamiento del Throughput para distintas velocidades de transmisión.

De la figura 5 observamos que de acuerdo con la velocidad básica (1v), el throughput es bajo en la región de bajo tráfico, este comportamiento es debido a que hay pocos terminales móviles transmitiendo y no porque haya demasiada interferencia entre terminales móviles. Si aumentamos al doble la velocidad de transmisión 2v la ganancia de procesado se reduce a la mitad $G_p/2$, esto es posible porque no tiene caso que se tenga una G_p alta cuando hay pocos terminales móviles intentando transmitir.

Nos damos cuenta además, que cuando se aumenta la velocidad de transmisión el throughput disminuye, esto es porque el throughput está en función de la probabilidad de recibir correctamente un paquete P_c y ésta a la vez de la probabilidad de error en el bit, por lo que si se aumenta la velocidad de transmisión la probabilidad de recibir correctamente un paquete disminuye y esto se ve reflejado en el throughput. De la misma manera sucede cuando se sigue aumentando la velocidad. Lo importante de esto es que se pueda usar distintas velocidades dependiendo del tráfico en el canal, por lo que se necesitan algoritmos que maneje dinámica las velocidades de transmisión y que se adapten a las condiciones del tráfico.

Ahora es necesario evaluar otro parámetro importante, que es el retardo. El retardo es el tiempo (en slots) que tarda el terminal móvil en generar un paquete y transmitirlo con éxito. Este comportamiento lo observamos en la figura 6.

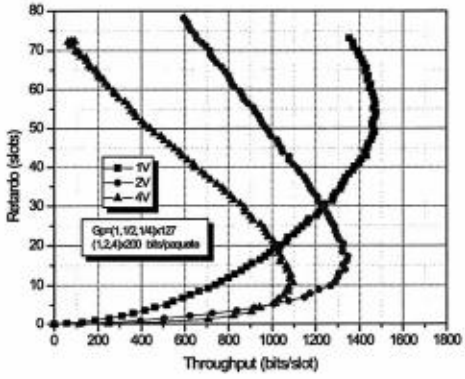


Figura 6. Comportamiento del retardo para distintas velocidades de transmisión.

En la figura 6 observamos el retardo es alto en la región de bajo tráfico cuando se utiliza una velocidad 1v, esto es porque se presentan demasiadas colisiones. Hay que precisar que las colisiones ocurren cuando dos o más terminales móviles poseen el mismo código. Al aumentar la velocidad se aumenta el número de bits en la transmisión, pero lo importante para este caso, es que el retardo disminuye

debido a que como hay pocos terminales móviles, por lo tanto, la probabilidad de que asignen códigos iguales disminuye. Con respecto a la interferencia multiusuario, se puede decir que no afecta demasiado ya que se ha mencionado que hay pocos terminales móviles.

7.2 ALGORITMO CONTROLADO POR EL TERMINAL MÓVIL – MS

Hasta este momento solo se ha mencionado el comportamiento del throughput y el retardo para diferentes velocidades de transmisión. Pero el objetivo es obtener una técnica MAC (Medium Access Control) que sea adaptable a las condiciones del tráfico en el canal radio. Usando el algoritmo MS, el terminal móvil aumenta o disminuye su velocidad de transmisión dependiendo el número de éxitos o fallos consecutivos. El comportamiento del throughput para distintos casos se muestra en la figura 7.

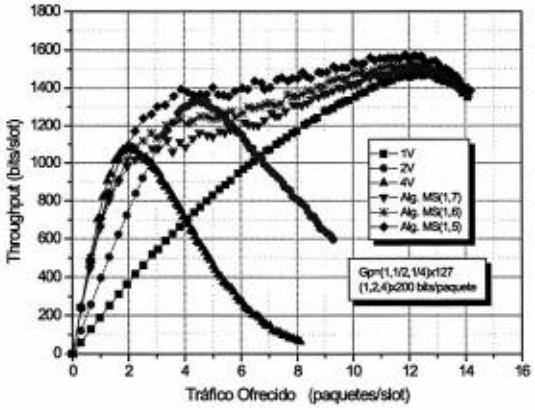


Figura 7. Comportamiento del throughput con el algoritmo MS.

Tres casos son tratados para el algoritmo MS (figura 7): MS(1,7), MS(1,6) y MS(1,5). La dupla MS(max_tx, min_suc) nos indica para max_tr cuantas fallas consecutivas debe haber antes de disminuir la velocidad de transmisión, mientras que min_suc indica cuantos éxitos consecutivos debe haber antes de aumentar la velocidad de transmisión. Hay que mencionar que la velocidad máxima es 4v y la mínima 1v. En la misma figura observamos que la dupla M(1,5) tiene el mejor comportamiento ya que se adapta mejor a las condiciones de tráfico. Una ventaja de este algoritmo es que es manejado por el terminal móvil, por consiguiente no necesita ninguna información de la estación base para aumentar o disminuir su velocidad de transmisión. Pero la desventaja es que se necesita hacer varias

pruebas para encontrar la mejor alternativa para la dupla M(max_tx, min_suc). Al incrementar max_tx (>1) lo que produce es retrasar la capacidad del algoritmo. Reduciendo min_suc (<5) resulta elevar el valor óptimo, mientras con min_suc>5 la adaptación al tráfico sería menor. Se ha obtenido una repuesta en el throughput adecuada, pero es necesario obtener el comportamiento del retardo para determinar si el algoritmo cumple con lo propuesto: adaptabilidad de la carga al sistema y un retardo bajo. Por lo que en la figura 8 se muestra el comportamiento obtenido del retardo para el caso de MS(1,5).

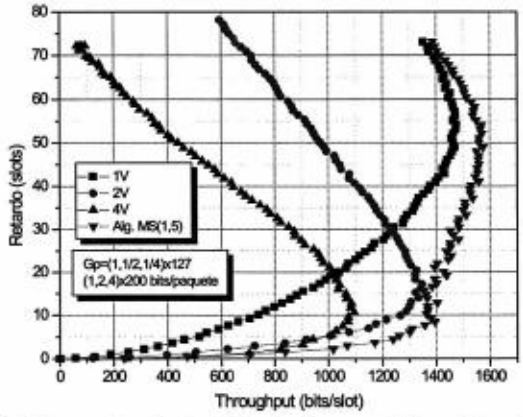


Figura 8. Comportamiento del retardo con el algoritmo MS(1,5).

En la figura 8 observamos que el retardo permanece bajo en la región de bajo tráfico, por lo que se cumple con el objetivo de disminuir el retardo. El comportamiento que se obtiene es porque cuando hay pocos terminales móviles, el número de colisiones se reduce y la interferencia multiusuario es mínima, entonces transmiten con una velocidad alta y el propio algoritmo ajusta la velocidad dependiendo que sucede en la transmisión. Así que este algoritmo es fácil de implantar ya que solo se necesita un contador en el terminal móvil y éste controla su velocidad de transmisión.

VIII. CONCLUSIONES

En este trabajo se analizó el esquema S-Aloha/DS-CDMA donde se observaron prestaciones limitadas tanto en la técnica de acceso como en la carga del sistema. Al usar S-Aloha/DS-CDMA visualizamos que en la región de bajo tráfico donde hay pocos terminales móviles transmitiendo el throughput es bajo y el retardo alto, para mejorar esta respuesta e incrementar el throughput es necesario au-

mentar la velocidad de transmisión manteniendo un ancho de banda constante pero disminuirla cuando se generen demasiadas colisiones, por lo que se es necesario adicionar al sistema el manejo de multi-Velocidades para obtener mejores prestaciones y un sistema adaptable a las condiciones del tráfico.

Para el manejo de distintas velocidades de transmisión dependiendo del tráfico en el canal se necesitan algoritmos que manejen de manera dinámica las velocidades de transmisión y que se adapten a las condiciones del tráfico.

En este trabajo se analizó el algoritmo controlado por la *Estación Móvil* el cual permite aumentar o disminuir la velocidad en función del número de éxitos o fallos consecutivos de paquetes transmitidos.

Con este trabajo se puede apreciar las posibilidades para modificar el comportamiento del esquema S-Aloha/DS-CDMA y así lograr una mayor eficiencia, ya que se trataría de conseguir que las prestaciones del esquema estuviesen limitadas por la carga que soporta y no por el protocolo de acceso en sí.

REFERENCIAS

1) LIU, Z. and El Zarki, M "Performance Analysis of DS-CDMA with Slotted Aloha Random Access for Packet PCNs" Proceedings of the 51st IEEE International Symposium on Personal Indoor an Movile Radio Communications Vol 4,pp.1034-1039,1994.
2) LIU, Z. and El Zarki, M "Performance Analysis of DS-CDMA with Slotted Aloha Random Access for Packet PCNs" Wireless Network ,pp.1-16. 1995.
3) Ottoson , T. and Svensson, A. "Multirate Schemes for Multimedia Aplications in DS-CDMA Systems", Proceedings of the Conference on Rafio Sciences and Telecommunications, 1996
4) Ottoson , T. and Svensson, A. "Multi-Rate Schemes in DS-CDMA Systems", Proceedings of the 45th IEEE Vehicular Technology Conference , Vol. 2 , pp 1006-1010 , 1995
5) Oh, S-J. and Wasserman K. M., "Dynamic Spredaing Gain Control in Multiservice CDMA Networks,"*IEEE Journal on Selected Areas in Communications*, Vol. 17, No. 5, pp.918-927, May 1999.
6) Lyu, D., et. al., "Performance Evaluation of a Variable Processing Gain DS/CDMA System,"*IEICE Transactions Fundamentals*, vol. E80-A, no. 2, pp. 393-399, February 1997.
7) I., C-L., and Sabnani, K. K., "Variable spreading gain CDMA with adaptive control for integrated traffic in wireless networks,"*Proceedings of the 45th IEEE Vehicular Technology Conference*, vol. 2, pp. 794-795, 1995.

8) I, C-L. and Gitlin, R. D., "Multi-Code Wireless Personal Communications Networks,"*Proceedings of the International Conference on Communications*, vol. 2, pp.1060-1064, 1995.

9) I., C-L., et. al., "Performance of Multi-Code CDMA Wireless Personal Communications Networks,"*Proceedings of the 45th IEEE Vehicular Technology Conference*, vol. 2, pp. 907-911, 1995.

10) Schotten, H. D., et. al., "Adaptive Multi-Rate Multi-Code CDMA Systems," *Proceedings of the 48th IEEE Vehicular Technology Conference*, vol. 2, pp. 782-785, 1998.

11) Pursley, M. B., "Performance Evaluation for Phase-Coded Spread-Spectrum Multiple Access Comunications- Part I; System Analysis", IEEE Transactions on Communications, Vol 25. No. 8,pp 795,799, August 1977.

12) Pursley, M. B. and Sarwate , D. V., "Performance Evaluation for Phase-Coded Spread-Spectrum Multiple Access Comunications- Part II; Code Sequence Analysis", IEEE Transactions on Communications, Vol 25. No. 8,pp 800,803, August 1977.

6. Contribución a la Provisión de Servicios Multimedia con Calidad de Servicio Extremo a Extremo en Entornos Inalámbricos.

Tania Y. Guerrero Melendez, Luis J. de la Cruz Llopis

Parte I: "Entorno wireless basado en la norma IEEE-802.16: VoD"

Abstract

In the last few years multimedia services have had rapid growth, such as Voice over IP, video conferencing, video on demand and online gaming. The most important factor behind this rapid growth is the increasing availability of broadband access to commercial and residential environments. At the same time, the population was familiar with wireless and mobile devices. Nowadays, users all over the world have become more accustomed to the availability of broadband wireless access (BWA). This phenomenon has increased the use of multimedia applications in wireless and wired networks. Among all these services offered by the communication networks, currently, video on demand (VoD) is one of the main multimedia services used by the customers. All the multimedia communications have strict network requirements and to give a good service it is necessary to satisfy these requirements. The multimedia traffic is time-sensitive and has wide bandwidth requirements.

The current trend in the development of multimedia (real-time) internet applications and the rapid growth of mobile systems indicate that the future internet architecture will need to support several applications with different quality of service (QoS) requirements.

As mentioned above, the successful deployment of multimedia services requires accessing networks which provide the strictest service requirements ne-

cessary to this kind of applications. One of the challenges for broadband wireless access networks is to provide QoS for services with different characteristics. IEEE 802.16 technology intends to provide wireless broadband connectivity in a metropolitan environment, for mobile and fixed users, using a well defined quality of service framework. For this reason, the 802.16 or WiMAX technology will be addressed in this document as a suitable access technology.

1. CAPITULO INTRODUCTORIO. TRANSMISIONES INALÁMBRICAS

Las nuevas tecnologías de acceso a la red han aportado la movilidad, flexibilidad y comodidad que en un principio era impensable, debido a las limitaciones propias del cable. Por esta misma razón es que durante los últimos años se ha presentado un gran despliegue de este tipo de tecnologías, el cual ha permitido la generación de nuevos modelos de acceso a internet.

Existen múltiples tecnologías de transmisiones inalámbricas, las cuales pueden ser clasificadas en base al rango de cobertura, como redes de área personal (PAN), de área local (LAN), de área metropolitana (MAN) y de área extensa (WAN).

Las redes inalámbricas poseen dos topologías de trabajo, (1) centralizadas o basadas en infraestructura y (2) distribuidas o ad-hoc. Una red basada en infraestructura tiene una instalación planeada de dispositivos de red, en donde un nodo se conecta a la red mediante un punto de acceso o una estación base. Generalmente las redes inalámbricas operan con una topología de infraestructura. En estas redes todos los nodos que se encuentran dentro del rango de cobertura del punto de acceso/estación base se conectan a él, y es a través de éste que tienen acceso al backbone de la red. Esto significa que todas las comunicaciones desde o para un punto dentro del rango de transmisión de la estación base pasarán a través de él. En las redes con esta topología de trabajo, los puntos de acceso o estaciones base son de crítica importancia en el aspecto de conectividad para éste tipo de redes, pues el mal funcionamiento o la ausencia repentina de alguno de estos nodos provocarían una fractura en la red.

Por otro lado se encuentran las redes distribuidas o ad-hoc, a diferencia de las redes centralizadas, las redes ad-hoc no cuentan con una instalación de dispositivos planificada, es decir, en estas redes los dispositivos inalámbricos se conectan dinámicamente entre ellos. La principal ventaja en las redes ad-hoc es la flexibilidad y la robustez que presentan, pues el fallo o la ausencia de un nodo no presenta problemas de conectividad alguno.

Las redes inalámbricas, al igual que el resto de redes de comunicaciones están normalizadas por diferentes estándares que facilitan la interacción entre dispositivos de distintos fabricantes.

1.1 NORMAS IEEE

Brindar movilidad y conectividad total se ha convertido en el objetivo de muchas investigaciones a nivel mundial. Esto ha provocado que las tecnologías inalámbricas existentes desarrollen otras capacidades para brindar diferentes servicios además de los tradicionales. Así mismo, han ido surgiendo tecnologías para competir con las existentes o complementar sus capacidades y alcances.

El Instituto de Ingenieros Eléctricos y Electrónicos (IEEE), organismo de reconocido prestigio a nivel internacional en el área de las telecomunicaciones, tiene en su haber dos de los estándares mayormente divulgados dentro del área de las redes inalámbricas tanto a nivel local como metropolitano. Dichos estándares son el 802.11 y el 802.16 respectivamente, en sus múltiples versiones, estándares a los cuales haremos referencia a lo largo de este documento.

NORMAS IEEE PARA REDES INALÁMBRICAS A NIVEL METROPOLITANO

En el marco de las redes de área metropolitana surge el estándar IEEE-802.16, concebido como tecnología de acceso de banda ancha inalámbrica de última milla, cuyas expectativas estaban orientadas hacia la provisión de cobertura en zonas rurales y con pocos accesos. A diferencia del 802.11, este 802.16 surge con un diseño que le permite afrontar el soporte de parámetros QoS mediante mecanismos de gestión y control [8].

De esta tecnología existen dos vertientes, la versión fija y la móvil ver (Figura 1). La versión fija fue ratificada por el instituto de ingenieros eléctricos y electrónicos (IEEE) en junio del año 2004. Actualmente IEEE-802.16 en esta versión fija, puede alcanzar velocidades de 72 Mbps y distancias de 50 o 15 kilómetros en línea de vista o sin línea de vista respectivamente. La versión móvil o 802.16e, está preparada para funcionar en ambientes vehiculares con velocidades de movimiento de hasta 75 millas/h (±120km/h) [14],[10], [3]. En los entornos móviles unos de los aspectos más significativos a tomar en consideración es el consumo de energía. Consientes de ello, además de la movilidad soportada por esta versión, contempla el aspecto de ahorro de energía con el propósito de extender el tiempo de vida de la batería de los dispositivos móviles.

Existe una entidad no lucrativa denominada WIMAX Forum, formada en el año 2003 por empresas de diversos sectores de las comunicaciones (operadores, fabricantes de hardware, entre otros). El objetivo de este foro es promover la interoperabilidad entre los diferentes productos de BWA (Broadband Wireless Access), así como también acelerar el despliegue de soluciones inalámbricas. Para cumplir con estos objetivos, este foro ha creado un certificado que avala la compatibilidad entre dispositivos de banda ancha. Para que determinado dispositivo obtenga este certificado es necesario que cumpla con el estándar IEEE-802.16.

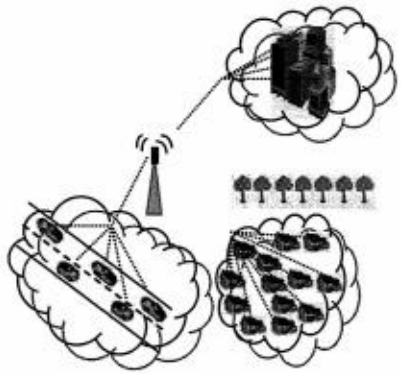


Figura 1. Aplicaciones del 802.16.

El competidor más cercano de Wimax en Europa viene de sistemas inalámbricos ampliamente desplegados como UMTS (Universal Mobile Telecommunications Service), quien proporciona además de la telefonía móvil de tercera generación (3G) el servicio de acceso DSL [15]. La ubicación de Wimax frente a otras tecnologías de acceso se muestra en (Figura 2), en donde se puede observar su situación en un rango intermedio entre movilidad y velocidad.

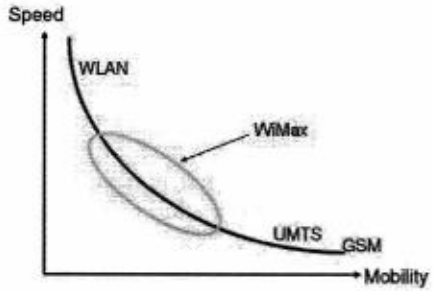


Figura 2. Situación WiMAX. Fuente: [15]

WiMAX no es una tecnología, sino, una marca de certificación dada a los equipos que cumplen con la interoperabilidad de la familia de estándares IEEE-802.16. Algo similar ocurre con WiFi para el estándar 802.11. Sin embargo, sus nombres han sido adoptados como de uso popular para hacer mención de dichas tecnologías.

Normativas IEEE para el modo mesh

En los últimos años ha surgido una nueva tecnología conocida como Wireless Mesh Networks (WMN). Esta tecnología pretende crear una interacción real entre las diferentes redes de comunicaciones. La arquitectura de red de estas WMNs está diseñada de tal manera que los nodos puedan comunicarse con otros vía multisalto o reenvío. El objetivo general de esta tecnología es extender el rango de cobertura de las redes actuales sin sacrificar capacidad de canal.

Las principales características que distinguen a estas redes son su dinamismo en la auto-configuración, auto-organización y auto-corrección [4]. Gracias a estas características es posible obtener una integración flexible, despliegue ágil, fácil mantenimiento y bajo coste, con lo cual además es posible garantizar el establecimiento y mantenimiento de la conectividad entre los nodos. Todo esto convierte a las redes mesh en una tecnología sumamente atractiva, la cual cuenta con la confianza de múltiples órganos de investigación tanto a nivel educativo como comercial. Universidades como la Carnegie-Mellon o el Instituto de Telecomunicaciones y Tecnologías de la Información de California, realizan investigaciones en este campo con la ayuda de bancos de pruebas. Así mismo múltiples compañías proveedoras de servicios brindan soluciones WMNs propietarias. Debido a los problemas de interoperabilidad que esto representa, la IEEE realiza grandes esfuerzos por crear un estándar para este tipo de redes en ambientes tanto LAN como MAN.

IEEE-802.16-mesh

El modo mesh a nivel metropolitano es incluido en el estándar IEEE 802.16-2004. En ésta tecnología existen dos tipos de nodos, llamados mesh-Base Station (mBS) y mesh Subscriber Station (mSS). Se pueden encontrar múltiples diferencias entre los modos PMP y mesh de este estándar. Una de estas diferencias es la manera en la cual se realizan las transmisiones. En el modo mesh, las transmisiones entre dos nodos son llevadas sobre enlaces bidireccionales, los cuales son establecidos durante la fase de inicialización de una nueva mSS [1]. En éste

modo mesh, cada mSS se comunica con sus vecinos sin la ayuda de una mBS. Generalmente uno o más nodos mSS toman el papel de BS para conectar la red mesh al exterior.

Cuando un nodo quiere unirse a la red mesh, necesita pasar por un proceso de anexión y autoconfiguración. Estos procesos los realiza obteniendo información de la red contenida en mensajes que han sido enviados por los nodos que conforman la red [2],[9]. Estos mensajes contienen información de configuración de la red, un ejemplo de ello es el ID de la mesh BS y el canal en uso. En las comunicaciones entre dos nodos mesh, la calidad de servicio es provista en base a mensajes, y cada mensaje contiene los parámetros de servicio en el encabezado. Los parámetros asociados con cada mensaje son comunicados junto con el contenido del mensaje mediante el MAC-SAP[1]. Un ejemplo de red 802.16 mesh se presenta en (Figura 3), en donde gracias al despliegue del modo mesh se proporciona servicio a un sector del poblado sin necesidad de instalar otra estación base.

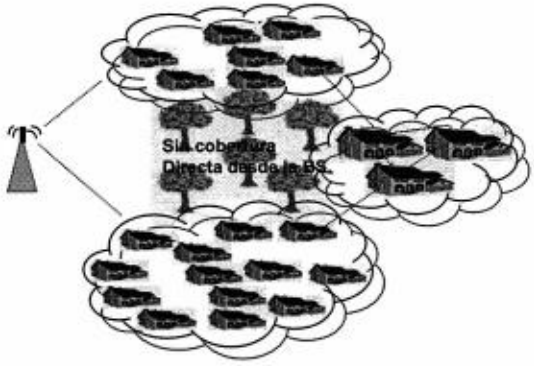


Figura 3. Despliegue mesh para proporcionar cobertura de red a una zona geográfica no cubierta por la BS.

1.1 SOPORTE DE CALIDAD DE SERVICIO

Contar con calidad de servicio en las redes de comunicaciones ya no es considerado como un valor añadido, sino, que ahora es una característica de considerable importancia. Podemos definir calidad de servicio como un concepto que mediante la acción de señalar o indicar los parámetros de transmisión pueden ser garantizadas las características pactadas para un servicio en particular. Sin embargo, la provisión de calidad de servicio es un problema que está siendo estudiado desde

1 Punto en una pila de protocolos donde los servicios de capas inferiores están disponibles para la capa superior inmediata.

algunos años atrás, tiempo durante el cual han surgido diferentes propuestas para darle solución.

Actualmente, existen dos principales maneras de proporcionar calidad de servicio. Una de ellas está basada en la reserva de recursos, mientras que la otra realiza una priorización de tráficos en base a su tipo. Estas propuestas son conocidas como Intserv y Diffserv respectivamente. Ambas propuestas han salido de los grupos de trabajo del organismo de estandarización IETF (Internet Engineering Task Force). Ambas propuestas afrontan la provisión de calidad de servicio desde el nivel de red. A este nivel se proporciona una conectividad extremo a extremo entre nodos que no se encuentran directamente conectados entre sí. Sin embargo es necesario afianzar bien el soporte de estos parámetros a un nivel inferior, es decir, proporcionar mecanismos que soporten la provisión de calidad de servicio a nivel MAC.

QoS EN ENTORNOS WIRELESS MAN

Al igual que en los estándares ethernet, la QoS en Wimax es implementada en la capa MAC. [5] Este estándar especifica dos modos de compartir el medio wireless, punto- multipunto (PMP) y Mesh. En el modo PMP, la BS sirve a un conjunto de SSs que se encuentran dentro de un mismo sector de la antena por medio de broadcast, en donde todas las SSs están recibiendo la misma transmisión de la BS. Por otro lado, en el modo Mesh, el tráfico puede ser encaminado a través de otras SSs, así como también se puede dar solamente entre SSs. Dichas conexiones son unidireccionales para el modo PMP, ya sea de Down-Link (en sentido BS a SS) o de Up-Link (de SS a BS), en donde las primeras (DL) pueden ser tanto unicast como multicast, mientas que las de UL solamente pueden ser unicast. Por otro lado en el modo mesh las conexiones están dadas de forma bidireccional, y son establecidas durante la fase de inicialización de un nuevo nodo mesh SS.

Debido a que no es factible tratar los requerimientos de QoS de todas las aplicaciones de la misma manera, wimax ofrece cuatro diferentes clases de servicio implementados en la capa MAC. Estas son: UGS, rtPS, nrtPS y BE [1]. Dentro de dichas clases la categorización se realiza en base a tres puntos característicos: (1) Requerimientos de servicio (QoS), (2) patrón de llegadas de los paquetes y (3) mecanismos de envío de peticiones de ancho de banda a las BS. Existe una quinta clase definida dentro del apartado e del estándar 802.16, denominada ertPS (extended real-time Polling Service) [11], la cual soporta flujos de servicio en tiempo real como VoIP.

La tipificación de estas clases de servicio facilita la distribución del ancho de banda entre múltiples usuarios, ya que cada conexión de UL tiene asociado un tipo de servicio. En base a estos parámetros, durante el establecimiento de una conexión se negocian las necesidades del servicio, de esta manera, una BS asigna a cada SS el ancho de banda necesario de acuerdo a sus requerimientos de retardo. Esto es, mediante una cabecera MAC la SS notifica a la BS el tamaño de cola para una conexión en específico, esperando para ser enviados. Sin embargo, mientras que esta petición de recursos se realiza por conexión, la BS los asigna a una SS. Debido a esto, una SS necesita implementar un algoritmo local que le permita redistribuir los recursos otorgados entre sus conexiones de la manera que mejor le convenga.

Tanto las estaciones base (BS) como las estaciones cliente (SS) proporcionan QoS de acuerdo al conjunto de parámetros definidos por el flujo de servicio. Cada uno de estos flujos de servicio es identificado por un SFID (Service Flow IDentifier), y si éste se encuentra activo, incluye un identificador de conexión (CID). Dentro de la capa MAC, los paquetes incluyen un CID de manera que son asignados al flujo de tráfico pertinente, en función de los parámetros QoS del servicio (UGS, ertPS, rtPS, nrtPS, BE). Estas cinco clases o tipos de servicio son detallados en (Tabla 1), en donde se hace mención al tipo de aplicación para los cuales están planeados.

Tabla 1. Clases de servicio para 802.16.

Clase de servicio	Aplicación
UGS	Voz (paquetes de tamaños fijos)
ertPS	VoIP (flujos de servicio en tiempo real)
rtPS	VoD (paquetes de tamaños variables)
nrtPS	ftp, http
BE	e-mail

Al tener una arquitectura centralizada, la asignación de ancho de banda es controlada por la estación base. Existen diferentes métodos para realizar la petición-asignación de recursos, principalmente mediante sondeo y/o petición. En ambos métodos las peticiones pueden ser incrementales o agregadas. Cuando una BS recibe una petición incremental simplemente agrega la cantidad de ancho de banda solicitada. Por otro lado, si la petición es agregada, la estación base reemplazará la cantidad de ancho de banda que está entregando por la solicitada

en esta nueva petición. Tanto los incrementos como las reducciones de ancho de banda son necesarios para todas las clases de servicio, exceptuando UGS, en donde los tamaños asignados son fijos. Debido a que ertPS toma características de UGS, las concesiones recibidas por la BS son de manera no solicitada. Sin embargo, mientras que en UGS las asignaciones son fijas en tamaño, en ertPS son dinámicas.

2. Estudio de factibilidad de transmisiones multimedia sobre redes 802.16

Todos hemos sido testigos del rápido crecimiento que han presentado las aplicaciones multimedia durante los últimos años. Uno de los principales factores que ha impulsado este mencionado fenómeno es el aumento en la disponibilidad de acceso de banda ancha. A la par de este fenómeno, la población se ha familiarizado con el uso de dispositivos inalámbricos y móviles, lo que ha dado pie a la generación de otros servicios.

Las comunicaciones multimedia tienen estrictos requerimientos de calidad de servicio, los cuales deben ser satisfechos para proveer un servicio aceptable [13], [17], [11], [16]. El tráfico multimedia es caracterizado por tener una fuerte sensibilidad al tiempo, así como también unos requerimientos de ancho de banda inelásticos. Desde el punto de vista de las comunicaciones, las aplicaciones multimedia pueden ser divididas en dos grupos: (1) aplicaciones interactivas como las comunicaciones de voz o las videoconferencias, y (2) aplicaciones de difusión como audio y video bajo demanda. Una de las aplicaciones multimedia que actualmente está captando gran interés es el video bajo demanda (VoD).

El gran reto de las redes de acceso inalámbrico de banda ancha es proporcionar QoS simultáneamente cuando coexisten servicios con características y requerimientos muy variados. La norma IEEE-802.16 permite una distribución de recursos entre diferentes usuarios de la red de acuerdo a los requerimientos de cada uno de estos. Esto es posible gracias a que cada conexión está asociada a un tipo de servicio que a su vez está ligado a un conjunto de parámetros QoS. El estándar IEEE 802.16-2004 define cuatro clases de servicio: *UGS (Unsolicited Grant Service), rtPS (real-time Polling Service), nrtPS (non real-time Polling Service) y BE (Best Effort)*. Además en el apartado e de esta misma norma se especifica una quinta clase de servicio, denominada *ertPS (extended real-time Polling Service)*.

La característica base de estas redes para otorgar garantías de QoS es la manera de trabajar, esto es, la BS coordina todas las comunicaciones, tanto de UL

como de DL. En otras palabras, un algoritmo en la BS analiza los requerimientos de las SS dependientes de ella y los traduce a número de slots en cada trama 802.16. Sin embargo, el estándar 802.16 no especifica un algoritmo para realizar esta asignación de slots, así como tampoco especifica el scheduler a utilizar.

Como cada SS envía diferente cantidad de datos y paquetes de diferente tamaño, el scheduler en la BS debe reasignar los slots en cada trama y si se transmiten 400 tramas por segundo la BS deberá tomar 400 decisiones de asignación de slots por segundo. Por este motivo creemos que el mecanismo de scheduling a utilizar debe ser simple.

La BS utiliza colas separadas para las conexiones de DL y UL- virtuales, en las cuales se colocan las peticiones de recursos de las SS, ver Figura 4. El estado en las colas de DL cambia cada vez que un paquete llega o sale de la cola, mientras que en la cola virtual de UL el estado cambia cada vez que la BS recibe una petición de recursos de una SS. Una vez teniendo la información de los recursos demandados y los requerimientos de QoS de dichas peticiones la BS asigna los slots necesarios para cubrir las peticiones recibidas, esto es tanto de UL como de DL. Los recursos son asignados en forma de burst de datos, en donde cada uno de ellos está formado por un número entero de slots. Cada uno de estos slots tiene la misma duración, la cual es dependiente de la capa física. Una vez establecida la duración de la trama, la BS no puede cambiarla, en caso contrario todas las SS adheridas a ella deberán sincronizarse nuevamente.

El encargado de asegurar los requerimientos de conexión mediante la asignación del número adecuado de slots basado en el bandwidth demandado y los requerimientos de QoS de cada petición es la BS. Cada conexión con su cola asociada es tratada como una sesión separada y dado que no hay un scheduler específico para realizar la asignación de slots y considerando a su vez que todos los slots en una misma trama son del mismo tamaño, proponemos utilizar un gestor de colas que trabaje de forma WRR, en donde el peso asignado indique el numero de slots otorgado a cada cola y cada round sea una trama.

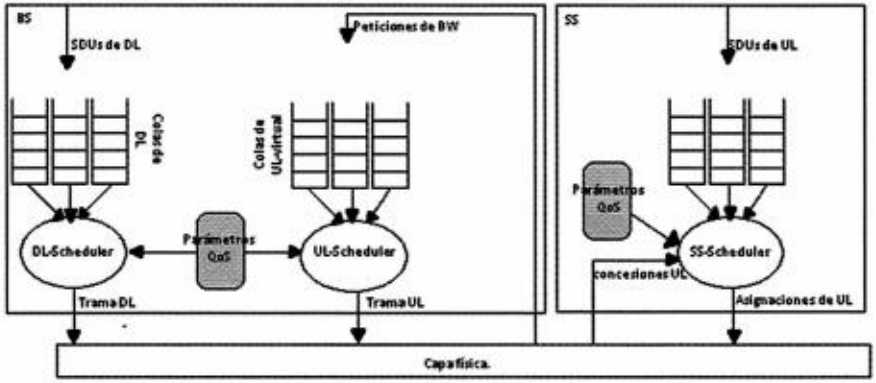


Figura 4. Mecanismos QoS dentro de las BS y SS.

2.1 TRANSMISIÓN DE MÚLTIPLES PELÍCULAS SOBRE UN MISMO CANAL

Gracias a la utilización de estas clases de servicio se encuentra viable satisfacer los requerimientos de calidad de servicio en aplicaciones multimedia como es el caso del video bajo demanda (VoD) en un entorno inalámbrico 802.16. Siguiendo en este ámbito, se ha experimentado el funcionamiento de una red basada en el estándar IEEE-802.16 cuando a través de ella se proporciona un servicio de características muy estrictas. El tráfico de estas transmisiones está compuesto por trazas reales de múltiples películas. Cabe mencionar que el proceso de preparación y mezcla de las trazas no es un proceso trivial y requiere de la consideración de múltiples factores [12]. Las herramientas de simulación son una importante manera de hacer investigación es diferentes áreas. En nuestro caso la herramienta utilizada es el simulador a nivel de dispositivo SCALEV [6][7].

El *Video Digital* está compuesto por cuadros o imágenes del mismo video que son mostrados a una determinada tasa. Existen diferentes sistemas alrededor del mundo y cada uno de ellos utiliza diferentes detalles técnicos como el número de cuadros por segundos o frames per second (fps, por sus siglas en ingles) y el número de líneas, entre otros. Los tres principales sistemas son NTSC, SECAM y PAL (National Television System Committe, Sequential Color with Memory y Phase Alternative Line respectivamente) y cada uno de ellos se utiliza en distintas zonas geográficas del mundo. NTSC es utilizado en la zona de América del Norte, Japón, Filipinas, Corea del Sur y Taiwan; SECAM utilizado en Francia, algunos países de Europa del Este y otros países de África y PAL para el resto de Europa.

Las tasas de transmisión de cuadros son distintas entre ellos, para efectos del presente trabajo el sistema que utilizamos es PAL, por lo cual se transmiten 25 fps, lo que significa que el tiempo de llegada entre paquetes es de 40 ms para una película, 20ms para dos películas y así sucesivamente, ver Figura 5.

El flujo generado cuenta con una codificación MPEG-4, el cual está formado por GoPs (Group of Pictures) que mantienen una estructura de cuadros o frames. Existen tres tipos de cuadros:

Cuadro I. Es el cuadro principal en la estructura MPEG así como también son los de mayor importancia en el proceso de decodificación, ya que sin éste se pierde todo el GoP. Son los encargados de llevar la información más relevante de las imágenes. Son codificados sin referencia a otros cuadros y solamente existe un cuadro I por cada GoP.

Cuadro P. Este tipo de cuadro contiene información relativa a las diferencias respecto al último cuadro de tipo I, su tamaño aproximado es de un 20 % respecto a un cuadro de tipo I promedio. Codifica redundancia temporal.

Cuadro B. Es el cuadro con menor importancia para el proceso de decodificación dentro del GoP. Codifica redundancia temporal y solamente lleva información relativa al último cuadro de tipo P. Su tamaño es de tan solo el 10% respecto al tamaño medio de in cuadro de tipo I.

Una secuencia codificada utilizando estos tres tipos de cuadros consigue un alto grado de compresión y una razonable accesibilidad, mientras que en una secuencia de codificación utilizando solamente cuadros de tipo I se consigue un alto grado de accesibilidad pero la compresión más baja.

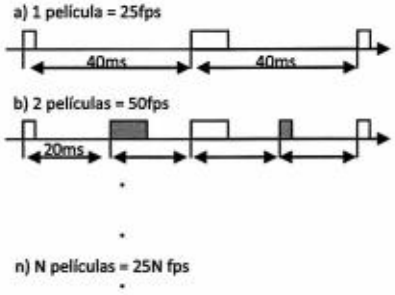


Figura 5. Tiempo entre llegadas de paquetes.

En esta parte del proyecto de investigación se realizo una serie de transmisiones enfocándose en una sección específica, esto es, las transmisiones no fueron hechas extremo a extremo, solo se consideró la sección de red entre el punto de distribución de contenidos y un dispositivo en la red. En este trabajo, se buscó obtener la capacidad requerida para proveer un servicio con fuertes restricciones como lo es la entrega de servicios multimedia. Los resultados obtenidos son mostrados en Tabla 2 y Tabla 3.

Estas transmisiones se realizan sobre un canal basado en la norma 802.16 que puede contar con una capacidad de hasta 70 Mb, a través del cual se transmiten diferentes películas en modo bajo demanda desde un servidor de video instalado en un SS. Este trabajo se realiza con el objetivo de observar las necesidades de recursos que se van teniendo conforme el número de películas demandadas aumenta. Esto es visto desde la situación en la cual el servicio es satisfecho desde un mismo servidor.

Como se menciona anteriormente, el escenario simulado es la transmisión de diferentes películas, en modo video bajo demanda,

Cada cuadro de las películas utilizadas es fragmentado y encapsulado agregándole a cada fragmento los encabezados MAC correspondientes. La razón de fragmentar los frames de cada película es que dado que el tamaño de estos cuadros puede superar el tamaño de los slots, éstos tienen que ser segmentados para que se pueda realizar su transmisión sobre una trama 802.16, después de realizada la transmisión de los segmentos del frame, se vuelve a unir cada uno de esos fragmentos que forman el frame completo. En la Figura 6 se muestra la estructura de estas tramas, en ella se puede observar los encabezados que lleva cada MAC PDU.

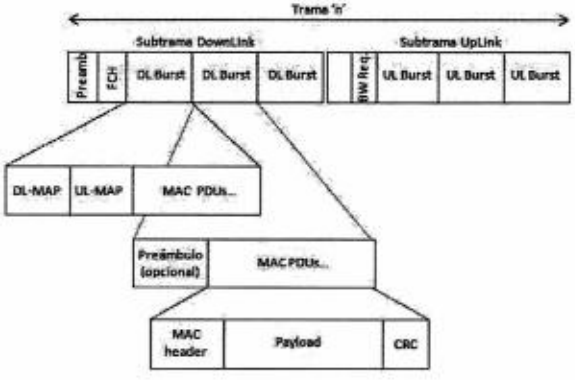


Figura 6. Estructura de las tramas 802.16.

Las transmisiones de estas películas mediante el simulador SCALEV, fueron hechas variando la capacidad de canal y el número de películas, esto con el objetivo de identificar las capacidades de canal necesarias para lograr unos tiempos de transmisión inferiores a los 100 y 50 milisegundos. En el conjunto de simulaciones realizadas se utilizaron diferentes bloques de películas, los cuales van desde 1 hasta 80 películas en un mismo bloque. En cada una de estas transmisiones se buscó obtener los tiempos de transmisión inferiores a los 100 milisegundos para el 90, 95 y 99 por ciento de los paquetes transmitidos utilizando diferentes capacidades de canal. Esto mismo fue realizado buscando los tiempos de transmisión de 50 milisegundos para los tres percentiles mencionados. Como resultado, en la Tabla 2 y Tabla 3 son mostradas las capacidades necesarias para obtener los tiempos de transmisión requeridos por la aplicación a utilizar.

Tabla 2. Capacidad de canal respecto al número de películas transmitidas con tiempos de transmisión inferiores a los 100ms para el percentil 99.

No. de películas	10	20	30	40	50	60	70	80
Capacidad de canal Mbps	7.3	13	18.25	24	29.25	34.5	38	45

Tabla 3. Capacidad de canal respecto al número de películas transmitidas con tiempos de 50ms.

No. de películas	1	5	10	20	30	40	50	60	70	80
P 90 Mbps.	1.35	3.4	6.48	12.1	17.45	22.75	28	33.7	39.05	44.26
P 95 Mbps	1.55	3.5	6.735	12.4	17.78	23.2	28.75	34.2	39.55	44.7
P 99 Mbps	1.9	3.9	7.35	13.05	18.35	24.05	29.5	34.85	40.15	45.48

Algunos de los resultados obtenidos son mostrados de manera grafica en la Figura 5 y Figura 6. En estas gráficas podemos observar el comportamiento que tienen los paquetes en una capacidad de canal determinada. Así mismo es fácilmente apreciable como esta capacidad tiene un crecimiento directamente relacionado al número de películas. En estas mismas figuras, así como en las Tabla 2 y Tabla 3 se puede apreciar que con una mínima diferencia en la capacidad de canal se pueden tener unos tiempos de transmisión adecuados o muy altos.

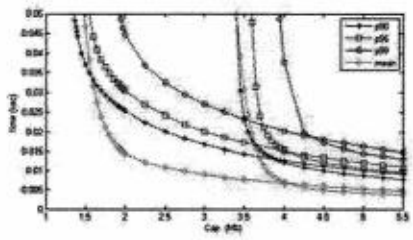


Figura 7. Transmisión de 1 y 5 películas.

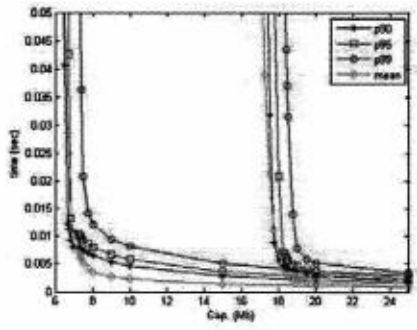


Figura 8. Transmisión de 10 y 30 películas.

Podemos mencionar que establecer transmisiones de video bajo demanda en este tipo de redes es totalmente factible, ya que la calidad de servicio requerida puede ser satisfecha gracias a la utilización de las adecuadas clases de servicio definidas dentro de la norma 802.16 para el modo PMP. Cabe mencionar que este estudio se encuentra en proceso y lo que se presenta en este documento es un resumen de los avances que se tienen hasta el momento de su presentación.

3. CONCLUSIONES

Las redes de comunicaciones inalámbricas basadas en la norma IEEE 802.16 proporcionan una oportunidad para ofrecer un amplio abanico de aplicaciones en diferentes entornos. Además de otros múltiples servicios, VoD es uno de los servicios que pueden ser provistos en zonas en donde las redes de acceso convencionales no existan. En este documento se ha abordado el tema de la transmisión de servicios de video bajo demanda en un escenario en el cual se emplea esta tecnología. Dicha tecnología representa una considerable opción gracias a sus versátiles características de diferenciación en cuanto a servicios se refiere.

Dado que al momento de proveer un servicio de este tipo la calidad percibida por el usuario final es de suma importancia, en este documento adoptamos la calidad de servicio como la habilidad para asegurar al cliente una buena experiencia, es decir, el tiempo que tarda en tener el video completo así como la fluidez con que se observa y escucha.

Existen múltiples mecanismos y arquitecturas con las cuales optimizar los tiempos de transmisión de video bajo demanda. Thouin y Coates en [13] publican su trabajo en referencia a estas arquitecturas. Sin embargo en nuestro trabajo se observó un sistema de transmisión enfocándonos en un dispositivo en concreto.

REFERENCIAS BIBLIOGRÁFICAS.

[1] IEEE 802.16-2004, IEEE Standard for Local and Metropolitan Area Networks Part 16: Air Interface for Fixed Broadband Wireless Access Systems. IEEE, Oct. 2004.

[2] IEEE 802.11e Standard for Information Technology Telecommunications and Information Exchange between Systems - Local and Metropolitan Area Networks - Specific Requirements. Part 11: Wireless LAN Medium Access Control (MAC) and Physical (PHY) Specifications. Amendment 8: Medium Access Control (MAC) Quality of Service Enhancements. IEEE Computer Society, Nov. 2005.

[3] Z. Abichar, Y. Peng and J-M. Chang, "WiMAX: The Emergence of Wireless Broadband", IEEE Computer Society, Jul/Aug. 2006.

[4] I. Akyildiz, X. Wang and W. Wang, "Wireless Mesh Networks: a survey". Computer Networks 47. 2005.

[5] C. Cicconetti, C. Eklund, L. Lenzini and E. Mingozzi, "Quality of Service Support in IEEE 802.16 Networks",IEEE Network Magazine, vol.20, no.2, Mar. 2006.

[6] L. de la Cruz, "SCALEV Lite v2.3, Manual de Usuario", [online], 2007. Available from: http://globus.upc.es/~ljcruz/

[7] L. de la Cruz y E. Sanvicente, "Scalev: Herramienta Software para la Evaluacion de Algoritmos de Scheduling". VI Jornadas de Ingeniería Telemática Jitel'07, Malaga, España. Pp. 237-244. ISBN 978-84-690-6670-6.

[8] C. Eklund, R. Marks, K. Stanwood and S. Wang, "IEEE Standard 802.16: A Technical Overview of the Wireless MANTM Air Interface for Broadband Wireless Access". IEEE Communications Magazine, Jun. 2002.

[9] T. Guerrero and L. De la Cruz. "IEEE 802.11e: Control de acceso al medio wireless con garantías de QoS". Reporte interno, Universidad Politécnica de Catalunya, 2006.

[10] Juniper Research [online], 2006. Available from: http://www.juniperresearch.com

[11] N. Loutfi, "WiMAX Technology for Broadband Wireless Access". England, John Wiley & Sons Ltd., 2007. ISBN 978-0-470-02808-7.

[12] Patrick Seeling, Martin Reisslein and Besan Kulapala, "Network Performance Evaluation Using Frame Size and Quality Traces of Single-Layer Video: A Tutorial", *IEEE Communication Surveys, Third Quarter 2004*, vol.6, no.3. Video traces available from: http://trace.eas.asu.edu

[13] F. Thouin and M. Coates, "Video-on-Demand Networks: Design Approaches and Future Challenges", IEEE Network, Mar/Apr. 2007.

[14] H-Y. Wei, S. Ganguly, R. Izmailov and Z-J. Haas, "Interference-Aware IEEE 802.16 WiMax Mesh Networks", In Proceedings of 61st IEEE Vehicular Technology Conference (VTC´05), Stockholm, Sweden, May 29 – Jun 1, 2005.

[15] WiMAX Spectrum Owners Alliance [online], 2006. Available from: http://www.wisoa.com

[16] Q. Zhang, W. Zhu and Y-Q. Zhang, "End-to End QoS for Video Delivery Over Wireless Internet", Proceedings of the IEEE, Vol. 93, No. 1. Jan. 2005.

[17] Y. Zhang, J. Luo and H. Hu, "Wireless Mesh Networking: Architectures, Protocols and Standards". U.S.A., Auerbach Publications, 2007. ISBN 0-8493-7399-9.

7. Obtención de Parámetros de una Red Telefónica para la distribución optima del tráfico en la Universidad Autónoma de Tamaulipas.

Miguel Angel Walle Vázquez, Carlos del Río Bocio,
Marco Antonio Panduro Mendoza

1 Capitulo introductorio

En este capítulo se abordan los conceptos básicos, las terminologías propias de la ingeniería de tráfico.

1.1 Análisis mediante teoría de colas

Una cola es una línea de espera, la teoría de colas es un conjunto de modelos matemáticos que describen sistemas de líneas de espera particulares. El objetivo de un análisis mediante teoría de colas nos ayuda en el caso de los sistemas telefónicos a determinar si el servicio que se presta es estable a parte que nos ayuda a determinar la capacidad de servicio apropiado [16].

Las redes telefónicas pueden ser muy frecuentemente modeladas según el esquema siguiente:

Fig. 1.1 Modelo simple de un sistema de colas

Este esquema aparece de forma natural al estudiar las redes telefónicas. Muestra una fuente de llamadas, una cola de espera o almacenamiento temporal

a la espera de que las unidades que en ella se acumulen sean atendidas por un servidor.

De este modelo, deben destacarse dos aspectos fundamentales: La disciplina con que se generan las llamadas (λ) y la disciplina con que se atienden (μ). El término disciplina hace referencia a la estadística de las unidades de información.

En el caso de la disciplina de generación (también podemos referirnos a ella como disciplina de llegada de unidades), se trata de la estadística (momentos) de los tiempos de llegada de las unidades.

Es muy importante notar que λ y μ corresponden al promedio de estas estadísticas, pero no aportan más información sobre la forma en que se genera la información (ráfagas, uniforme, etc.).

Una línea de transmisión puede ilustrar un ejemplo: la cola modela el retardo de transmisión (con posibles variaciones, asociadas al tamaño de la cola), y las tasas λ y μ corresponden a la velocidad de entrada y salida de la información de dicha línea, que podría tener pérdidas de información.

Si se obtiene una tasa de llegada de llamadas al sistema de $\lambda = 5$ llamadas/seg. (En promedio se recibe una llamada cada $1/\lambda = 0{,}2$ segundos. En este caso se determina que en función del valor de la tasa de servicio, se obtiene bajo razonamiento simple:

Donde $\mu > \lambda$, el sistema atiende las llamadas en la cola a un ritmo inferior al que llegan.

Lo que nos indica, que el sistema no es capaz de atender llamadas que se reciben a razón de 5 por segundo en promedio, por lo cual el tamaño de la cola dependerá de las estadísticas de las llamadas que arriban a la cola, y si se da permanente tenderá al infinito.

Donde $\mu < \lambda$, el sistema es capaz de atender llamadas que se reciben a razón 5 por segundo en promedio, por lo cual el tamaño de la cola tendrá tamaño finito. El tamaño en general no será nulo, porque aunque las llamadas tengan una media λ, podrían arribar llamadas en ráfagas.

Por último cuando $\mu = \lambda$, en este caso el sistema se encuentra al límite de estabilidad.

Resumiendo,

$$\mu \begin{cases} < \lambda \Longrightarrow \text{Cola} \longrightarrow \infty \\ = \lambda \Longrightarrow \text{Limite estabilidad} \\ > \lambda \Longrightarrow \text{Estable} \end{cases} \qquad (1.1)$$

Esto nos lleva a la necesidad de evitar sistemas cuyas colas de espera estén muy ocupadas, donde se pueden emplear valores de $\mu > \lambda$. Esta apreciación no es incorrecta, pero no siempre es posible considerarla.

Para el diseño de las redes de telefonía, es razonable prever que el sistema pueda sufrir de congestión y que este, sea rentable económicamente.

Por otra parte, los diseños sobre dimensionados favorecen en mucho la calidad del servicio en atención al cliente, pero el costo-beneficio se ve afectado a la hora de amortizarlos.

Así bien el dimensionamiento deberá estar sujeto a un diseño que garantice ciertos niveles de congestión, con cuotas mínimas de calidad (retardo y pérdidas de llamadas).

Con esto se define el parámetro utilización ó intensidad de tráfico en el enlace como la relación entre la tasa de llegadas y la de servicio. Siendo,

$$\rho = \lambda/\mu \tag{1.2}$$

Donde, rehaciendo la relación 1.1, se puede escribir que:

$$\rho \begin{cases} < 1 \Longrightarrow \text{Cola} \longrightarrow \infty(\text{inestable}) \\ = 1 \Longrightarrow \text{Limite} \quad \text{estabilidad} \\ > 1 \Longrightarrow \text{Estable} \end{cases} \tag{1.3}$$

Cualquier sistema que se desee analizar deben considerar los siguientes aspectos:

a) La estadística de las llamadas que llegan al sistema
b) La estadística de las llamadas que son atendidas en la cola
c) Cuantos llamadas que pertenecen a la misma cola, pueden ser atendidas simultáneamente
d) Cuantos clientes generan llamadas hacia la cola
e) Con cual disciplina operan las colas, desde el punto de vista de almacenamiento de las llamadas y su entrega a los servidores para que sean atendidas. Algunas como FIFO (*first in first out* — primera en entrar, primera en salir —) y LIFO (*last in firstout* — última en entrar, primera en salir —)

1.2 PROCESOS DE POISSON

En la sección anterior se ha mencionado la importancia de la forma (estadística) en que las unidades se reciben en el sistema y en la que son atendidas. Es lo que se denomina disciplina de llegadas y de servicio[16].

En esta sección veremos una de las más usadas por ser simple y con propiedades y características generales: los procesos de Poisson.

1.2.1 DEFINICIÓN DE PROCESO DE POISSON

Considérense las siguientes hipótesis:

a) Un proceso en el cual cada llegada sea independiente de cuando se produjo la anterior. Denominémosle *sin memoria*
b) Población infinita. Es decir, que el número de fuentes sea tan grande que se pueda considerar que la tasa promedio de llegada de unidades no depende de la ventana temporal y es por tanto una constante, cuyo valor es λ
c) Que la probabilidad de que se produzca una llegada sea proporcional al tiempo Δ t, es decir, que sea $\lambda \cdot \Delta$ t+O(Δ t), donde O(Δ t) es una O de Landau y hace referencia a los términos de orden superior a Δt (tienden hacia 0 más rápido que Δ t, conforme Δt tiende a 0)

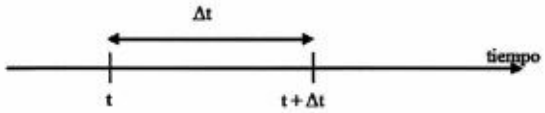


Fig. 1.2 Llegadas consecutivas

Bajo estas hipótesis, se demostrará más adelante que la probabilidad de que se produzcan n llegadas de unidades en un tiempo T (o Δ t) es:

$$P_n(T) = \frac{(\lambda T)^n}{n!} e^{-\lambda T} \tag{1.4}$$

Se puede demostrar fácilmente que está normalizado. Esto es,

$$\sum_{n=0}^{\infty} P_n(t) = 1 \tag{1.5}$$

De esta expresión se puede obtener la probabilidad de que no se produzca ninguna llegada en un tiempo *t* mediante *n*=0, *T*=*t*, es decir,

$$P_0(t) = \frac{(\lambda t)^0}{0!} e^{-\lambda t} = e^{-\lambda t} \tag{1.6}$$

La probabilidad de tener alguna llegada en tiempo t se puede obtener de dos formas:

a) Bien como la suma de tener 1 llegada, 2 llegadas, 3 llegadas, etc. hasta infinito
b) Bien, de forma más simple (recordando que las probabilidades están normalizadas como indica la expresión 1.5) como 1 menos la probabilidad de no tener ninguna

Por ambas vías, se obtiene que:

$$P_{n\neq 0}(t) = 1 - e^{-\lambda t} \tag{1.7}$$

1.2.2 PROPIEDADES

El promedio de unidades en el sistema en un intervalo de tiempo t se puede evaluar según la expresión

$$E[n] = \sum_{n=0}^{\infty} n \cdot P_n(t) \tag{1.8}$$

Por lo que,

$$E[n]=\sum_{n=0}^{\infty} n\cdot\frac{(\lambda t)^n}{n!}e^{-\lambda t}=e^{-\lambda t}\sum_{n=1}^{\infty}\frac{(\lambda t)^n}{(n-1)!}=e^{-\lambda t}\cdot\lambda t\cdot e^{\lambda t}=\lambda t \tag{1.9}$$

Para realizar el calculo de la sumatoria, hay que recordar que

$$e^x=\sum_{k=0}^{\infty}\frac{x^k}{k!} \tag{1.10}$$

A la vista de que E[n] = λ•t se puede deducir - tal como puede esperarse que λ es la velocidad de las llegadas por unidad de tiempo, ya que λ = E[n] / t.

De un modo parecido puede evaluarse la varianza de las llegadas de un proceso de Poisson:

$$\sigma^2=E[n^2]-[E[n]]^2 \tag{1.11}$$

Este cálculo puede efectuarse de fácilmente teniendo en cuenta que

$$E[n^2]-[E[n]]^2=E[n(n-1)]+E[n]-[E[n]]^2 \tag{1.12}$$

Finalmente se obtiene que

$$\sigma^2=\lambda\cdot t \tag{1,13}$$

En resumen, los procesos de Poisson cumplen que

$$E[n]=\lambda t \quad \text{y} \quad \sigma^2=\lambda\cdot t \tag{1.14}$$

Según esta propiedad, se definen los siguientes tipos de tráficos, en función de la relación entre la varianza y la media

$$\text{Si}\ \frac{\sigma^2}{E[n]}\begin{cases}>1\Rightarrow \text{Tráfico de pico}\\ =1\Rightarrow \text{Tráfico de Poisson}\\ <1\Rightarrow \text{Tráfico suavizado}\end{cases} \tag{1.15}$$

1.2.3 DISTRIBUCIÓN DE LAS LLEGADAS EN UN PROCESO DE POISSON

Hasta aquí se han estudiado cuántas llegadas se producen en un determinado intervalo de tiempo en una estadística de un proceso de Poisson. En esta sección se estudiará cuánto tiempo transcurre entre llegadas consecutivas mediante la evaluación analítica y obtendremos la función densidad de las llegadas para un proceso de Poisson.

Considérese para ello un intervalo de tiempo con un origen de tiempo arbitrario, al final del cual se produce la llegada de la siguiente unidad.

Fig. 1.3 Llegada de una unidad en $t=\tau$

En la situación de la figura 1.3, se tiene que no se recibe ninguna unidad en el intervalo de tiempo comprendido en $(0, \tau)$.

Por lo tanto, la probabilidad de no tener ninguna llegada en el intervalo $(0, \tau)$ es exactamente la de que τ sea mayor a t. Es decir,

$$P(\tau > t) = P_0(t) = e^{-\lambda t} \tag{1.16}$$

Y por tanto, $P(\tau \leq t)$ $1 - e^{-\lambda t}$. Nótese que $P(\tau \leq t)$ denota la función distribución $F_\tau(t)$, por lo que por simple derivación puede obtenerse la función densidad:

$$f_\tau(t) = \frac{dF_\tau(t)}{dt} = \lambda \cdot e^{-\lambda t} \tag{1.17}$$

En resumen, en un proceso de Poisson, las llegadas siguen una función densidad exponencial.

A partir de esta función densidad, puede calcularse el tiempo medio entre llegadas:

$$E[t] = \int_0^{\infty} t \cdot f_\tau(t) \cdot dt = \int_0^{\infty} t \cdot \lambda \cdot e^{-\lambda t} \cdot dt = \frac{1}{\lambda} \tag{1.18}$$

En consecuencia, el tiempo promedio de llegadas es de 1/ , lo cual es un resultado completamente esperado a la vista de las hipótesis de partida.

Otros momentos estadísticos son:

$$E[t^2] = \frac{2}{\lambda^2} \tag{1.19}$$

$$\sigma^2 = \frac{1}{\lambda^2} \tag{1.20}$$

1.2.4 PROPIEDAD DE SUPERPOSICIÓN

Supóngase dos procesos de Poisson independientes. Se puede enunciar la propiedad de la superposición del siguiente modo:

La suma de procesos de Poisson es otro proceso de Poisson cuya tasa es la suma de las tasas tributarias.

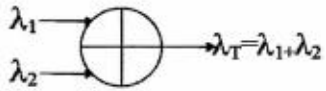


Fig. 1.4 Suma de procesos de Poisson

La demostración se basa en justificar el caso de dos fuentes de Poisson, ya que para el caso de tres, basta con asociar primeramente dos de ellos para demostrar que también se cumple. Del mismo modo se puede proceder para demostrar cualquier otro número de fuentes.

1.2.5 PROPIEDAD DE DESCOMPOSICIÓN

Supóngase un caudal λ_T de Poisson. Al aplicar una función que descomponga este caudal en caudales más pequeños de forma aleatoria de acuerdo a una probabilidad P_i, los caudales resultantes tendrán una tasa $\lambda_i = \lambda_T . P_i$, y serán también de Poisson.

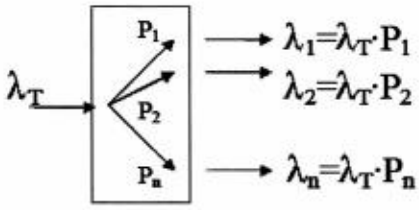


Fig. 1.5 Descomposición de procesos Poisson

La demostración puede efectuarse para el caso de separar 2 de ellos, y el caso general será la descomposición de éstos hasta conseguir los n flujos deseados.

1.3 CADENAS DE MARKOV

Ahora veremos una herramienta de análisis muy completa, basada en la definición de estados en los cuales se pueda encontrar un sistema, con el fin de buscar las probabilidades ubicarlo en uno de estos estados. Ya con esto se pueden calcular parámetros que permiten caracterizarlo[16].

Si tenemos un sistema con diversos estados. Denominado E_i al estado *i*. El estado E_i puede representar el estado asociado a que i usuarios estén en un instante dado efectuando una llamada telefónica. Si hubiera *n* circuitos en total para cursar las llamadas, habría que definir desde un estado E_o hasta un estado E_n.

En esta situación, se denota la probabilidad de estar en el estado E_m en el instante t_i como $P_r[E_m(t=t_i)]$, que de forma abreviada puede escribirse como $P_m(t_i)$.

Se define el *Vector de Estado* del sistema como:

$$\mathbf{P(t_i) = [P_0(t_i), P_1(t_i), P_2(t_i), \ldots]} \tag{1.21}$$

Nótese que la notación del vector se puede distinguir porque en ella no aparece el subíndice.

Se trata de un *vector estocástico*, puesto que por definición se verifica que:

$$\begin{aligned} &0 \le P_m(t_i) \le 1 \\ &\sum_{\forall m} P_m(t_i) = 1 \end{aligned} \tag{1.22}$$

Diremos que tenemos una *cadena* por disponer de un conjunto de estados que pueden representarse gráficamente enlazados entre ellos mediante flechas de transición entre unos estados y otros.

En general, la evolución de un sistema puede depender de todos los estados pasados, es decir, que, la $P_r[E_m\ (t=t_i)]$, puede depender de los estados anteriores $E_n(t=t_i)$, $E_p(t=t_{i-1})$, $E_q(t=t_{i-2})$, etc.

cumpliéndose que $t + t_i > t_i > t_{i-1} > t_{i-2}$ > etc. En el caso de que únicamente dependa del estado presente, $E_n(t=t_i)$, se puede escribir que:

$$P_r[E_m(t=t_{i+1}) \mid E_n(t=t_i), E_p(t=t_{i-1}), E_q(t=t_{i-21}),\ldots] = P_r[E_m(t=t_{i+1}) \mid E_n(t=t_i)]$$

(1.23)

En este caso, diremos que estamos ante un proceso sin memoria, un proceso de Markov.

Según las posibles transiciones entre los estados, queda definida la cadena de Markov, tal como muestra la figura 1.6. Nótese que no es necesario que las flechas alcancen todos los posibles estados.

Cada flecha va asociada a una probabilidad de transición entre estados que debe ser definida.

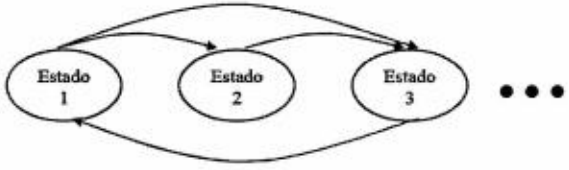


Fig. 1.6 Ejemplo simple de cadena de Markov

1.3.1 SISTEMAS DE TIEMPO DISCRETO Y SISTEMAS DE TIEMPO CONTINUO

La clasificación de los sistemas según si consideran transiciones de estado en instantes de tiempo determinados o indefinidos conduce a definir los sistemas como de *tiempo discreto* o *continuo* respectivamente.

Las señales de reloj de los sistemas digitales son un ejemplo de sistema discreto. La llegada del público a la entrada de un cine es un ejemplo de sistema continuo, puesto que se efectúan sin ningún instante de tiempo predeterminado.

De ahora en adelante nos vamos a interesar únicamente por los sistemas de tiempo continuo.

1.3.2 CADENAS DE MARKOV DE TIEMPO CONTINUO

Dado que nos interesan los sistemas cuyas llegadas y salidas se produzcan en tiempo continuo, de ahora en adelante nos ocuparemos únicamente de cadenas de este tipo.

En este caso, la notación que anteriormente habíamos expresado en general como $P_r\,[E_m\ \ (t = t_{i-1})]$ ahora se puede expresar simplemente como $P_r\,[E_m\ \ (t)]$, o más abreviadamente $P_m\ \ (t)$.

El vector de estado quedará escrito para tiempo continuo como

$$P(t) = [P_0(t),\ P_1(t),\ P_2(t),\ \ldots] \tag{1.24}$$

Cumpliéndose que

$$0 \le P_m(t) \le 1$$
$$\sum_{\forall m} P_m(t) = 1 \tag{1.25}$$

En este caso, que sea una cadena de Markov conduce a que la notación sea

$$P_r[E_n(t) \mid E_m(u),\ E_p(v),\ E_q(w),\ldots] = P_r[E_n(t) \mid E_m(u)], \quad \text{donde } t>u>v>w>\ldots \tag{1.26}$$

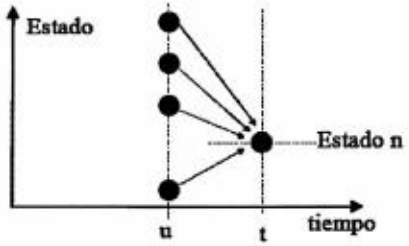
que abreviadamente expresaremos como

$$P_r[E_n(t) \mid E_m(u)] = P_{mn}(u,t) \tag{1.27}$$

Esta expresión se interpreta como la probabilidad de pasar del estado m al n desde el instante u al t.

Dado que la probabilidad de estar en el estado *n* en el instante *t* puede descomponerse según todos los caminos procedentes de cada uno de los estados hasta *n*, podemos escribir que:

$$P_n(t) = \sum_m P_m(u) \cdot P_{mn}(u,t) \tag{1.28}$$



Fig, 1.7 Transiciones desde todos los estados hasta uno determinado

Disponiendo todos los valores $P_{mn}(u,t)$ en forma matricial se define $P(u,t)$. Con esta nueva notación, puede expresarse que

$$P(t) = P(u) \cdot P_{mn}(u,t) \tag{1.29}$$

Esta expresión sintetiza todos los aspectos relacionados con las cadenas de Markov de tiempo continuo estudiados en esta sección.

1.3.3 ECUACIÓN DE FUTURO

De acuerdo a la figura 1.8, que muestra la evolución del sistema desde un estado *m* en el instante *u* hasta un estado *n* en $t+\Delta t$, definimos la ecuación de futuro como:

$$P_{mn}(u, t+\Delta t) = \sum_p P_{mp}(u,t) \cdot P_{pn}(t, t+\Delta t) \tag{1.30}$$

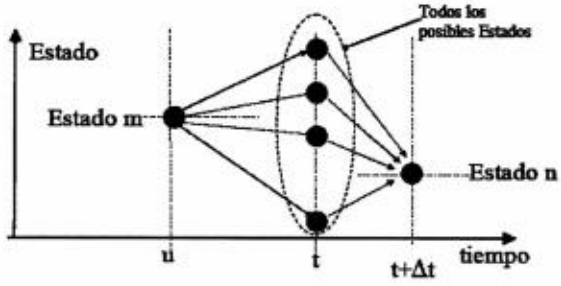


Fig. 1.8 Evolución temporal genérica de los estados

Extrayendo el término $p=n$ de la sumatoria y restando $P_{mn}(u,t)$ a la expresión 1.30:

$$P_{mn}(u,t+\Delta t)-P_{mn}(u,t)=\left[\sum_{p\neq n}P_{mp}(u,t)\cdot P_{pn}(t,t+\Delta t)\right]+P_{mn}(u,t)\cdot P_{nn}(t,t+\Delta t)-P_{mn}(u,t)$$

(1.31)

Dividiendo por Δt y tendiendo al límite $\Delta t \to 0$, podemos reconocer la expresión de la derivada, con lo cual resulta la siguiente ecuación de futuro:

$$\frac{\partial P_{mn}(u,t)}{\partial t}=\left[\sum_{p\neq n}P_{mp}(u,t)\cdot q_{pn}(t)\right]+P_{mn}(u,t)\cdot q_{nn}(t)$$

(1.32)

En esta expresión se ha definido la velocidad de transición $q_{pn}(t)$ y de permanencia $q_{nn}(t)$ de la siguiente manera:

$$q_{pn}(t)=\lim_{\Delta t\to 0}\frac{P_{pn}(t,t+\Delta t)}{\Delta t}$$

(1.33)

$$q_{nn}(t)=\lim_{\Delta t\to 0}\frac{P_{pn}(t,t+\Delta t)-1}{\Delta t}$$

(1.34)

Se puede aplicar la siguiente condición inicial: en instante de tiempo 0 se parte del estado 0, por lo cual $p_{on}(0,t)=p_n(t)$. Entonces la ecuación de futuro puede ser rescrita de la siguiente manera:

$$\frac{dP_n(t)}{dt} = \left[\sum_{p \neq n} P_p(t) \cdot q_{pn}(t)\right] + P_n(t) \cdot q_{nn}(t) \tag{1.35}$$

Además, puede demostrarse fácilmente que

$$\sum_n q_{pn}(t) = 0 \tag{1.36}$$

Si se define la matriz *Q(t)* como la formada por $[q_{pn}]$, se puede escribir la siguiente ecuación que resume lo contado hasta el momento:

$$\frac{dP(t)}{dt} = P(t) \cdot Q(t) \tag{1.37}$$

1.3.4 PROCESOS DE NACIMIENTO Y MUERTE

Hasta el momento, se han estudiado las cadenas de Markov en las cuales desde cualquier estado se puede ir a cualquier otro en el siguiente instante de tiempo.

Vamos a continuación a restringir esta situación y tomar como hipótesis que únicamente se puede pasar en el siguiente instante de tiempo a un estado inmediatamente vecino, esto es, desde el estado E_n se puede pasar al E_n, al E_{n-1} o permanecer en E_n. Este escenario define los procesos de nacimiento (cuando se pasa a un estado superior) y muerte (a un estado inferior).

En este caso, únicamente no serán nulas las probabilidades $p_{m, m+1}$, $p_{m, m}$ y $p_{m, m-1}$.

Como ejemplo, puede tomarse las llegadas a la cola de un cine. Cuando llega alguien, se pone a la cola. Incluso, aunque llegue un grupo, se puede considerar que cada persona se pone en cola con un diferencial de tiempo entre cada una de ellas.

En este escenario, la ecuación de futuro queda reducida a la siguiente expresión:

$$\frac{dP_n(t)}{dt} = P_{n-1}(t) \cdot q_{n-1,n}(t) + P_{n+1}(t) \cdot q_{n+1,n}(t) + P_n(t) \cdot q_{nn}(t) \tag{1.38}$$

Se define $q_{n-1, n}(t) = \lambda_{n-1}(t)$ y se le conoce como *velocidad de nacimiento*.

Se define $q_{n-1, n}(t) = \mu_{n-1}(t)$ y se le conoce como *velocidad de muerte*.

En este caso, considerando 1.36, se cumplirá que $q_{nn}(t) = -(q_{n,n+1}(t) + q_{n,n-1}(t))$. Usando la nueva notación, resulta que $q_{nn}(t) = -(\lambda_n(t) + \mu_n(t))$, donde hay que prestar una especial cuidado a los subíndices.

Con ello, la ecuación de futuro va tomando la forma buscada:

$$\begin{cases} \dfrac{dP_n(t)}{dt} = \mu_{n+1}(t)\cdot P_{n+1}(t) + \lambda_{n-1}(t)\cdot P_{n-1}(t) - (\lambda_n(t) + \mu_n(t))\cdot P_n(t) & ;n > 0 \\ \dfrac{dP_0(t)}{dt} = \mu_1(t)\cdot P_1(t) - \lambda_0(t)\cdot P_0(t) & ;n = 0 \end{cases} \tag{1.39}$$

Constituye un sistema de ecuaciones diferenciales. Para su resolución, por motivos de linealidad del sistema, siempre deberá despreciarse una y tomar otra ecuación que sea linealmente independiente.

Una que resulta adecuada para este fin es:

$$\sum_{n=0} P_n(t) = 1 \tag{1.40}$$

Y puede tomarse la siguiente representación gráfica:

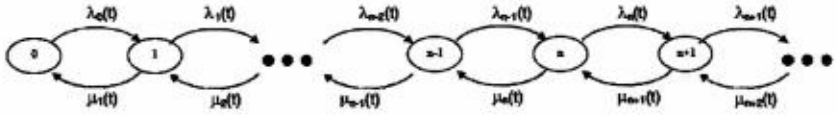


Fig. 1.9 Cadena de Markov de tiempo continuo para procesos de nacimiento y muerte

1.3.5 EJEMPLO

Considérese las siguientes hipótesis:

a) Proceso homogéneo. En este caso, las tasas de nacimiento y muerte no dependerán del estado $\lambda_{n\_}\lambda(t)$, $\forall$n:
b) Población infinita. Un conjunto de fuentes tan "alta" permite garantizar que la velocidad del sistema será constante. En este caso, $\lambda(t) = \lambda$ (= constante), $\forall$t
c) Asumamos también que sea un proceso de nacimiento puro, es decir, no hay muertes ($\mu(t) = 0$)
d) Finalmente, consideremos que sea de Markov, y por tanto no tenga memoria

Una primera reflexión nos conduce a pensar que éstas son precisamente las características de los procesos de Poisson. Veamos, pues, con las herramientas de que disponemos, qué podemos obtener.

La cadena de Markov asociada a estas hipótesis es la mostrada en la figura 1.10.

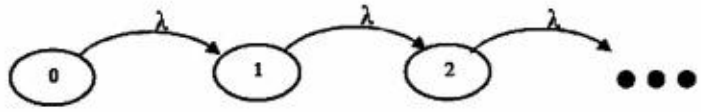


Fig. 1.10 Cadena de Markov para un proceso de Poisson

Substituyendo éstos en las hipótesis del ejemplo en la ecuación de futuro:

$$\begin{cases} \dfrac{dP_n(t)}{dt} = 0 + \lambda \cdot P_{n-1}(t) - (\lambda + 0) \cdot P_n(t) & ;n > 0 \\ \dfrac{dP_0(t)}{dt} = 0 - \lambda \cdot P_0(t) & ;n = 0 \end{cases} \tag{1.41}$$

Es decir,

$$\begin{cases} \dfrac{dP_n(t)}{dt} = \lambda \cdot P_{n-1}(t) - \lambda \cdot P_n(t) & ;n > 0 \\ \dfrac{dP_0(t)}{dt} = -\lambda \cdot P_0(t) & ;n = 0 \end{cases} \tag{1.42}$$

Con la condición inicial

$$P_k(0) = \begin{cases} 1 & ;k = 0 \\ 0 & ;k \neq 0 \end{cases} \tag{1.43}$$

donde la condición inicial tiene en cuenta que en el instante inicial el sistema se encuentra en el estado 0 (sin ninguna unidad).

Este sistema es fácilmente resoluble:

$$\underline{P_0(t)}: \quad \frac{dP_0(t)}{dt} = -\lambda \cdot P_0(t) \rightarrow P_0(t) = A \cdot e^{-\lambda \cdot t}$$

Mediante la condición inicial, se pueden obtener las constante de integración que irán apareciendo en el sistema. Para el caso de A de forma inmediata se obtendrá su valor 1.

$$P_0(t) = e^{-\lambda \cdot t}$$

$$P_1(t): \quad \frac{dP_1(t)}{dt} = \lambda \cdot P_0(t) - \lambda \cdot P_1(t) \rightarrow \frac{dP_1(t)}{dt} = \lambda \cdot e^{-\lambda \cdot t} - \lambda \cdot P_1(t) \rightarrow P_1(t) = \lambda \cdot t \cdot e^{-\lambda \cdot t}$$

Siguiendo la resolución del sistema para P2(t), P3(t), etc. se puede obtener una expresión general:

$$P_n(t) = \frac{(\lambda \cdot t)^n}{n!} \cdot e^{-\lambda \cdot t} \tag{1.44}$$

Se reconoce la expresión que al principio de este capítulo se usó para definir los procesos de Poisson.

Por lo tanto, este ejemplo permite justificarla.

1.3.6 PROCESOS DE NACIMIENTO Y MUERTE EN RÉGIMEN PERMANENTE

Cuando se desea estudiar el régimen permanente de un sistema, éste ha alcanzado una situación en la cual ya no hay dependencias temporales. Las derivadas respecto al tiempo pasan a ser nulas.

En el caso de los procesos de nacimiento y muerte, el régimen permanente supone que dPn(t)/dt =0 y el sistema de ecuaciones diferenciales 1.39 queda reducido a:

$$\begin{cases} 0 = \mu_{n+1} \cdot P_{n+1} + \lambda_{n-1} \cdot P_{n-1} - (\lambda_n + \mu_n) \cdot P_n & ;n > 0 \\ 0 = \mu_1 \cdot P_1 - \lambda_0 \cdot P_0 & ;n = 0 \end{cases} \tag{1.45}$$

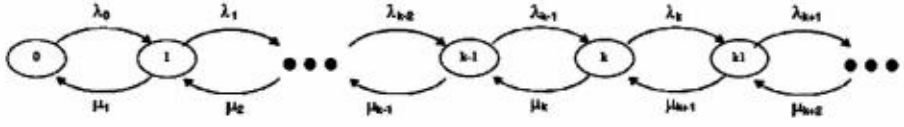


Fig. 1.11 Cadena de Markov para un proceso de nacimiento y muerte

1.3.7 ESTUDIO MEDIANTE FLUJOS

De estas ecuaciones y cadena representada se puede concluir que estos procesos pueden estudiarse en régimen permanente como flujos entrantes y salientes.

Los valores que pueden encontrarse son los siguientes, asociados a los estados iniciales, finales y el flujo entre ellos:

Estado inicial	Estado final	Flujo
0	1	$P_0 \cdot \lambda_0$
1	2	$P_1 \cdot \lambda_1$
k	k+1	$P_k \cdot \lambda_k$
1	0	$P_1 \cdot \mu_1$
2	1	$P_2 \cdot \mu_2$
k	k-1	$P_k \cdot \mu_k$

Dado que el sistema no acumula unidades, podemos afirmar que la suma de flujos entrantes debe coincidir con el de salientes,

$$\sum Flujos \quad Entrantes = \sum Flujos \quad Salientes \tag{1.46}$$

A partir de este simple razonamiento, se obtienen también el sistema de ecuaciones para los procesos de nacimiento y muerte, de forma simple y sin tener que recordarlas.

1.3.8 CÁLCULO DE LAS PROBABILIDADES DE ESTADO DE LOS PROCESOS DE NACIMIENTO Y MUERTE

Directamente podemos resolver las ecuaciones, empezando desde el estado 0 y llegando hasta el estado *k*, para obtener una expresión general:

n=0: Directamente, $P_1 = \frac{\lambda_0}{\mu_1} P_0$

n=1: $\mu_2 \cdot P_2 = P_1 \cdot (\lambda_1 + \mu_1) - \lambda_0 \cdot P_0 = \frac{\lambda_0}{\mu_1} P_0 \cdot (\lambda_1 + \mu_1) - \lambda_0 \cdot P_0 = \frac{\lambda_0 \cdot \lambda_1}{\mu_1} P_0 \rightarrow P_2 = \frac{\lambda_0 \cdot \lambda_1}{\mu_1 \cdot \mu_2} P_0$

siguiendo la iteración y llegando a $n=k$, fácilmente se puede deducir que:

$$P_k = P_0 \frac{\lambda_0 \cdot \lambda_1 \cdot \ldots \cdot \lambda_{k-1}}{\mu_1 \cdot \mu_2 \cdot \ldots \cdot \mu_k} = P_0 \cdot \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} \tag{1.47}$$

Para obtener el valor de *P0*, se puede despejar de la expresión

$$P_0 = \frac{1}{1 + \sum_{k=1}^{\infty} \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}} \tag{1.48}$$

La expresión obtenida para *Pk* y de *P0* son especialmente útiles para el estudio de los sistemas que abordaremos en las siguientes secciones.

Nótese que *P0* es la probabilidad de que no haya ninguna unidad en el sistema. Por consiguiente, el valor 1 - *P0* es la probabilidad de tener alguna unidad en el sistema.

1.4 CONCEPTOS BÁSICOS DE INGENIERÍA DE TRÁFICO

1.4.1 CONCEPTOS FUNDAMENTALES

Acarreo de tráfico. Es el volumen de tráfico que pasa por un conmutador, tráfico ofrecido es la cantidad de tráfico para un conmutador.

Para dimensionar una trayectoria de tráfico o el tamaño de un intercambio telefónico se debe conocer la intensidad de tráfico representativa de una temporada ocupada. El tráfico es muy aleatorio por naturaleza. Una consistencia certera puede ser encontrada en un horario de trabajo normal, a través de un día típico la variación es más que un periodo de 1- hora que se puede ver que es el más grande.

Hora ocupada. La hora ocupada refiere al volumen del tráfico o número de intentos de llamada y es continuo en un período de un intervalo en el que es cuantiado.

Pico de hora ocupada. Es la hora ocupada cada día; esta no es usualmente igual al número de días.

Tiempo consistente de hora ocupada. Es el periodo de 1 – hora que empieza al mismo tiempo cada día por el cuál el promedio del tráfico o intento de llamada es mayor que los días en consideración.

El periodo de ingeniería. es definido como la hora ocupada de temporada ocupada, la cual es la hora más ocupada del día más ocupado de la semana.
El promedio de la hora ocupada de la sesión ocupada. es usado para grupos de troncales y siempre tiene un criterio de servicio aplicado.
Tráfico. El flujo de tráfico a través de una central se define como el producto del número de llamadas y su duración promedio durante un periodo de observación de una hora. Es decir,

$$A = CT$$

Donde A = Flujo de tráfico
C = No. de llamadas originadas en una hora
I = Tiempo promedio de llamada

Ejemplo: 200 llamadas con una duración promedio de 2 min. son generadas durante un periodo de una hora por los suscriptores de una colonia de la ciudad

$A = 200 * 2 = 400$ minuto llamada

La intensidad de tráfico. es el flujo de tráfico expresado en horas-llamadas. Y representa el numero promedio de llamadas simultaneas. Para el ejemplo anterior,

$Ai = 400/60 = 6.67$ horas-llamada

La densidad de tráfico. representa el número de llamadas simultáneas en un instante dado.
Tráfico transportado. es el volumen de tráfico manejado por la central, y se obtiene de mediciones.
Tráfico ofrecido. es una cantidad no medible, correspondiente al tráfico transportado más el tráfico bloqueado o perdido (si lo hay).

1.4.2 UNIDADES DE TRÁFICO TELEFÓNICO

Erlang. A la unidad internacional de tráfico telefónico se le denomina Erlang en reconocimiento al matemático danés A. K. Erlang, fundador de la teoría de tráfico telefónico. Un Erlang representa un circuito ocupado por una hora. La intensidad de tráfico expresada en erlangs representa:

1. El número promedio de llamadas en progreso simultáneamente durante el periodo de una hora.
2. El número promedio de llamadas originadas durante un periodo de tiempo igual al promedio de llamada normal.
3. El tiempo total, expresado en horas, para transportar todas las llamadas.

Ejemplo:

1. *El número promedio de troncales ocupadas es 9.*
2. *En promedio se originan 9 llamadas cada tres minutos, ó tres llamadas por minuto, ó un total de l80 llamadas originadas en una hora (9/3 *60).*
3. *El tiempo total ocupado para transportar las 180 llamadas es de 9 horas (180 * 3/60).*

Cien-segundos-llamada. Los términos "unidad de llamada" UC ("Unit call") ó su sinónimo 'Cien-segundos-llamada" CCS ("Hundred-call-seconds") son de uso mas o menos generalizado. Y corresponde al número de circuitos ocupados en observaciones de cada 100 segundos. La relación de los ccs con el Erlang es:

1 Erlang=36 CCS

En un ejemplo precedente, la suma de 36 observaciones es 36*9 =324 CCS

Bloqueo, llamadas pérdidas y grado de servicio. Asumiendo que los intercambios telefónicos son para 5 000 suscriptores y que no más del 10% de los suscriptores desean el servicio simultáneamente de cualquier manera el intercambio es dimensional con suficiente equipo para completar las 5 000 conexiones simultáneas. Cada conexión puede ser entre cualquiera de los 5 000 suscriptores. Si el suscriptor 501 intenta hacer una llamada no puede por que todo el equipo está ocupado de cualquier manera la línea con la que él desea hacer comunicación podría estar ocupada también. Esta llamada de suscriptor 501 se denomina llamada pérdida o llamada bloqueada. La probabilidad de tener un bloqueo es un parámetro importante en la ingeniería de tráfico en los sistemas de telecomunicaciones, si las condiciones de congestión son introducidas a un sistema de comunicaciones se puede esperar que estas funcionen en un ahora ocupada.
Un conmutador es dimensionado para soportar la carga de la hora ocupada pero se podría sobre dimensionar la capacidad de este sistema pero sería muy redundante y por lo tanto poco económico.

Variaciones en el tráfico telefónico. Para determinar el dimensionamiento de las instalaciones telefónicas en concordancia con las necesidades de los subscriptores, se requiere la comprensión de la naturaleza del tráfico telefónico y su distribución con respecto al tiempo y destino. Los volúmenes de tráfico varían de estación a estación, de mes a mes, de día a día, de hora a hora y aún de minuto a minuto dentro de una misma hora.

La duración de las conversaciones es otra importante variable a considerar. Aunque la duración de llamada puede variar considerablemente entre centrales y temporadas del año, se ha encontrado por mediciones reales, que tiempos de conversación de uno a tres minutos son relativamente frecuentes, en tanto que diez ó más minutos son mas ocasionales.

La hora de mayor ocupación. Es el período interrumpido de 60 minutos durante el cual el tráfico es máximo. Tradicionalmente la planta telefónica es dimensionada de acuerdo a la intensidad de tráfico de la hora de mayor ocupación.

Grado de servicio. El término grado de servicio define la proporción de las llamadas que se permite fallar durante la hora de mayor ocupación debido a la limitación, por razones económicas, del equipo de conmutación de las plantas. En una oficina central con varias etapas de conmutación, existen grados de servicio para cada uno de dos que van desde 1 pérdida en 100 llamadas hasta 1 en 1,000. El grado de servicio total es aproximadamente igual a la suma de los grados de servicio parciales.

Grado de servicio = (número de llamadas perdidas) / (número total de llamadas ofrecidas)

El tráfico en una red de comunicaciones. Se refiere al acumulado de todas Las solicitudes de los usuarios que la red está atendiendo. En lo que a la red se refiere, las solicitudes de servicio arriban aleatoriamente y usualmente requieren tiempos de servicio impredecible. El primer paso del análisis de tráfico es la caracterización de los arribos de tráfico y tiempos de servicio en un marco probabilístico. A partir de lo cuál la red pueda ser evaluada en términos de cuánto tráfico transporta bajo cargas normales o promedio y con que frecuencia el volumen de tráfico excede la capacidad de la red.

La impredecible naturaleza del trafico telefónico es el resultado de dos procesos aleatorios subyacentes: El arribo de llamadas y los tiempos de retención. El arribo de un usuario particular se considera por lo general que ocurre completamente al azar y que es totalmente independiente del arribo de otros usuarios

Así que el número de arribos durante un intervalo de tiempo particular es indeterminado. En la mayoría de los casos los tiempos de retención también se distribuyen aleatoriamente. En algunas aplicaciones este crecimiento de aleatoriedad se puede sustituir por considerar tiempos de retención constantes. En cualquier caso la carga de tráfico presentada a una red depende fundamentalmente tanto de la frecuencia de arribos como de los tiempos promedios de retención de cada arribo.

2 Redes Neuronales Artificiales

En este capítulo se abordan los conceptos fundamentales y las terminologías propias de las Redes Neuronales Artificiales (RNA) [17].

2.1 Introducción

Muchos avances han ayudado al desarrollo de los sistemas inteligentes, algunos inspirados en redes neuronales. Los investigadores de muchas disciplinas científicas, han diseñado redes neuronales artificiales (RNAs siglas en español), para resolver una gran variedad de problemas en reconocimiento de patrones, predicciones, optimización, memoria asociativa, y control.

Se tiene que proponer aproximaciones convencionales para resolver estos problemas. Las aplicaciones acertadas que pueden satisfacer a la ciencia deben de estar bajo ambientes controlados, que no son flexibles para su ejecución y rendimiento fuera de este dominio. Las RNAs proveen alternativas, y muchas aplicaciones pueden beneficiarse de estas.

2.2 ¿Porqué las Redes Neuronales Artificiales (RNAs)?

En el curso de la evolución se han obtenido muchoas características deseables del cerebro humano en Von Neumann o la computadora paralela moderna, Esta incluye:

- paralelismos masivo,
- representación distribuida y calculo de operaciones,
- habilidad de aprender,
- habilidad generacionalización,
- adaptatividad,
- inherente procesamiento de información contextual,
- tolerancia a fallas, y

- bajo consumo de energía.

Esto ha ayudado a alcanzar que los dispositivos basados en redes neuronales biológicas, posean algunas de estas características deseables.

En el campo de la inteligencia artificial se refiere habitualmente de forma más sencilla como redes de neuronas o redes neuronales, las redes de neuronas artificiales, son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

Las Redes Neuronales Artificiales (ANNs de Artificial Neural Networks) fueron originalmente una simulación abstracta de los sistemas nerviosos biológicos, formados por un conjunto de unidades llamadas "neuronas" o "nodos" conectadas unas con otras. Estas conexiones tienen una gran semejanza con las dendritas y los axones en los sistemas nerviosos biológicos.

El Primer modelo de red neuronal fue propuesto en 1943 por McCulloch y Pitts en términos de un modelo computacional de "actividad nerviosa". El modelo de McCulloch-Pitts es un modelo binario, y cada neurona tiene un escalón o umbral prefijado. Este primer modelo sirvió de ejemplo para los modelos posteriores de Jhon Von Neumann, Marvin Minsky, Frank Rosenblatt, y muchos otros.

Una primera clasificación de los modelos de ANNs podría ser, atendiendo a su similitud con la realidad biológica:

1. Los modelos de tipo biológico. Este comprende las redes que tratan de simular los sistemas neuronales biológicos así como las funciones auditivas o algunas funciones básicas de la visión.
2. El modelo dirigido a aplicación. Estos modelos no tienen porque guardar similitud con los sistemas biológicos. Sus arquitecturas están fuertemente ligadas a las necesidades de las aplicaciones para las que son diseñados.

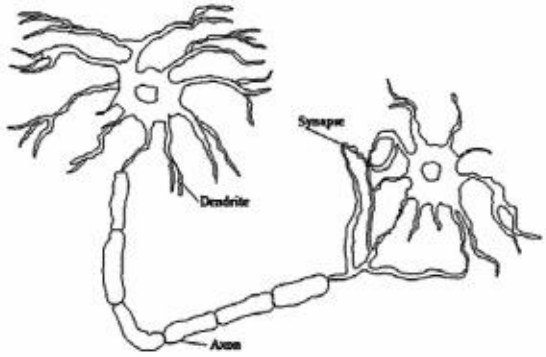
2.2 Redes Neuronales de tipo Biológico

Se estima que el cerebro humano contiene más de cien mil millones 1011 de neuronas y 1014 sinápsis en el sistema nervioso humano. Estudios sobre la anatomía del cerebro humano concluyen que hay más de 1000 sinápsis a la entrada y a la salida de cada neurona. Es importante notar que aunque el tiempo de conmutación de la neurona (unos pocos milisegundos) es casi un millón de veces menor que en

las actuales elementos de las computadoras, ellas tienen una conectividad miles de veces superior que las actuales supercomputadoras.

El objetivo principal de de las redes neuronales de tipo biológico es desarrollar un elemento sintético para verificar las hipótesis que conciernen a los sistemas biológicos.

Las neuronas y las conexiones entre ellas (sinápsis) constituyen la clave para el procesado de la información. Observe la figura:



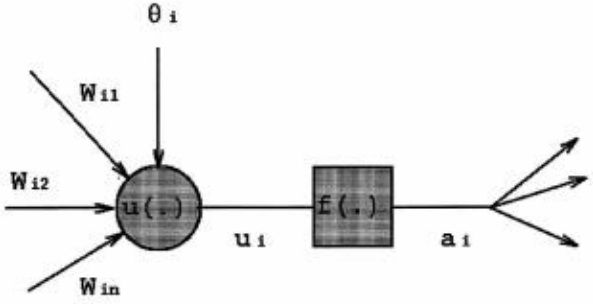
La mayor parte de las neuronas poseen una estructura de árbol llamadas dendritas que reciben las señales de entrada que vienen de otras neuronas a través de la uniones llamadas sinápsis. Algunas neuronas se comunican solo con las cercanas, mientras que otras se conectan con miles.

Hay tres partes en una neurona:

1. el cuerpo de la neurona,
2. ramas de extensión llamadas dendrítas para recibir las entradas, y
3. un axón que lleva la salida de la neurona a las desdirías de otras neuronas.

La forma que dos neuronas interactuan no está totalmente conocida, dependiendo además de cada neurona. En general, una neurona envía su salida a otras por su axón. El axón lleva la información por medio de diferencias de potencial, u ondas de corriente, que depende del potencial de la neurona. Este proceso es a menudo modelado como una regla de propagación representada por la función de red u(.). La neurona recoge las señales por su sinápsis sumando todas las influencias excitadoras e inhibidoras. Si las influencias excitadoras positivas dominan, entonces la neurona da una señal positiva y manda este mensaje a otras neuronas por sus sinápsis de salida. En este sentido la neurona puede ser modelada como

una simple función escalón f(.). Como se muestra en la próxima figura, la neurona se activa si la fuerza combinada de la señal de entrada es superior a un cierto nivel, en el caso general el valor de activación de la neurona viene dado por una función de activación f(.).



2.3 REDES NEURONALES PARA APLICACIONES CONCRETAS

Las ANNs dirigidas a aplicación están en general poco ligadas a las redes neuronales biológicas. Ya que el conocimiento que se posee sobre el sistema nervioso en general no es completo, se han de definir otras funcionalidades y estructuras de conexión distintas a las vistas desde la perspectiva biológica. Las características principales de este tipo de ANNs son los siguientes:

1. Auto Organización y Adaptatividad: utilizan algoritmos de aprendizaje adaptativo y auto organización, por lo que ofrecen posibilidades de procesado robusto y adaptativo (véase entrenamiento adaptativo y redes auto organizativas).
2. Procesado No Lineal: aumenta la capacidad de la red de aproximar, clasificar y su inmunidad frente al ruido.
3. Procesado paralelo: normalmente se usa un gran número de células de procesado por el alto nivel de interconectividad.

Estas características juegan un importante papel en las ANNs aplicadas al procesado de señal e imagen. Una red para una determinada aplicación presenta una arquitectura muy concreta, que comprende elementos de procesado adaptativo masivo paralelo combinadas con estructuras de interconexión de red jerárquica.

2.4 TAXONOMÍA DE LA REDES NEURONALES

Existen dos fases en toda aplicación de las redes neuronales: la fase de aprendizaje o entrenamiento y la fase de prueba. En la fase de entrenamiento, se usa un conjunto de datos o patrones de entrenamiento para determinar los pesos (parámetros de diseño) que definen el modelo neuronal. Una vez entrenado este modelo, se usará en la llamada fase de prueba o funcionamiento directo, en la que se procesan los patrones de prueba que constituyen la entrada habitual de la red, analizándose de esta manera las prestaciones definitivas de la red.

- Fase de Prueba: los parámetros de diseño de la red neuronal se han obtenido a partir de unos patrones representativos de las entradas que se denominan patrones de entrenamiento. Los resultados pueden ser tanto calculados de una vez como adaptados iterativamente, según el tipo de red neuronal, y en función de las ecuaciones dinámicas de prueba. Una vez calculados los pesos de la red, los valores de las neuronas de la última capa, se comparan con la salida deseada para determinar la validez del diseño.

- Fase de Aprendizaje: una característica de las redes neuronales es su capacidad de aprender. Aprenden por la actualización o cambio de los pesos sinápticos que caracterizan a las conexiones. Los pesos son adaptados de acuerdo a la información extraída de los patrones de entrenamiento nuevos que se van presentando. Normalmente, los pesos óptimos se obtienen optimizando (minimizando o maximizando) alguna "función de energía". Por ejemplo, un criterio popular en el entrenamiento supervisado es minimizar el least-square-error (error cuadrático medio) entre el valor del maestro y el valor de salida actual.

Las aplicaciones del mundo real deben acometer dos tipos diferentes de requisitos en el procesado. En un caso, se requiere la prueba en tiempo real pero el entrenamiento ha de realizarse "fuera de linea". En otras ocasiones, se requieren las dos procesos, el de prueba y el de entrenamiento en tiempo real. Estos dos requisitos implican velocidades de proceso muy diferentes, que afectan a los algoritmos y hardware usados.

Atendiendo al tipo de entrenamiento, una posible taxonomía de las redes neuronales es:

Fijo	No supervisado	Supervisado
Red de Hamming	Mapa de características	Basadas en decisión
Red de Hopfield	Aprendizaje competitivo	Perceptrón
		ADALINE (LMS)
		Perceptrón Multicapa
		Modelos Temporales Dinámicos
		Modelos Ocultos de Markov

2.5 Redes Neuronales Supervisadas y no Supervisadas

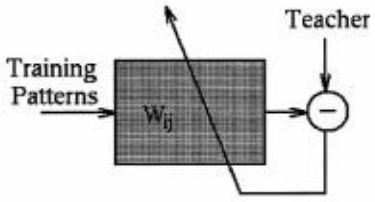
Las redes neuronales se clasifican comúnmente en términos de sus correspondientes algoritmos o métodos de entrenamiento: redes de pesos fijos, redes no supervisadas, y redes de entrenamiento supervisado. Para las redes de pesos fijos no existe ningún tipo de entrenamiento.

Reglas de entrenamiento Supervisado

Las redes de entrenamiento supervisado han sido los modelos de redes más desarrolladas desde inicios de estos diseños. Los datos para el entrenamiento están constituidos por varios pares de patrones de entrenamiento de entrada y de salida. El hecho de conocer la salida implica que el entrenamiento se beneficia la supervisión de un maestro. Dado un nuevo patrón de entrenamiento, por ejemplo, (m+1)-ésimo, los pesos serán adaptados de la siguiente forma:

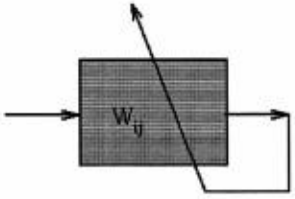
$$w_{ij}^{(m+1)} = w_{ij}^{(m)} + \Delta w_{ij}^{(m)}$$

Se puede ver un diagrama esquemático de un sistema de entrenamiento supervisado en la siguiente figura:



REGLAS DE ENTRENAMIENTO NO SUPERVISADO

Para los modelos de entrenamiento No Supervisado, el conjunto de datos de entrenamiento consiste sólo en los patrones de entrada. Por lo tanto, la red es entrenada sin el beneficio de un maestro. La red aprende a adaptarse basada en las experiencias recogidas de los patrones de entrenamiento anteriores. Este es un esquema típico de un sistema "No Supervisado":



Ejemplos típicos son La Regla de Aprendizaje de Hebb, y la Regla de Aprendizaje Competitiva. Un ejemplo del primero consiste en reforzar el peso que conecta dos nodos que se excitan simultáneamente.

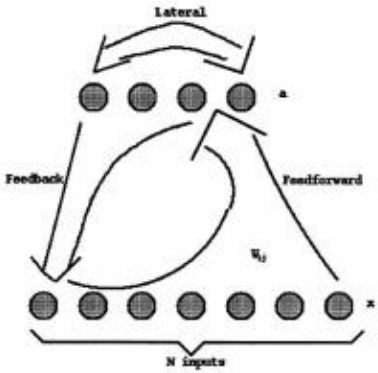
Como ejemplo de aprendizaje competitivo, si un patrón nuevo se determina que pertenece a una clase reconocida previamente, entonces la inclusión de este nuevo patrón a esta clase matizará la representación de la misma. Si el patrón de la entrada se determinó que no pertenece a ninguna de las clases reconocidas anteriormente, entonces la estructura y los pesos de la NN serán ajustados para reconocer la nueva clase.

2.6 ESTRUCTURA DE REDES NEURONALES ARTIFICIALES

Los aspectos más característicos de las estructuras son la estructura de conexión, el tamaño de la red y la elección entre ACON y OCON.

Estructuras de conexión de atrás hacia delante

Una red neuronal se determina por la neurona y la matriz de pesos. El comportamiento de la red depende en gran medida del comportamiento de la matriz de pesos. Hay tres tipos de capas de neuronas: la de entrada, las ocultas y la de salida. Entre dos capas de neuronas existe una red de pesos de conexión, que puede ser de los siguientes tipos: Hacia delante, hacia atrás, lateral y de retardo, tal como puede verse en la siguiente figura:



1. Conexiones hacia delante: para todos los modelos neuronales, los datos de las neuronas de una capa inferior son propagados hacia las neuronas de la capa superior por medio de las redes de conexiones hacia adelante.
2. Conexiones hacia atrás: estas conexiones llevan los datos de las neuronas de una capa superior a otras de la capa inferior.
3. Conexiones laterales. Un ejemplo típico de este tipo es el circuito "el ganador toma todo" (winner-takes-all), que cumple un papel importante en la elección del ganador.
4. Conexiones con retardo: los elementos de retardo se incorporan en las conexiones para implementar modelos dinámicos y temporales, es decir, modelos que precisan de memoria.

Las conexiones sinópticas pueden ser total o parcialmente interconectadas, como muestra la figura. También es posible que las redes sean de una capa con el modelo de pesos hacia atrás o bien el modelo multicapa hacia adelante. Es posible así mismo, el conectar varias redes de una sola capa para dar lugar a redes más grandes.

TAMAÑO DE LAS REDES NEURONALES

En una red multicapa de propagación hacia delante, puede haber una o más capas ocultas entre las capas de entrada y salida. El tamaño de las redes depende del numero de capas y del número de neuronas ocultas por capa.

- Número de capas: en una red multicapa, hay una o más capas de neuronas ocultas entre la entrada y la salida. El número de capas se cuenta

a menudo a partir del número de capas de pesos (en vez de las capas de neuronas).

- Número de unidades ocultas: El número de unidades ocultas está directamente relacionado con las capacidades de la red. Para que el comportamiento de la red sea correcto (esto es, generalización), se tiene que determinar apropiadamente el número de neuronas de la capa oculta.

APROXIMACIONES ACON FRENTE A OCON

Abordamos el problema de cuantas redes son necesarias para la clasificación en multicategorias. Típicamente, cada nodo de salida se usa para representar una clase. Por ejemplo, en un problema de reconocimiento alfanumérico, hay 36 clases; así que en total habrá 36 nodos de salida. Dado un patrón de entrada en la fase de prueba, el ganador (i.e., la clase que gana) es normalmente el nodo que tiene el valor más alto a la salida.

Dos posibles tipos de arquitectura son "All-Class-in-One-Network" (ACON), esto es, todas las clases en una red y "One-Class-in-One-Network" (OCON), esto es, una red para cada clase. En la aproximación ACON, todas las clases son reconocidas dentro de una única súper red. En algunos casos es ventajoso descomponer esta macro red en varias subredes mas pequeñas. Por ejemplo, una red de 36 salidas se puede descomponer en 12 subredes, cada una responsable de tres salidas. La descomposición mas extrema es la llamada OCON, donde una subred se dedica para una sola clase. Aunque el número de subredes en la estructura OCON es relativamente largo, cada subred individual tiene un tamaño menor que la red ACON.

2.7 APLICACIONES

Con el fin de llegar al entendimiento global de ANNs, adoptamos la siguiente perspectiva, llamada top-down que empieza por la aplicación se pasa al algoritmo y de aquí a la arquitectura:

Esta aproximación a las ANNs está motivada por la aplicación, basada en la teoría y orientada hacia la implementación. Las principales aplicaciones son para el procesado de señal y el reconocimiento de patrones. La primera etapa algorítmica representa una combinación de la teoría matemática y la fundamentación heurística por los modelos neuronales. El fin último es la construcción de neurocomputadores digitales, con la ayuda de las tecnologías VLSI y el procesado adaptativo, digital y paralelo.

Desde el punto de vista de la aplicaciones, la ventaja de la ANNs reside en el procesado paralelo, adaptativo y no lineal. Las ANNs han encontrado muchas aplicaciones con éxito en la visión artificial, en el procesado de señales e imágenes, reconocimiento del habla y de caracteres, sistemas expertos, análisis de imágenes médicas, control remoto, control de robots, inspección industrial y exploración científica. El dominio de aplicación de las ANNs se puede clasificar de la siguiente forma: asociación y clasificación, regeneración de patrones, regresión y generalización, y optimización.

ASOCIACIÓN Y CLASIFICACIÓN

En esta aplicación, los patrones de entrada estáticos o señales temporales deben ser clasificadas o reconocidas. Idealmente, un clasificador debería ser entrenado para que cuando se le presente una versión distorsionada ligeramente del patrón, pueda ser reconocida correctamente sin problemas. De la misma forma, la red debería presentar cierta inmunidad contra el ruido, esto es, debería ser capaz de recuperar una señal "limpia" de ambientes o canales ruidosos. Esto es fundamental en las aplicaciones holográficas, asociativas o regenerativas.

- Asociación: de especial interés son las dos clases de asociación: autoasociación y heteroasociación. El problema de la autoasociación es re-

cuperar un patrón enteramente, dada una información parcial del patrón deseado. La heteroasociación es recuperar un conjunto de patrones B, dado un patrón de ese conjunto. Los pesos en las redes asociativas son a menudo predeterminados basados en la regla de Hebb. Normalmente, la auto correlación del conjunto de patrones almacenado determina los pesos en las redes auto asociativas. Por otro lado, la correlación cruzada de muchas parejas de patrones se usa para determinar los pesos de la red de Heteroasociación.

- Clasificación no Supervisada: para esta aplicación, los pesos sinápticos de la red son entrenados por la regla de aprendizaje no supervisado, esto es, la red adapta los pesos y verifica el resultado basándose únicamente en los patrones de entrada.
- Clasificación Supervisada: esta clasificación adopta algunas formas del criterio de interpolación o aproximación. En muchas aplicaciones de clasificación, por ejemplo, reconocimiento de voz, los datos de entrenamiento consisten de pares de patrones de entrada y salida. En este caso, es conveniente adoptar las redes Supervisadas, como las bien conocidas y estudiadas redes de retropopagación. Este tipo de redes son apropiadas para las aplicaciones que tienen una gran cantidad de clases con límites de separación complejos.

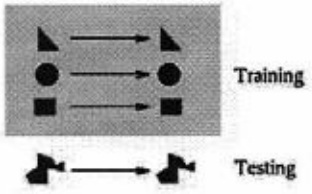
COMPLETAR LOS PATRONES

En muchos problemas de clasificación, una cuestión a solucionar es la recuperación la información, esto es, recuperar el patrón original dada sola una información parcial. Hay dos clases de problemas: temporales y estáticos. El uso apropiado de la información con textual es la llave para tener éxito en el reconocimiento.

GENERALIZACIÓN

La Generalización. Se puede extender a un problema de interpolación. El sistema es entrenado por un gran conjunto de muestras de entrenamiento basados en un procedimiento de aprendizaje supervisado. Una red se considera que esta entrenada con éxito si puede aproximar los valores de los patrones de entrenamiento y puede dar interpolaciones "suaves" para el espacio de datos no entrenado. El objetivo de la Generalización es dar una respuesta correcta a la salida para un estímulo de entrada que no ha sido entrenado con anterioridad. El sistema debe inducir la característica saliente del estímulo a la entrada y detectar la regularidad. Tal habilidad para el descubrimiento de esa regularidad es crítica en muchas apli-

caciones. Esto hace que el sistema funcione eficazmente en todo el espacio, incluso ha sido entrenado por un conjunto limitado de ejemplos.



2.8 Aplicaciones para el Análisis de la Capacidad de Tráfico Telefónico

El objetivo principal es pronosticar la capacidad de red (tráfico) del sistema telefónico. En este proyecto el algoritmo de propagación de retroalimentación de la ANN se refiere a entrenar y probar los datos alimentados. Esta técnica es utilizada por muchos investigadores para muchas aplicaciones y reportes con resultados satisfactorios. En este caso la información que se analiza, es según los días de la semana, según el fin de semana, los días festivos y aquellos días donde el medio ambiente tiene influencia. La redes con tres y cuatro entradas son probadas y el resultado obtenido es comparado en términos de análisis de errores.

En todas las redes de telefonía, existen problemas de congestión, por sobre carga de llamadas o por que los sistemas están saturados o desactualiazados, o simplemente el diseñar para ampliar o implementar nuevas redes requiere del uso de muchos modelos matemáticos lineales, para establecer las capacidades actuales.

Las ANNs son modelos electrónicos crudos basados en la estructura neuronal del cerebro. Le cerebro básicamente aprende de la experiencia. Esta es la ventaja natural, donde algunos problemas pueden ser alcanzados a través de las computadoras, resolviendo eficientemente y con poco recurso de energía. Este modelo del cerebro también promete el uso de menos técnicos para el desarrollo de las maquinas que den la solución.

3 Investigación y resultados del análisis

3.1 Investigación

El establecimiento de sistemas de medición automatizado de Grados de Servicio [1], obedece a la necesidad de medir la habilidad de los usuarios para acceder a un sistema de troncales telefónicas, que bien pueden ser en cuanto a la capacidad de disponibilidad para acceder al sistema de telefonía pública o al propio sistema de troncales privado. Los proveedores de telecomunicaciones saben del Grado de Servicio requerido, estos deberán asegurar los suficientes circuitos de telecomunicaciones o rutas disponibles dentro de los acuerdo de los nivel de servicio especificados, que al final se verán reflejados por la Calidad de Servicio [2].

Los términos usados en la ingeniería de tráfico son de uso estándar. La taza de arribo es el número de llamadas que llegaron con facilidad en un determinado periodo de tiempo finito. La utilización de circuito es también llamada la eficiencia de circuito, se define como la proporción de tiempo que un circuito esta ocupado o el tiempo proporcional en promedio que cada circuito en un grupo esta ocupado. El bloqueo ocurre cuando el número de llamadas, de entrada o salida, excede el número de facilidades (líneas, troncales, agentes o operadoras) disponibles para soportarlas. Una llamada bloqueada es dada por una señal de ocupado, donde se requiere que la llamada se desconecte y se intente de nuevo. En un sistema de comunicaciones, es el trascurso de un periodo de 60 minutos donde ocurre la carga máxima de tráfico de todas las 24 horas del día llamada hora de ocupación. Un Erlang es definido como una unidad sin dimensionamiento de intensidad de tráfico. En sistemas de telecomunicaciones, la intensidad de tráfico es medida como el promedio de ocupación de una facilidad durante un específico periodo de tiempo [1].

El estudio del rendimiento durante el ciclo de desarrollo de un sistema de conmutación puede ser usado para predecir la capacidad del sistema, comparada en contraste con arquitecturas refinadas ad-hoc y estimando el impacto de nuevos servicios. En el estudio normalmente se emplea el modelo de colas, por cuestiones de rendimiento esenciales, que son inmediatamente relacionadas a la competencia por los recursos. Desde la perspectiva de rendimiento de un sistema de conmutación, las llamadas telefónicas son asociadas dentro de un conjunto de estímulos que deben ser procesadas en moda secuencial y en orden para establecer una conexión entre pares. Una construcción de un modelo teórico de encolamiento que caracterice el tiempo de respuesta a este estimulo, este se obtiene asumiendo que el estimulo es generado como un proceso de Poisson [3].

Designar modelos estocásticos de tiempo discreto para abarcar varias mediciones secuenciales de tráfico telefónico son derivados de suposiciones clásicas de teorías de colas. Los modelos son lineales y sus coeficientes dependen de la longitud del ciclo de medición normalizado por el tiempo de espera [4].

Los algoritmos usados para seleccionar las rutas para las llamadas en redes telefónicas representan un importante mecanismo para el control en tiempo real del rendimiento de la red, especialmente cuando la red pasa por situaciones anormales, tales como sobre carga, errores y cargas pronosticadas y las fallas suceden en porciones de la red [5].

El análisis de capacidad de tráfico de una línea telefónica, puede ser representado en una aplicación de Red Neuronal Artificial, el objetivo es pronosticar la capacidad de tráfico de la red telefónica identificando los flujos. Las técnicas de propagación regresivas de una red neuronal artificial, son seleccionadas por un gran número de investigadores para en muchas aplicaciones y reportes [6].

Cuando se planea la futura expansión de las redes telefónicas, el aprovisionamiento oportuno de los servicios o cuando se pretende estudiar el rendimiento de las mismas, es necesario ofrecer la cantidad de tráfico necesario para que las comunicaciones entre cada par destino-origen se puedan realizar, para esto se establece el uso de un modelo de interpolación de tráfico iterativo el cual estimara el tráfico ofertado, para el tamaño de los enlaces se usa el modelo de dimensionamiento basado en las demandas futuras. También una pronostico de matriz de tráfico ofertado para un específico periodo de planeación, es usado por el modelo anterior para obtener los datos de dimensionamiento [7].

Las llamadas no completadas en un conmutador telefónico con una estructura óptima son consideradas fallas, dentro de las implicaciones en la calidad de servicio y el degradamiento en la satisfacción del cliente. Las observaciones son enfocadas solamente en las llamadas internas. En un corto historial, las facilidades comunes de llamadas y los servicios adicionales son descritos como conductas conocidas de los usuarios durante la administración de la llamada. La atención particular se centra en encontrar las razones de la baja completación de las llamadas [8].

El grado de servicio de una ruta alterna esta normalmente basado en el promedio de varias horas de probable bloqueo de rutas alternativas. Los métodos contemporáneos de grado de servicio de rutas alternas requieren ajustes para las cargas de tráfico promedio con apropiados factores de corrección en orden para asegurar la ganancia suficiente [9]-[13].

3.2 Resultados del análisis

Después de analizar documentos, publicaciones arbitradas, libros y demás, la aplicación de modelos matemáticos lineales [1]-[5],[7]-[13],[16], en mucho nos ayudan a obtener parametrización de un sistema como la red telefónica de la Universidad Autónoma de Tamaulipas, y siguen siendo los métodos tradicionales para ello, además de permitir medir el desempeño, considerando todos los factores necesarios para implementar, expandir u optimizar los servicios que ella ofrece.

Por otro lado, el uso algoritmos de propagación de retroalimentación de Redes Neuronales Artificiales (RNAs) [14][15][17], nos permitieron también obtener resultados favorables para la parametrización, además de poder pronosticar el comportamiento de la red telefónica y las capacidades del propio sistema, con resultados sorprendentes, la ventaja de estos algoritmos es que no son programados, son entrenados, retroalimentandose de información real y pronosticando situaciones futuras de lo sistemas.

4 Conclusiones y líneas futuras

4.1 Conclusiones

Se describen a continuación las conclusiones y aportaciones del presente trabajo, así como las futuras líneas trabajo que pudieran dar seguimiento a la investigación.

Se ha realizado un estudio de los modelos matemáticos lineales para poder parametrizar los sistemas telefónicos.

Se ha realizado un estudio de los algoritmos que pueden implementados a través de las Redes Neuronales Artificiales.

Se evaluaron las ventajas y desventajas de los dos tipos de modelos. Así también se realizó una revisión bibliográfica de los posibles modelos que en la actualidad se utilizan para parametrizar las redes telefónicas.

Estamos iniciando los trabajos para utilizar la herramienta de MATLAB 6.5 con el ANN toolbox, con la base de datos actual de la red telefónica de la Universidad Autónoma de Tamaulipas.

4.2 LÍNEAS FUTURAS

Como una de las líneas futuras se pretende que este trabajo sea utilizado como fase inicial para la medición y parametrización de la red telefónica de la Universidad Autónoma de Tamaulipas.

La utilización de la herramienta de MATLAB 6.5 y el lenguaje de programación JAVA, nos permitirá implementar aplicaciones que no solo sirvan a la Universidad Autónoma de Tamaulipas, si no para cualquier institución o empresa.

BIBLIOGRAFÍA

[1] M Inayatullah, F Ullah and A Khan, "An Automated Grade Of Service Measuring System". 2nd International Conference on Emerging Technologies, Peshawar, Pakistan, November 2006.

[2] H Ali, M Inayatullah and S Rotenstreich, "Resource Allocation and QoS in Mobile Ad Hoc Networks", Proceedings of the 2004 ACM international symposium, 2004.

[3] J. S. Kaufman, M. S. Karlsson and J. S. Willie, "Workload Characterization in PBX Performance Studies"

[4] J Filipiak and P Chemouil, "Modeling and Prediction of Traffic Fluctuations in Telephone Networks", IEEE Transactions on Comunications, Vol. Com-35, No. 9, September 1987

[5] K. R. Krishnan, "Markov Decision Algorithms for Dynamic Routing", IEEE Communications Magazine, October 1990

[6] D. M. Ali, N. Yaacob, M. F. Bin Osman, "Traffic Capacity of a Telephone Line Using Artificial Neural Network (ANN) ", 2004 RF and Microwave Conference, October 5-6, Subang, Selangor, Malaysia, Universiti Teknologi MARA, 2004

[7] A.P. Engelbrecht, "A Model for the Estimation of Offerted Traffic from Measured Traffic Parameters", University of Stellenbosh

[8] D. Jevtic, "Unsucessful Calls in PBX - Human Factor Considerations", Zagreb, CROATIA HR-10000, 1998

[9] L. Lee, Application of Equivalent Trunks Technique to Alternative Route Engineering, IEEE Transaction on Comunications, July 1973

[10] K. R. Krishnan, T. J. Ott, "Routing of Telephone Traffic as a Controlled Harkov Process", Proceedings of 23rd Conference on Decision and Control, Las Vegas, NV, December 1984

[11] J. Regnier and W. H. Cameron, "State-Dependent Dynamic Traffic Management for Telephone Networks" IEEE Communications Magazine, October 1990

[12] N. T. Koussoulas, "Performance Analysis of Circuit-Switched Networks with State-Dependent Routing", IEEE Transactions on Communications, Vol. 41, No. 11,November 1993

[13] K. Chan and T. P. Yum, "The Maximun Mean Time to Blocking Routing in Circuit-Switched Networks", IEEE Journal on Selected Areas in Communications, Vol. 12, No. 2, February 1994

[14] W. T. Beyene, "Application of Artificial Neural Networks to Statistical Analysis and Nonlinear Modeling of High-Speed Interconnect Systems", IEEE Transactions on Computer-Aided Design Of Integrated Circuits and System, Vol. 26,No. 1,January 2007

[15] H. Salazar, R. gallego and R. Romero, "Artificial Neural Networks and Clustering Techniques Applied in the Reconfiguration of Distribution Systems", IEEE Transactions on Poer Delivery, Vol. 21, No.3, July 2006

[16] X. Hesselbach and J. A. Bosh, "Análisis de redes y sistemas de comunicaciones", Ediciones UPC, Primera Edición, Octubre 2002

[17] D. A. De la Fuente and A. Vega, "Tutorial de Redes Neuronales",UPM Señales Sistemas y Radiocomunicaciones, Edición en Español, Octubre 2003

8. Análisis de Mensajes SIP Mediante Esquemas de Interoperabilidad en Hardware y Software.

Everardo Huerta, Aldo Luis Méndez Pérez

1. Componentes y Funcionamiento de una Red VoIP.

1.1 Definición de VoIP

VoIP viene de las palabras en ingles Voice Over Internet Protocol. Como dice el término, VoIP intenta permitir que la voz viaje en paquetes IP y obviamente a través de Internet. La telefonía IP conjuga dos mundos históricamente separados: la transmisión de voz y la de datos. Se trata de transportar la voz previamente convertida a datos, entre dos puntos distantes. Esto posibilitaría utilizar las redes de datos para efectuar las llamadas telefónicas, y por ende desarrollar una única red convergente que se encargue de cursar todo tipo de comunicación, ya sea voz, datos, video o cualquier tipo de información.

La VoIP por lo tanto, no es en sí mismo un servicio sino una tecnología que permite encapsular la voz en paquetes para poder ser transportados sobre redes de datos sin necesidad de disponer de los circuitos conmutados convencionales conocida como la PSTN, que son redes desarrolladas a lo largo de los años para transmitir las señales vocales.

La PSTN se basaba en el concepto de conmutación de circuitos, es decir, la realización de una comunicación requería el establecimiento de un circuito físico durante el tiempo que dura ésta, lo que significa que los recursos que intervienen en la realización de una llamada no pueden ser utilizados en otra hasta que la primera no finalice, incluso durante los silencios que se suceden dentro de una conversación típica.

En cambio, la telefonía IP no utiliza circuitos físicos para la conversación, sino que envía múltiples conversaciones a través del mismo canal (circuito virtual) codificadas en paquetes y en flujos independientes. Cuando se produce un si-

lencio en una conversación, los paquetes de datos de otras conversaciones pueden ser transmitidos por la red, lo que implica un uso más eficiente de la misma.

Según esto, son evidentes las ventajas que proporciona las redes VoIP, ya que con la misma infraestructura podrían prestar mas servicios y además la calidad de servicio y la velocidad serian mayores; pero por otro lado también existe la gran desventaja de la seguridad, ya que no es posible determinar la duración del paquete dentro de la red hasta que este llegue a su destino y además existe la posibilidad de pérdida de paquetes, ya que el protocolo IP no cuenta con esta herramienta.

1.2. ENCAPSULAMIENTO DE UNA TRAMA VoIP

Una vez que la llamada ha sido establecida, la voz será digitalizada y entonces transmitida a través de la red en tramas IP. Las muestras de voz son primero encapsuladas en RTP (Protocolo de Transporte en tiempo Real) y luego en UDP o TCP antes de ser transmitidas en una trama IP. La siguiente figura muestra un ejemplo de una trama VoIP sobre una red LAN y WAN.

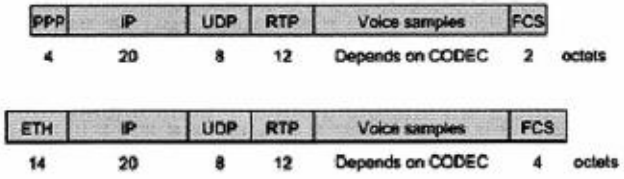


Figura 1. Ejemplo de una trama VoIP sobre una red LAN y WAN.

1.3. SESSION INITIATION PROTOCOL

Session Initiation Protocol (SIP o Protocolo de Inicialización de Sesiones) es un protocolo de señalización simple, utilizado para telefonía y videoconferencia por Internet. Basado en el Protocolo de Transporte de correo simple (SMTP) y en el Protocolo de Transferencia Hipertexto (HTTP) fue desarrollado por el IETF MMUSIC Working Group con la intención de ser el estándar para la iniciación, modificación y finalización de sesiones interactivas de usuario donde intervienen elementos multimedia como el video, voz, mensajería instantánea, juegos online y realidad virtual. SIP es uno de los protocolos de señalización para voz sobre IP, acompañado por H.323. SIP es definido completamente en la RFC 2543 y en la RFC 3261.

SIP es un protocolo de la capa de aplicación independiente de los protocolos de paquetes subadyacentes (TCP, UDP, ATM, X.25). SIP esta basado en una arquitectura cliente servidor en la cual los clientes inician las llamadas y los servidores responden las llamadas. Es un protocolo abierto basado en estándares, ampliamente soportado y no es dependiente de un solo fabricante de equipos.

SIP es un protocolo más nuevo que H.323 y no tiene madurez y soporte industrial al mismo tiempo. Sin embargo, por su simplicidad, escalabilidad, modularidad y comodidad con la cual integra con otras aplicaciones, este protocolo es atractivo para uso en arquitecturas de voz paquetizados. SIP puede establecer sesiones de dos partes (llamadas ordinarias), de múltiples partes (en donde todos pueden oír y hablar) y de multidifusión (un emisor, muchos receptores). Las sesiones pueden contener audio, video o datos. SIP solo maneja establecimiento, manejo y terminación de sesiones.

Algunas de las características claves que SIP ofrece son:

a) Resolución de direcciones, mapeo de nombres y redirección de llamadas.
b) Descubrimiento dinámico de las capacidades media del endpoint, por uso del Protocolo de Descripción de Sesión (SDP).
c) Descubrimiento dinámico de la disponibilidad del endpoint.
d) Origen y administración de la sesión entre el host y los endpoints.

Beneficios de SIP

Algunos de los beneficios claves de SIP son:

- SIMPLICIDAD: SIP es un protocolo muy simple. El tiempo de desarrollo del software es muy corto comparado con los productos de telefonía tradicional. Debido a la similitud de SIP a HTTP y SMTP, el rehúso de código es posible.
- EXTENSIBILIDAD: SIP ha aprendido de HTTP y SMTP y ha construido un exquisito grupo de funciones de extensibilidad y compatibilidad.
- MODULARIDAD: SIP fue diseñado para ser altamente modular. Una característica clave es su uso independiente de protocolos. Por ejemplo, envía invitaciones a las partes de la llamada, independiente de la sesión misma.
- ESCALABILIDAD: SIP ofrece dos servicios de escalabilidad:

o PROCESAMIENTO DE SERVIDOR: SIP tiene la habilidad para ser Stateful Stateless.

- o ARREGLO DE LA CONFERENCIA: Puesto que no hay requerimiento para un controlador central multipunto, la coordinación de la conferencia puede ser completamente distribuida o centralizada.
- INTEGRACION: SIP tienen la capacidad para integrarse con la Web, E-mail, aplicaciones de flujo multimedia y otros protocolos.
- INTEROPERABILIDAD: porque es un estándar abierto, SIP puede ofrecer interoperabilidad entre plataformas

1.4. DISEÑO DEL PROTOCOLO

SIP es un protocolo de capa de aplicación y puede ejecutarse sobre UDP o TCP. Los clientes SIP usan el puerto 5060 en TCP (Transmission Control Protocol) y UDP (User Datagram Protocol) para conectar con los servidores SIP, en caso de utilizar un protocolo seguro como SIPS el puerto a utilizar es el 5061, este punto se analizará más adelante cuando hable de TLS (Transport Security Protocol).

SIP es usado simplemente para iniciar y terminar llamadas de voz y video. Todas las comunicaciones de voz/video van sobre RTP (Real-time Transport Protocol).

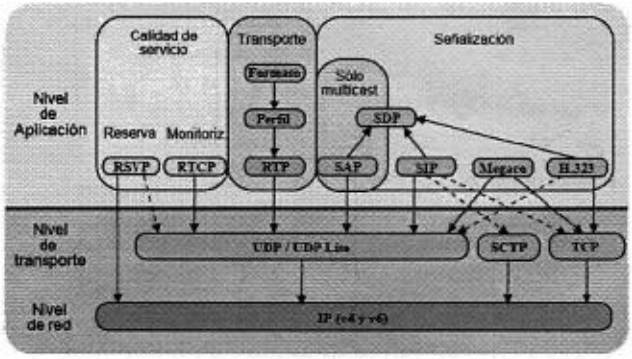


Figura 2. Diseñó del protocolo

La primera versión propuesta para estándar (SIP 2.0) fue definida en el RFC 2543. El protocolo aclarado en el RFC 3261, aunque muchas implementaciones están usando todavía versiones en fase de borrador. Hay que fijarse en que el número de versión sigue siendo 2.0.

1.5. Capa de transporte en SIP

SIP puede utilizar en su capa de transporte (Nivel 4 en el modelo OSI) tanto UDP, TCP como TLS Transport Layer Security (Refiriéndonos a TLS sobre TCP). TLS es utilizado para dar un cierto nivel de seguridad, encriptando la información que usualmente es vulnerable a ataques ya que se envía en texto plano.

La utilización de SIP sobre TCP sin encriptación está tendiendo a desaparecer en usos no pagos o de VoIP en Internet debido a la sencillez de UDP, la creciente confiabilidad de las redes y a la inútil necesidad de retransmisión en una conexión de voz o de media donde esta presente la transmisión en tiempo real.

De todas maneras es importante que un Agente Usuario (UA) de alto rendimiento como un sipphone por ejemplo, soporte tanto TCP como UDP como protocolos de transporte, ya que si un UA trata de establecer una sesión TCP con su par, y esté no soporta TCP en su capa de transporte, la sesión no se podrá establecer desembocando en un mensaje ICMP de "Not Supported" o un reset de la conexión TCP, donde el extremo llamante deberá cambiar el protocolo de transporte de su mensaje de pedido sobre UDP para crear compatibilidad en la red y establecer la conexión. Siendo el caso más óptimo la compatibilidad al primer intento para aprovechar la capacidad y recursos de la red.

Otro punto crucial en el momento de decidir el protocolo que se utilizará en la capa de transporte es el tamaño máximo de segmento, el cual está involucrado directamente con el codec a utilizar, tomando en cuenta la notable diferencia de compresiones entre, por ejemplo, G.729, G.711, etc. En la RFC 3261 está definido el uso de UDP y TCP obligatoriamente, este último en caso de ser necesario algún tipo de fragmentación del paquete que exceda la MTU.

La negociación de codecs, puertos y servicios multimedia se realiza en el protocolo SDP (Session Description Protocol) embebido en SIP, donde comúnmente los puertos utilizados de SIP son el 5060 en texto plano (UDP y TCP) y el puerto 5061 en caso de TLS. Sin embargo, en la práctica se puede presentar el uso de puertos comprendidos entre el 5060 hasta el 5070.

1.6 Elementos SIP de red

Los terminales físicos conocidos como agentes usuarios (UA) pueden ser dispositivos en sí o softwares instalados en una PC, con el aspecto y/o funcionalidad de teléfonos tradicionales, pero que usan SIP y RTP para la comunicación. Están disponibles comercialmente gracias a muchos fabricantes. Algunos de ellos usan numeración electrónica (ENUM) o DUNDi para traducir los números existentes

de teléfono a direcciones SIP usando DNS (Domain Name Server), así llaman a otros usuarios SIP saltándose la red telefónica, con lo que el proveedor de servicio normalmente actúa de pasarela hacia la red pública conmutada de telefonía para los números de teléfono tradicionales (cobrando por ello).

SIP hace uso de elementos llamados servidores proxy para ayudar a enrutar las peticiones hacia la localización actual del usuario, autenticar y autorizar usuarios para darles servicio, posibilitar la implementación de políticas de enrutamiento de llamadas, y aportar capacidades añadidas al usuario. También aporta funciones de registro que permiten al usuario informar de su localización actual a los servidores proxy.

Aunque dos terminales SIP puedan comunicarse sin intervención de infraestructuras SIP (razón por la que el protocolo se define como punto-a-punto), este enfoque es impracticable para un servicio público. Hay varios implementaciones de softswitch (de Cisco, Sonus, Linksys y muchas más) que pueden actuar como proxy y elementos de registro. Otras empresas, como Ubiquity Software y Dynamicsoft tienen productos cuya implementación está basada en estándares, construidos sobre la especificación Java JAIN.

1.7. MENSAJES DEL PROTOCOLO SIP

a) DIRECCIONES SIP: SIP trabaja en una premisa simple de operación cliente servidor. Los clientes o endpoints son identificados por direcciones únicas definidas como URL's, es decir las direcciones vienen en un formato muy similar a una dirección de correo electrónico, a fin de que las paginas Web puedan contenerlos, lo que permite hacer click en un vinculo para iniciar una llamada telefónica. Las direcciones SIP siempre tienen el formato de user@host. El user puede ser: nombre, número telefónico. El host puede ser: dominio (DNS), dirección de red (IP).

b) MENSAJES SIP: SIP usa mensajes para la conexión y control de llamadas. Hay dos tipos de mensajes SIP: mensajes de peticiones y respuestas. Los mensajes SIP son definidos como sigue:

   I. INVITE: Solicita el inicio de una llamada. Los campos de la cabecera contienen:

      I. Dirección origen y dirección destino.
      II. El asunto de la llamada.

III. Prioridad de la llamada.
IV. Peticiones de enrutamiento de llamada.
V. Preferencias para la ubicación de usuario.
VI. Características deseadas de la respuesta.

II. TRYING: Indica que el servidor Proxy esta tratando de establecer la comunicación.
III. RINGING: Indicación de aviso de llamado.
IV. BYE: Solicita la terminación de una llamada entre dos usuarios.
V. REGISTER: Informa a un servidor de registro sobre la ubicación actual del usuario.
VI. ACK: Usado para facilitar un intercambio confiable de mensajes entre los pares. Confirmación de diferentes campos del mensaje INVITE.
VII. CANCEL: Cancela una solicitud pendiente.
VIII. OPTIONS: Solicita información a una Host acerca de sus propias capacidades. Se utiliza antes de iniciar la llamada a fin de averiguar si ese host tiene la capacidad de transmitir VoIP, etc.
IX. 200OK: Sirve para enviar confirmaciones satisfactorias de diferentes sucesos.
X. INFO: Usada para señalización de sesiones de media.

1.8. LLAMADA DE PC A PC SOBRE TCP

- Para establecer una llamada, el llamante crea una conexión TCP con el llamado.
- La conexión se realiza utilizando un acuerdo de tres vías.
- Envía un mensaje INVITE en un paquete TCP, indicando la dirección de destino, la capacidad, los tipos de medios y los formatos del llamante.
- Si el llamado acepta la llamada, responde con un código de respuesta tipo http (200 para aceptación). Opcionalmente también puede proporcionar información sobre sus capacidades, tipos de medios y formatos.
- El llamante responde con un mensaje ACK para terminar el protocolo y confirmar la recepción del mensaje 200.
- En este punto, pueden comenzar el flujo de datos utilizando el protocolo RTP.
- El flujo de datos se controla mediante el protocolo RTCP.

- Cualquiera puede solicitar la terminación de la llamada enviando un mensaje BYE.
- Cuando el otro lado confirma su recepción, se termina la llamada.

1.9. SIP LLAMADAS Y TRANSACCIONES

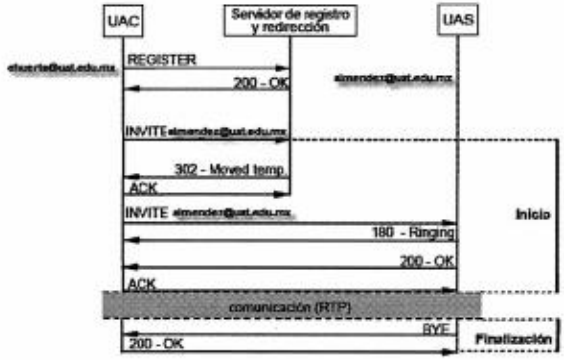


Figura 3. Llamadas y Transacciones.

1.10. REAL-TIME TRANSPORT PROTOCOL

RTP son las siglas de Real-time Transport Protocol (Protocolo de Transporte de Tiempo real). Es un protocolo de nivel de aplicación (no de nivel de transporte, como su nombre podría hacer pensar) utilizado para la transmisión de información en tiempo real, como por ejemplo audio y video en una video-conferencia.

Está desarrollado por el grupo de trabajo de transporte de Audio y Video del IETF, publicado por primera vez como estándar en 1996 como la RFC 1889, y actualizado posteriormente en 2003 en la RFC 3550, que constituye el estándar de Internet STD 64. Inicialmente se publicó como protocolo multicast, aunque se ha usado en varias aplicaciones unicast. Se usa frecuentemente en sistemas de streaming, junto a RTSP, videoconferencia y sistemas push to talk (en conjunción con H.323 o SIP). Representa también la base de la industria de VoIP.

La RFC 1890, obsoleta por la RFC 3551 (STD 65), define un perfil para conferencias de

audio y vídeo con control mínimo. La RFC 3711, por otro lado, define SRTP (Secure Real-time Transport Protocol), una extensión del perfil de RTP para conferencias de audio y vídeo que puede usarse opcionalmente para proporcionar confidencialidad, autenticación de mensajes y protección de reenvío para flujos

de audio y vídeo. Va de la mano de RTCP (RTP Control Protocol) y se sitúa sobre UDP en el modelo OSI.

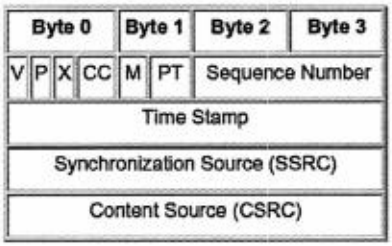


Figura 4. Estructura de Encabezado.

- RTP número de versión (V - versión number): 2 bits. La versión definida por la
- especificación actual es 2.
- Relleno (P - Padding): 1 bit. Si el bit del relleno está colocado, hay uno o más bytes al final del paquete que no es parte de la carga útil. El byte más último en el paquete indica el número de bytes de relleno. El relleno es usado por algunos algoritmos de encriptación.
- La extensión (X - Extensión): 1 bit. Si el bit de extensión está colocado, entonces el encabezado fijo es seguido por una extensión del encabezado. Este mecanismo de la extensión posibilita implementaciones para añadir información al encabezado RTP.
- Conteo CSRC (CC): 4 bits. El número de identificadores CSRC que sigue el encabezado fijo. Si la cuenta CSRC es cero, entonces la fuente de sincronización es la fuente de la carga útil.
- El marcador (M - Marker): 1 bit. Un bit de marcador definido por el perfil particular de media.
- La carga útil Type (PT): 7 bits. Un índice en una tabla de perfiles de media que
- describe el formato de carga útil. Los mapeos de carga útil para audio y video están especificados en el RFC 1890.
- El número de Secuencia: 16 bits. Un único número de paquete que identifica la posición de este en la secuencia de paquetes. El número del paquete es incrementado en uno para cada paquete enviado.
- Timestamp: 32 bits. Refleja el instante de muestreo del primer byte en la carga útil. Varios paquetes consecutivos pueden tener el mismo timestamp si son lógicamente generados en el mismo tiempo - por ejemplo, si son todo parte del mismo frame de video.

- SSRC: 32 bits. Identifica la fuente de sincronización. Si la cuenta CSRC es cero, entonces la fuente de carga útil es la fuente de sincronización. Si la cuenta CSRC es distinta a cero, entonces el SSRC identifica el mixer (mezclador).
- CSRC: 32 bits cada uno. Identifica las fuentes contribuyentes para la carga útil. El número de fuentes contribuyentes está indicado por el campo de la cuenta CSRC; Allí puede haber más de 16 fuentes contribuyentes. Si hay fuentes contribuyentes múltiples, entonces la carga útil son los datos mezclados de esas fuentes.

2. OBJETIVO.

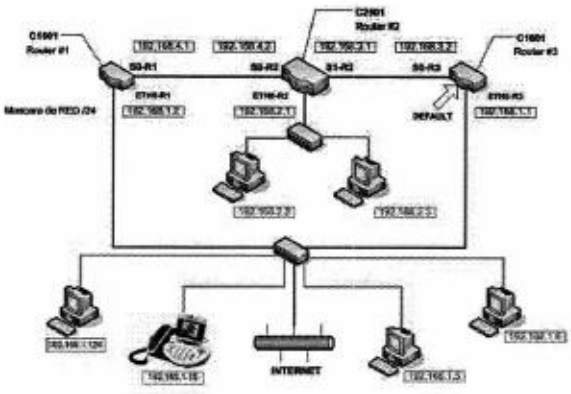
2.1 ALCANCE DEL PROTOCOLO SIP Y SU FUNCIONAMIENTO.

- Análisis del paquete IP, capa de red, capa de transporte y capa de aplicación (SIP).
- Análisis del establecimiento de una llamada. (SDP, etc.)
- Mensajes del protocolo SIP:

    - INVITE
    - Trying
    - Ringing
    - 200 ok
    - BYE
    - REGISTER
    - ACK
    - CANCEL
    - OPTIONS

- Verificar en los ensayos realizados el funcionamiento del protocolo.

MAQUETA



En el siguiente gráfico podemos observar el diagrama de la maqueta utilizada para realizar las capturas.

2.2 ESCENARIOS

1. *Llamada exitosa desde el softphone al videophone.*
   Objetivo: observar y analizar el establecimiento, transcurso y desconexión de una comunicación con ambos equipos disponibles.

2. *Llamada exitosa desde el videophone al softphone.*
   Objetivo: observar y analizar el establecimiento, transcurso y desconexión de una comunicación con ambos equipos disponibles. Diferencias con el caso anterior.

3. *Llamada exitosa desde el softphone 1 al softphone 2 y viceversa.*
   Objetivo: observar y analizar el establecimiento, transcurso y desconexión de una comunicación con ambos equipos disponibles. Diferencias respecto a los casos anteriores.

4. *Llamada desde el videophone al softphone en modo DND (Do Not Disturb).*
   Objetivo: observar y analizar el establecimiento fallido en una comunicación con el softphone no disponible.

5. *Llamada desde el softphone al videophone ocupado.*
   Objetivo: observar y analizar el establecimiento fallido en una comunicación con el videophone en uso.

6. *Llamada fallidas desde el softphone 1 al softphone 2.*
   Objetivo: observar y analizar las causas por las cuales no fue posible realizar el establecimiento de las comunicaciones.

ESCENARIO 1. LLAMADA EXITOSA DESDE EL SOFTPHONE AL VIDEOPHONE

Objetivo:

- Observar y analizar el establecimiento, transcurso y desconexión de una comunicación con ambos equipos disponibles.

Desarrollo:

Para lograr el objetivo planteado se configuró el softphone (Figura 6) en modo peer to peer (Figura 7), esto es debido a que las pruebas realizadas fueron dentro de una misma red sin la intervención de un SIP proxy. Esta configuración se debe a que el dispositivo debería estar registrado a un Proxy (en este caso el teléfono SIP) para realizar o recibir llamadas. La dirección IP configurada en la PC en la cual se instaló el Linksys era 192.168.1.128, como luego se verá en la captura es quien inicia la comunicación. Los puertos configurados en el software fueron del 5060 al 5062 (Figura 8) y los códecs habilitados eran G.711, G.729, etc.

Figura 6. Softphone.

Figura 7. Softphone en modo peer to peer

Figura 8. Puertos configurados 5060 al 5062

La configuración en el videophone fue básicamente la asignación de la dirección IP 192.168.1.15, en forma estática. Los codecs no son configurables en el teléfono utilizado, el único codec que soporta es G.711 A-law. Además, para realizar comunicaciones internas a la red el dispositivo no debe estar registrado en ningún Proxy.

EJECUCIÓN:

Para realizar la prueba se generó una llamada a la dirección IP del teléfono SIP desde el Linksys. Cuando se detectó la llamada entrante fue aceptada. Luego de unos segundos desde el softphone se finalizó la llamada. El siguiente gráfico muestra el intercambio de mensajes en la captura realizada desde la PC en que se encontraba el softphone.

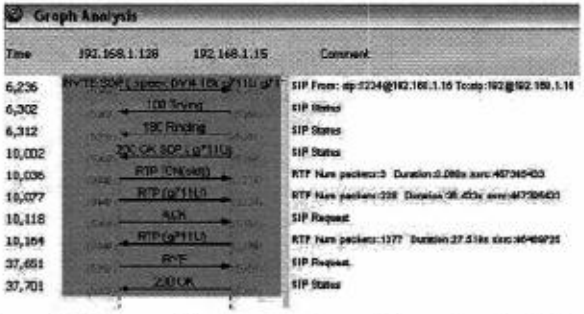


Figura 9. Intercambio de mensajes en la captura desde la PC donde se encuentra el softphone.

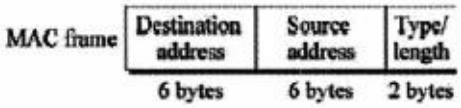
Como se puede apreciar, la comunicación sigue el intercambio teórico de mensaje antes mencionados. El establecimiento de la llamada comienza con un mensaje Invite por parte del softphone, donde se envía encapsulado sobre SIP, un mensaje de SDP (session description protocol), con el fin de negociar códec, puertos, etc. En la siguiente captura se analizará este protocolo más detalladamente. Luego el siguiente mensaje es un Trying enviado por el teléfono SIP en respuesta al Invite recibido, luego en el mismo sentido se envía un mensaje de Ringing, para confirmar que se está dando aviso de la llamada entrante. Al aceptar la comunicación (simplemente levantando el tubo en el videoteléfono) éste envía un mensaje de 200 OK con un mensaje de SDP, encapsulado sobre SIP, proponiendo codec a utilizarse (G.711 U-Law). A continuación se envían paquetes RTP (real-time transport protocol), con payload Comfort noise, este paquete sirve para ahorrar ancho de banda durante los silencios de voz, emulando un ruido en el otro extremo, para que el usuario no tenga la sensación de interrupción de la comunicación al existir silencios.

El teléfono SIP confirma la negociación con un mensaje ACK y luego comienza el intercambio de audio por medio de RTP el protocolo que se utiliza para el intercambio de media, en este sentido es importante aclarar que SIP es el protocolo utilizado para la señalización y en esta etapa de la comunicación no aparece. Finalmente se termina la comunicación desde el lado del softphone y envía un mensaje BYE de SIP, al cual el otro extremo responde con un 200 ok para confirmar el fin del llamado.

El análisis comienza con el primer mensaje enviado que es el llamado INVITE, a continuación se muestra la captura del paquete correspondiente:

Figura 10. Comienzo de análisis INVITE

Figura 11. Aquí se ve en el recuadro superior de color azul el encabezado a nivel de capa de enlace, recordemos su estructura:



Como se ver en la captura el protocolo utilizado es Ethernet, por lo tanto el tercer campo indica a que protocolo corresponde el próximo header. En el caso de ser 802.3 este campo refleja la longitud del campo de Data real. Decimos real porque 802.3 se encarga de controlar la longitud mínima de trama, rellenando si hace falta el campo de Data. Por eso es necesario indicar la longitud real del campo de Data, para poder descartar el relleno si lo hubo. Si este campo tiene un valor mayor a 1536 se trata de Ethernet, de otra forma estaríamos hablando de 802.3.

Es posible ver la dirección MAC de destino remarcada en rojo: 00:16:B6:4D:EB:13, donde los primeros 24 bits indican el fabricante del dispositivo, en este caso sería Linksys, por ser el fabricante.

La dirección MAC de origen remarcada en azul es: 00:12:F0:5C:64:5E, donde HP es el fabricante de la placa de red de la pc en la cual se instaló el software. Luego en el siguiente campo, remarcado en verde, podemos ver que el protocolo de red que encapsula es ip.

A continuación se analiza el encabezado de Ip recuadrado en color rojo en la captura anterior.

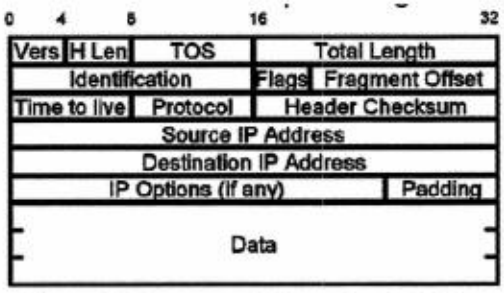


Figura 12. Estructura del encabezado de ip.

- En este caso podemos ver que se trata de un datagrama de Ip versión 4 y de una longitud de 20 bytes (campos remarcados en amarillo), es importante aclarar que el valor de este último campo corresponde a palabras

de 32 bits, por lo tanto el valor que aparece es 5. Es decir 5 x 32 bits=160 bits=20 Bytes, a partir de este valor podemos inferir que se trata de un datagrama sin opciones, de otra manera el encabezado tendría un tamaño mayor a 20 Bytes.

- A continuación vemos que el campo de Type of Service (remarcado en lila) tiene un valor nulo, no se hace ninguna distinción para este datagrama en cuanto a confiabilidad, prioridad, retardo o throughput.

- El siguiente campo (resaltado en celeste) es el tamaño total del datagrama que tiene un valor de 903 Bytes.
- Luego el campo de identificación (en gris) del paquete muestra un valor de 5525 que sirve para distinguirlo de otros paquetes, esto es necesario para identificar los fragmentos correspondientes a un datagrama que ha sido fragmentado.
- Vemos que el campo de flags (en verde) tiene un valor nulo, es decir que el datagrama puede ser fragmentado (Don´t fragment=0) y que o bien no ha sido fragmentado o es el último fragmento.
- Luego el siguiente campo de fragment offset (en lila) se encuentra en cero, esto quiere decir que es el primer fragmento o no ha sido fragmentado. Por éstos dos últimos valores inferimos que el datagrama no fue fragmentado.
- El campo de Time to Live (en amarillo) tiene un valor de 128, es decir que este paquete podría atravesar 128 redes como máximo hasta llegar a destino, éste valor se decrementa por cada salto, si éste llegara a 1, se descartaría el paquete.
- El próximo campo indica el protocolo que contiene el payload, aquí se ve que el protocolo utilizado en la capa de transporte es UDP, el valor es de 9 para este protocolo.
- A continuación tenemos el checksum (en verde), que sirve para control de errores del header, en este caso vemos que el encabezado no contiene errores.
- Los siguientes campos son las direcciones ip de fuente y destino respectivamente. En este caso se tiene que 192.168.1.128 es quien generó el datagrama y 192.168.1.15 es el destino de dicho datagrama. (en celeste y gris respectivamente)

El análisis continúa con el próximo encabezado, el de transporte, que en este caso como se dijo anteriormente se utilizará UDP. Recordemos su estructura.

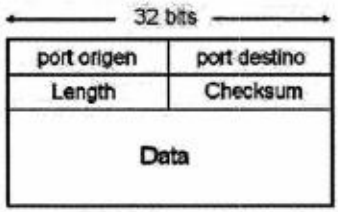


Figura 13. Encabezado de transporte.

- El primer campo indica el puerto origen, en este caso por tratarse de sip era esperable que sea 5060 al igual que el puerto de destino (remarcados en celeste y azul respectivamente).
- Luego vemos que el campo siguiente (en rojo) es el que indica el tamaño del segmento UDP incluyendo el header, como era de esperar tiene un valor de 883 Bytes, 20 Bytes menos que del datagrama, el tamaño del header de éste.
- Y por último (en gris) tenemos el checksum del segmento, este es opcional y cuando no se lo calcula, este campo se pone en cero. Aquí vemos que se ha utilizado y no hubo errores.

En la capa de aplicación tenemos el mensaje encapsulado en sip y sus protocolos dependientes, en los próximos escenarios se analizarán estos protocolos más detalladamente.

El próximo mensaje que se envía en respuesta al INVITE anterior es un TRYING, a continuación podemos ver la captura:

Figura 14. Mensaje TRYING

El análisis de este mensaje es muy similar al anterior, sólo se indicarán las diferencias más importantes en cada nivel.

A nivel de enlace podemos ver que los campos de direcciones mac se invierten dado que la trama es enviada en sentido inverso, desde el videoteléfono hacia el softphone. Otra vez en el campo de type vemos que se tiene IP como próximo header. A nivel de red se ve también que las direcciones IP han sido invertidas al igual que las mac. Como diferencia más importante podemos ver que el tamaño del datagrama disminuyó dado que este mensaje no encapsula ningún protocolo auxiliar.

En cuanto al encabezado de UDP, se puede comprobar nuevamente que difiere en 20 bytes respecto al de IP.

Luego se comprueba que el mensaje a nivel de aplicación es el TRYING antes mencionado.

El próximo mensaje es el RINGING, podemos ver su captura, aquí es importante tener en cuenta que el sentido de este datagrama es el mismo que el anterior (trying), coincidiendo con lo que se anticipó en forma teórica.

Figura 15. Captura Mensaje RINGING

Aquí se verifica en los campos de direcciones mac e ip que el sentido del mensaje es otra vez desde el teléfono SIP hacia el Linksys. A nivel de aplicación podemos comprobar que se trata de un mensaje ringing.

El siguiente mensaje es un 200 OK, enviado desde el softphone. La captura se muestra a continuación:

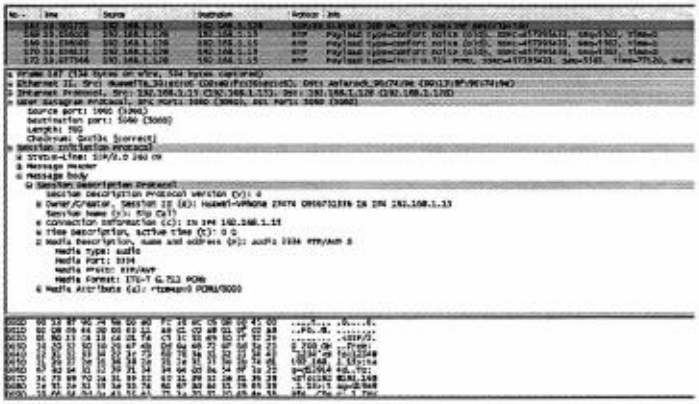


Figura 16. Captura de mensaje 200 OK

Este mensaje tiene la función de confirmar la aceptación del llamado. Aquí se ve que se encuentra un mensaje de SDP encapsulado en SIP con el fin de confirmar el codec que será utilizado en la comunicación. En este caso se confirma G.711 U-Law como anticipamos anteriormente. Esto será analizado detalladamente más adelante.

Los siguientes mensajes son de RTP, el primero contiene un mensaje de Comfort Noise, esto sirve para acomodar los parámetros de los filtros FIR que emularán un leve ruido al detectar silencios para confort del usuario. Aquí se ve este tipo de mensaje dado el codec que se está utilizando, G.711 no viene por default con esta opción, por lo tanto se realiza esta facilidad.

Figura 17. Mensaje de Comfort Noise

El siguiente mensaje es enviado desde el soft al videoteléfono y se trata de un ACK, cuya función es confirmar diferentes valores de los campos enviados

en el mensaje INVITE. Veamos a continuación como estos campos coinciden en ambos mensajes.

Figura 18. Mensaje enviado desde el soft al videoteléfono y se trata de un ACK.

A partir de ahora comienza el intercambio de audio sobre RTP. Como se ve en el primer gráfico del intercambio de mensajes, sólo hay paquetes desde el videoteléfono hacia el softphone, esto se debe a que en el momento de efectuar las capturas no disponíamos de un micrófono en la pc, por lo tanto no se enviaron paquetes en este sentido. A continuación se muestra sólo un paquete de RTP a modo de ejemplo.

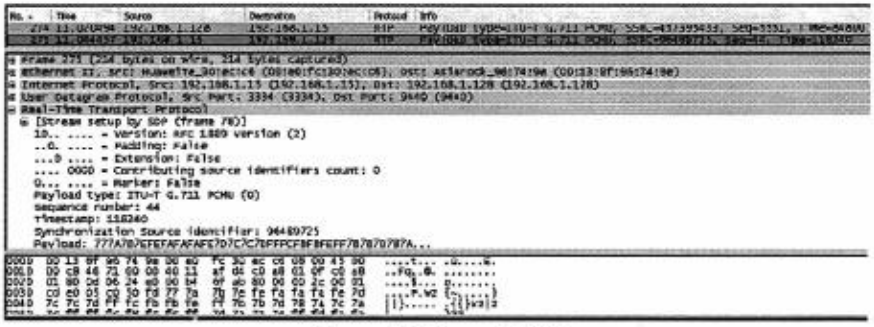


Figura 19. Paquete RTP

En este ejemplo se finalizó la comunicación desde el softphone. Aquí se ve, dado que es ese extremo quien envía el mensaje BYE. A continuación podemos apreciar la captura de este paquete.

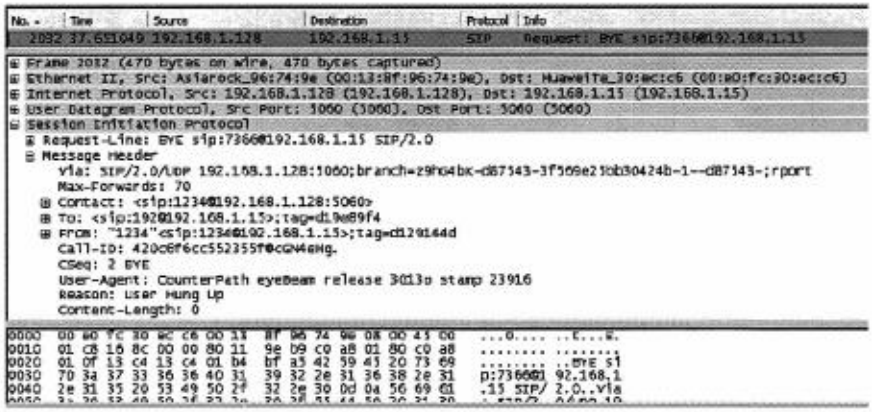


Figura 20. Finalización de la comunicación desde el softphone.

Como se vio anteriormente este mensaje notifica la terminación de una llamada. Del otro extremo este mensaje es contestado con un 200 OK confirmando el término de la llamada.

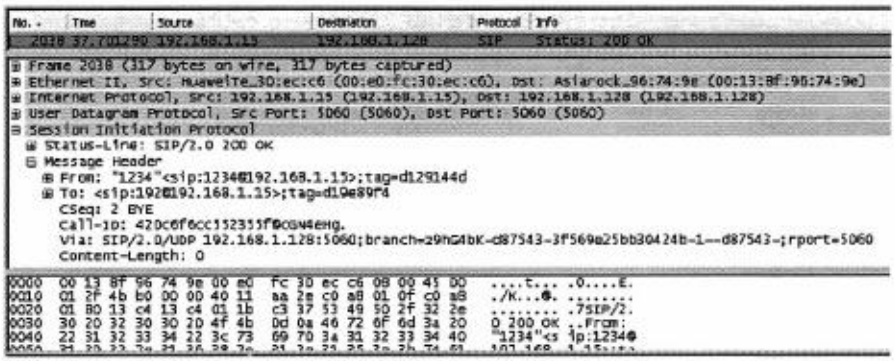


Figura 21. Respuesta mensaje 200 OK

Para finalizar con este escenario podemos concluir que se cumplió con el marco teórico presentado anteriormente. La comunicación se estableció dentro de los parámetros normales, exitosamente. No se encontraron mensajes no esperados. Los campos de nivel de enlace, red y transporte que fueron analizados en cada caso coincidieron con los resultados esperados.

ESCENARIO 2. LLAMADA EXITOSA DESDE EL SOFTPHONE AL VIDEOPHONE

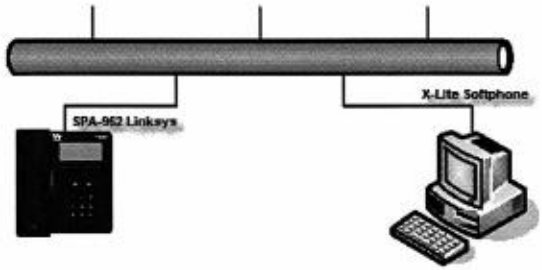
Objetivo:

- Observar y analizar el establecimiento, transcurso y desconexión de una
- comunicación con ambos equipos disponibles.
- Diferencias con el caso anterior.

Ejecución:

El escenario en este caso es el siguiente:

a) El usuario A (teléfono Linksys) llama al usuario B (softphone X-Lite).
b) Luego de dejar sonar el Softphone, la llamada es atendida.
c) Se espera algunos segundos. El usuario B finaliza la llamada.



La siguiente figura muestra el call flow de dicha llamada:

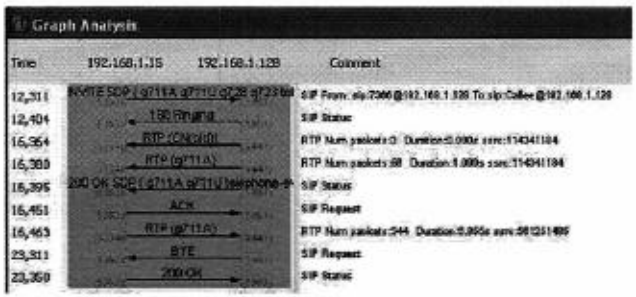


Figura 23. Call flow de llamada.

2.1 INVITE

La llamada es iniciada por el Usuario A, utilizando el método INVITE. Es interesante ver los bytes del mensaje como son presentados por el ethereal:

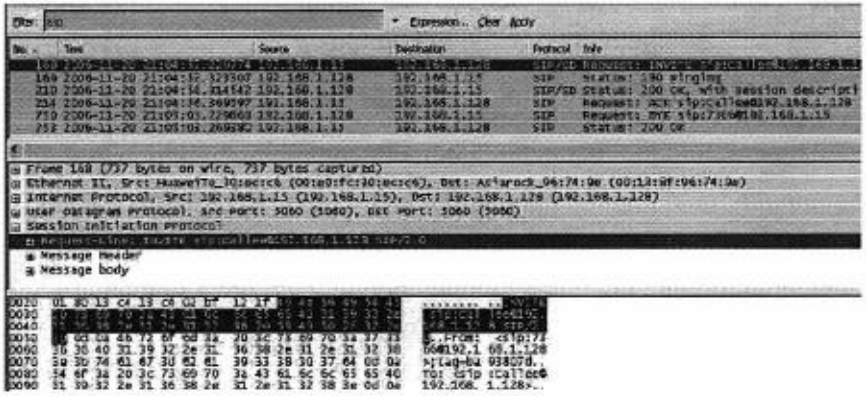


Figura 24 . Bytes de mensaje en método INVITE

Se puede ver que el mensaje se encuentra codificado en ASCII, lo que facilita la decodificación por parte de los usuarios y administradores. El costo de enviar los mensajes en texto plano es un mayor uso de ancho de banda, pero como veremos a continuación, la señalización en condiciones normales, requiere pocos mensajes para establecer la llamada.

Como puede verse en la figura, se identifican 3 segmentos dentro del mensaje INVITE:

- Request Line
- Message Header
- Message Body

Request Line: INVITE sip:Callee@192.168.1.128 SIP/2.0

En esta sección del mensaje se puede ver principalmente el Método de SIP utilizado (INVITE, TRYING, etc.) y la versión de SIP. A diferencia de H.323, SIP no asegura retrocompatibilidad entre las versiones, por lo que podría ser que un método soportado en 1.0 no se siga usando en una versión posterior.

En el caso del mensaje INVITE, se puede ver el destinatario (TO) del mensaje: Callee@192.168.1.128 SIP Message Header:

Figura 25. En la siguiente Figura se puede ver el header de SIP

El encabezado de SIP esta conformado por los siguientes campos:

- From
- To
- CSeq
- Call-ID
- Via
- Contact
- Max Forwareds
- Allow
- Content-Length
- content-Type

Cada uno de estos campos esta separado delimitado usando los ASCII 13-10 (CR-CF)

From:
El campo FROM es el identificador lógico del User Agent que genera el pedido (UA Client). Esta compuesto por el URI (Uniform Resource Identificator) y opcionalmente el parámetro DISPLAY Name, el cual es el nombre que se presentara si el usuario tiene el servicio de caller ID activo.

Adicionalmente, dentro del FROM se encuentra el parámetro TAG, el cual es un identificador generado por el user agent client en el momento de hacer el pedido. Este identificador, junto con el parámetro con el identificador de la llamada (Call-

ID), sirven para identificar el dialogo entre el User Agent Client y el User Agent Server (UAS).

Figura 26. Campo FROM capturado, dentro del INVITE:

Es interesante notar que en este caso el URI: 7366@192.168.1.128 esta utilizando el identificador del Videophone (7366), pero la IP que se utilizó para formar dicho mensaje es la del softphone. Esto es porque los UAC conforman el URI utilizando la IP del SIP-Proxy, y en este escenario, como se configuro el terminal para trabajar en modo peer to peer, se utiliza la ip del destinatario.

El tag que se generó para establecer este diálogo es: BA93807d.

TO:

En el campo TO se indica el destinatario lógico del pedido. El UAC genera el campo TO a partir de lo ingresado por el usuario del videophone. En este caso, la llamada se generó marcando la IP del destino, por lo que el UAC cliente generó el URI automáticamente:

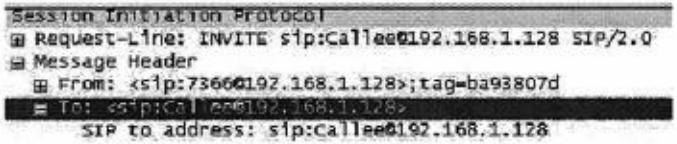


Figura 27. Mensaje que muestra el Campo TO

El URI generado automáticamente utiliza el string "Callee@<IP marcada>". Notar que dentro del TO no se está utilizando el TAG.

CSeq:

Este parámetro define el orden de las transacciones. Consiste de un número de secuencia y de un método. Este número de secuencia se va incrementando de a 1 por vez. En el caso de esta captura, este parámetro tiene el valor: CSeq: 6 INVITE. El método utilizado para conformar el CSeq debe ser el mismo método utilizado para generar el diálogo.

Call-ID:

Este parámetro es un identificador único que agrupara una serie de mensajes. Es obligatorio que sea el mismo durante todos los mensajes que intercambiados entre UAC y UAS. Para asegurar que el identificador sea único se recomienda utilizar RFC 1750. (Uso de cryptographically random identifiers)

El call-ID capturado es:

Call-ID: 5213e9a3785258c01d1b1cc0ba93807d@192.168.1.15

Resulta interesante notar que el UAC del Videophone utiliza su propia IP como parte del new call ID.

VIA:

El parámetro VIA identifica el protocolo de transporte y la ubicación a donde se debe enviar la respuesta. El UAC debe insertar este parámetro obligatoriamente siempre que se genera el request. Es importante notar que este parámetro está presente en los mensajes enviados por el UAC solamente.

Dentro del campo VIA se incluye el parámetro Branch, el cual se utiliza para identificar la transacción dentro del UAC. Este parámetro debe ser único.

La RFC 3261 especifica que el valor que tomará por defecto el Branco ID comienza con z9hG4bK.

En la captura se puede ver que el protocolo de transporte es UDP, y la ip y el puerto del mismo es la del Videophone.

Via: SIP/2.0/UDP 192.168.1.15:5060;branch=z9hG4bKba93807da

Contact:

Este parámetro se utiliza para identificar la instancia específica del UA a donde se puede enviar la requests, fuera del dialogo en curso.

En esta captura, este parámetro tiene el valor:

Contact: <sip:7366@192.168.1.15>

Max Forwards:

Este parámetro sirve para limitar la cantidad de saltos que un request puede transitar.

Por defecto este valor se fija en 70 saltos. Si este parámetro llega a 0, la llamada es terminada con código 483 (too many hops).

Allow:

Indica todos los métodos soportados por el UA. Si este parámetro no esta presente, no significa que el UA no soporta ningún método, sino que le mismo no los proveyó. Este mensaje busca optimizar la cantidad de mensajes que se necesitan para concluir.

En el caso del mensaje de invite, este parámetro incluyó:

Allow: INVITE, ACK, CANCEL, OPTIONS, BYE, REGISTER, PRACK, UPDATE, INFO

Content Length:

Indica la cantidad de bytes del message body. Si no hay información en el cuerpo del mensaje, este parámetro debe ser seteado en 0. Este parámetro puede ser abreviado utilizando"l:".

En este caso el largo del mensaje es de:

Content-Length: 274

Content-Type:

Este parámetro indica el tipo de información contenida en el body. Algunos ejemplos son:

Content-Type: application/sdp

c: text/html; charset=ISO-8859-4

En este caso, el protocolo que se utilizó:

Content-Type: application/sdp

Message Body:

El cuerpo del mensaje SIP, como se vio en el header, esta usando SDP:

Figura 28. Cuerpo del mensaje SIP

El protocolo Session Description Protocol está definido en la RFC 2327. La ventaja de usar un protocolo adicional para establecer los parámetros que definirán el establecimiento de la sesión es que esto permite a SIP adaptarse fácilmente tanto para comunicaciones de voz, como de aplicaciones multimedia.

El protocolo SDP consiste en diferentes tags, cada uno de los cuales describe un parámetro en particular de la sesión. Se puede ver los siguientes parámetros:

v - SDP Protocol Version

o – Owner- Creator, Session Id: Dentro de estos parámetros se puede ver el identificador de la session, la IP y el owner:

Figura 29. Parámetros Owner/creator, Session id

c – Connection Information:

En este parámetro se presenta la IP donde se debe enviar el stream de audio. En este caso es la misma IP del VideoPhone: 192.168.1.15.

t- Time description: Es el tiempo que lleva activa el stream de audio.

m- Media Description:

Este tag tiene la descripción de todos los codecs soportados por el UAC que inició la conversación. Se especifica el media type, que informa que el contenido de la session es audio, el media port, el cual define el puerto UDP que se asignó para recibir el stream de RTP. Los codecs de audio (video o dtmfs) se identifican utilizando valores preestablecidos estándar: 8 G.711A, 0 G711U, 4 G.723, etc.

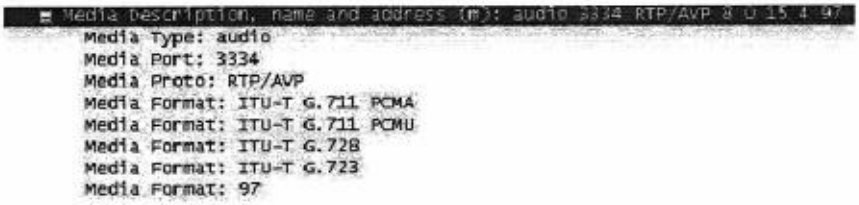


Figura 30. Codecs que soporta el UAC

Conforme sea necesario, cada codec adapta sus parámetros utilizando media attributes:

a – Media attributes:

En la figura se puede ver como se especifican los diferentes parámetros para cada códec en particular. Los atributos se relacionan con el codec a través del parámetro "Media Format".

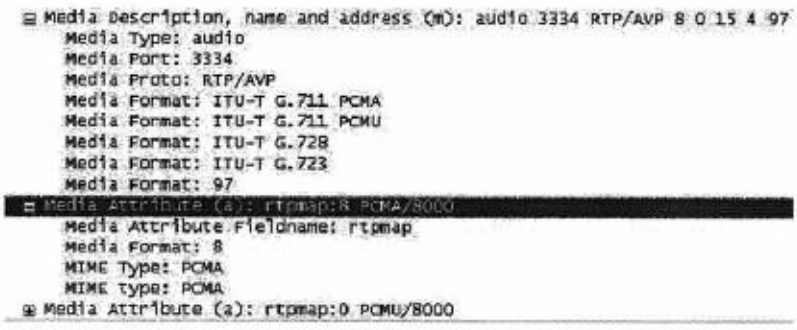


Figura 31. Despliegue de todos los códecs que tiene el UAC

2.2 RINGING

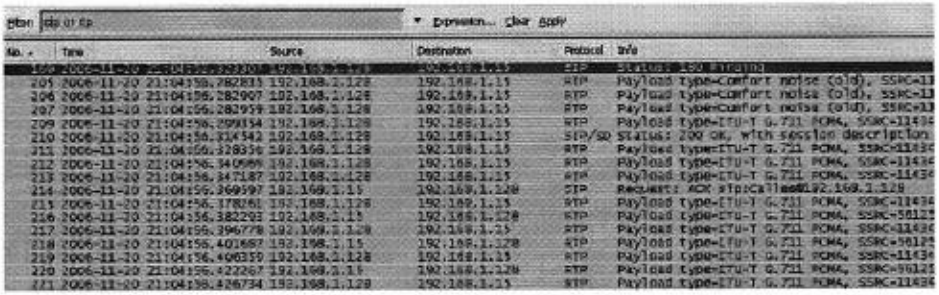
Este mensaje es enviado por el UAS (en este caso, el softhpone) para indicar que el usuario está siendo alertado sobre la invitación. Es importante notar que en comparación al escenario 1, no se está enviando el mensaje TRYING. Esto es propio de la implementación del softphone, ya que no hay una interfase a excitar, como por ejemplo en un teléfono en el que hay una interfase con una línea analógica, la que tarda un tiempo hasta que comienza a sonar.

Figura 32. Mensaje Ringing

Mirando en detalle este mensaje, y comparándolo contra los campos vistos previamente en el INVITE, se puede observar que:

a) EL FROM y el TO enviados en el header del INVITE SE MANTIENEN, es decir, el UAS no cambia los valores provistos por el UAC. La única diferencia es que el UAS agrega el TAG en el TO, con lo que la llamada queda identificada por el Call-ID, tag provisto por el UAC, tag provisto por UAS.
b) El CSeq mantiene el identificador provisto en el INVITE.
c) El parámetro VIA mantiene los mismos valores que en el INVITE, y lo mismo es válido para el Branch.
d) El parámetro CONTACT es actualizado con los valores propios del UAS.

Es interesante ver un detalle de lo que ocurre una vez recibido el mensaje de Ringing:



Se puede ver que el softphone, una vez que envío el mensaje de Ringing, como tiene la información de SDP del Videophone, ya puede comenzar a enviar audio (en un sentido). Antes de enviar el 200 OK con la información necesaria para que el videophone abra el canal en la dirección contraria, envía paquetes de RTP con comfort noise, para que el usuario no quede escuchando un canal mudo, lo que daría la sensación de que la comunicación fallo.

Luego, el sofphone envía el mensaje 200OK con el SDP para que el UAC puede abrir el canal, en el sentido inverso.

2.3 200 OK

Este mensaje conserva los mismos valores en los parámetros que el RINGING.

En el Message Body se puede ver que se está enviando como protocolo preferido G.711 Mu (8), y además se utilizará RTP/AVP (101) para los eventos del teléfono (dtmfs). El audio será dirigido al puerto UDP 9440.

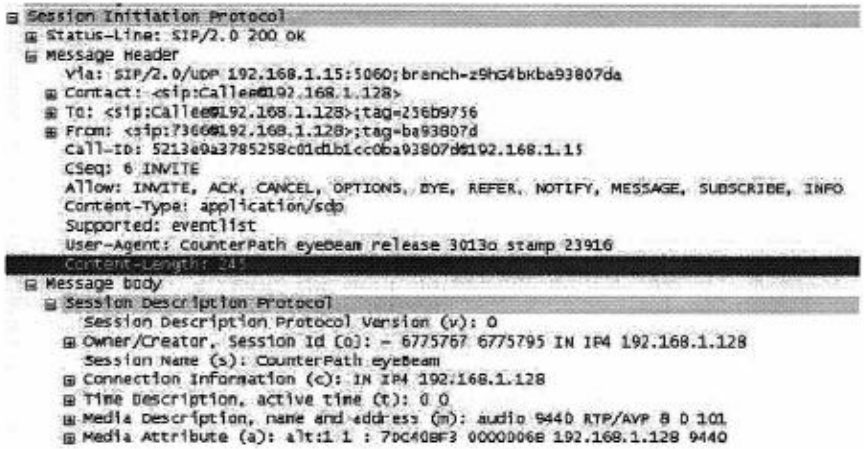


Figura 34. Parámetros de Mensaje 200 OK

2.4 ACK VIDEOPHONE -> SOFTPHONE

Este mensaje confirma que el Videophone recibió la notificación 200 OK del UAS, y que el puerto fue abierto con la información provista en dicho mensaje. Se puede observar que luego de este mensaje, el Videophone comienza a enviar RTP utilizando G.711 al puerto 9440.

2.5 BYE SOFTPHONE VIDEOPHONE

Lo primero que resulta interesante es que el mensaje BYE es generado por el Softphone en lugar del videophone, quien era el que había iniciado la conversación.

Como en este caso el UAC es el softphone, se invierte el FROM y el TO, y el VIA se actualiza con los valores correspondientes al Softphone.

Sin embargo, tanto los tags como el call-ID se mantienen como en los otros mensajes.

Figura 35. Mensaje BYE en Softphone

2.6 200 OK VIDEOPHONE SOFTPHONE

Este mensaje confirma que la llamada fue desconectada del lado del Videophone. Resulta importante destacar que en este caso el Content-Length está en 0 porque no se incluye información de SDP.

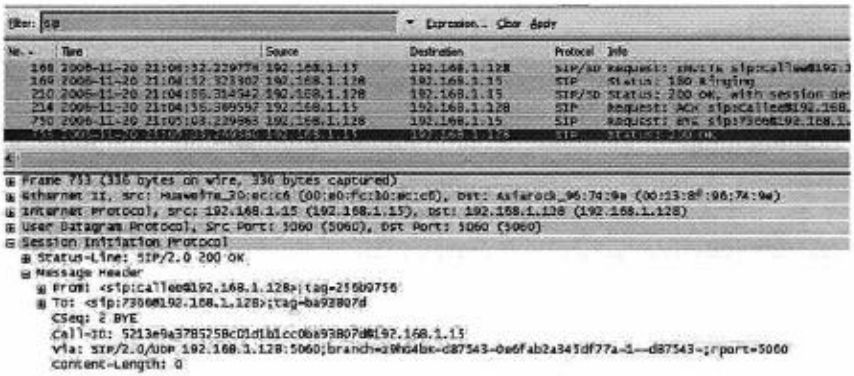


Figura 36. Confirmación de llamada con mensaje 200 OK

CONCLUSIONES

Del análisis anterior resulta interesante notar que como los parámetros FROM, TO, VIA y CONTACT se adaptan de acuerdo al rol que esta cumpliendo cada Terminal en cada momento (UAC o UAS).

También resulta interesante destacar como se realiza el control de flujo de los mensajes de SIP, lo que justifica porque la mayoría de los fabricantes utiliza UDP como protocolo de transporte en lugar de TCP.

ESCENARIO 3. LLAMADA EXITOSA DESDE EL SOFTPHONE 1 AL SOFTPHONE 2 Y VICEVERSA

En el presente escenario se realiza una comunicación exitosa entre dos softphones. El fin de esta captura es verificar que al no existir un sip proxy entre los dispositivos no aparece ningún mensaje del tipo trying. Para efectuar la comunicación nuevamente se tuvo que configurar ambos lados como peer-to-peer. Vemos a continuación el intercambio de mensajes.

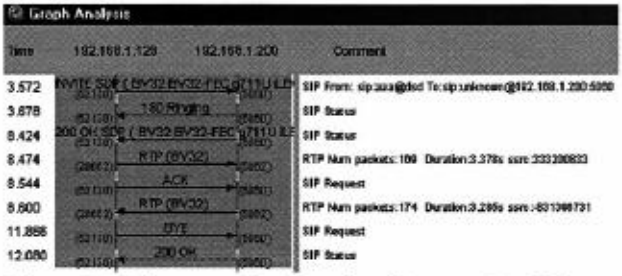


Figura 37. Diagrama de comunicación entre 2 softphone

Aquí constatamos que luego del primer mensaje INVITE el otro extremo responde directamente con un mensaje del tipo RINGING obviando el TRYING. Esto se debe nuevamente a que al no existir un Proxy no tiene sentido en una conexión peer to peer el envío de este mensaje. A continuación vemos la captura.

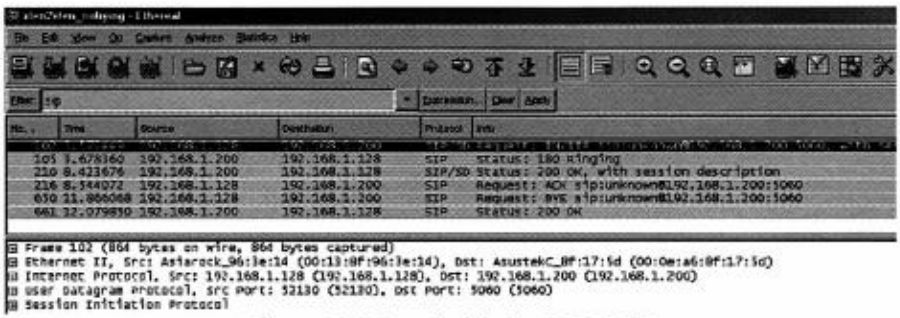


Figura 38. Mensaje del tipo RINGING

Esta comunicación se desarrolla en términos normales igual que los anteriores dos escenarios. En este caso podemos mencionar que aunque no se disponían de micrófonos en ninguno de los dos extremos hay intercambio de audio por RTP. Esto se debe a que al haberse negociado el codec G.711 U-Law no se implementa la supresión de silencios, es decir que hay intercambio de audio pero sin ningún contenido.

En conclusión, se verifica la hipótesis respecto del Trying anteriormente planteada.

ESCENARIO 4. LLAMADA DESDE EL VIDEOPHONE AL SOFTPHONE EN MODO DND (DO NOT DISTURB)

A continuación se muestra los mensajes intercambiados.

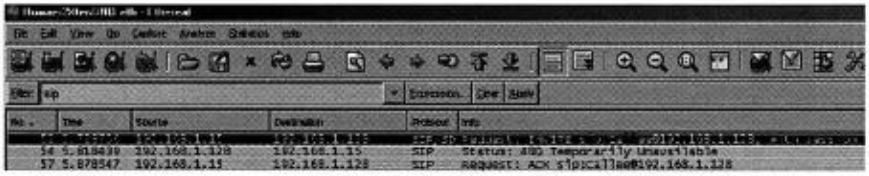


Figura 39. Llamada modo DND

Figura 40. SoftPhone configurado en modo Do not Disturb

Como toda comunicación se inicia con un INVITE desde el VideoPhone, quien inicia la llamada, hacia el Softphone. Este paquete es de características idénticas a los anteriores.

La diferencia fundamental con los escenarios 1, 2 y 3 donde se establecían las llamadas es que al estar el softphone en modo DnD, responde con el mensaje 480 Temporarily Unavailable solicitando la finalización de la llamada a lo que responde con el mensaje ACK finalizando de esta forma la llamada.

MENSAJE 480 TEMPORARILY UNAVAILABLE.

Este mensaje es dado cuando el otro equipo está conectado correctamente pero no está en condiciones de responder la llamada. Por ejemplo: cuando no está logueado, cuando está logueado pero en un estado que no permite el ingreso de otra comunicación o se encuentra en modo Do not Disturb (DnD). En el teléfono que realizó la llamada generalmente aparece un mensaje diciendo que el teléfono destino no se encuentra disponible, intente más tarde.

Figura 41. Captura de mensaje 480

En la línea de Status Line se visualiza que el mensaje enviado es el 480, los campos que contiene son los mismos a los del resto de los mensajes SIP.

Cada vez que se termina una llamada o no se concreta por algún motivo especial se envía un mensaje de error en respuesta a los distintos fallos detectados.

A continuación se adjunta una tabla con los principales mensajes de error.

Como se ha indicado anteriormente corresponde con las respuestas de la clase:

4xx - Respuestas de fallo de método.

5xx - Respuestas de fallos de servidor.

6xx - Respuestas de fallos globales.

Estos errores se corresponden con los mensajes de error Q.931 o DSS1 y suponen el mapeo de los eventos SIP con los códigos de error de la RTC (Red telefónica conmutada)

Evento SIP	Valor decimal (DSS1)	Valor hexadecimal (DSS1)	Valor transmitido en el canal D	Detalle
400 Bad request	127	7f	ff	Interworking, unspecified
401 Unauthorized	57	39	b9	Bearer capability not authorized
402 Payment required	21	15	95	Call rejected
403 Forbidden	57	39	b9	Bearer capability not authorized
404 Not found	1	01	81	Unallocated (unassigned) number
405 Method not allowed	127	7f	ff	Interworking, unspecified
406 Not acceptable	127	7f	ff	Interworking, unspecified
407 Proxy authentication required	21	15	95	Call rejected
408 Request timeout	102	66	e6	Recover on Expires timeout
409 Conflict	41	29	a9	Temporary failure
410 Gone	1	01	81	Unallocated (unassigned) number
411 Length required	127	7f	ff	Interworking, unspecified
413 Request entity too long	127	7f	ff	Interworking, unspecified
414 Request URI (URL) too long	127	7f	ff	Interworking, unspecified
415 Unsupported media type	79	4f	cf	Service or option not available
420 Bad extension	127	7f	ff	Interworking, unspecified
480 Temporarily unavailable	18	12	92	No user response
481 Call leg does not exist	127	7f	ff	Interworking, unspecified
482 Loop detected	127	7f	ff	Interworking, unspecified
483 Too many hops	127	7f	ff	Interworking, unspecified
484 Address incomplete	28	1c	9c	Address incomplete (invalid number format)
485 Address ambiguous	1	01	81	Unallocated (unassigned) number
486 Busy here	17	11	91	User busy
487 Request cancelled	127	7f	ff	Interworking, unspecified
488 Not acceptable here	127	7f	ff	Interworking, unspecified
500 Internal server error	41	29	a9	Temporary failure
501 Not implemented	79	4f	cf	Service or option not implemented
502 Bad gateway	38	26	a6	Network out of order
503 Service unavailable	63	3f	bf	Service or option unavailable
504 Gateway timeout	102	66	e6	Recover on Expires timeout
505 Version not implemented	127	7f	ff	Interworking, unspecified
580 Precondition Failed	47	2f	af	Resource unavailable, unspecified
600 Busy everywhere	17	11	91	User busy
603 Decline	21	15	95	Call rejected
604 Does not exist anywhere	1	01	81	Unallocated (unassigned) number
606 Not acceptable	58	3a	ba	Bearer capability not presently available

Tabla I. Principales mensajes de error SIP

ESCENARIO 5. LLAMADA DESDE EL SOFTPHONE AL VIDEOPHONE OCUPADO

A continuación se muestra los paquetes intercambiados entre el Soft Phone y el Video Phone.

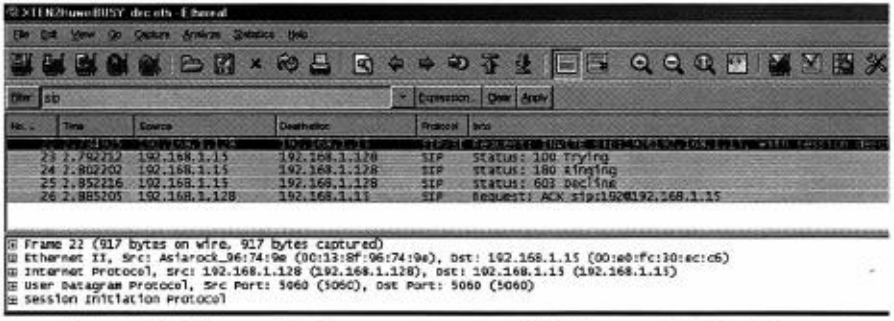


Figura 42. Intercambiados paquetes entre Soft Phone y el Video Phone.

En este caso, se realizó una llamada desde el Softphone hacia el Videophone, el cual se encontraba ocupado. El videophone responde al mensaje INVITE con un TRYING y RINGING, pero al detectar que no está en condiciones de recibir la llamada, envía el mensaje de error 603 DECLINE o llamada rechazada, a lo que el softphone responde con el ACK terminando la conexión.

MENSAJE 603 DECLINE

El mensaje 603 DECLINE indica que el teléfono al que se está llamando está correctamente conectado pero el destinatario explícitamente no quiere atender la llamada.

En este caso en el Video Phone se presionó la tecla Cancel mientras timbraba, terminando así la llamada. En el display del llamante aparece la leyenda "Intente más tarde".

A continuación se muestra la captura.

Figura 43. Mensaje 603 DECLINE

En el campo Status Line se ve que este paquete es código 603, el cual no posee contenido adicional describiendo el motivo, por lo que tiene características similares a la del resto de los mensajes SIP.

CONCLUSIONES ESCENARIOS 4 Y 5

La aparición de varios mensajes no usuales tales como DECLINE, Request Terminated, etc., amplia los escenarios básicos que podrían aparecen en una conexión SIP (analizar tabla de eventos SIP) (RFC 3261).

ESCENARIO 6. LLAMADA FALLIDA DESDE EL SOFTPHONE 1 AL SOFTPHONE 2.

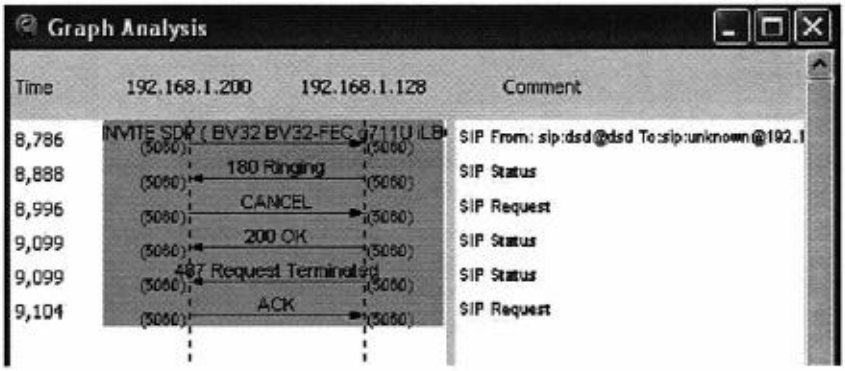


Figura 44. Esquema de paquetes intercambiados entre los softphones

En esta comunicación el Softphone IP: 192.168.1.200 genera una llamada hacia el Softphone IP: 192.168.1.128, el cual responde con un Ringing indicando que se encuentra conectado y está timbrando.

Antes de que el receptor atienda la llamada, el softphone que la generó la termina, por lo que este envía un mensaje CANCEL al softphone indicando que se cancela la sesión INVITE y pidiendo al otro softphone que deje de ringuear, a lo que responde con un mensaje de confirmación 200 OK. Luego el Softphone 128 envía el mensaje correspondiente para la finalización de la llamada, en este caso envía el mensaje 487 Request Terminated, el cual indica que recibió un CANCEL. A lo que la otra parte responde con el ACK finalizando la llamada.

Los paquetes INVITE y RINGING ya fueron explicadas por lo que no se muestran.

MENSAJE CANCEL:

El mensaje CANCEL es enviado cuando el que inició la llamada la quiere finalizar antes de que el destinatario la haya atendido o finalizado, o sea, cancelar el pedido de INVITE.

El envío de este mensaje es similar a pedirle a la otra parte que deje de llamar, a lo que la otra parte responde enviando el 200 OK y el mensaje 487, permitiendo así la finalización de la llamada. En la captura vemos que la línea Req Line describe un paquete CANCEL, el cual no presenta diferencias sobre el resto de los paquetes, debido a que no contiene campo de aclaraciones o contenidos.

Figura 45. Línea Req Line arrojando el mensaje CANCEL

487 REQUEST TERMINATED:

Este mensaje es enviado como respuesta a un CANCEL o BYE permitiendo la finalización de la llamada.

A continuación se muestra la captura.

Figura 46. Respuesta 487 Request Terminated

En la línea Status Line vemos que el mensaje enviado es el 487 Request Terminated. Los campos de este mensaje son similares a los anteriores debido a que tampoco poseen campo de aclaraciones o contenido.

CONCLUSIONES

Podemos observar en este análisis que hay un acuerdo con la descripción teórica detallada anteriormente con respecto al mensaje CANCEL, donde el UA con la dirección IP: 192.168.1.128 envía una petición de cancelación posterior a su petición INVITE, concluyendo de esta manera con la comunicación.

BIBLIOGRAFIA CONSULTADA

Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, E. Schooler, "SIP : Session Initiation Protocol ", RFC 3261, June 2002.

B. Roach, "Specific Event Notification", RFC 3265, June 2002.

J. Lennox, H. Schulzrinne, "Call Processing Language Framework and Requirements", RFC 2824, May 2000.

S. Donovan, "The SIP INFO Method", RFC 2976, October 2000

J. Rosenberg, H. Schulzrinne, "Reliability of Provisional Responses in the Campbell, J. Rosenberg, H. Schulzrinne, C. Huitema, D. Gurle, "SIP Extension for Instant Messaging ", RFC 3428 , December 2002.

M. Handley, V. Jacobson, "SDP: Session Description Protocol", RFC 2327, April 1998.

9. Algoritmos y Métodos aplicados en el Reconocimiento de Voz.

Brenda Valdez Reyna, Carlos del Rio Bocio,
Marco A. Panduro Mendoza

Capítulo Introductorio

Si nos paramos a pensar un poco sobre este tema, descubriremos que la cantidad de información que se necesita para realizar este trabajo es demasiada, empezando por los modelos básicos, el tema se bifurca en muchos caminos y es prácticamente imposible obtener una visión completa de todo. Debido a su complejidad matemática, física, informática y lingüística tendríamos que hacer un estudio previo de todos estos campos, lo que requiere mucho tiempo de estudio.

Me he limitado a describir los procesos de reconocimiento de voz de una manera superficial e intuitiva, haciendo hincapié en algunos los algoritmos y métodos más exitosos hasta ahora ya que, el reconocimiento de la voz constituye una parte importante del tratamiento del habla. Las técnicas de reconocimiento más desarrolladas son aquellas comúnmente usadas para el idioma inglés, las cuales incluyen las Redes Neuronales, el Análisis de Predicción Lineal (LPC) y el Alineamiento Temporal (DTW), estos algoritmos han tenido éxito habiendo sido sometidos a pruebas bajo diversos ambientes.

Me ahorrare fórmulas matemáticas complejas ya que considero este trabajo como una introducción teórica para luego poder profundizar, en alguna otra ocasión, en cualquiera de los temas expuestos aquí. (Daré por sabido, conceptos de muestreo y demás.) Tocaré temas de probabilidad, pero como he dicho antes, lo haré de forma teórica, intuitiva y superficial. Considero ésta la mejor propuesta, porque para saber de algo en concreto habrá que empezar desde el principio; no habrá ningún momento en que el lector se pierda, espero, por tanto, aclarar el nivel de dificultad de este trabajo.

¿Qué es reconocimiento de voz? Al hablar de reconocimiento de voz, podemos imaginarnos varios campos de aplicación. Desde la demótica hasta la inteligencia artificial.

¿Emplearemos reconocimiento de palabras aisladas o del habla continua? ¿Será o no será dependiente del locutor? ¿Tendrá una gramática restringida? Todo depende de la aplicación que queramos. Por ejemplo, si queremos un sistema que reconozca un número limitado de palabras para poder apagar o encender las luces de nuestra casa, está claro que grabando unos cuantos ejemplos que servirán de patrones a identificar con las entradas, bastará para poder satisfacer nuestras necesidades.

Imaginemos que en vez de 10 palabras queremos tratar un vocabulario completo y no sólo eso, queremos poder hablar con naturalidad y que el sistema identifique las palabras, las frases y el significado. Es decir, queremos que un robot nos entienda, para ello el nivel de complejidad se eleva a un nivel casi impensable.

Un sistema de reconocimiento de voz podrá operar identificando:

- Palabras aisladas
- Fonemas (mayor complejidad)

Éste último podrá utilizarse para reconocer palabras, frases, etc. Además de su entendimiento.

Si nos interesa un sistema simple de reconocimiento de palabras, actualmente se venden módulos que funcionan mediante comparación de patrones. Necesitaremos almacenar en una memoria dichos patrones y luego se compararán las entradas con éstos dando una salida de tipo binario (1 ó 0).

El método de funcionamiento se podrá comprender más adelante, ya que conociendo lo difícil se intuye lo fácil. Por ello me voy a ceñir en el reconocimiento de fonemas, que es actualmente el sistema más perseguido por los más ambiciosos investigadores.

Como veremos, no se analiza fonema por fonema, sino que se divide la señal (en función del tiempo) en pequeñas ventanitas de unos 20 mseg. y se van analizando las frecuencias además de sus variaciones.

La dificultad empieza a nacer cuando nos damos cuenta de que al pronunciar las palabras: "siete" y "nueve" hay cuatro letras señaladas en negrita que parecen ser la misma, pero lo mejor de todo es que la pronunciación, en al menos dos de ellas, es diferente, aunque sea la "e" depende mucho de dónde la coloquemos, qué es lo que la precede y en qué estado de ánimo la pronunciamos. Es decir,

necesitamos predecir de alguna manera qué tipo de "e" es. Aquí entra en juega la probabilidad, pero retornemos al principio explicando todo paso por paso.

2. Panorama General del Reconocimiento de Voz

2.1 Historia del Reconocimiento de Voz

La historia del reconocimiento de voz, se remonta en el tiempo, en 1870 Alexander Graham Bell quería construir un sistema/dispositivo que hiciera el habla visible a las personas con problemas auditivos. De lo que resultó el teléfono.

En 1880 Tihamir Nemes: Solicita permiso para una patente para desarrollar un sistema de trascripción automática que identificara secuencias de sonidos y los imprimiera (texto). Pero fué rechazado como "Proyecto no Realista".

30 años después AT&T Bell Laboratorios construyó la primera máquina capaz de reconocer voz (basada en Templates) de los 10 dígitos del inglés. Requería un extenso reajuste a la voz de una persona, pero una vez logrado tenía un 99% de certeza. Por lo tanto surge la esperanza de que el reconocimiento de voz sea simple y directo.

En 1950 se da un avance con múltiples paradigmas de trabajo y resultados, inclusive muchas de las técnicas utilizadas con éxito debieron esperar más de 10 años para pasar de la teoría a la práctica, inclusive en laboratorios. Algunos de los principales fueron en 1952, Bell Labs, con su reconocimiento aislado de dígitos, medición de resonancia espectral en vocales, con rangos de 50 al 100%, 1959 reconocimiento de vocales y algunas consonantes, con analizador de espectro y comparadores de patrones con resultados del 93%, ambos dispositivos de hardware y exclusivamente en laboratorio.

- En los 60's se comenzó a experimentar con normalización temporal según la detección de los puntos de comienzo y fin de las palabras, utilizando en general hardware específico e inicios de uso de computadoras. A mediados de los 60's, la mayoría de los investigadores reconoce que era un proceso mucho más intrincado y sutil de lo que habían anticipado. Por lo tanto empiezan a reducir los alcances y se enfocan a sistemas más específicos:

    - Dependientes del Locutor.
    - Flujo discreto de habla (con espacios / pausas entre palabras).
    - Vocabulario pequeño (menor o igual a 50 palabras).

Estos sistemas empiezan a incorporar técnicas de normalización del tiempo (minimizar diferencia en velocidad del habla). Además, ya no buscaban una exactitud perfecta en el reconocimiento, algunos años después IBM y CMV trabajan en reconocimiento de voz continuo, pero no se ven resultados hasta la década de 1970's, donde hubo avances significativos en reconocimientos de palabras aisladas, y comienzos de experimentación en reconocimiento independiente del locutor (speaker independent). Se advierte que las fuentes de información semántica, sintáctica y contextual, ayudan a mejorar la calidad de los sistemas.

A principios de 1970, se produce el 1er Producto de reconocimiento de voz, el VIP100 de Threshold Technology Inc. (utilizaba un vocabulario pequeño, dependiente del locutor, y reconocía palabras discretas). Gana el U.S. National Award en 1972.

Nace el interés de ARPA del U.S. Department of Defense, y gracias al lanzamiento de grandes proyectos de investigación y financiamiento por parte del gobierno se precipita la época de la inteligencia artificial. El proyecto financiado por ARPA busca el reconocimiento de habla continua de vocabulario grande. Impulsa que los investigadores se enfoquen al entendimiento del habla.

Los sistemas empiezan a incorporar módulos de:

- Análisis léxico (conocimiento léxico).
- Análisis sintáctico (Estructura de Palabras).
- Análisis semántico (Significado).
- Análisis pragmático (Intención).

Este proyectos (el mas grande es de los años 70's) termina en 1976 con el resultado de que CMU, SRI, MIT crearon sistemas para el proyecto ARPA SUR (Speech Understanding Research). El reconocimiento de una sentencia completa de gramática acotada requería de 50 computadoras (HARPY system del Carnegie Mellon University).

En los 80's se aplicaron los conceptos de dynamic time warping. Se produce un importante cambio de paradigma de comparación de plantillas hacia el modelado estadístico/probabilística como un gran avance de aproximación al reconocimiento de voz. A mitad de los 80's se hizo masiva una técnica que revolucionó el campo de reconocimiento se trata de los modelos ocultos de Markov o HMM que obtuvo excelentes resultados en el modelado de señales de voz y virtualmente indispensable hoy en día.

Se reintroduce el uso de redes neuronales (ANN) que habían vencido algunos obstáculos de tipo conceptual y de recursos necesarios para su implementación. También se comenzó a experimentar con reconocimiento continuo de vocabularios largos independientes del locutor.

En los 90's se comenzó a hacer énfasis en interfaces de lenguaje natural, y recuperación de la información en grandes documentos de voz, continuó la investigación de reconocimiento continuo en vocabularios grandes y a usarse masivamente a través de redes telefónicas, también en el estudio de sistemas en condiciones de ruido. Antes y durante la mitad de los 90's se dio la investigación de sistemas híbridos HMM-ANN, que también han dado excelentes resultados siendo la excelencia de los motores de reconocimiento de voz de hoy en día (2).

Es interesante que en los años 80's a 90's, surgen los sistemas de vocabulario amplio, que ahora son la norma (más de 1000 palabras) y bajan los precios de estos sistemas.

Las empresas más importantes actualmente: Philips, Lernout & Hauspie, Sensory Circuits, Dragon Systems, Speechworks, Vocalis, Dialogic, Novell, Microsoft, NEC, Siemens, Intel (apoyo / soporte técnico), entre otros.

2.2 Concepto de Voz

La voz es el sonido producido por el aparato fonador humano. Hay dos mecanismos básicos de producción de voz: 1) la vibración de las cuerdas vocales, que da lugar a sonidos "sonoros" (vocales, semivocales, nasales, etc.), y 2) las interrupciones (totales o parciales) en el flujo de aire que sale de los pulmones, que dan lugar a los sonidos "sordos" (fricativas, plausivas, etc.). Adicionalmente hay combinaciones de ambos mecanismos, como las oclusivas sonoras (en español b, d y g). Los sonidos así producidos luego se matizan por la configuración del resto del tracto vocal.

2.2.1 Sistema Fonador Humano

Onda de presión acústica originada voluntariamente a partir de los movimientos de la estructura anatómica del sistema fonador. Los distintos sonidos se producen al pasar el aire emitido por los pulmones, a través de todo el sistema de producción, en una determinada posición de cada parámetro articulatorio.

Este sistema puede modelarse como un filtro, cuya función de transferencia depende del sonido articulado. La entrada al filtro se puede modelar mediante una señal de excitación, que se corresponde con el paso del aire generado por los

pulmones a través de la tráquea y las cuerdas vocales, y también será dependiente del sonido generado.

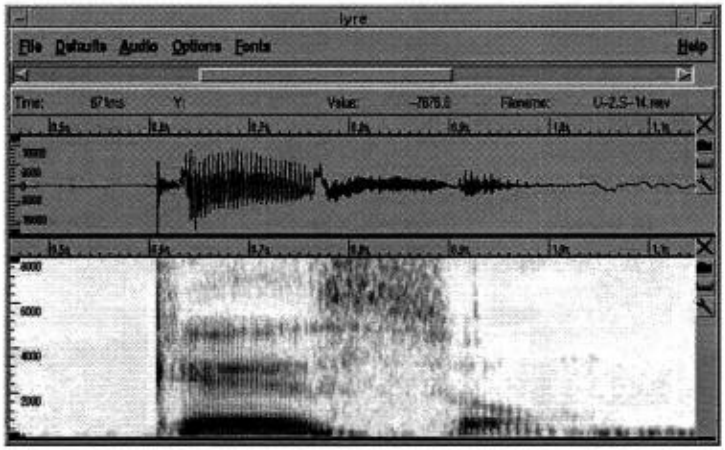
2.2.1.1 Aparato Fonador

Partiendo del conocimiento de la producción de sonidos por las cuerdas vocales, hay que tener en cuenta qué es lo que nos hace distinguir las letras y las consonantes. Si acudimos a referencias fonéticas, nos explicarán que hay ciertas consonantes oclusivas, otras son fricativas, etc. Y esto influye muchísimo en el traspaso del dominio del tiempo a la frecuencia.

Como ejemplo, citemos la "s", en un espectrograma veríamos ruido a altas frecuencias, sin embargo la "a" tiene ciertas componentes frecuencia les de alta energía. La posición de la lengua, la abertura de la boca, los labios, todo un conjunto fonético que consigue emitir infinidad de sonidos. En nuestro idioma se acotan dichos sonidos para poder construir un lenguaje ordenado. Otros idiomas recogen otros sonidos producidos por el aparato fonador que difieren bastante del castellano.

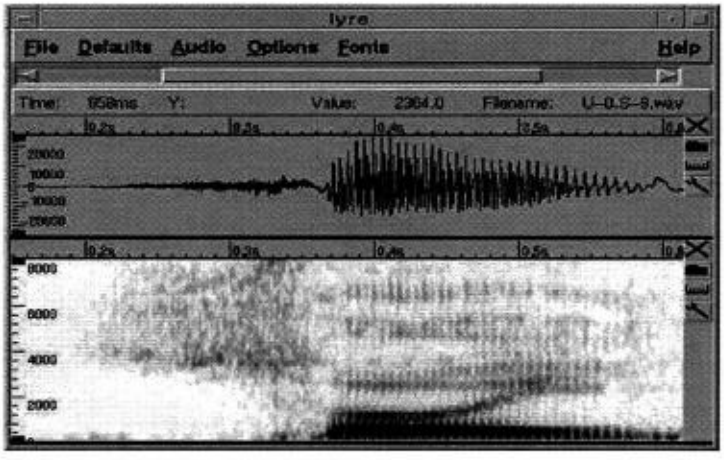
Si emitimos un sonido constante y sólo movemos la lengua, nos daremos cuenta de que en cierta manera producimos el mismo sonido pero cambiamos las distribuciones armónicas... Los Formantes.

2.2.1.2 Formantes



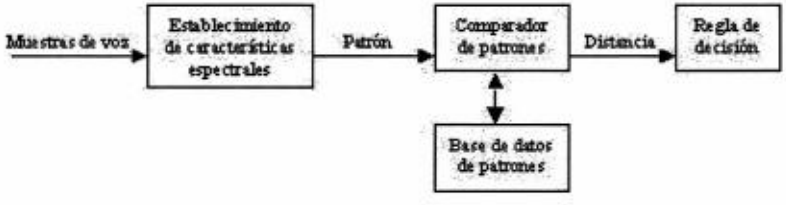
Son frecuencias que entran en resonancia en las cavidades nasales y orales, saliendo hacia el exterior como la información más importante del habla. En re-

conocimiento de voz solemos anular la frecuencia fundamental y nos quedamos con los dos primeros formantes. Este es el ejemplo de la palabra "queso" se visualiza perfectamente las altas frecuencias debidas a la "s" y los dos primeros formantes de la "e", así como el tono fundamental.



Esta es la palabra "soy". Podemos ver perfectamente que el primer formante se desplaza en frecuencia por el efecto del corrimiento de la lengua en el fonema /oi/ (4).

2.2.1.3 Reconocimiento del habla empleando técnicas de comparación de patrones:

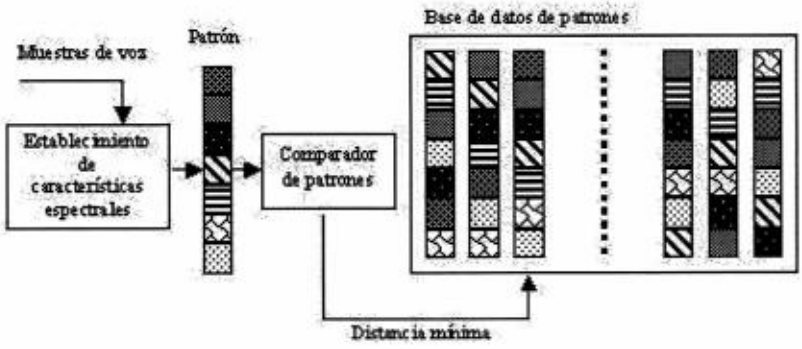


Su principal ventaja inmediata reside en que no es necesario descubrir características espectrales de la voz a nivel fonético, lo que evita desarrollar etapas complejas de detección de formantes, de rasgos distintivos de los sonidos, tono de voz, etc.

Esto está muy bien para un número finito de palabras, cuyo número no sea muy grande. Si queremos implementar esto para un completo entendimiento de

nuestro lenguaje, a ver quien se atreve a coger un diccionario y grabar palabra por palabra. Sería una auténtica locura, además de ser inútil porque si por ejemplo pronunciásemos la palabra "queso" tendríamos que hacerlo exactamente igual que en la grabación. Tendríamos que decir "queso" con la misma velocidad, con el mismo tono... etc.

Si nuestro "queso" se pronuncia muy rápido, habría que ajustar los tiempos de inicio y de final, pero si el sistema no está seguro de que sea esa palabra... adiós muy buenas. Necesitaríamos un sistema que aprenda por sí mismo éstas posibles deficiencias y se atreva a estipular qué palabra es.



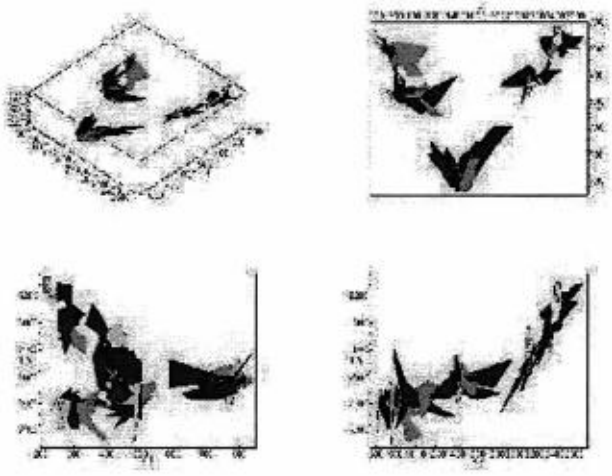
La base de datos que necesitaríamos sería tan grande, que harían falta sistemas toscos de almacenamiento. Cuando el sistema intente buscar la palabra, al menos basará su funcionamiento en el establecimiento de una distancia matemática entre vectores, de tal manera que se puede calcular lo cercano que se encuentra cada patrón. De todos modos, existe la necesidad de aplicar este sistema única y exclusivamente a ciertos casos donde el número de palabras necesarias sea pequeño.

También se puede constituir los grupos de patrones por unidades tales como sonidos básicos (fonemas y demás clasificaciones de sonidos cortos). Al grabar estos sonidos en la base de datos, se obtendrán sus características espectrales (suele hacerse con los parámetros LPC, de los cuales hablaremos después).

Por último mencionar que por mucho que se mejore éste sistema, siempre existirá el error al normalizar en tiempo y amplitud éstas señales de entrada para que coincidan con el patrón.

2.2.1.4 Estudio basado en la posición de los formantes

Para obtener una información detallada de los tres primeros formantes hay que recurrir a otra solución. Un simple espectrograma nos da información... pero no tanta. ¿Qué podríamos hacer para poder diferenciar unas personas de otras?, ¿qué haríamos para poder ver las relaciones entre formantes de una manera más precisa? La solución está en construir gráficas donde F1 (primer formante) se situé, por ejemplo, en el eje de abscisas, F2 en el eje de ordenadas, y así probando todas las combinaciones. También podemos establecer gráficos de 3 dimensiones donde intervienen los tres formantes.



Esta gráfica muestra lo dicho. Cada color representa una persona diferente, si nos fijamos en la de arriba a la derecha (F2 eje x, F1 eje y), podremos observar las vocales (i, e, a, o, u) de derecha a izquierda. Cuanto más a la derecha esté la información, mayor frecuencia tendrá F2. Cuanto más arriba, menor frecuencia tendrá F1. Las demás gráficas se usan para complementar casi siempre a la que hacemos referencia. De esta forma podremos distinguir entre quién habla y qué sonido produce dicha persona.

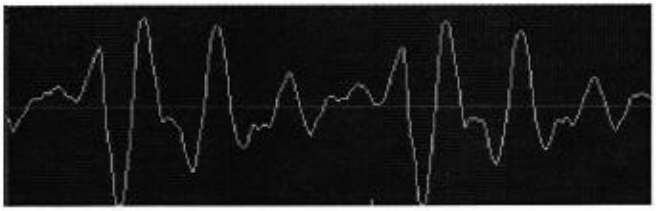
Pero una aplicación recóndita reside aquí. ¡Podemos diferenciar si una vocal es adyacente a una consonante bilabial (por ejemplo, la "b"), o si la misma vocal es adyacente a una consonante velar, o a una dental/interdental... Lo que quiero expresar con esto es la solución a la identificación de las consonantes adyacentes a una misma vocal. Bien antes hemos dicho que una misma vocal puede pronunciarse de diferentes maneras según sus consonantes adyacentes.

En estas gráficas podríamos apreciar cómo F1 y F2 bajan en frecuencia (desplazadas hacia la izquierda y hacia arriba) en el caso de tener adyacente una consonante bilabial. F1 baja y F2 sube en el caso de tener una consonante velar, este caso se ve claramente porque es la parte coloreada de azul en la gráfica de arriba a la derecha, donde podemos observar que está desplazada de las demás hacia arriba y hacia la derecha (F1 baja y F2 sube).

Un análisis mucho más profundo revelaría grandes detecciones sobre la evolución de los formantes (4).

2.2.2 Clasificación de Sonidos

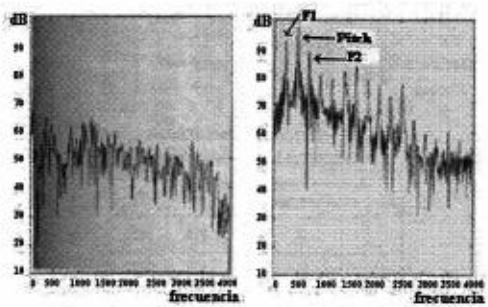
- Señales sonoras: En ellos las cuerdas vocales vibran y el aire pasa a través del tracto vocal sin impedimentos importantes.



- Señales sordas también conocidas como fricativas: En ellos las cuerdas vocales no vibran y existen restricciones importantes al paso del aire que proviene de los pulmones, por lo que son de amplitud menor y normalmente de naturaleza más ruidosa que los sonoros.

2.2.3 Análisis Frecuencial (I)

La señal de voz es limitada en banda, a unos 8 kHz. Sin embargo, la mayor parte de la información se encuentra en los primeros 4 kHz, que es aproximadamente el ancho de banda utilizado en las comunicaciones por vía telefónica.



Espectrograma de /i/ Espectrograma de /u/

2.2.4 ANÁLISIS FRECUENCIAL (II)

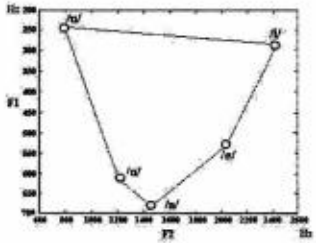
De las figuras de antes se pueden sacar varias conclusiones:

- La periodicidad del fonema /u/. (Hay que recordar que la presencia de armónicos en el espectro indica cierta periodicidad de la señal).

- El margen habitual del valor del pitch para locutores masculinos adultos del valor del pitch es de 50 a 250 Hz, mientras que para locutoras se encuentra entre 120 y 500 Hz.

- Existencia de resonancias o zonas enfatizadas (formantes), en el espectro de los sonidos sonoros, por ejemplo las vocales, esto permite identificar a la vocal a partir de sus formantes.

Nota: para formar el triángulo vocálico solo se requiere dos formantes.



Triángulo vocálico para el castellano

2.3 Procesamiento Digital

2.3.1 Fase 1ª – Digilitación de Voz

El procesamiento digital de señal mediante un DSP, ordenador, etc., requiere previamente la conversión de la señal acústica a eléctrica mediante un micrófono, y la conversión de la señal analógica resultante a señal digital. Por otra parte, para restaurar o generar señal audible a partir de un sistema digital, será necesaria la conversión digital a analógica, su amplificación, y su radiación mediante un altavoz.

Etapas de digitalización:

- Recogida de información mediante un transductor.
- Filtrado antialiasing.
- Muestreo (Teorema de Nyquist).

2.3.2 Fase 2ª – Codificación De Voz

Las técnicas de codificación de voz pretenden reducir el volumen de información necesario para almacenar o transmitir una señal de voz, de forma que la pérdida de calidad de la señal decodificada respecto a la señal sin comprimir sea lo menor posible. Por supuesto, deberá mantenerse la inteligibilidad del mensaje, y existirá un compromiso calidad versus tabla de compresión, complejidad computacional, etc.

Tipos de codificación:

A) Codificación de forma de onda: intentan reproducir fielmente la forma de la onda de la señal a codificar.

B) Codificación paramétrica (*): se basan en un modelo de producción del habla, e intentan reproducir en el proceso de decodificación una señal que al escucharla se parezca a la original, aunque existan distorsiones en la forma de onda generada.

(*)Nota: en el reconocimiento de voz, la codificación paramétrica es ampliamente utilizado CVoiceControl/kVoiceControl

2.4 RECONOCIMIENTO DE VOZ

2.4.1 INTRODUCCIÓN

El reconocimiento de la voz constituye una parte importante del tratamiento del habla. Las técnicas de reconocimiento más desarrolladas son aquellas comúnmente usadas para el idioma inglés, las cuales incluyen el Análisis de Predicción Lineal (LPC) y el Alineamiento Temporal (DTW).

2.4.1.1 Tipos de enfoque en el reconocimiento

- Reconocer palabras aisladas: las palabras se pronuncian entre pausas pequeñas de tal forma que el procesamiento se realiza teniendo como unidades lingüísticas las palabras de un vocabulario específico.
- Reconocer palabras conectadas: las palabras se pronuncian sin pausas (habla normal) de tal forma que el reconocimiento se lleva a cabo basándose en la coincidencia de palabras aisladas de referencia.
- Reconocer fonemas y difonos (reconocimiento continuo de voz): basada en la separación de la señal de voz en estas unidades lingüísticas, para su posterior análisis.

2.4.1.2 Ventajas / Desventajas

De los diferentes tipos de reconocedores:

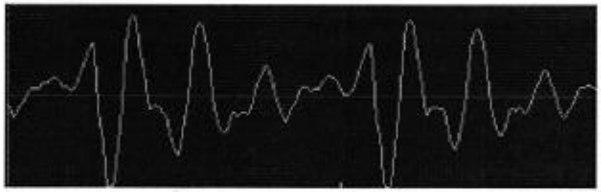
- La complejidad de los reconocedores mediante fonemas es bastante mayor en comparación con los de palabras aisladas.
- Los reconocedores de palabras aisladas no permiten vocabularios medianamente extensos (<50 palabras), debido principalmente al alto coste de memoria, que esta técnica requiere. Mientras que los reconocedores de fonemas permiten una mayor extensión del vocabulario a reconocer.
- Palabras del vocabulario castellano: 300000 palabras
- Fonemas existentes en el castellano: 37 fonemas
- Una limitación del reconocedor de palabras aisladas es tener que "dictar", de forma aislada, cada palabra del texto a reconocer.

2.4.2 Modelado del Tracto Vocal

2.4.2.1 Modelado del Tracto Vocal (I)

Existe un modelo que describe el proceso del habla clasificando las señales en dos tipos:

- Señales sonoras: se caracterizan por tener alta energía y contenido frecuencial en el rango de los 300 Hz a 4000 Hz, las cuales se generan por intermedio de las cuerdas vocales y además presentan cierta periodicidad.



- Señales sordas también conocidas como fricativas: se caracterizan por tener baja energía y componente frecuencial uniforme presentando aleatoriedad en forma de ruido blanco (3).

2.4.2.2 Modelado del Tracto Vocal (Ii)

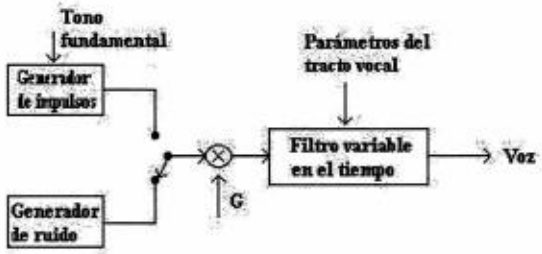
El tracto vocal modelado se manifiesta como un filtro variable en el tiempo cuyos parámetros varían en el tiempo en función de la acción consciente que se realiza al pronunciar una palabra.

El filtro variable en el tiempo tiene dos posibles señales de entrada que dependerán del tipo de señal, sonora o sorda (no sonora). Para señales sonoras la excitación será un tren de impulsos de frecuencia controlada, mientras que para las señales no sonoras la excitación será ruido aleatorio.

La combinación de estas señales modelan el funcionamiento de la glotis. El espectro de frecuencias de la señal vocal puede obtenerse a partir del producto del espectro de la excitación por la repuesta en frecuencia del filtro.

El tracto vocal manifiesta un número muy grande de resonancias, sin embargo se consideran solo las tres o cuatro primeras que toman el nombre de 'formantes' y cubren un rango de frecuencias entre 100 y 3500 Hz. Esto es debido a que las resonancias de alta frecuencia son atenuadas por la característica frecuencial del tracto que tiende a actuar como un filtro pasa bajo con una caída de aproximadamente -12 dB por octava.

Este modelo es una simplificación del proceso del habla. Los sonidos fricativos no se filtran por el tracto con la misma extensión en que lo hacen las señales sonoras por lo que el modelo no es muy preciso para este tipo de señales. Además el modelo supone que las dos señales pueden separarse sin considerar ninguna interacción entre ellas, lo cual no es cierto ya que la vibración de cuerdas vocales es afectada por las ondas de presión dentro del tracto. Sin embargo estas consideraciones pueden ser ignoradas resultando el modelo lo suficientemente adecuado (3).



Modelo del tracto vocal

2.4.3 REPROCESAMIENTO DE LA SEÑAL VOCAL PREÉNFASIS

Se hace necesario para el análisis realizar un pre-procesamiento de la señal vocal. Esto se realiza a través de técnicas que permitan extraer la información acústica directamente a partir de la señal vocal emitida. Esto se realiza mediante la técnica de preénfasis y la aplicación de una ventana de Hamming.

La etapa de preénfasis se realiza con el propósito de suavizar el espectro y reducir las inestabilidades de cálculo asociadas con las operaciones aritméticas de precisión finita. Además se usa para compensar la caída de -6 dB que experimenta la señal al pasar a través del tracto vocal. Se usa un filtro digital de primer orden cuya función de transferencia es:

$$H(z)=\frac{Y(z)}{X(z)}=1-az^{-1},\ a=0.95$$

Y la ecuación en diferencias correspondiente a la función de transferencia es:

$$y[n]=x[n]-ax[n-1]$$

Y la representación de esta ecuación en un diagrama de bloque es:

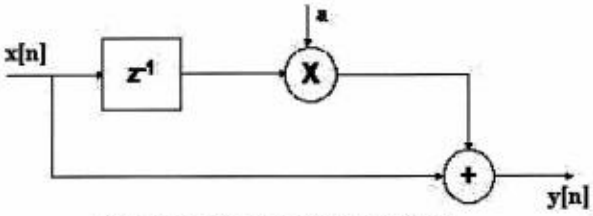
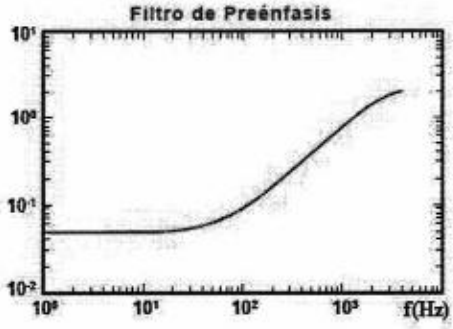


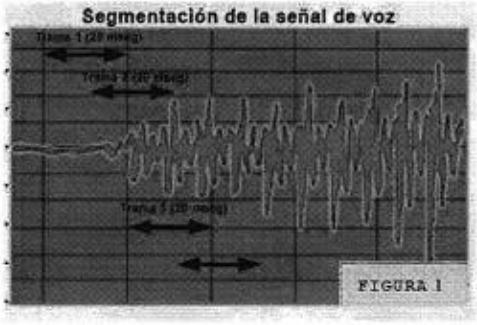
Diagrama de bloques del sistema



Por último solo queda mostrar la gráfica en el dominio frecuencial del filtro de preénfasis.

2.4.4 PREPROCESAMIENTO DE LA SEÑAL VOCAL SEGMENTACIÓN Y ENVENTANADO

Una vez que la señal se ha digitalizado y filtrado con el filtro preénfasis, la señal de voz se segmenta en tramas de 20 ó 30 mseg. Con un desplazamiento cuyo valor típico es 10 mseg. Por ejemplo, imaginemos que tenemos la señal de la figura 1. Las rayas verticales de la gráfica se corresponden con 20 mseg de tiempo.



Para analizar este trozo de voz se procederá de la siguiente manera, la primera trama de voz es la indicada en la FIGURA 1. La segunda trama no comenzará en el siguiente segmento indicado por el trazo vertical, sino que estará desplazado 10 mseg. Respecto del comienzo de la trama anterior, y así sucesivamente. En la FIGURA 1, la punta de la flecha y su longitud indica comienzo y duración de la trama. Es por esto que se habla de duración de las tramas y desplazamiento.

Cada trama de 20 mseg. se procesa de la siguiente forma: Después de la segmentación se aplica una ventana Hamming, la cual elimina los problemas causados por los cambios rápidos de la señal en los extremos de cada trama de voz. Es por eso por lo que se utiliza la segmentación con un desplazamiento para conseguir transiciones suaves entre tramas. En la práctica es deseable normalizar la ventana para que la potencia de la señal sea aproximadamente igual a la potencia de la señal antes del enventanado.

La teoría de la ventana fue un tema activo de investigación en el procesado digital de señal, hay muchos tipos de ventana: rectangular, Hamming, Hanning, Blackman, Bartlett y Kaiser. Hoy en día se utiliza exclusivamente la ventana Hamming para el reconocimiento de voz, que es un caso específico de la Hanning.

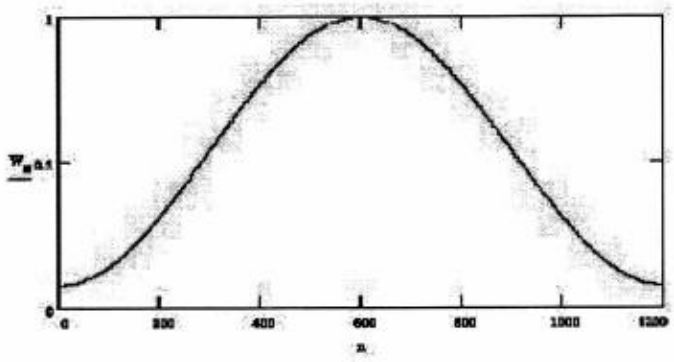
Una ventana Hamming se define como:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N_s - 1}\right)$$

Ns es el número de muestras de la trama.

$0 \leq n \leq Ns$

Cuya representación gráfica es:



EVENTANADO. Al periodizar el bloque de análisis aparecen discontinuidades. Estas discontinuidades producen componentes espectrales que no existen en la señal original.

Para eliminar las discontinuidades, se debe multiplicar las señal por otra señal (ventana).Actualmente existen varios tipos de ventanas: triangular, rectangular, Hanning, Hamming, etc.

2.4.5 ANÁLISIS DE PREDICCIÓN LINEAL (LPC).

ECUACIÓN DEL FILTRO FIR (I)

Una de las técnicas mas usadas en el procesamiento de señales de voz viene a ser el análisis de predicción lineal. Esta técnica ha probado ser muy eficiente debido a la posibilidad de parametrizar la señal con un número pequeño de patrones con los cuales es posible reconstruirla adecuadamente.

Los parámetros obtenidos mediante este método se caracterizan por variar en forma lenta durante las ventanas de tiempo de análisis. Mediante esta técnica podemos representar a la señal vocal mediante parámetros que varían en el tiempo los cuales están relacionados con la función de transferencia del tracto vocal y las características de la fuente sonora. Otra ventaja es que no requiere demasiado tiempo de procesamiento, lo cual es importante a la hora de la implementación en un computador.

El modelo matemático expuesto establece que el tracto vocal puede modelarse mediante un filtro digital siendo los parámetros los que determinan la fun-

ción de transferencia. El problema consiste en, dado un segmento de palabra, extraerle sus parámetros que en este caso vienen a ser los coeficientes del filtro. El análisis de predicción lineal permite aproximar una señal a partir de señales pasadas. En este caso se trata de predecir señales de voz mediante un filtro FIR (filtro de respuesta impulsiva finita), cuya función de transferencia se deduce a partir de:

$$y(n) = -\sum_{k=1}^{p} a_k \cdot G \cdot y(n-k) \cdot x(n)$$

Como se puede observar la señal de voz se representa por medio de señales anteriores y x(n) viene a ser la entrada del filtro, el cual será un tren de impulsos periódicos o una fuente de ruido aleatorio. El tren de impulsos producirá señales sonoras mientras la fuente de ruido aleatorio producirá señales no sonoras a la salida del filtro. De esta manera el filtro viene a representar un modelo del tracto vocal.

2.4.5.1 Análisis De Predicción Lineal (Lpc).

FUNCIÓN DE TRANSFERENCIA (II)

La función de transferencia del filtro se obtiene sacando la transformada z a la relación anterior con lo que se obtiene:

$$H(z) = \frac{G}{1 + \sum_{k=1}^{p} a_k \cdot z^{-k}}$$

Donde G viene a ser la ganancia del filtro y dependerá de la naturaleza de la señal. Dada la señal y(n), el problema consiste en determinar los coeficientes de predicción a_k y la ganancia G. Serán los coeficientes de predicción los que se usarán como parámetros de reconocimiento de palabras. Su determinación se realiza minimizando el error que se comete cuando se intenta realizar la aproximación de la señal (3).

2.4.5.2 Análisis de Predicción Lineal (Lpc).

AUTOCORRELACIÓN (III)

Desarrollando el error mediante mínimos cuadráticos, obtenemos:

$$R(i) = \sum_{n=0}^{N-1-j} y(n) \cdot y(n+i) \rightarrow i \geq 0$$

A continuación se procede a realizar un análisis de auto correlación. La función de auto correlación proporciona una medida de la correlación de la señal con una copia desfasada en el tiempo de si misma, donde p es el orden de análisis. De aquí se extraen los p coeficientes de auto correlación, valores típicos de p pueden ser entre 10 y 15. Podemos identificar los coeficientes de auto correlación en las ecuaciones que minimizan los errores en la estimación de la señal predicha. Para resolver este conjunto de ecuaciones se recurre al algoritmo de Levinson-Durbin el cual permite resolver el sistema de ecuaciones de una forma eficiente. Debido a la complejidad no se va a poner el algoritmo, solo decir que es útil para calcular los coeficientes de predicción del filtro.

Teniendo los coeficientes del filtro a_k se dispone, para la ventana de análisis, la función de transferencia del modelo del tracto vocal en ese instante, es decir se dispone con la forma con la que la cavidad vocal se comporta y que junto con la señal de excitación se obtiene el sonido emitido en ese momento (3).

2.4.6 Alineamiento Temporal (Dtw)

La siguiente etapa del análisis viene a ser la que se encarga de realizar la comparación de patrones acústicos. El éxito de este tipo de sistemas dependerá de la aplicación de una técnica conocida como Alineamiento Temporal (Dynamic Time Warping) la cual tiene en cuenta la variación en la escala del tiempo de dos palabras a comparar. El problema que se presenta cuando se pronuncia una palabra es que esta no siempre se realiza a la misma velocidad, lo que produce importantes distorsiones temporales. Estas distorsiones afectan no sola a la palabra considerada sino también a sus componentes acústicos. Las variaciones temporales no son generalmente proporcionales a la velocidad de locución y podrán variar de locutor a locutor.

Es por esto que se hace necesario un procedimiento que permita comparar dos palabras, sin considerar las distorsiones temporales. Los métodos que se usan

para realizar lo expuesto se basan en algoritmos de programación dinámica. Para dos palabras a comparar dichos algoritmos proporcionan una medida de disimilitud que puede ser aprovechada en el reconocimiento de palabras aisladas

2.4.7 CUANTIFICACIÓN VECTORIAL

Una parte importante en cualquier tipo de procesamiento de voz viene a ser la optimización de los algoritmos en cuanto a velocidad y almacenamiento. La técnica que a continuación se expone permitirá un ahorro en memoria lo que a su vez permitirá que los algoritmos se ejecuten a mayor velocidad ya que no tendrán que hacer uso de dispositivos externos de memoria.

Las técnicas de parametrización de la señal vocal se realizan tomando una secuencia de ventanas de tiempo, cada una de las cuales es representada por un número p de parámetros. Podemos apreciar que cada ventana de tiempo se puede representar por un vector de p dimensiones. Cuando se almacenan los parámetros lo que generalmente se realiza es cuantificar cada parámetro separadamente usando un número determinado de bits. Esto se conoce como cuantificación escalar y no es la forma más económica de almacenamiento ya que implica la ocurrencia uniforme de las ventanas de la señal vocal en el espacio vectorial. En aplicaciones de codificación y reconocimiento resulta más conveniente y económico el empleo de una técnica conocida como cuantificación vectorial.

La idea principal de la cuantificación vectorial es particionar el espacio vectorial en sectores, cada uno de los cuales será representado por un solo vector que puede ser el centroide. El conjunto de centroides viene a ser el libro índice (codebook) que conforman los niveles de cuantificación y a cada uno se le asigna una dirección o etiqueta. Como la parametrización se realiza por ventanas de tiempo pequeños representados por vectores de 14 dimensiones, a cada ventana se le asignará un vector. Para efectuar la cuantificación de un vector de entrada lo que se realiza es asignarle la dirección del vector del libro índice más cercano evaluado mediante una medida de disimilitud que puede ser la distancia cepstral euclídea o cualquier otra como la distancia de Itakura.

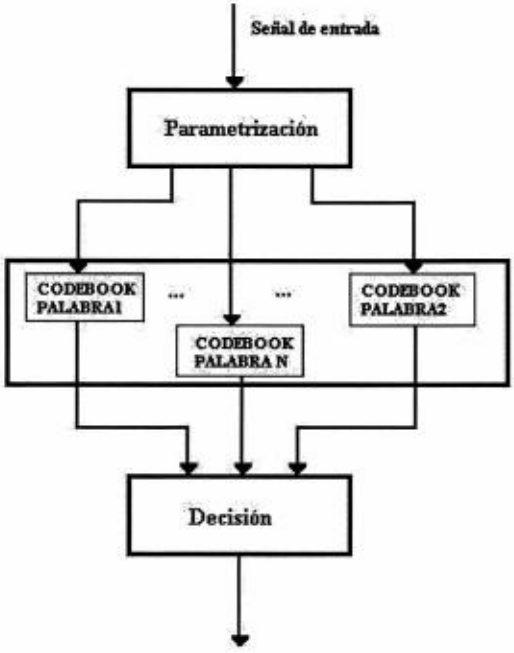
Un aspecto muy importante de cualquier sistema de cuantificación vectorial es la obtención del libro índice (codebook), el espacio vectorial debe ser dividido en sectores los cuales se hallan partiendo de vectores de entrenamiento. Dichos vectores deben representar fielmente el espacio de interés. El libro índice se obtiene empleando un algoritmo conocido como LBG, cuyo nombre se deriva de los creadores Yoseph Linde, Andrés Buzo y Robert Gray.

El algoritmo LBG debe partir de un libro índice inicial ° con el cual se compara cada vector del espacio a cuantizar con cada componente del libro índice.

Se compone una partición de distorsión mínima P (¬) clasificando cada vector mediante la distancia mínima con los vectores del libro índice. La suma de las distancias se compara con el umbral de distorsión, si resulta mayor se vuelve a calcular una nueva partición hasta que la de distancia total sea inferior al umbral.

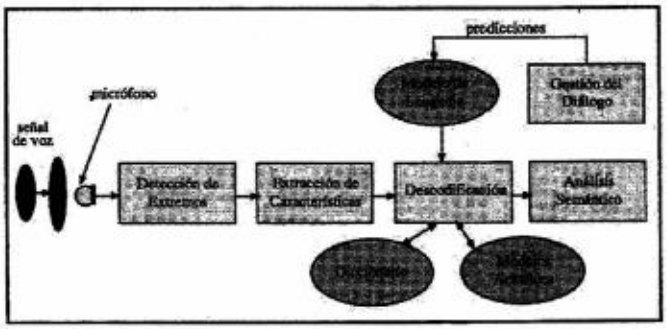
2.4.7.1 Esquema de un Reconocedor de Cuantificación Vectorial

Actualmente está en desuso. Sin embargo, es un método que requiere pocos cálculos. Si las unidades elementales son palabras, se calcula un modelo para cada palabra, consistente en un codebook. El proceso de reconocimiento consiste en codificar la palabra desconocida, y escoger el codebook que proporcione la menor distancia de codificación. El codebook llevará asociada cuál es la palabra que está modelando. La figura muestra de forma esquematiza un sistema reconocedor:



Reconocimiento de palabras mediante cuantificación vectorial

2.4.8 DIAGRAMA DE BLOQUES DE UN RECONOCEDOR DE VOZ



En la figura que se muestra a continuación se presenta el esquema típico de un reconocedor de voz, al que se le han añadido los bloques de análisis semántico y de gestión de dialogo.Estos dos últimos bloques mejoran la tasa de reconocimiento gracias a la información que tienen sobre el estado del dialogo, por lo que se puede considerar que forman parte del reconocimiento. Vamos a ilustrar esto con un ejemplo: supongamos que el usuario pronuncia una frase del tipo: "querría hacer una llamada". El analizador semántico extraerá el significado de la frase, que en este caso se reduce a la acción que desea realizar el usuario ("llamar") y el gestor del dialogo pedirá al mismo que diga el número de teléfono o el nombre de la persona o empresa a la que desea llamar.

En estas circunstancias, lo lógico es que el usuario conteste con un número de teléfono o un nombre, por lo que si el gestor de dialogo proporciona esta información al reconocedor, se reduce la posibilidad de que este falle.

En el proceso de reconocimiento se emplean cuatro tipos de información:

- Los modelos acústicos. que permiten que el reconocedor identifique los sonidos pues proporcionan información sobre las propiedades y características de los mismos.
- El diccionario. que indica que conjunto de sonidos forma cada palabra del vocabulario.
- El modelo del lenguaje. que tiene información de cómo se deben de combinar las palabras para formar frases.
- Los sistemas de dialogo o conversiónales. el reconocedor suele disponer de predicciones sobre el contenido de la siguiente frase que pronunciara el locutor.

Así pues el funcionamiento del reconocedor completo es: La señal de voz entra por el micrófono y se convierte en una señal eléctrica analógica que es posteriormente digitalizada (convertida en secuencia de unos y ceros). Esta señal pasa al detector de extremos, que es el encargado de detectar la presencia de voz y de pasar dicha voz al siguiente bloque del reconocedor. El extractor de características calcula una serie de parámetros de la señal de voz que tienen información relevante para el proceso de reconocimiento. Estos parámetros se pasan al decodificador, el cual se apoya en los modelos acústicos, los modelos del lenguaje, las predicciones y el diccionario para generar la frase reconocida. Posteriormente, el analizador semántico extraerá el significado de la frase, que será utilizado por el gestor del diálogo para, en función del estado de la conversión, tomar la decisión mas adecuada y hacer una predicción sobre la siguiente interacción con el usuario.

Como el problema general del reconocimiento de voz no esta totalmente resuelto, existen muchos tipos de reconocedores especializados en resolver problemas concretos. Por este motivo, no se pueden comparar dos reconocedores si no están especializados en la misma tarea, y aun en este caso, habrá que asegurar que las condiciones de la prueba son idénticas para ambos, antes de pronunciarse sobre la calidad de cada uno de ellos: comparar dos reconocedores sin tener en cuenta su especialización es como comparar un coche de carreras con uno familiar. No hay uno mejor que otro simplemente son distintos. Como este es un punto muy importante que da lugar a mucha confusión, a continuación se presenta una clasificación de los sistemas de reconocimiento atendiendo a varios criterios:

- Según el número de locutores que pueden reconocer:
  - Dependientes del locutor: solo reconocen a la persona para la que han sido entrenados.
  - Multilocutor: reconocen a un conjunto pequeño de personas.
  - Independientes del locutor: reconocen a cualquier persona.

- Según el tamaño del vocabulario que reconocen:
  - Reconocedores de vocabulario pequeño: hasta 40 palabras.
  - Reconocedores de vocabulario mediano: hasta 400 palabras.
  - Reconocedores de vocabulario grandes: hasta 4000 palabras.
  - Reconocedores de vocabulario muy grandes: hasta 40000 palabras.
  - Reconocedores de vocabulario limitado: más de 40000 palabras.

- Según el canal:
  - Reconocedores a través de micrófono.

- Reconocedores para la red telefónica (fija, móvil analógica o móvil digital).

- Según el tiempo de respuesta:

  - Reconocedores de tiempo real: son reconocedores que dan la respuesta lo suficientemente deprisa como para que el usuario puede interaccionar con ellos.
  - Resto: reconocedores en los que el tiempo de respuesta no es un factor importante (por ejemplo: sistemas de reconocimiento empleados para la transcripción de informes).

2.4.9 PROBLEMAS DEL ANÁLISIS Y RECONOCIMIENTO DE VOZ.

El análisis de la señal de voz y su posterior reconocimiento deben superar algunos problemas que en principio parecen triviales ya que son superados de forma sencilla por los seres humanos, algunos de ellos son, la correcta elección y extracción de las características de la señal de voz, tratar con corrección las variaciones inherentes a género, velocidad de emisión, pronunciación y acentos, tamaños de los vocabularios a reconocer, ruido y distorsión de los entornos donde se utilizan, inclusive hasta el estado de ánimo del locutor.

Pese a las dificultades, se ha logrado gracias a múltiples corrientes independientes de investigación y desarrollo, sistemas de uso real, en los cuales la exactitud es superior al 90%, siempre considerando tareas acotadas de una u otra manera.

Por ejemplo reconocimiento de dígitos, para un solo locutor, en canales sin ruido, se logran niveles de más de 99% de exactitud. Sistemas comerciales para vocabularios grandes, obtienen de 90% a 95%, cayendo a 87% para diferentes locutores y diferentes canales.

2.4.10 SITUACIÓN ACTUAL DE RECONOCIMIENTO DE VOZ DE HABLA CONTINÚA

Principalmente se enfocan al reconocimiento de palabras completas y mencionaré algunas cuestiones sobre producción de la voz, referente a los tramos sonorizados y los tramos no sonorizados.

2.4.10.1 Reconocimiento de Palabras Completas

La mayoría de los sistemas automáticos de reconocimiento de voz identifican fonemas simples o sílabas y unen estos en palabras. Por el contrario el reconocimiento de "palabras completas" está basado en el análisis de una palabra completa, o incluso en el de una secuencia de palabras.

En esta aproximación, el espectrograma de voz de una palabra, por ejemplo su energía, se considera como una fotografía y se analiza mediante métodos conocidos de procesamiento de imagen de manera específica, los sistemas bidimensionales de Fourier calculan la transformación de la energía de la voz (o una transformación equivalente).

Tan sólo se retienen los componentes de baja frecuencia de la imagen transformada, haciendo así el análisis independiente de fluctuaciones limitadas en la velocidad y forma. Un total de 49 coeficientes de Fourier tienen unos resultados excelentes en el reconocimiento de voz en sistemas de alta definición y telefónicos.

2.4.10.2 Algunas Cuestiones Sobre Producción de la Voz

Los sonidos vocales son producidos por la acción del aire que impulsan los pulmones sobre el tracto vocal, en la laringe existen dos membranas (cuerdas vocales) que permiten variar el área de la tráquea por la cual circula (glotis). Durante el habla, dichas membranas permanecen en continuo movimiento (abrir y cerrar) lo que origina una mezcla de características en la generación de la voz, que se dan en llamar "voiced" y "unvoiced", o sonorizado y no sonorizado.

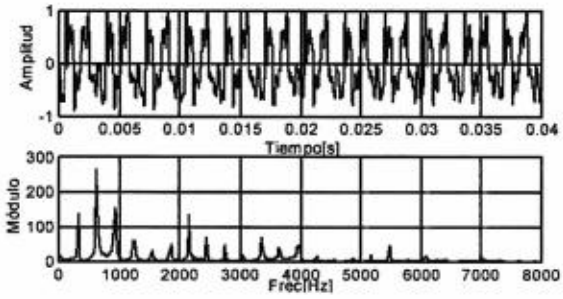
En los tramos sonorizados, las cuerdas vocales están normalmente cerradas y vibran al ser recorridas por la corriente proveniente de los pulmones, su frecuencia (pitch) se determina por el largo y tensión de las membranas y está en el orden de 50 a 400 Hz. El efecto de este continuo abrir y cerrar de la glotis aparece como un tren de pulsos casi continuo.

En los no sonorizados, las cuerdas vocales permanecen abiertas y el aire pasa al resto del tracto vocal en forma de turbulencia. Queda así establecido que el resultado del habla en la región de la glotis será, una sucesión de pulsos o ruido blanco, según sea sonorizado o no, pero lo que termina de moldear el aspecto de la voz es el resto del tracto vocal que da "forma" al sonido actuando como un filtro que impone su propia respuesta en frecuencia.

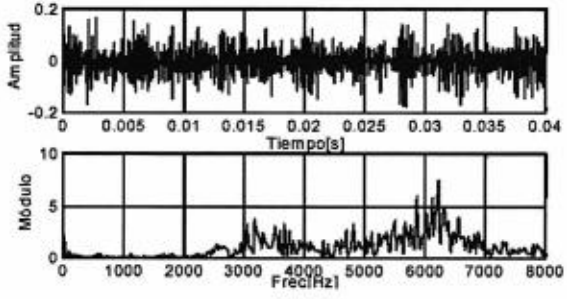
Para los sonorizados, como las vocales, el tracto vocal actúa como cavidad resonante, que produce picos en el espectro resultante conocidos como "formants" o formantes (Normalmente cerca de los 500 Hz), que en sí mismos contienen la

mayoría de la información de la señal y están formados de impulsos correspondientes a la naturaleza vibratoria de las cuerdas vocales. La forma y ubicación de los formantes depende de forma general de su tamaño y características particulares, y de forma particular del tracto vocal, posición de la lengua, labios, mandíbulas, etc. es decir todo lo que forma parte de la articulación de los sonidos.

La figura muestra un tramo de señal de voz sonorizada y su espectro, en el que se observan 3 formants" (a 750Hz y armónicos impares).



A continuación se observa un segmento no sonorizado, con su correspondiente espectro, sobresale su característica similar al ruido y su espectro tipo pasa altos.



3.1 TIPOS DE RECONOCIMIENTO DE VOZ.

Muchas veces suelen confundirse las condiciones bajo las cuales se habla de un sistema basado en reconocimiento de voz. Debido a la mala utilización de los calificativos que se utilizan para describir estos sistemas y, algunas veces, a la mala traducción hecha a partir de otros idiomas, se ha generado una ambigüedad o

concepción errónea de lo que en realidad es el término "reconocimiento de voz". Para esclarecer esta situación es necesario distinguir los tipos de reconocimiento de fonemas, lexemas o sonidos en general que son objeto de investigación y, algunos, desarrollo en la actualidad [TLDP2] y [TLDP3].

3.1.1 Lenguaje Natural

Este tipo de reconocimiento busca identificar el significado de las palabras u ordenes emitidas mediante el habla no ensayada, continua y sin atenerse a una expresión específica predeterminada (aunque sí a un idioma en específico), incluso tomando en cuenta factores como pequeños tartamudeos y rectificaciones hechas por el usuario al hablar. Se trata, entonces, de que la persona que emplea el reconocedor pueda expresarse como quiera, con frases cortas o largas, empleando las palabras que desee. En resumen, se pretende que el hablante se exprese como lo haría con otra persona, en vez de con un sistema automático. Es el reconocedor más avanzado y por tanto el más potente, novedoso y complejo de realizar ya que involucran generalmente (si no siempre) la implementación de un sistema de inteligencia artificial.

Cabe mencionar que de esta forma se elimina en gran parte el rechazo de los usuarios hacia este tipo de sistemas, al permitir que el cliente se exprese libremente, en lugar de estar sujeto a un diálogo preestablecido, fuertemente controlado y limitado. La clase de aplicaciones que permite desarrollar este tipo de reconocedor pueden ser mucho más completas que las demás y con mayor funcionalidad, pues no se ven limitadas al uso de menús prefijados, sino que interpretan el mensaje del usuario, permitiendo un servicio mucho más dinámico e interactivo. Los sistemas de reconocimiento de lenguaje natural están en la actualidad en una etapa de desarrollo prematura, aunque logrando importantes avances a futuro.

3.1.2 Habla Continua

A diferencia del tipo de reconocimiento mencionado en el apartado previo, este tipo de reconocimiento sí utiliza "palabras reservadas" (más concretamente un número de fonemas que, enlazados, forman una palabra reconocible) para interactuar con la computadora, lo cuál implica el establecimiento de un conjunto de elocuciones predefinidas con anterioridad basadas en la combinación de estas palabras y las cuales serán, única y exclusivamente, las que el sistema reconocerá como parte de los comandos o instrucciones válidas al momento de interactuar con el usuario.

La característica particular de este tipo de sistemas y la cual lo diferencia del tipo que se presentará en el apartado posterior es que no necesita que exista una pausa entre las elocuciones pronunciadas por el usuario (generalmente palabras); en otros términos, esto quiere decir que no es necesario que el hablante indique al sistema la terminación de una orden o comando mediante la utilización de una mínima pausa intencional, sino que el sistema estará en la capacidad de poder identificar cada una de las órdenes pronunciadas dentro de la muestra gracias a la implementación de algún método especial que determine los límites de ellas, lo cuál permite al usuario hablar casi naturalmente si el conjunto de palabras reconocibles y los comandos formados por éstas así lo permiten.

3.1.3 Voz

Se debe resaltar el hecho de que todos los tipos de reconocimiento, en su forma más general, son reconocedores de voz y es por eso que este término es utilizado, hasta cierto punto de manera errónea, para hacer referencia a cada uno de ellos; sin embargo es necesario precisar, en los casos en los que no se esté discutiendo el tipo de reconocimiento tratado en el presente sub-apartado, específicamente de cual de los tipos de reconocimiento se está hablando, para así evitar incurrir en una ambigüedad de términos que podrían prestarse a confusiones por parte de las personas que no están lo suficientemente familiarizadas con ellos o que simplemente los desconocen.

a) **Elocuciones aisladas**. Los reconocedores de voz basados en elocuciones aisladas usualmente requieren que cada elocución carezca de una señal de audio, diferente de la elocución misma, tanto al inicio como al final de la muestra de audio utilizada para el reconocimiento; más específicamente, es necesario que exista un silencio antes y después del comando asociado a la acción que el usuario desea que el programa ejecute.
   Esto no significa que únicamente acepte comandos representados por una única palabra, sino que se requiere que las órdenes o comandos sean pronunciados uno a la vez, a diferencia de los tipos de reconocedores que automáticamente identifican las diferentes elocuciones en una misma muestra.
   Muy a menudo, este tipo de reconocedores tienen estados marcados como "Escucha/No Escucha", los cuales permiten indicar al usuario del sistema los momentos en los que está facultado a emitir una elocución que será procesada; además estos estados le brindan al sistema la oportunidad de procesar la muestra en los tiempos de espera o pausas que

tiene que realizar obligatoriamente el usuario cuando el estado es "No Escucha".

b) **Elocuciones conectadas**. Es similar al reconocimiento de elocuciones aisladas salvo que éste tipo de reconocimiento sí admite una concatenación de elocuciones dentro de una misma muestra, casi como el reconocimiento del habla continua pero a menor escala y teniendo como única restricción que el usuario deberá emitir una breve pausa entre las órdenes para demarcarlas apropiadamente y facilitar así que el sistema pueda reconocerlas.

3.1.4 Verificación O Identificación De Hablante

La intención primordial de este tipo de reconocimiento no suele ser la de detectar un conjunto de comandos específicos, lo cuál no lo imposibilita para que esto pueda formar parte de sus funciones; su propósito último es confirmar, basado en la señal de voz, que la persona realmente es quien dice ser.

La complejidad de esta disciplina radica en la alta variabilidad de la señal de voz. Una misma persona un día puede habar más rápido o venir con algunos problemas en sus cuerdas vocales. Otro problema muy complejo es la robustez a diferentes ambientes en los que puede estar trabajando.

3.2 Clasificación

Los beneficios de interfaces de usuario manejadas por voz han sido apoyados por varios años. La voz es una forma natural de comunicación que es persuasiva, eficiente y puede ser usada a distancia. Sin embargo, la aceptación amplia de interfaces humano-computadora con voz es un hecho todavía por ocurrir. Tomando esto en cuenta, varios esfuerzos de investigación se han iniciado para enfocarse sobre la voz como un canal de entrada auxiliar en ambientes multimodales.

Los sistemas de reconocimiento de voz pueden clasificarse de acuerdo a diferentes criterios, actualmente la mayoría de software con reconocimiento de voz esta basada en los siguientes criterios: si necesitan entrenamiento, independencia del hablante, nivel de ruido y de acuerdo a la cantidad de palabras. Fue por eso que para esta investigación se han tomado dicha clasificación.

3.2.1 NECESITAN ENTRENAMIENTO

Junto con las características técnicas de sistemas de reconocimiento de voz, es importante entender los factores humanos de voz como una modalidad de la interfaz. La más significante es que la voz es temporal y diferente para cada persona. Una vez pronunciada la información, ya no se dispone más de ella. Esto puede representar una carga adicional para la memoria del usuario y limita severamente la habilidad de repasar, revisar y la información de referencias cruzadas. La voz puede ser usada a distancia lo cual la hace ideal para situaciones de manos y ojos ocupados. Es omnidireccional y por lo tanto puede comunicarse a múltiples usuarios.

Los sistemas dependientes de la persona que habla deben ser entrenados para cada usuario individual, pero típicamente tienen más altas tasas de exactitud. Los sistemas adaptables a la persona que habla, un enfoque híbrido, inicia con plantillas independientes de la persona que habla y las adapta a usuarios específicos sobre el tiempo sin entrenamiento explícito, sin embargo para estos casos las tasas de exactitud disminuyen.

3.2.2 INDEPENDENCIA DEL HABLANTE

Los sistemas dependientes del hablante se diseñan alrededor de un hablante en específico. Son generalmente más exactos para el hablante correcto, pero mucho menos exactos para otros. Asumen que el hablante hablará con voz uniforme sin algún cambio de tono. Los sistemas independientes del altavoz se diseñan para una variedad de hablantes. Cabe mencionar, que los sistemas adaptantes comienzan como sistemas independientes y utilizan generalmente técnicas del entrenamiento para adaptarse al hablante para aumentar su exactitud del reconocimiento. Aunque para hablar de este tipo de reconocimiento de voz, se tienen que mencionar los tres tipos de modelado en los que esta basado el reconocimiento de voz:

- **Un modelo del lenguaje** que está basado en una gramática de estados finitos. En esta gramática se definen las posibles frases que serán utilizadas para la interacción humano-computadora.

- **Un modelo de pronunciación** formado por el vocabulario necesario para definir las frases de la gramática, donde la pronunciación se representa como la secuencia de fonemas correspondiente a cada palabra.

- **Un modelo acústico** basado en Hidden Harkov Models (HMM). Se modelan unidades independientes del contexto para identificar los fonemas del lenguaje.

De esta manera, para que un sistema de reconocimiento se considere independiente del hablante, el modelo acústico de los fonemas del lenguaje se lleva a cabo utilizando el tipo de modelado independiente del contexto y el entrenamiento de los mismos se puede hacer utilizando los modelos ocultos de Markov.

3.2.3 Nivel de Ruido

En este apartado, cabe aquel tipo de software que esta diseñado para captar el habla de una manera casi natural, o que esta diseñada para operar bajo canales de transferencia que no sean del todo estables, es decir que tengan algún nivel de distorsión (por ejemplo, en teléfonos celulares) y posiblemente en un entorno donde exista ruido u otras personas que estén hablando simultáneamente.

3.2.4 Cantidad de Palabras

Los sistemas de voz continuos pueden reconocer palabras habladas en un ritmo natural mientras que los sistemas de palabras aisladas requieren de una pausa deliberada entre cada palabra. No obstante más deseable, la voz continua es más difícil de procesar por la dificultad en detectar los límites de cada palabra.

El tamaño del vocabulario puede variar de 20 palabras a más de 40,000 palabras. Los grandes vocabularios causan dificultades en mantener exactitud, pero los pequeños pueden imponer restricciones no deseadas sobre la naturalidad de la comunicación. A menudo el vocabulario debe ser restringido por reglas gramaticales las cuales identifican como las palabras pueden ser habladas en el contexto.

4. Algoritmos

4.1 Algoritmo General

Hay que distinguir, en un principio, el algoritmo más general de reconocimiento de voz que es utilizado en la gran mayoría de aplicaciones que implementan esta forma de interacción como interfaz con el usuario (sean estas de reconocimiento de lenguaje natural, habla o voz), para poder comprender en qué etapas de dicho

algoritmo se encuentran las diferentes técnicas (algunas de ellas reconocidas también como algoritmos) y a que tipo de técnicas pertenecen. Básicamente el algoritmo general de un sistema de reconocimiento de voz puede dividirse en tres grandes etapas o módulos: procesamiento de la señal en el front-end, modelado acústico y modelado del lenguaje. El algoritmo más general funciona ejecutando los siguientes pasos [TLDP1]:

1. **Grabar el audio y detectar la elocución**. Este paso puede ser cubierto en un gran número de formas diferentes. Los puntos de inicio de las elocuciones pueden encontrarse haciendo una comparación de los niveles de audio del ambiente con la muestra que acaba de grabarse. Los puntos de finalización son más difíciles de encontrar debido a muchos factores como los ecos o sonidos producidos por el hablante (como la respiración) lo cuál propicia frecuentemente identificaciones inexactas de ellos.

2. **Pre-Filtrado**. La forma en la que se lleva a cabo este paso depende grandemente de las otras características del sistema de reconocimiento. Los métodos más comunes son el método de "Bancos de Filtros" (basados en FFT) el cuál utiliza una serie de de filtros de audio preestablecidos para preparar la muestra, y el método de Codificación Lineal Predictiva (LPC) el cuál utiliza una función de predicción para calcular las diferencias (errores). También se utilizan diferentes formas de análisis espectral. En esta etapa se realiza el pre-énfasis, la normalización, la separación en bandas, etc.

3. **Enmarcado y Segmentación en ventanas**. Este paso consiste en transformar la muestra de entrada en varias muestras con un formato utilizable para su reconocimiento, básicamente su función principal es la de cortar la señal y separarla en muestras de un tamaño y formato específico (ventanas o segmentos). También incluye la preparación de los límites de la muestra para el análisis (remover los extremos tanto inferiores como superiores de la señal entre otras cosas). Este paso generalmente es agregado al 2 o al 4 y depende directamente de la técnica a utilizar en el 5.

4. **Filtrado**. El filtrado adicional no siempre está presente. Es la preparación final para cada uno de los segmentos o ventanas antes de su reconoci-

miento y usualmente consiste en alinear temporalmente cada ventana y normalizarla.

5. **Comparación y emparejamiento (Reconocimiento)**. Existe una enorme cantidad de técnicas disponibles para llevar a cabo este paso. La mayoría incluyen comparar la ventana actual con muestras conocidas. Existen métodos que utilizan Hidden Markov Models (HMM, que se mencionaron anteriormente y se especificarán en un apartado posterior), análisis de frecuencia, análisis diferencial, técnicas y atajos de álgebra lineal, distorsión espectral, y métodos de distorsión temporal. Todos estos métodos son usados para generar una probabilidad y nivel de concordancia entre la ventana y las muestras.

6. **Ejecución de la acción**. La acción puede ser, en principio, casi cualquier cosa que el desarrollador desee y esté facultado a hacer.

Es importante mencionar que se deben distinguir en esta solución general al menos cuatro capas [UGRANPH: p.3]:

- Capa Acústica: en la que se extraen las características.
- Capa Fonética: en la que se determinan las unidades sonoras básicas.
- Capa Sintáctica: en la que se aplican las reglas gramaticales.
- Capa Semántica: comprender el mensaje y eliminar aquellas interpretaciones que carezcan de sentido.

4.2 ALGORITMO EM

Dadas unas observaciones $y = (y_1, \ldots, y_n)$, realización de un modelo aleatorio (paramétrico), y, con densidad $g(y|\theta)$, la estimación máxima verosímil se basa en la maximización de la función de verosimilitud $L(\theta|y) = g(y|\theta)$

En ocasiones $L(\theta|y)$ es difícil de manejar, pero se pueden aumentar los datos, y, con datos "artificiales", z, para crear una verosimilitud $L(\theta|y,z) = f(y,z|\theta)$ mas fácil de tratar.

Lógicamente, $g(y|\theta) = \int f(y,z|\theta)dz$ g. Denotaremos por $k(z|\theta,y)$ la densidad de Z condicionada por y, es decir,

Tenemos $k(z|\theta,y) = \dfrac{f(y,z|\theta)}{g(y|\theta)}$. entonces que:

$$\log L(\theta|y) = \log L(\theta|y,z) - \log k(z|\theta,y).$$

Al tratar de maximizar $\log L(\theta|y,z)$ nos encontramos con el problema de que z no está disponible. Una posible solución es sustituir $\log L(\theta|y,z)$ por su valor esperado con la distribución condicional de Z dado y ^{y} un valor inicial de los parámetros, θ_0 :

$$Q(\theta|\theta_0,y) := \mathrm{E}_{\theta_0|y} \log L(\theta|y,z) = \int \log L(\theta|y,z)k(z|\theta,y)dz.$$

Ahora maximizamos $Q(\theta|\theta_0,y)$ y tomamos como valor actualizado del parámetro:

$$\hat{\theta}_{(1)} = \underset{\theta}{\operatorname{argmax}} \quad Q(\theta|\theta_0,y).$$

El procedimiento puede aplicarse de forma iterativa, repitiendo las dos etapas anteriores: partiendo de la estimación $\hat{\theta}_{(j)}$,

Etapa E (Calculo de esperanzas), calculamos:

$$Q(\theta|\hat{\theta}_{(j)},y) := \mathrm{E}_{\theta_{(j)}|y} \log L(\theta|y,z) = \int \log L(\theta|y,z)k(z|\theta_{(j)},y)dz.$$

Etapa M (Maximización): optimizamos $Q\left(\theta\middle|\hat{\theta}_{(j)},y\right)$ para obtener:

$$\hat{\theta}_{(j+1)} = \underset{\theta}{\operatorname{argmax}} \quad Q(\theta|\theta_{(j)}, y).$$

La propiedad clave del algoritmo EM es que:

$$L(\hat{\theta}_{(j+1)}|y) \geq L(\hat{\theta}_{(j)}|y).$$

Bajo ciertas condiciones de regularidad, la igualdad se da si y sólo si $\hat{\theta}_{(j+1)} = \hat{\theta}_{(j)}$ y la secuencia $\left\{\hat{\theta}_{(j)}\right\}_j$ se aproxima hacia el estimador máximo verosímil $\hat{\theta}$. En la práctica se emplea interactivamente el algoritmo EM hasta que la diferencia $L\left(\hat{\theta}_{(j+1)}|y\right) - L\left(\hat{\theta}_{(j)}|y\right)$ es menor que cierta tolerancia prefijada.

Para comprender por que se da denotaremos:

$$H(\theta|\theta', y) = \int \log k(z|\theta, y)\, k(z|\theta', y) dz,$$

Es decir, $H(\theta|\theta', y)$ es el valor esperado de $\log k(z|\theta, y)$ con la distribución condicionada por y y parámetro θ'. De la ecuación (3.3) se obtiene (teniendo en cuenta que $\log L(\theta|y)$ no depende de z) que:

$$\log L(\theta|y) = Q(\theta|\theta', y) - H(\theta|\theta', y)$$

Y, por lo tanto:

$$\begin{aligned} \log L(\hat{\theta}_{(j+1)}|y) - \log L(\hat{\theta}_{(j)}|y) \;=\; & \left(Q(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, y) - Q(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, y)\right) \\ & - \left(H(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)}, y) - H(\hat{\theta}_{(j)}|\hat{\theta}_{(j)}, y)\right) \end{aligned}$$

$$
\begin{aligned}
H(\hat{\theta}_{(j+1)}|\hat{\theta}_{(j)},y) - H(\hat{\theta}_{(j)}|\hat{\theta}_{(j)},y) &= \int \log \frac{k(z|\hat{\theta}_{(j+1)},y)}{k(z|\hat{\theta}_{(j)},y)} k(z|\hat{\theta}_{(j+1)},y)\,dz \\
&\leq \log \int \frac{k(z|\hat{\theta}_{(j+1)},y)}{k(z|\hat{\theta}_{(j)},y)} k(z|\hat{\theta}_{(j+1)},y)\,dz \\
&= \log \int k(z|\hat{\theta}_{(j+1)},y)\,dz = \log 1 = 0.
\end{aligned}
$$

El primer termino en el lado derecho de esta igualdad es positivo como consecuencia de $(Q(\theta|\theta_{(j)},y))$ alcanza su mayor valor posible en $\theta_{(j+1)}$. Por otra parte, por la concavidad de la función $\log$ (y la desigualdad de Jensen),

La conclusión es que $\log L\left(\hat{\theta}_{(j+1)}|y\right) - \log L\left(\hat{\theta}_{(j)}|y\right) \geq 0$, y esto prueba.

4.2.1 ESTIMACIÓN MÁXIMO VEROSÍMIL EN MODELOS DE MARKOV OCULTOS

En los modelos de Markov ocultos la aplicación de las ideas anteriores aparece como algo natural, puesto que hay una parte de los "datos" que no está disponible: la cadena oculta. Comprobaremos en que se traduce el algoritmo EM en la estimación por máxima verosimilitud de un MMO-multinomial. Al algoritmo EM en este contexto particular se le llama habitualmente algoritmo de Baum-Welch.

La Log-verosimilitud completa, es decir, la basada en las observaciones $s_1,...,s_n$ y en los estados ocultos no observados $i_1,...,i_n$ es:

$$
\begin{aligned}
\log L_T^c(\delta,\Gamma,\Pi;\mathbf{s},\mathbf{i}) &= P(S_1 = s_1,...,S_T = s_T, C_1 = i_1,\ldots,C_T = i_T) \\
&= \sum_i o_i \log \delta_i + \sum_{i,j} n_{i,j} \log \gamma_{i,j} + \sum_{j,k} m_{j,k} \log \pi_{k,j},
\end{aligned}
$$

Donde o_i es el indicador de ocupación inicial del estado oculto i, n_{ij} es el numeró de transiciones del estado i al j en la cadena oculta y $m_{j,k}$ es el número de observaciones con valor a_k y estado oculto j.

$$Q(\delta, \Gamma, \Pi|\delta^{(l)}, \Gamma^{(l)}, \Pi^{(l)}) = \sum_i e_i \log \delta_i + \sum_{i,j} c_{i,j} \log \gamma_{i,j} + \sum_{j,k} d_{j,k} \log \pi_{k,j}.$$

$$e_i = E_{\delta^{(l)},\Gamma^{(l)},\Pi^{(l)}|s}(o_i) = P\left(C_1 = i/s_1, ..., s_T\right) = \frac{\alpha_1(i)\beta_1(i)}{L_T}$$

$$c_{i,j} = E_{\delta^{(l)},\Gamma^{(l)},\Pi^{(l)}|s}(n_{i,j}) = \sum_{t=1}^{T-1} P\left(C_t = i, C_{t+1} = j/s_1, ..., s_T\right) = \frac{\gamma_{i,j}}{L_T} \sum_{t=1}^{T-1} \alpha_t(i)\pi_{s_{t+1},j}\beta_{t+1}(j)$$

$$d_{j,k} = E_{\delta^{(l)},\Gamma^{(l)},\Pi^{(l)}|s}(m_{j,k}) = \sum_{t=1}^{T} P\left(S_t = k, C_t = j/s_1, ..., s_T\right) = \frac{1}{L_T} \sum_{t:s_t=k} \alpha_t(j)\beta_t(j).$$

Fijados los valores $(\delta^{(l)},\Gamma^{(l)},\Pi^{(l)})$, en el paso E del algoritmo EM debemos sustituir $\log L_T^c$ por su valor esperado condicionado por s:

$$\delta_i^{(l+1)} = \frac{e_i}{\sum_i e_i}$$

$$\gamma_{i,j}^{(l+1)} = \frac{c_{i,j}}{\sum_j c_{i,j}}$$

$$\pi_{k,j}^{(l+1)} = \frac{d_{j,k}}{\sum_k d_{j,k}}.$$

Posteriormente, en el paso M maximizamos $Q(\delta,\Gamma,\Pi|\delta^{(l)},\Gamma^{(l)},\Pi^{(l)})$ para obtener $(\delta^{(l+1)},\Gamma^{(l+1)},\Pi^{(l+1)})$ (7):

4.3 ALGORITMO DE VITERBI

En reconocimiento de voz y otras aplicaciones tiene interés determinar los estados de la cadena oculta que con mayor verosimilitud han originado la secuencia observada.

A este problema se le suele llamar "problema de decodificación". Existen distintas versiones del problema. Una de ellas es la decodificación global. Esto consiste en determinar la secuencia de estados $\hat{i}_1,...,\hat{i}_t$ que maximizan la probabilidad condicional.

$$P\left(C_1 = i_1, ..., C_T = i_T/S_1 = s_1, ..., S_T = s_T\right).$$

La decodificación global se puede efectuar mediante el método conocido como algoritmo de Viterbi. Describimos este algoritmo a continuación:

Si $\lambda = (A, B, \pi)$ es el modelo inicial, y $\lambda = (\bar{A}, B, \bar{\pi})$ reestimado. Se puede demostrar entonces que:

1. El modelo inicial λ, define un punto crítico de la función de probabilidad, en cuyo caso $\lambda = \lambda$, o

2. El modelo λ es más probable que λ en el sentido de que $P(0 | \bar{\lambda}) > P(0 | \lambda)$, ej., hemos encontrado un nuevo modelo λ, a partir del cual es más probable que la secuencia de observación se haya producido. Por tanto, podemos mejorar la probabilidad de que O sea observado a partir del modelo, si utilizamos iterativamente λ en lugar de λ y repetimos la reestimación hasta que se alcance algún punto restrictivo. El modelo resultante es conocido como el HMM con máxima probabilidad (7).

4.5 ALGORITMO K-MEANS

1. INICIALIZACIÓN: Arbitrariamente elegimos M vectores o palabras de código, codewords, como el grupo inicial del codebook.

2. BÚSQUEDA DEL MÁS CERCANO: Por cada vector de observación, se busca el codeword en el codebook que es el más cercano (en términos de distancia), y asigna a ese vector a la celda correspondiente.

3. ACTUALIZACIÓN DEL CENTROIDE: actualiza el codeword en cada celda o sector usando el centroide de los vectores de entrenamiento asignados a un sector.

4. ITERACIÓN: Repite los pasos 2 y 3 hasta que la distancia media caiga debajo de un umbral prefijado.

La forma de cada sector o celda o partición es muy dependiente de la medida de distorsión espectral y las estadísticas de los vectores en el grupo de entrenamiento.

Este método es el más simple y por tanto existen numerosas modificaciones y mejoras, algunos de sus puntos débiles son:

1. Los resultados dependen en forma muy acentuada de los valores iniciales elegidos como palabras de código.

2. También hay gran dependencia del número de sectores M así como de la implementación de la "distancia" usada.

3. Puede suceder que algunos de los sectores resulten vacíos (7).

4.6 Algoritmo LBG.

Se analizará con algún detalle debido a su buen desempeño, para eso comenzaremos por el algoritmo fundamental LBG.

El algoritmo LBG, lleva su nombre debido a sus autores Y. Linde, A. Buzo y R. M. Gray, en él se elige 1 codeword inicial de entre los vectores de datos a clasificar, luego se utiliza el algoritmo de división binaria para duplicar el número de codewords, los vectores de observación se agrupan en torno a los codewords que les presentan menor distancia, se recalculan los codewords como la media multidimensional de cada sector y se agrupan nuevamente los datos, el proceso se detiene cuando el codebook no presenta variación significativa y al llegar al número de codewords deseados.

Este algoritmo de gran popularidad (que utiliza el algoritmo k-Means) produce codebooks que logran un mínimo local en la función de error por distorsión.

Para generar un codebook de M sectores o palabras de código:

En primer lugar designando un inicial para luego utilizando una técnica de división llegar a obtener un codebook inicial, luego iterando la misma técnica de división en los codewords hasta que llegamos a obtener el número de codewords igual a M que va a ser el tamaño del codebook deseado.

El procesamiento se denomina división binaria:

1. Designar 1 vector del codebook o codeword inicial, éste resulta ser el centroide del grupo de los vectores de entrenamiento.

2. Calcular la media del grupo de entrenamiento:

$$Cw_i \frac{t}{N} \sum_{n=1}^{N} x_n$$

2.1. Calcular el error o distancia media entre el codeword inicial y los vectores de entrenamiento:

$$D = \frac{t}{N}\sum_{n=1}^{N}\|x_n - Cw_t\|^2$$

3. Duplicar el tamaño del codebook mediante la división de cada codeword:

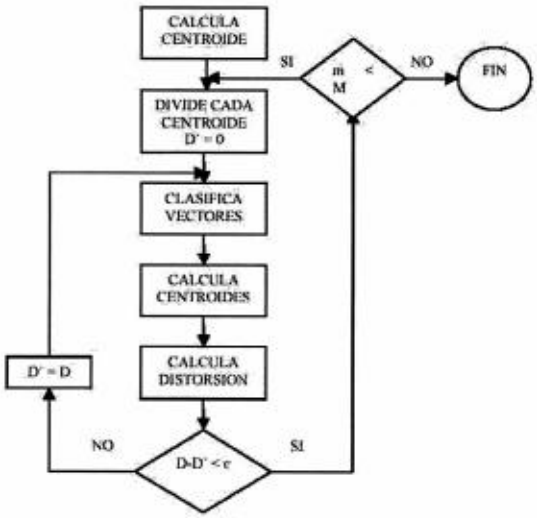
$$Cw_i^+ = Cw_i(1+\epsilon)$$
$$Cw_i^- = Cw_i(1-\epsilon) \qquad 0 < \epsilon << 1$$

4. Usar el algoritmo K-Means para tomar el mejor grupo de centroides para la separación del codebook.

5. Iterar pasos 3 y 4 hasta llegar a un codebook de tamaño M.

Una de las causas que motivo el uso de un VQ fue la suposición que, en el límite, el codebook debería idealmente tener 36 vectores, uno por cada fonema, suposición que es incorrecta.

En el esquema de la figura se representa el algoritmo que sintetiza el proceso de generación del codebook mediante el método LGB.



Algoritmo de generación del codebook por división binaria

4.7 ALGORITMO LBG-U

$E(D,C) = \sum_{x \in D} d(x, Cw_{i(x)})$ siendo D : Conjunto de vectores de datos x

C : Codebook

$Cw_{i(x)}$: Codeword representativo del sector i que le corresponde al vector x

Existe una modificación que permite obtener una sustancial mejora, y el que la logra es el algoritmo LBG-U. Si existiera algún parámetro que describiera la influencia de cada codeword al error total, podría tratar de modificarse el que fuera más importante, es decir podría situarse otro codeword muy cercano a él de forma de reducir el error. El problema es ahora encontrar el codeword que debemos tomar de nuestro codebook para realizar esta tarea, la solución sería contar con otro parámetro que informara sobre la contribución de cada codeword a la reducción del error, esto se logra gracias a la medida de utilidad que da el nombre al método.

La utilidad de cada codeword se obtiene si, teniendo el error por distorsión total restamos el error por distorsión que resulta de un codebook al que se le extrae únicamente el codeword representativo i. Refresquemos entonces la idea de error por distorsión E :

Se trata de suma las distancias de todos los vectores a cuantificar, a su codeword representativo. La medida de utilidad U para cada sector se logra comparando la distorsión del codebook C, con la que tendría si el codeword a evaluar no existiera:

$$U(Cw_i) = E(D, C - Cw_i) - E(D, C)$$

La remoción de un codeword afecta la distorsión sólo para aquellos vectores de dicho sector, que en su ausencia se incorporarán al sector cuyo representante sea el segundo en cuanto a su distancia, mientras que el resto no se verá afectado. Entonces se puede reescribir U como:

$$U(Cw_i) = \sum_{x \in V_i} d(x, Cw_{i'(x)}) - d(x, Cw_i) \quad \text{con } Cw_{i'(x)} \text{ codeword secundario para el vector } x$$

V_i región i

El codeword que presente menor utilidad será el más apto para ser removido pues será el que ejerza menor incidencia sobre el error, marca lo útil que es cada codeword para el codebook.

El error correspondiente a cada sector $E(Cw_i)$:

$$E(Cw_i) = \sum_{x \in V_i} d(x, Cw_i)$$

Se trata de la sumatoria de las distancias de los vectores de una determinada región V_i, a su vector representativo. A esta altura se cuenta con información para elegir; el codeword menos útil Cw_a, que será eliminado, y el que produce mayor contribución al error Cw_b que debe ser reforzado por otro cercano. La medida de que tan cerca estará se puede aproximar calculando la desviación Standard de los elementos del sector V_b, tomando una longitud mucho menor que ésta y la dirección del vector será tomada aleatoriamente. En resumen se ejecuta el algoritmo LBG y se obtiene un codebook, de este se elimina el codeword de menor utilidad, que será reemplazado por otro ubicado junto al de mayor error desplazado una longitud y dirección especificadas:

$$Cw_a = \arg\min_{Cw \in C} U(Cw)$$

$$Cw_b = \arg\max_{Cw \in C} E(Cw)$$

$$Cw_a = Cw_b + \left(\varepsilon \sqrt{E(Cw_b)/N}\right) u \quad \text{con} \quad 0 < \varepsilon << 1$$

N siempre será mayor que *std* de sector Vb

u vector n dimen. aleatorio

El proceso se itera mientras que se produzca una disminución en el error por distorsión. En la práctica se obtiene una reducción del error de más del 10%, a costa de un incremento del tiempo de procesamiento de 3 a 7 veces.
Por último la gran ventaja de éste método se pone de manifiesto al manejar datos con grandes diferencias de densidad entre grupos, en los que es capaz de mover codewords de forma óptima.

4.7.1 UTILIZACIÓN DEL CUANTIFICADOR Y DEL CODEBOOK.

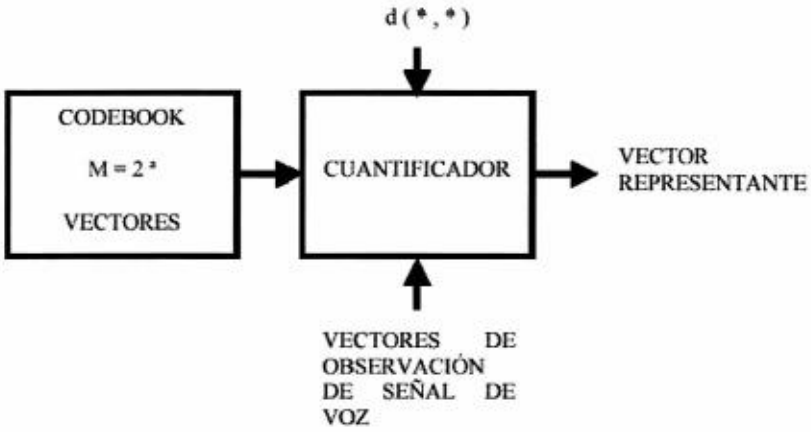
Una vez construido el codebook, el procedimiento para cuantificar vectores es básicamente realizar una búsqueda completa a través del codebook para encontrar el mejor representante.

Si anotamos los vectores del codebook, de tamaño M, como $Cw, 1 \leq w \leq M$, y tomamos al vector de observación a ser cuantificado como V, luego el vector representante o codeword, Vm^* es:

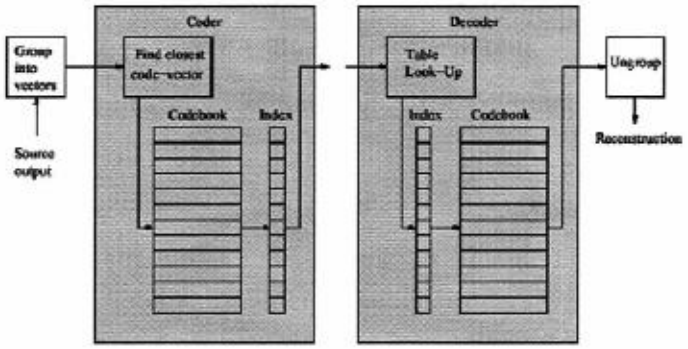
$$Vm^* = \arg\min d(V, Cw)$$

$$1 \leq w \leq M$$

Un procedimiento de cuantificación para señal de voz elige el vector más cercano del codebook al vector de observación y utiliza ese vector denominado codeword, como la representación resultante para etapas posteriores. Se refiere como al vector "vecino" más cercano, toma como entrada, vectores de señal de voz y da como respuesta, a su salida, el vector que mejor representa esa entrada.



4.7.2 APLICACIONES



Esquema de transmisión de señales o datos

1. Transmisión de señales y datos con el resultado de una reducción en el ancho de banda debido a que solamente se transmite un índice al vector representante, en el que se necesita un codebook en el lado transmisor como codificador y otro codebook, idéntico al primero, en la función decodificadora en el lado del receptor obteniendo a su salida, además, el error por distorsión.

2. Compresión de imágenes, tomando como ejemplo la figura, vemos que se divide la imagen en rectángulos fijos, conteniendo cada uno de ellos muchos píxeles, y que van a ser representados, en la cuantificación, por el índice al vector representante o codeword contenido en el codebook. Cada sector de la imagen se transmite usando el índice correspondiente al codeword del codebook.

0	0	1	0
2	2	2	0
3	2	2	0
0	4	0	0

Imagen dividida y cuantificada

0
1
2
3
4

Índice del *codebook*

4.7.3 EN RECONOCIMIENTO DE VOZ

1. El uso de múltiples codebooks en los cuales cada codebook se crea separadamente (e independientemente) para cada una de las representaciones de la señal de voz (espectral o temporal). Por ejemplo se podría crear un codebook para las representaciones de los parámetros del cepstrum y otro en forma separada conteniendo las representaciones de las derivadas del cepstrum.

2. K-tuples cuantificadores en el cual K-tramas de señal de voz se codifican a la vez, en lugar de una única trama como es común. La idea es utilizar las correlaciones en el tiempo entre los sonidos vocales puros y los que tienen componente vocal. La desventaja ocurre cuando los sonidos donde la correlación a través del cuantificador es baja, como son los sonidos transitorios y consonantes.

3. Cuantificación de matrices en las cuales se crea un codebook de sonidos o palabras de secuencia de longitud variable. El concepto es manejar la variación temporal vía algunos tipos de procedimientos dinámicos y por

medio de eso crear un codebook de secuencias de vectores que representan sonidos típicos.

4. Modelos ocultos de Markov en los cuales ambas reducciones, en tiempo y espectro, se usan para cuantificar la emisión completa de la voz de una manera definida y eficiente.

CONCLUSIÓN

Los algoritmos, métodos y técnicas para el análisis y reconocimiento de voz empleadas actualmente son eficientes, hasta cierto grado, se manifiesta que las operaciones que se manejan en los algoritmos y métodos poseen un alto nivel de abstracción y refleja una alta complejidad, que requiere de estudios más profundos y tiempos prolongados para evitar errores en el reconocimiento de un comando hablado o reconocimiento de voz; esto porque los problemas que deben resolver en un sistema de este tipo están relacionados a una gran variedad de disciplinas como la acústica, el procesamiento de señales, reconocimiento de patrones, fonética, ciertamente ciencias de la computación y otras más que a veces no se pueden (o no se quieren) tomar en cuenta debido a la inexperiencia en el ámbito por parte de los desarrolladores.

Hasta no encontrar los algoritmos o métodos ideales para el reconocimiento de voz que eliminen en un ciento por ciento el riesgo de verse afectadas por las causas de error se describirán brevemente a continuación, no será posible tener un sistema que utilice esa técnica en el que se pueda confiar plenamente.

a) Diferencias entre los hablantes; las variaciones entre las voces de los usuarios de un sistema de reconocimiento de voz son muchísimas y con frecuencia muy difícil de obviar.

b) Condiciones del ambiente; los lugares donde el reconocimiento de voz representa una aplicación importante (oficinas, centros de estudio, etc.) a menudo presentan condiciones muy adversas para la aplicación de esta técnica (ruido, señales distorsionadoras, atenuación de la entrada, etc.).

c) Entonación y timbre; incluso una misma persona puede pronunciar con diferente entonación y timbre de voz una misma palabra dependiendo de una gran variedad de factores, incluso la rapidez misma con la que se pronuncia la palabra puede interferir en su reconocimiento.

d) Complejidad del lenguaje; las características particulares de los fonemas que pertenecen al lenguaje del cual se desean reconocer los patrones de voz, para posteriormente obtener los comandos u órdenes que deberá ejecutar el computador, conllevan una posible fuente de error dependiendo de su complejidad.

Por último, la normalización de las señales es un proceso muy importante dentro del reconocimiento de voz y especialmente para obtener una correlación cruzada (lo que permite la comparación de las dos señales) que sea lo más precisa posible.

BIBLIOGRAFÍA

La mayoría de la información encontrada aquí, así como de las gráficas, dibujos y espectrogramas, han sido extraídas de:

Libro: Reconocimiento de Voz y Fonética Acústica. Autores: Jesús Bernal Bermúdez, Jesús Bobadilla Sancho, Pedro Gómez Vilda. Editorial: Ra-Maâ

[TLDP1] "Speech Recognition HOWTO", Stephen A. Cook, University of Toronto, Canada.
http://www.tldp.org/HOWTO/Speech-Recognition-HOWTO/inside.html#RECOGNIZERS

[TLDP2] "Speech Recognition HOWTO", Stephen A. Cook, University of Toronto, Canada.
http://www.tldp.org/HOWTO/Speech-Recognition-HOWTO/introduction.html#TYPES

[TLDP3] "Speech Recognition HOWTO", Stephen A. Cook, University of Toronto, Canada.
http://www.tldp.org/HOWTO/Speech-Recognition-HOWTO/introduction.html#BASICS

[UDLAP] "Reconocimiento de Voz", Ingrid Kirschning de Ayala, Universidad de Las Américas, Puebla, México.
http://ict.udlap.mx/people/ingrid/Clases/IS412/index.html

[UGRANPH] "Fundamentos De Reconocimiento de Voz", Antonio M. Peinado Herreros, Departamento de Electrónica y Tecnología de los Computadores, Universidad de Granada, España.

10. Contribución a la Integración de Servicios Multimedia en Redes de Sensores.

Vicente P. Saldivar, Luis J. de la Cruz Llopis

I. Introduccion

Las redes de sensores inalámbricos (Wireless Sensor Networks, WSN), tienen como principal objetivo realizar estimaciones/detecciones de fenómenos físicos de manera confiable y eficiente. Están formadas por una gran cantidad de nodos multifuncionales, limitados en energía, procesamiento, capacidad de memoria, de bajo costo y de pequeñas dimensiones. Estos nodos se comunican a corta distancia y realizan un trabajo colectivo para obtener información del entorno, y suelen ser utilizados para obtener lecturas, detectar eventos, localización de objetivos y control.

Existen muchos estudios sobre las WSN para solventar los problemas de restricciones en los nodos, la gran mayoría de ellos están relacionados con la comunicación de datos convencionales y están diseñados para ser eficientes en el consumo de energía, por ser esta una de las principales limitaciones de los nodos. Sin embargo, existen propuestas donde el uso de aplicaciones con soporte para audio y video (comúnmente denominadas como aplicaciones multimedia) se hace necesario para complementar las actuales soluciones implementadas en las WSN e incluso para una serie de nuevas aplicaciones. A las WSN con esta capacidad se les conoce como redes de sensores multimedia inalámbricos (Wireless Multimedia Sensor Networks, WMSN). Para visualizar la necesidad de las WMSN, imaginemos un campo de batalla donde son desplegados varios nodos con sensores de movimiento que, cuando detecten algún movimiento informarán al centro de mando para iniciar un medida. Esta información por sí sola no sería de gran utilidad para el centro de mando ya que no se contaría con información suficiente para emitir una respuesta a este acontecimiento. Por ejemplo, si los nodos detectan el movimiento y se ordena un ataque sin ninguna otra información,

este ataque podría ser lanzado a tropas aliadas o civiles. Por el contrario, si los sensores además de poseer un sensor de movimiento fuesen dotados de un dispositivo de captura de imágenes serían capaces de, en el momento de detectar el movimiento, iniciar el envió de imágenes al centro de mando. Una vez que el centro de mando recibe la información puede tomar una acción correcta dependiendo de la información visualizada.

El soporte del tráfico multimedia impone una serie de problemas que deben ser solventados para que este tipo de comunicaciones sea factible. Por ejemplo, el tráfico multimedia no requiere el 100% de confiabilidad, pero es muy exigente en cuanto a requerimientos de retardo, variación del retardo (*jitter)*, ancho de banda y los cambios en las tasas de transmisión [Gur05, Mis07]. A estos requerimientos se les conoce comúnmente como QoS y es importante recalcar que no sólo son importantes para la transmisión de datos multimedia sino también para datos escalares. Por ejemplo, si una estación de monitoreo de movimientos sísmicos necesita enviar una alerta cuando se detecte un movimiento superior a determinada escala. Si dicho evento se produce, pero la información llega con retardo a la estación, es posible que no se tomen las medidas oportunas para enfrentar el cataclismo. Como resultando de la necesidad de QoS en ciertas aplicaciones, se crea un complejo diseño donde la pila de protocolos debe operar lo más eficientemente posible para dar soporte a estas tecnologías, tomando en cuenta tanto las necesidades del tráfico multimedia y de datos así como las limitaciones del nodo, en especial la energía.

Dentro del modelo por capas OSI (Figura 1), la capa que más repercusión tiene en el consumo de energía del nodo es la capa de Enlace. Más específicamente, el diseño de la subcapa MAC ejerce una gran influencia en el consumo de energía ya que a través de ella se controla la radio, dispositivo que contribuye enormemente al consumo de energía total que ronda aproximadamente entre 36mW y 57mW para transmitir y 12mW a 63mW para recibir. Por su parte, la capa de Red es una de las principales capas para dar soporte a la QoS, algo que es elemental en las aplicaciones multimedia así como en ciertas aplicaciones que utilizan sólo datos escalares. También, esta capa es la intermediaria entre la capa de Aplicación y la capa de Enlace para el intercambio de parámetros de rendimiento.

Figura 1. Pila de protocolos modelo OSI.

Además de los factores de QoS, existe otro que también puede degradar el servicio ofrecido a los usuarios, que es la relativa facilidad con la cual los nodos pueden sufrir un ataque de seguridad. Es importante garantizar al usuario que, además de que la información le llegara a tiempo y sin errores ocasionados por la red, la información estará segura y será íntegra. Esto lo lograremos a través de mecanismos de seguridad que, como en el caso de las capas MAC y Red, sean energéticamente viables.

II. WIRELESS SENSOR NETWORKS

Las WSN son un tipo de redes inalámbricas compuestas de pequeños dispositivos con limitaciones en energía y capacidad. Su función consiste en informar a un usuario sobre eventos de importancia que se llevaban a cabo en el lugar donde son desplegados. Aunque en un principio esta tecnología se pensaba para uso militar, sus características han despertado el interés de sectores como la investigación ambiental, la medicina e incluso de sectores comerciales como por ejemplo el agrícola. El uso de esta tecnología promete impactar la forma en la que se realizan diferentes tareas. Sin embargo, es importante señalar que aun existen diferentes factores que evitan el uso de estas redes de manera común (el consumo de energía, principalmente). Aunque, existe una amplia investigación en torno a estos factores y sus posibles soluciones.

Las WSN comparten en muchos aspectos similitudes con las redes Ad-hoc móviles (MANET), sin embargo, tienen significativas diferencias. Dentro de estas podemos mencionar, la densidad de nodos, la topología y la movilidad, si bien las más importantes son la limitación de recursos y los modelos de comunicaciones [Lan07].

Los dispositivos o nodos de las WSN están compuestos principalmente de las siguientes unidades: módulo de procesamiento, de energía, de comunicación y detección. Además de estos módulos pueden existir otros sistemas que complementen al nodo como por ejemplo sistemas de movilidad, localización y generación de energía. Estos componentes están detallados en la Figura 2 [Aky02]. Antes de mencionar algunas características de los nodos, es importante señalar que cada vez más se utiliza una red en capas en lugar de una red plana. Dependiendo del tipo de red (homogénea o heterogénea) los nodos tendrán las mismas características o existirán algunos nodos con mayores capacidades. Los nodos vienen equipados generalmente con un procesador de 8 o 16 bits. Aunque existen algunos como el Imote que incorporan un procesador de 32 bits. La velocidad del reloj oscila entre 1-10 MHz. En cuanto a la memoria, la Flash (utilizada para

almacenar el código de programa) es generalmente suficiente. No así la memoria RAM donde la cantidad de memoria (4-10Kb) obliga a minimizar el impacto en memoria del software. Por otro lado, el ancho de banda manejado por la mayoría de los nodos comercializados está en los 250 Kbps. Además, la calidad y el rango del enlace son muy pobres [Lan 07].

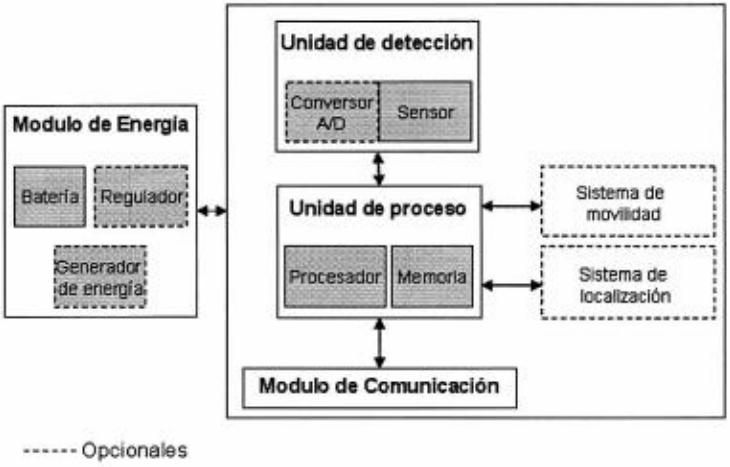


FIGURA 2. COMPONENTES DE UN NODO DE LAS WSN.

Las WSN no tienen una arquitectura de red estandarizada pero se suele utilizar el modelo empleado por Akyildiz [Aky02] el cual está representado en la Figura 3. El modelo utiliza cinco capas (Física, Enlace, Red, Transporte y Aplicación, y también emplea tres planos de administración. La capa Física es la responsable de las tareas de selección de frecuencia, generación de frecuencia portadora, detección de señal, etc. Comúnmente en este tipo de redes, se utiliza la banda industrial, científica y médica (ISM). La capa de Enlace es la responsable de la multiplexación de flujos de datos, detección de tramas, control de acceso y control de errores. Dentro de ella se encuentra la subcapa MAC, de la cual abundaremos más en el siguiente apartado. La siguiente capa (Red) se encarga de la elección de la ruta para la transmisión de la información y la entrega fiable de datos, así como de establecer la estructura de red. La capa de Transporte ayuda a mantener el flujo de los datos. Por último, dependiendo de las tareas a realizar, diferentes tipos de software puede ser implementado en la capa de Aplicación. Por otra parte, los planos de administración ayudan a la coordinación de las tareas de los nodos así como al control del consumo de energía total.

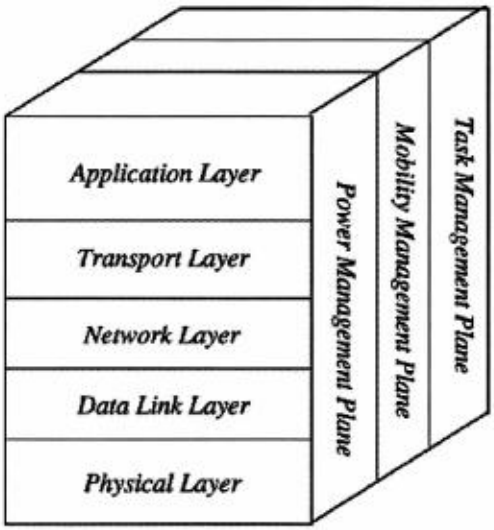


FIGURA 3. PILA DE PROTOCOLOS DE LAS WSN.

Dependiendo de la aplicación, diferentes arquitecturas y diseños han sido considerados para las WSN. Siendo que el funcionamiento de las capas MAC y de Red están ligados a la arquitectura empleada, mencionaremos algunas de las características que pueden ser tomadas en cuenta para el diseño de estas arquitecturas [Wan06, You04]:

- El tipo de despliegue de la red.- La forma de despliegue dependerá del tipo de aplicación, pero generalmente suele ser manual o aleatoria. Esta característica afectará al protocolo de encaminamiento, pues de ella depende la manera en que están agrupados los nodos.

- Características del Hardware.- En los primeros trabajos sobre WSN se asumía que todos los nodos eran homogéneos. Sin embargo, dependiendo de la aplicación, un nodo puede tener mayores capacidades y ser elegido para una función especial. Por ejemplo, muchos de los protocolos de encaminamiento designan a un nodo como clusterhead (CH) y suele ser el nodo con mayores capacidades.

- La dinámica de la red.- Se puede presentar por diferentes factores, desde el cambio dinámico en la topología hasta el provocado por la na-

turaleza poco fiable de las comunicaciones inalámbricas. La topología puede cambiar ya sea por el despliegue de nuevos nodos, su destrucción, movilidad o fallo. También, el evento a detectar puede ser dinámico o estático dependiendo de la aplicación, por ejemplo, si es necesario el seguimiento/detección el evento es dinámico. Los eventos dinámicos requieren del envió continuo de datos, mientras que, los estáticos sólo cuando se detecta el evento.

- Comunicaciones del nodo.- Durante la creación de la infraestructura, el proceso de poner en marcha las rutas de encaminamiento están influenciadas por las consideraciones de energía. El empleo de encaminamiento multisalto consume menos energía que la comunicación directa (un solo salto). Sin embargo, el encaminamiento multisalto introduce un overhead significativo.

- Modelos de entrega de datos.- Dependiendo de la aplicación, pueden existir cuatro tipos de entrega de datos. El primero es la entrega continua de datos, donde los nodos envían información constantemente a la estación base. En los modelos basados en eventos y basados en peticiones, el envió de datos empieza cuando un evento ocurre o una petición es realizada por el usuario. En algunas WSN se emplea un modelo hibrido, el cual emplea una combinación entre el envío continuo y los basados en peticiones o eventos. El tipo de modelo empleado afecta a las capas de Red y MAC en lo que se refiere a consumo de energía y estabilidad de las rutas.

- La redundancia de datos.- La mayoría de las WSN genera datos redundantes. En algunas aplicaciones, esta característica se puede explotar agregando los datos y reduciendo el número de transmisiones. Existen dos técnicas principales, (1) el agregado de datos y (2) la fusión de datos. En la primera, se combinan los datos de diferentes fuentes usando técnicas como: supresión, min, max y promedio. Algunas de estas funciones se pueden llevar acabo parcial o completamente cada nodo. La segunda técnica el agregado de datos se realiza mediante técnicas de procesamiento de señal como beamfroming. Aunque se reduce la cantidad de transmisiones, el agregado de datos complica el diseño de la capa MAC y el soporte para QoS, además, introduce latencia.

TABLA 1.
CARACTERÍSTICAS EN EL DISEÑO DE LA ARQUITECTURA.

Tema de diseño	Factores
Dinámica de la red	Cambio en la topología Tipo de evEnto detectado
Despliegue del nodo	Manual Aleatoria
Comunicación	Multi-hop Single-hop
Entrega de datos.	Continuo Basado en petición Basado en evento Hibrido
Capacidades del nodo	Multifunción Homogéneo Heterogéneo
Redundancia de datos	Utilizar o no agregado o fusión de datos.

Es importante señalar que la información multimedia influye en los factores de diseño anteriores. Dentro de los modelos de entrega de datos, el modelo continuo parece ser el menos adecuado para las WSN debido a que una transmisión multimedia constante requiere un importante consumo de energía. Es preferible la utilización de métodos basados en peticiones o eventos. También, el agregado de datos no se puede realizar mediante métodos como supresión, min, max y promedio, complicando, por lo tanto el diseño de la arquitectura [Gur05].

A. Calidad de servicio en las WSN

La mayoría de las investigaciones realizadas hasta ahora en las WSN se han enfocado principalmente en la reducción del consumo de energía. Conceptos como la latencia, throughput y la variación de retardo no eran las principales preocupaciones de los trabajos de investigación [You04]. Sin embargo, el incremento del interés por las aplicaciones en tiempo real aunado a la inclusión de transmisiones de voz, video e imágenes traen consigo un nueva valoración de los parámetros hasta ahora desestimados. Poco trabajo existe hasta ahora que ofrezca QoS a las WSN y los existentes solo toman en consideración una capa o entorno de aplicación [You04, Wan06].

La forma de dar soporte a la QoS en las redes cableadas es generalmente sobre dotando de recursos o con ingeniería de tráfico. Proporcionar un extra en los recursos es algo complicado en las WSN tomando en cuenta las limitaciones de las mismas. Con ingeniera de tráfico, las aplicaciones o usuarios son clasificados en clases, cada una de la cual recibirá un trato dependiendo de la clasificación dada. Dos de los esquemas usados por la ingeniera de tráfico en las rede IP son los (1) basados en reservas (Interserv) y (2) los no basados en reserva (Diffserv). En el primer método, los recursos son asignados de acuerdo a las peticiones de la aplicación y de acuerdo a una política de administración. El segundo método emplea diferentes estrategias como control de admisión, clases de tráfico, etc. Sin embargo, estos métodos no pueden ser directamente empleados en las WSN principalmente por el limitado ancho de banda, la posibilidad de que exista movilidad en algunas aplicaciones y la calidad del enlace. De igual forma, las soluciones implementadas en redes inalámbricas tradicionales son difíciles de implementar sobre todo por la diferencia entre capacidades de los dispositivos [Che04].

Por otra parte, soluciones implementadas en las MANETS (donde se comparten similitudes) emplean esquemas más complejos que incluyen modelos de QoS, señalización para reserva de recursos, encaminamiento y control de acceso al medio con QoS. Sin embargo, estas soluciones no son apropiadas para las WSN por varias razones, entre ellas las diferencias entre estas dos redes y las características de diseño de las arquitecturas en las WSN.

Aunque las WSN tienen muchos de los problemas de las redes inalámbricas para dar soporte a la QoS, requieren de un tratamiento especial debido a sus características. En [You04] se mencionan las consideraciones de diseño para gestionar la QoS en las WSN, las cuales son:

- Limitación del ancho de banda.
- Redundancia de los datos.
- La compensación entre energía y retardo.
- Limitaciones del buffer.
- Soporte de diferentes tipos de tráfico.

En [Che04] han realizado una clasificación de las aplicaciones de acuerdo a la entrega de datos presentes en las WSN y sus requerimientos. Como hemos mencionado, existen cuatro tipos de modelo de entrega de datos:

- Continuo.- Aquí, los nodos envían su información continuamente a una tasa pre-especificada. Dentro del tráfico que puede existir están los datos que requieren ser transmitidos en tiempo real y aquellos que no lo re-

quieren. Las de tiempo real requieren un cierto ancho de banda y retardos restringidos.

- Basado en eventos.- La mayoría de las aplicaciones basadas en eventos son interactivas, en tiempo real, misión crítica y no requieren de un rendimiento extremo a extremo.
- Basado en peticiones.- tienen los mismos requerimientos que los basados en eventos. Sin embargo, mientras que en los basados en eventos el trafico fluye de los nodos al sink (o punto de acceso, clusterhead), en este modelo el flujo de datos es iniciado por el sink.
- Hibrido.- En este modelo, dos de los modelos anteriormente descritos pueden coexistir requiriendo un mecanismo que soporte diferentes tipos de tráfico con requerimientos de QoS.

III. Capa Mac

Como se ha mencionado, la limitación de energía es uno de los principales factores que afectan al nodo. Su escasez conlleva un diseño enfocado a reducir su consumo por parte de los diferentes dispositivos que componen a un nodo. El diseño de los protocolos MAC también se ve afectado por esta limitante, además, el transceptor (transceiver) es la dispositivo con mayor consumo de energía del nodo. Por tal motivo, los protocolos MAC para las WSN son diseñados con el objetivo de reducir el consumo de energía.

Un protocolo MAC provee diferentes funcionalidades dependiendo de los requerimientos de la red, el dispositivo y las capas superiores. Por lo general, las funciones que un protocolo MAC debe ser capaz de proveer son [Kre07]:

- Construcción de la trama.- El protocolo MAC definirá el tamaño de la trama y realizara el encapsulado y desencapsulado de los datos.
- Acceso al medio.- En esta capa se controla qué dispositivos participan en la comunicación y en qué momento. Esta es la función principal en los protocolos MAC para redes inalámbricas debido a que puede existir fácilmente colisiones.
- Fiabilidad.- Generalmente, mediante acuses de recibo (ACK) y retransmisiones (la capa MAC) se asegura que la transmisión entre dispositivos sea exitosa.
- Control de flujo.- Se utiliza para prevenir perdidas de tramas por sobrecarga en los "buffers" del dispositivo receptor.

- Control de errores.- Necesario para controlar la cantidad de errores presentes en las tramas mediante códigos de detección y/o corrección de errores.

Existen muchos trabajos sobre protocolos MAC en redes inalámbricas. Sin embargo, hay una serie de aspectos que limitan o nulifican el uso en las WSN de los protocolos MAC diseñados para otras redes inalámbricas. La mayoría de estos aspectos tiene que ver con las limitaciones de los nodos y la ineficiencia de los protocolos para solventarlas. Por esto, se han desarrollado nuevos protocolos específicos para las WSN. Para el diseño de estos protocolos, se han tomado en cuenta una serie de factores que son determinantes para estas redes, entre los principales están [Ye04]:

- Evitar las colisiones. Esta es una de las funciones principales de la capa MAC. Aunque no siempre es posible evitarlas, es necesario reducir la frecuencia con que se presentan.
- La eficiencia energética. Los nodos tienen una energía limitada. Por tal motivo es necesario el uso de algoritmos que hagan eficiente el uso de este recurso.
- La escalabilidad y adaptabilidad. Las WSN están pensadas para ser desplegadas densamente. Además, puede existir cambios en la topología. Es importante para un protocolo MAC adaptarse a estos cambios y condiciones manteniendo un correcto funcionamiento.
- Utilización del canal. También referido como capacidad de canal. Refleja que tan bien es utilizado el ancho banda total del canal.
- Latencia.- Nos referimos a este factor como el retardo extremo a extremo, es decir, el tiempo que transcurre desde que se envía el paquete hasta que se recibe con éxito. La importancia de la latencia depende de la aplicación.
- Caudal efectivo (throughput). Es la cantidad de datos transferidos con éxito (a menudo expresado en bits o bytes por segundo).
- Equidad (fairness). En las redes tradicionales esto es un atributo importante ya que todos los nodos desean acceder al canal. Este reparto debe ser equitativo para garantizar que todos los nodos pueden recibir y transmitir información. Sin embargo, en las WSN todos los nodos cooperan para llevar a cabo una tarea.

A. RAZONES DE PÉRDIDA DE ENERGÍA

La eficiencia energética es uno de los factores que más influye en el diseño de los protocolos MAC. Por tal motivo, es necesario identificar las causas de la pérdida de energía. En [Ye04, Dem06] se mencionan varios orígenes, dentro de ellos están: las colisiones, causadas por la transmisión de dos paquetes dentro del mismo periodo de tiempo. Estos paquetes son descartados y es necesaria su retransmisión lo que conlleva un consumo de energía extra. Otra fuente de pérdida de energía es causada por la sobrecarga en la red de paquetes de control (control packet overhead). Ya que, aunque los paquetes de control no llevan datos también consumen energía. Esto también perjudica el caudal útil (goodput).

Cuando un nodo recibe paquetes que no están destinados a él (overhearing), se produce una pérdida de energía ya que el nodo recibe información innecesaria. Por otra parte, si se envía información al nodo antes de que esté listo para recibirla (overmitting) se producirá una gasto de energía por parte del transmisor. Por último, la escucha inactiva (idle listening) es otro origen del malgasto de energía. Esta ocurre cuando un nodo no sabe cuando recibirá información y mantiene su radio encendida para escuchar el canal y recibir esta posible información.

Una de las formas que se utiliza para el ahorro de energía es poner el transceptor en modo "sleep". Aunque este método proporciona una ayuda al ahorro de energía, esto solo será sustancial cuando el transceptor esté en modo "sleep" durante una gran cantidad de tiempo [Lan07]. Sin embrago, el tiempo que transcurre entre el cambio del modo "sleep" al modo de recepción/transmisión afecta el envió de los datos, incrementando la latencia extremo a extremo.

B. CLASIFICACIÓN DE PROTOCOLOS MAC

Existen protocolos enfocados a las características de las WSN. Estos protocolos se pueden catalogar en libres de contienda (Contention Free), basados en contienda (Contention-Based) e híbridos (Tabla 2). Los protocolos libres de contienda intentan organizar los nodos cercanos para que la comunicación ocurra de manera ordenada y evitar las colisiones, sobre-escucha y la escucha inactiva. Aunque a primera vista el uso de protocolos libres de contienda parece atractivo, estos también conllevan dificultades en especial, la complejidad de implementación, sobrecarga en la red de paquetes de control y la falta de escalabilidad.

En los protocolos basados en contienda cada estación trata de acceder al canal cuando tiene datos para emitir, y por tanto puede haber conflictos al usar el canal (colisiones). A estos protocolos también se les denomina de acceso alea-

torio (random protocols), puesto que el patrón que van a emplear las estaciones para intentar ocupar el medio es impredecible (desde el punto de vista de la capa de acceso al medio). Las ventajas de usar los esquemas basados en contienda son: (1) la baja complejidad de implementación, (2) la naturaleza ad-hoc y (3) la flexibilidad en las fluctuaciones de tráfico. Sin embargo, el principal problema de este tipo de protocolos es la energía gastada debido a colisiones, idle listening y el overhearing [Lan05].

Ambos tipos de protocolos (basado en contienda y libre de contienda) no resuelven por completo los problemas de las WSN, en algunos casos, se resuelven algunos de los problemas a expensas de introducir nuevos [Lan07]. Esto condujo a los investigadores a desarrollar protocolos híbridos. Estos protocolos suponen el uso de los esquemas utilizados en los protocolos basados en contienda y los libres de contienda. En general, el período de transmisión consta de dos fases: (1) la fase de reserva basada en contienda y (2) la fase de transmisión basada en libres de contienda.

- Libres de contienda.
  Los protocolos libres de contienda están generalmente basados en TDMA, FDMA o CDMA, aunque más frecuentemente se utiliza el TDMA. Estos protocolos tratan de reducir el consumo de energía coordinando los nodos y para así evitar las colisiones, el idle listening y el overhearing. Sin embargo, el problema de estos protocolos es la necesidad de un nodo central que regule el acceso al medio, la cantidad de mensajes que pueden llegar a ser intercambiados para sincronizarse y la falta de escalabilidad.

Tabla 2
Protocolos MAC para las WSN.

Basados en contienda	LIBRES DE CONTIENDA	Híbridos
B-MAC	TRAMA	Z-MAC
X-MAC	FLAMA	PMAC
Preamble sampling	PEDAMACS	Crankshaft
WiseMAC	LMAC	
CSMA-MPS	SMACS	
STEM	BMA	
RATE EST	BitMAC	
Sift	SS-TDMA	
SEESAW	PACT	
f-MAC		
CSMA/ARC		
PicoRadio		
LPL		

LMAC es un protocolo propuesto por [Hoe04] que utiliza un mecanismo distribuido de selección de ranura basado en la información a dos saltos. Cada nodo tiene un tamaño de ranura fija donde siempre transmite un encabezado y opcionalmente puede transmitir los datos. Dentro del encabezado se encuentran diferentes campos incluyendo el destino y el tamaño de los datos a transmitir. LMAC no utiliza un sistema de acuses de recibo para evitar el cambio de la radio (transmisión/recepción). Esta función se le deja a las capas superiores.

Para facilitar la entrada de nuevos nodos a la red, cada cabecera incluye un conjunto de bits detallando que ranuras están ocupadas por los vecinos a un salto (del nodo que envía). Sumando este conjunto de bits de todas las cabeceras en una trama un nuevo nodo puede conocer que ranuras están libres, selecciona una aleatoriamente y comienza a enviar cabeceras para reclamar la ranura. Si dos nodos intentan acceder a la misma ranura, existirá una colisión. Para resolverla, un nodo vecino enviara un *broadcast* con las ranuras envueltas en la colisión como parte de su cabecera, señalándole a los nuevos nodos que establezcan un tiempo de *back-off* y lo intenten de nuevo.

A pesar de ser un protocolo libre de contienda, LMAC elimina la necesidad de un nodo central que controle el acceso al medio. Sin embargo el mecanismo de selección de ranuras tiene un problema, el número de nodos en el vecindario (a dos saltos) no puede exceder el número de ranuras disponibles en una trama. Seleccionando un número grande de ranuras puede acarear una pérdida de ranuras (sobre aprovisionamiento) y *overhead*. Si se selecciona un número pequeño de ranuras se imposibilita la adición de nuevos nodos a la red.

- Basados en contienda.
  A diferencia de los protocolos libres de contienda, los protocolos basados en contienda no dividen el canal o preasignan el canal para cada nodo. Por el contrario, un solo canal es compartido por todos los nodos y el acceso a él es bajo demanda. Debido a esto, los protocolos basados en contienda son más escalables. De igual manera proveen gran flexibilidad para manejar diferentes densidades de nodos y cargas de tráfico. También ofrecen una gran adaptabilidad a los cambios dinámicos en la red, no necesitan formar *clusters* y no es necesaria la sincronización de los nodos [Ye04, Lan07]. La principal desventaja de los protocolos basados en contienda es el uso ineficiente de la energía, normalmente, tienen todas las causas de pérdida de energía mencionadas en el inciso *A*).

Dentro de estos protocolos podemos mencionar el protocolo f-MAC [Roe06], el cual no utiliza ni la detección de colisiones ni el tiempo de sincronización entre los nodos y está basado en la propuesta "*framelet*" (de ahí su nombre). En este tipo de propuestas se utiliza un tamaño de paquetes pequeño y fijo (llamados *framelet*). El mensaje es transmitido en varias ocasiones usando estos paquetes. En la Figura 4 obtenida de [Roe06] se muestra la transmisión de un mensaje usando la propuesta *framelet.* La duración de de la transmisión de un **framelet** está indicada por *d*. Cada transmisión consiste en r *framelets* y estos *framelets* son enviados con una frecuencia de $f=1 / t$.

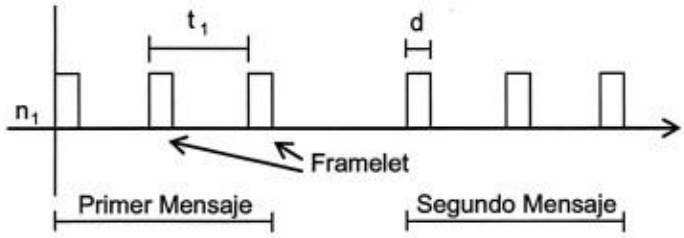


Figura 4. Transmisión Framelet.

Las políticas utilizadas por el protocolo f-MAC son las siguientes: (1) el tamaño del framelet d está definido por $d = \delta / 2$, donde δ es el tiempo de transmisión de un paquete. El número de framelets r por mensaje está definido por el número de nodos N dentro del rango de transmisión. Cada nodo ni debe de usar una periodicidad de transmisión específica fi para su framelet denotada por:

$f_i = 1 / t_i = 1 / (k_i \cdot \delta)$

Donde $k_i \in N^+$ debe de satisfacer la siguiente ecuación:

$$k_i . (r - 1) < M.C.M. (k_i, k_j) \quad \square k_i <, k_j \; 1 \le ij \le N$$

Después de transmitir un mensaje un nodo debe esperar un tiempo t' antes de poder enviar el siguiente mensaje. El tiempo t' se calcula como:

$$k_{max} = \max_{0 \le i \le N} (k_i)$$

$$t' = (k_{max} \cdot (r - 1) + 1) \cdot \delta$$

A partir de estas formulas, los autores deducen que un nodo ni necesita en el peor caso un tiempo Ti para transmitir un mensaje, donde:

$$T_i = (r - 1) \cdot t_i + t'$$

Dos características importantes de f-MAC son el retardo y que no utiliza eficientemente el canal. La utilización del canal se reduce para poder ofrecer garantías de retardo para el mensaje.

- Híbridos. Basados en contienda limitada
  Estos protocolos implementan funciones utilizadas en los protocolos basados en contienda y libres de contienda para tratar de obtener un mejor uso de los recursos de las WSN. Aquí se imponen alguna estructura donde a los nodos se les permita comunicarse pero mantienen la flexibilidad para adaptarse a las fluctuaciones de tráfico y los cambios en la topología.

El protocolo Zebra MAC (Z-MAC) diseñado por [Rhe05] combina el uso de los protocolos basados en contienda y libres de contienda intercambiándose entre ellos dinámicamente. El protocolo empieza con un algoritmo de asignación distribuido que toma en cuenta los nodos a dos saltos para obtener una distribución de ranuras libre de conflictos. En seguida, los nodos deben de pelear por el acceso cuando quieran enviar un mensaje. Cualquier nodo puede contender por cualquier ranura, sin embargo, los nodos previamente asignados tienen preferencias. Si un nodo detecta mucha perdida de paquetes, envía un *broadcast* notificando a los nodos que se cambien a un modo de contienda más alto. En este modo, los nodos ya no contienden por las ranuras previamente asignadas, lo que previene las colisiones. Después de 10 segundos, los nodos regresan a su operación normal [Lan07].

Los esquemas híbridos son escalables, soportan muchos tipos de tráfico, y sólo requiere un conocimiento local de la red. Sin embargo, requieren de sincronización y aunque mínimos, sigue existiendo *overhead* y retransmisión de paquetes. Según [Mis07], de las tres clasificaciones este esquema parece el más adecuado para dar soporte a las comunicaciones en tiempo real.

C. QoS A NIVEL DE MAC EN LAS WSN

Como se ha mencionado, existen muchos protocolos de nivel MAC para las WSN. Sin embargo, estas soluciones están más enfocadas a mejorar la conservación de

energía y no consideran cuestiones como *scheduling* y garantías en tiempo real en combinación con el manejo eficiente de la energía [Lan07, You04, Che04]. Según [Wan06] los requerimientos de QoS para la capa MAC son: rango de comunicación, tasa de transferencia, fiabilidad en la transmisión y eficiencia en el consumo de energía Las investigaciones que existen para proveer QoS en la capa MAC se pueden clasificar en [Aky07]:

- Políticas de acceso al canal. Como hemos visto, este es el primordial punto en la capa MAC para evitar las causas de pérdida de energía.
- *Scheduling*. Difiere de las redes tradicionales debido a que además de escoger la disciplinan de la cola que ayude a mantener los límites requeridos de la latencia, se necesita incorporar un control de energía/tasas de tráfico y considerar las condiciones de error en el canal.
- Control de error. Esta es una tarea difícil ya que el enlace inalámbrico es muy inestable. Existen dos mecanismos principales para combatir la falta de fiabilidad del enlace: *Foward Error Correction* (FEC) y *Automatic Repeat Request* (ARQ).

Dentro de las propuestas, encontramos el protocolo desarrollado por [Cac02] donde implementan una estructura celular como arquitectura de red (Figura 5). Este protocolo utiliza la multiplexación por división de frecuencias (FDM) entre las celdas adyacentes para permitir comunicaciones simultáneas en diferentes celdas. El *scheduler* utilizado es el *implicit earliest deadline first* (EDF) y es usado dentro de cada celda. Existe un *router* en el centro de cada celda, estos están equipados con dos transceptores para poder recibir y transmitir al mismo tiempo.

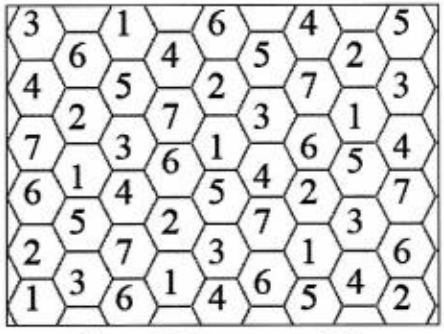


Figura 5. Estructura celular.

La comunicación *intracell* se realiza sin utilizar una estación base, mediante un esquema distribuido. La idea principal es replicar el *scheduled* EDF en cada nodo. Si los *schedulers* se mantienen idénticos, cada nodo conocerá quien tiene el mensaje con el tiempo para expirar más corto y este nodo tendrá el derecho a transmitir. Por otro lado, cuando un nodo escucha el canal, es capaz de conocer cuando un nodo ha terminado de transmitir y actualizar su *scheduled* para la siguiente ronda.

En la comunicación *intercell*, cada *router* transmite mensajes *intercell* usando el canal de la celda a la que pertenece y recibe mensajes usando el canal de la celda de la cual espera recibir. Los mensajes son ordenados de acuerdo al tiempo para expirar más corto por cada *router* y cada uno de ellos es capaz de alcanzar seis celdas vecinas a un salto.

Este protocolo provee un alto *throughput*, especialmente cuando existe una carga de trabajo considerable y baja latencia.

IV. Capa De Red

La capa de red es la responsable de de encaminar los paquetes, encontrando rutas que resulten eficientes energéticamente hablando para reducir el consumo global de energía y maximizar la vida operativa de la red. El encaminamiento en las WSN es muy desafiante debido a las características de este tipo de redes. En primer lugar, debido al número de nodos es casi imposible poder utilizar un esquema global de direcciones. Segundo, a diferencia de las comunicaciones típicas en las redes IP, donde la mayoría de las conexiones son punto a punto, las WSN tiene un flujo de datos de múltiples nodos a uno en particular (*sink*). También puede existir flujo en otras formas, pero suele ser un tráfico menor. Tercero, los nodos tienen muchas limitaciones tanto en energía, capacidad y almacenamiento. Cuarto, la mayoría de las aplicaciones espera que sus nodos estén estacionarios, salvo en casos donde solo algunos de los nodos tendrá movilidad. Quinto, las WSN son designadas por la aplicación, es decir, los requerimientos de diseño de la red cambian de acuerdo a la aplicación. Por último, la información recolectada por las WSN típicamente tiene redundancia, así que en algunas situaciones es posible la explotación de esta característica para mejorar la utilización del ancho de banda y la energía [Alk04].

Debido a estas diferencias con las redes tradicionales, los protocolos desarrollados hasta ahora no eran fácilmente adaptables a las WSN. Por tal motivo, muchos nuevos algoritmos han sido propuestos. Para lidiar con los problemas del consumo de energía, las técnicas de encaminamiento utilizadas emplean métodos

como agregado de datos, procesamiento en red, agrupamiento (*clustering*), diferente asignación de roles a los nodos y métodos *data-centric*.

Por otro lado, los protocolos de encaminamiento deben tener en consideración los siguientes factores de diseño [Alk04]:

- Heterogeneidad del enlace o nodo.
- Tolerancia a fallos.
- Escalabilidad.
- Medio de transmisión
- Conectividad.
- Cobertura.
- Calidad de servicio.

A. *Clasificación de los protocolos de red*

Existen varios métodos para clasificar los protocolos de encaminamiento en las WSN: basándose en (1) la estructura de la red, (2) la operación del protocolo y (3) la forma de encontrar la ruta al destino [Akk05]. El primer método clasifica según la estructura de la red en encaminamiento plano, jerárquico y basado en localización. En el plano, todos los nodos tienen asignados roles o funcionalidades iguales. En el jerárquico, los nodos tienen diferentes roles. En el basado en localización las posiciones de los nodos son explotadas para encaminar los datos. En el segundo método, los nodos son clasificados en multitrayectoria, basado en peticiones, negociación, QoS y coherente. En el tercer método, se clasifican en proactivo, reactivo e híbrido. Aquí, utilizaremos la clasificación basada en la estructura de red.

Tabla 3.
Clasificación de los protocolos de red.

Estructura de red	Operación del protocolo	Forma de encontrar la ruta
Plana	Negociación	Proactivo
Jerárquica	Multitrayectoria	Reactivo
Basada en localización	Peticiones	Híbrido
	QoS	
	Coherente	

- Protocolos de encaminamiento plano.
  Aquí, cada nodo tiene el mismo papel y todos los nodos colaboran juntos para realizar las tareas de observación. No es posible la utilización de una identidad global para cada nodo debido a la gran cantidad de ellos. Esto ha llevado al uso del encaminamiento céntrico de los datos, donde un nodo que necesita información envía peticiones a una región y espera el envió de esta por parte de los sensores localizados en dicha región. El uso de nombramiento basado en atributos es necesario para especificar las propiedades de la información requerida. Algunos de los primeros trabajos de este tipo fueron [Akk05]: protocolos de sensores para la información mediante negociación (SPIN) y difusión dirigida.

SPIN [Kul02] está diseñado para resolver las deficiencias del encaminamiento por inundación (*flooding*) y está basado en dos ideas básicas: (1) hacer eficiente el uso de recursos enviando peticiones con la descripción de la información necesaria antes de que los datos sean enviados, en lugar de solo enviar los datos y (2) evitar el envío de datos redundantes (solapar) y los mensajes duplicados transmitidos al mismo nodo (implosión).

El protocolo asigna un nombre a la información recolectada por el nodo (*meta-data*) y realiza la negociación usando tres tipos de mensajes, *ADV, REQ* y *DATA*. El protocolo inicia cuando se adquiere datos nuevos para compartir, entonces se envía un mensaje *ADV* el cual contiene el *meta-data*. Si un vecino está interesado en la información, contesta con un mensaje *REQ*. Después de esta negociación se envían los datos. La Figura 6 obtenida de [Akk05] muestra este proceso de negociación.

Existen muchos protocolos derivados del SPIN, entre ellos SPIN-BC, SPIN-PP, SPIN-EC, etc. Las ventajas de este protocolo radican en la disminución del consumo de energía, y la reducción de información redundante. Además, los cambios topológicos están confinados debido a que los nodos solo necesitan conocer a los vecinos a un solo salto. Sin embargo, no se garantiza la entrega de información.

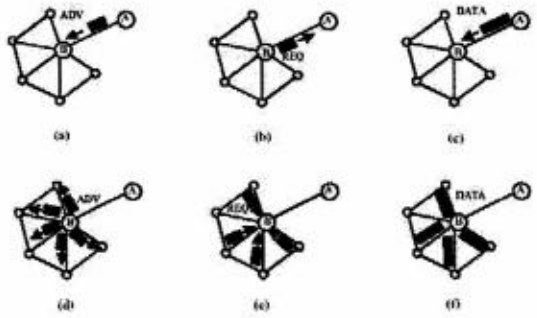


Figura 6. Proceso de negociación del protocolo SPIN.

Directed Diffusion [Int03] es un protocolo de encaminamiento céntrico que propone una técnica de agregación de datos, el cual se utiliza para resolver los problemas del solapamiento y la implosión. Este método es parecido a un multicast de árbol inverso. En la Figura 7, obtenida de [Aky02], se muestra un ejemplo de agregación de datos. El sink hace una petición de reporte, la información proveniente de los nodos es agregada si tiene los mismos atributos cuando alcanzan el mismo nodo en la trayectoria hacia el sink. Esto es aprovechado por el protocolo para eliminar la redundancia y minimizar el número de retransmisiones. Sin embargo existen diferentes factores que afectan la agregación de datos llevada a cabo en el protocolo, entre ellas están: el número de nodos, la posición y la topología.

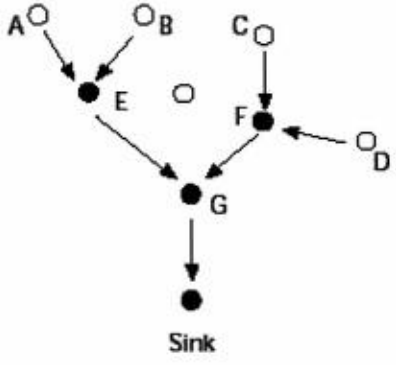


Figura 7. Ejemplo de agregado de datos

En *Directed Diffusion*, el *sink* hace una solicitud de información enviando *broadcast* con los *intereses*, los cuales son listas de pares, con los atributos y los valores que describen una tarea. Estos *intereses* se propagan salto por salto y son enviados por cada nodo a sus vecinos. A medida que el *interés* es propagado, se crea un enlace de respuesta (*gradiente*) para enviar la información que satisface la petición. Utilizando los *intereses* y los *gradientes*, se establecen las trayectorias entre el *sink* y los nodos. Debido a que se crean diferentes trayectorias, una de ellas es seleccionada de refuerzo. La Figura 8 tomada de [Akk05] muestra las diferentes fases del protocolo.

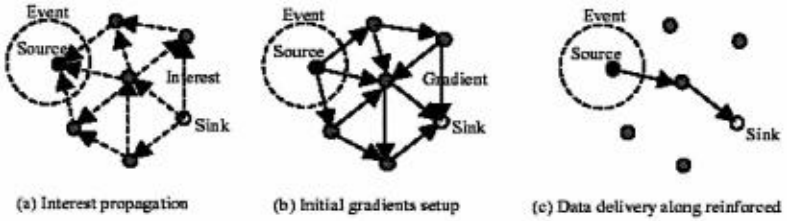


Figura 8. Fases del protocolo *Directed Diffusion*.

Existen muchos protocolos derivados o similares a *directed diffusion,* entre ellos:

- encaminamiento por rumor, principalmente diseñado para aplicaciones donde el encaminamiento geográfico no es factible.
- encaminamiento basado en gradientes en el cual se trata de obtener un distribución balanceada del tráfico de la red, incrementando así su vida útil.

Además de estos protocolos existen más basados en encaminamiento céntrico, algunos de ellos son: COUGAR, ACQUIRE, MCFA, entre otros.

- Protocolos de encaminamiento jerárquico.

Los protocolos jerárquicos se basan en la creación de agrupaciones llamados *clusters* y en la asignación de un nodo, generalmente el de mayores recursos energéticos, para realizar tareas especiales. A este nodo se le determina como CH, por esto, también son conocidos como protocolos de encaminamiento basado en agrupación.

Esta técnica le proporciona ventajas relacionadas con la escalabilidad y la eficiencia de comunicación. Además, la comunicación se realiza con multisalto y se realiza el agregado de datos y la fusión. El encaminamiento jerárquico se realiza en dos capas [Alk04], la primera es usada para la formación del *cluster* y la segunda para el proceso de encaminamiento.

El protocolo de agrupación jerárquica adaptable a baja energía (LEACH), es una de las primeras propuestas [Akk05] de este tipo. La operación de este protocolo se realiza por rondas. En la primera se decide la formación de *cluster* y después la transmisión de la información a la estación base.

Cuando el *cluster* comienza a ser creado, se escogen los CH. Cada nodo puede ser un CH, la decisión de serlo o no dependerá de él. Este, escoge un número entre 0 y 1, si el número es menor a un umbral T(n), el nodo será CH de esa ronda. El umbral está determinado por:

$$T(n)=\begin{cases}\dfrac{P}{1-P*(r \bmod \frac{1}{P})} & \text{si } n \in G \\ 0 & \text{de lo contrario}\end{cases}$$

donde P es el porcentaje deseado de CH, r es la ronda actual, y G es el conjunto de nodos que no han sido CH en las anteriores 1/p rondas. Después cada CH elegido envía un mensaje a los demás nodos. Dependiendo de la intensidad de este mensaje, los nodos que no son CH deciden a qué *cluster* pertenecer y reenvían esa decisión al CH al cual pertenece. Este intercambio de información se realiza con el protocolo CSMA. Después, los CH crean un esquema TDMA para el realizar la transmisión de información. Este ciclo se repite continuamente para balancear la carga de energía de cada nodo. En [Hei00] se especifica más a fondo el funcionamiento de LEACH.

Aunque LEACH mejora el tiempo de vida de la red, tiene varios supuestos [Al04] que necesitan ser revisados, entre ellos: el hecho de que un nodo puede transmitir con suficiente potencia para alcanzar la estación base, cada nodo tiene la capacidad computacional para soportar protocolos MAC, que los nodos siempre tienen información para transmitir y tienen la misma capacidad de energía en cada ronda de elección, entre otras. Además, no está dicho como el numero de CH predeterminado (P) pueda ser distribuido en toda la red. Existen varios protocolos basados en la idea de LEACH que implementan mejoras como TEEN, APTEEN, PEGASIS, entre otros.

- Protocolos de encaminamiento basado en localización.

Estos protocolos se basan en el conocimiento de la localización para el encaminamiento de la información. Cuando se conoce la ubicación de los nodos, se puede reducir el número de retransmisiones enviando una petición solo a la región donde se encuentren los nodos. Para conocer la localización de un nodo, se pueden intercambiar información entre vecinos ó utilizar un dispositivo, como el GPS, para obtener su emplazamiento. Sin embargo deben tenerse en cuenta las limitaciones de nodo en cuanto a energía.

GEAR es un protocolo diseñado para balancear la energía, motivado por las aplicaciones de las WSN [Yu01]. Utiliza la idea del protocolo de difusión dirigida, pero en lugar de enviar los intereses a toda la red, lo envía a cierta región y los paquetes se distribuyen dentro de ella. Cada nodo tiene un costo estimado y un costo aprendido para llegar a su destino. El costo estimado es la combinación de la energía residual y la distancia al destino. El costo aprendido es un refinamiento del costo estimado que, cuenta para el encaminamiento por los agujeros (cuando no hay vecinos cerca de la región destino) en la red [Akk05, Alk04]. El protocolo GEAR consta de dos pasos. En primer lugar, se envían los paquetes a la región. Si existe un nodo vecino que esté más cerca de la región destino, este es seleccionado como el siguiente salto. Si los vecinos están más lejos de la región que el nodo emisor, se envía la información a un nodo basándose en el costo aprendido. El segundo paso es diseminar los paquetes dentro de la región, lo cual se realiza de dos formas, utilizando el algoritmo de reenvió geográfico recurrente en la mayoría de los casos, o la inundación restricta cuando existe poca densidad de nodos.

- QoS a nivel de Red.

Los protocolo de encaminamiento deben proveer un balance entre consumo de energía y calidad en los datos. Principalmente, se deben satisfacer métricas de QoS como: retardo, ancho de banda, entre otras. Actualmente existen varios proyectos que están abordando los requerimientos de QoS. El trabajo publicado hasta ahora en este contexto puede ser catalogado en dos categorías. La primera se centra en la compensación entre energía y retardo. La segunda categoría trata de extender el tráfico para estimular el ancho de banda y reducir el retardo [You04].

El protocolo SAR (Sequential Assignment Routing) propuesto por [Soh00] es el primer protocolo para las WSN que emplea la noción de QoS. SAR es un protocolo multitrayectoria y toma la decisión de encaminamiento basado en tres

factores: energía, QoS de cada trayectoria y el nivel de prioridad de cada paquete. El protocolo crea las trayectorias en árbol, comenzando desde el nodo fuente hasta los nodos destino. Solo una de las trayectorias creadas se utiliza dejando las demás de respaldo. Si existe una falla, el procedimiento para crear la ruta se vuele a ejecutar. Aunque el protocolo es tolerante a fallos y se recupera fácilmente de los fallos, sufre de *overhead* para poder mantener las tablas y estados en cada sensor. SAR tampoco usa las rutas redundantes para dividir la carga y utilizar eficientemente el uso del ancho de banda.

Otro protocolo para dar soporte a la QoS es el propuesto por [Akk03]. Para calcular el costo del enlace, el protocolo utiliza una función que captura información de los nodos como: la cantidad de energía utilizada para transmitir, la tasa de error, la energía que le queda al nodo, entre otros parámetros de la comunicación. Para dar soporte al tráfico en tiempo real y al tráfico común, el protocolo emplea una cola basada en clases. El ancho de banda r, es definido como un valor inicial determinado por el *gateway* y representa la cantidad de ancho de banda disponible para el tráfico en tiempo real y el común en un enlace en particular. Como consecuencia, el *throughput* del tráfico normal no se ve disminuido ajustando el valor de r. El protocolo encuentra una lista de rutas con el menor costo utilizando un algoritmo extendido de Dijkstra. Una vez tiene esta lista, toma la ruta que cumple con el retardo extremo a extremo requerido. Aunque el protocolo funciona bien, debido a que el mismo valor de r es inicialmente indicado para todos los nodos, no permite un ajuste flexible del ancho de banda compartido por diferentes enlaces [You04].

SPEED [Tia03] es otro protocolo para las WSN que provee garantías extremo a extremo. El protocolo requiere que todos los nodos mantengan información acerca de sus vecinos y utiliza un método denominado *geographic forwarding* para encontrar las trayectorias. El protocolo trata de garantizar cierta velocidad para cada paquete en la red y con esto, cada estación puede estimar el retardo extremo a extremo de los paquetes dividiendo la distancia hacia el *sink* entre la velocidad del paquete. SPEED también puede evitar la congestión cuando la red está sobre cargada.

El modulo de encaminamiento de SPEED (SNFG) trabaja conjuntamente con otros 4 módulos. El mecanismo de intercambio de *beacons* recolecta información acerca de los nodos y su localización. La estimación del retardo en cada nodo es realizado calculando el tiempo que pasa para recibir un ACK después de enviarle un paquete a un vecino. Con estos valores de retardo el SNFG selecciona el nodo que cumple con los requerimientos de velocidad. Si el nodo no puede ser encontrado, se examina el *relay ratio*. El *relay ratio* es calculado por el modulo "*Neighborhood Feedback Loop*". Su cálculo se realiza observando la proporción

de pérdidas de los vecinos del nodo. Si la proporción es menor que un numero generado aleatoriamente entre 0 y 1, el paquete es descartado.

Por último, el modulo "*Backpressure-rerouting*" es utilizado para prevenir "huecos" cuando un nodo falla al tratar de encontrar el siguiente salto. Además elimina la congestión enviando los mensajes a la fuente para que encuentren nuevas rutas.

Según [Alk04, You04] SPEED funciona mejor que DSR y AODV en términos de retardo extremo a extremo y tasa de perdidas. Además, proporciona un menor consumo de energía debido a la simplicidad del algoritmo. Sin embargo, no considera ninguna métrica referente al consumo de energía en el encaminamiento.

V. Objetivos del Presente Proyecto

A. Objetivo general

El objetivo general de este proyecto es realizar el diseño de una arquitectura para las redes de sensores inalámbricas o WSN. Dicha arquitectura deberá de cumplir con los requerimientos de calidad de servicio para los tráficos tanto multimedia como escalar y ofrecer mecanismos de seguridad. La arquitectura que se plantea deberá estar basada en una optimización multicapa.

B. Objetivos particulares

Definir los compromisos entre calidad de servicio y el ahorro de energía, desarrollando un estudio de estos parámetros y su impacto en el tráfico a cursar.

Diseñar el modelo multicapa para hacer más eficientes los mecanismos de encaminamiento y acceso al medio.

Análisis de los problemas de seguridad que pueden presentarse y que degradarían el servicio prestado al usuario.

Definir y proponer mecanismos de seguridad.

Desarrollar los mecanismos para proveer calidad de servicio tanto a nivel MAC como de Red, considerando todos los factores que afectan tanto al consumo de energía como al tráfico.

C. *TRABAJO REALIZADO*

Durante los últimos meses se ha realizado una amplia investigación bibliográfica en torno al tema de las redes de sensores, en la cual se han tocado diversos aspectos de las mismas.

Esta tipo de redes son una tecnología recientemente emergida, la cual se encuentra en desarrollo. Esta tecnología destaca por sus múltiples y variadas aplicaciones, así como por las limitaciones de los dispositivos que las componen. Como primer punto, se ha realizado una revisión bibliográfica en busca de obtener unos conocimientos que sirvieran de base. Esta revisión fue referente a la situación actual de las WSN, sus características, requerimientos y aplicaciones, entre otros. De esta investigación se desprende la realización del reporte interno titulado "Wireless Sensor Networks" [Sal06].

En dicho reporte se mencionan las limitaciones del nodo, que pese a los actuales adelantos hechos por los principales fabricantes sigue sufriendo de falta de procesamiento, capacidad de almacenamiento y principalmente de energía comparados con los dispositivos de otras tecnologías. En la Tabla 4 se muestra un resumen de dichas características [Beu06].

Se destacaron también los factores de diseño en las WSN, dentro de los cuales podemos destacar los siguientes: Tolerancia a fallos, escalabilidad, hardware y costos de producción, despliegue y topología, heterogeneidad, autoconfiguración, optimización, adaptabilidad, seguridad y consumo de energía. Cada una de estos factores afecta de diferente modo al desarrollo de los protocolos que ha de ser aplicados en estas redes.

Por otro lado, también podemos mencionar la falta de un estándar para las WSN, aunque actualmente el 802.15.4 es el resultado de un primer esfuerzo por desarrollar un estándar para las WSN y estable ciertos parámetros para las capas tanto física como de enlace. En dicho reporte aparte de lo ya mencionado, también se muestran características sobre las diferentes capas tomando en cuenta un modelo dispuesto por Akyildiz en [Aky02].

Por otro lado, también se estudiaron varias herramientas de simulación para realizar las simulaciones de futuros trabajos. Entre las herramientas que se estudiaron, se encuentra: NCTUns [Wan06], una herramienta diseñada por el Prof. S.Y. Wang capaz de simular distintos protocolos usados en redes IP inalámbricas y cableadas. Dentro de sus ventajas destacan: la posibilidad de usarlo como emulador de terminales, la utilización de la pila de protocolos de Linux, cualquier programa desarrollado en Unix puede ser usado para generar tráfico, interfaces fácilmente utilizables, entre otras. Debido a su relativa "temprana edad" no existen tantos protocolos diseñados como en otros simuladores como NS-2. Además, no

se encuentran muchos estudios que realicen sus simulaciones en él. Hablando directamente de las WSN, esta herramienta no implementa mecanismos para simular el consumo de energía de un nodo, algo importante en un estudio de WSN ya que es esta su principal limitación. Es necesario el diseño de nodos que asemejen a un nodo de WSN. Por otra parte, el intentar modificar parámetros que no estén accesibles mediante la interfaz GUI implica recompilar el programa.

Tabla 4.
Características de los nodos disponibles para las WSN

	BTnode rev3	Mica2	Mica2Dot	Tmote Sky	Imote
Microcontroller	ATmega128l	ATmega128l	ATmega128l	MSP430F	ARM7
Architecture	8-Bit	8-Bit	8-Bit	16-Bit	32-Bit
Speed	7.3728 MHz	7.3728 MHz	4 MHz	8 MHz	12 MHz
Program Memory	128 kB	128 kB	128 kB	48 kB	512 kB
Data Memory	64 kB	4 kB	4 kB	10 kB	11 kB
Storage Memory	180 kB SRAM	512 kB	512 kB	1024 kB	–
External IO	40	51	18	16	30
On-Board Sensors	1	2	2	5	–
UI Components	4 LEDs	3 LEDs	1 LED	3 LEDs, 1 Button	1 LED
Size	1890 mm2	1856 mm2	492 mm2	2621 mm2	900 mm2

J-Sim [Tya05] es un simulador diseñado por la Universidad de Ohio basado en una arquitectura de componentes autónomos. Para la simulación de redes, el simulador tiene un modelo de red como parte de la arquitectura autónoma. El modelo define la estructura genérica de un nodo y los componentes de red pudiendo utilizar ambos como base para implementar protocolos en varias capas. Este simulador fue desarrollado en Java y contiene un paquete adicional desarrollado para las WSN, lo que lo convierte en una buena plataforma para el desarrollo de protocolos de estas redes.

Scalev [Cru07] es una herramienta desarrollada por el grupo de trabajo de Servicios Telemáticos del Departamento de Ingeniería Telemática de la Universidad Politécnica de Cataluña. Es una herramienta de fácil uso, ha sido diseñada en Java lo que le permite ser multiplataforma. Dentro de sus principales características destacan:

- caracterización de diferentes tráficos y sus características,
- diferenciación de servicios,
- diferentes tipos de schedulers.
- simulaciones simples y con barrido
- los resultados son fácilmente analizables.

También en el entorno de las WSN se estudió la integración entre WSN y redes TCP/IP. Aquí se observaron las principales formas de interconexión de estas tecnologías, donde destacan las basadas en la utilización de un Gateway generalmente a nivel de aplicación o en la superposición de las capas. Las arquitecturas basadas en la primera opción se subdividen en:

- uso del Gateway como simple repetidor,
- uso del Gateway como un Front-end.

El funcionamiento básico de estas propuestas es el siguiente: los nodos recolectan la información del fenómeno a estudiar y envían los datos a un Gateway. Este dispositivo se encarga de la interconexión entre redes. Además debido a que el Gateway es un dispositivo mucho más robusto que un nodo, este puede realizar funciones extra como agregado y compresión de los datos o la implementación de esquemas de políticas de seguridad. El Gateway reenvía la información a una base de datos. Esta base de datos se puede situar local o remotamente. El usuario puede acceder a la información mediante peticiones SQL, XML o a través de un servidor Web.

Este esquema permite escoger protocolos específicos para las redes de sensores. Sin embargo no existe interacción directa con los nodos y se crea un solo punto de falla. Una posible solución a estos problemas es emplear múltiples dispositivos de acceso y algún algoritmo de balanceo de carga.

Dentro de las soluciones de superposición se encuentran diferentes propuestas. Sin embargo el empleo de estos esquemas representa la modificación de la pila de protocolos de una u otra red dependiendo del tipo de superposición que se realice. Otros trabajos mencionan la posibilidad de usar TCP/IP directamente en los nodos [Dun04, May06]. Sin embargo esto aún no es viable debido principalmente a las limitaciones del nodo y diferencias entre arquitecturas.

REFERENCIAS

[Lan07] Langendoen, Koen. "Medium Access Control in Wireless Sensor Networks", Medium Access Control in Wireless Networks. Ed. H. Wu and Y. Pan, Nova Science Publishers, 2007.

[Ye04] Ye, W. and Heidemann, J. "Medium Access Control in Wireless Sensor Networks", Wireless Sensor Networks. Ed. C. S. Raghavendra, K. M. Sivalingam and T. Znati, Kluwer Academic Publishers, Norwell, MA, 2004, Pages 73-91.

[Lan05] Langendoen, Koen. "Energy-efficient Medium Access Control", Embedded Systems Handbook. Ed. R. Zurawski, CRC Press, 2005, Pages 34.1-34.29.

[Yu01] Yu, Yan et al, "Geographical and Energy Aware Routing: A Recursive Data Dissemination Protocol for Wireless Sensor Networks ", Technical Report UCLA/CSD-TR-010023, Computer Science Department, UCLA, Los Angeles, May 2001.

[Che04] Chen, D. and Varshney P. K., "QoS Support in Wireless Sensor Networks: A Survey." In Proceedings of the 2004 International Conference on Wireless Networks (ICWN 2004), Las Vegas, Nevada, USA, June 21-24 2004.

[Wan06] Wang, Yuanli, et al. "Requirements of Quality of Service in Wireless Sensor Network." In Proceedings of the International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies (ICNICONSMCL'06), IEEE Computer Society, 2006, Pages 116-121.

[You04] Younis, Mohamed et al. "On Handling QoS Traffic in Wireless Sensor Networks", In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04), IEEE Computer Society, Hawaii, 2004, Pages 292-301.

[Hoe04] L.F.W. van Hoesel and P.J.M Havinga, "A lightweight medium access protocol (LMAC) for wireless sensor networks: Reducing Preamble Transmissions and Transceiver State Switches", In Proceedings of the First International Conference on Networked Sensing Systems (INSS'04), Tokyo, 2004.

[Roe06] Roedig, Utz et al., "f-Mac: A deterministic media access control protocol without time synchronization", In Proceedings of the 3rd European Workshop on Sensor Networks (EWSN 2006), Volume 3868 of Lecture Notes in Computer Science, Springer, Berlin, February 2006. Pages 276–291.

[Rhe05] Rhee, Injong et al., "Z-MAC: a hybrid MAC for wireless sensor networks" In Proceedings of the 3rd international conference on Embedded networked sensor systems (SenSys'05), 2005, Pages 276–291.

[Cac02] Caccamo, Marco et al., "An Implicit Prioritized Access Protocol for Wireless Sensor Networks", In Proceedings of the 23rd I EEE Real-Time Systems Symposium (RTSS'02), Austin, Tx, USA, December 2002, Pages 39.

[Hei00] Heinzelman, W.R. et al., "Energy-efficient communication protocol for wireless microsensor networks", In Proceedings of the 33rd Annual Hawaii International Conference on System Sciences (HICSS'00), Hawaii, 4-7 Jan. 2000, Pages 10.

[Akk03] Akkaya, Kemal and Younis, Mohamed, "An Energy-Aware QoS Routing Protocol for Wireless Sensor Networks", In Proceedings of the 23rd International Conference on Distributed Computing Systems, IEEE Computer Society, 2003, Pages 710.

[Tia03] Tian, He et al., "SPEED: a stateless protocol for real-time communication in sensor networks", In Proceedings 23rd International Conference on Distributed Computing Systems, 2003, Pages 46-55.

[Aky02] Akyildiz, Ian F. et al., "Wireless Sensor Networks: a survey" Computers Networks, Volume 38, Issue 4, Pages 393-422. 2002

[Kre07] Kredo, K. and Mohapatra, P., "Medium access control in wireless sensor networks" Computer Networks, Volume 51, Issue 4, Pages 961-994, 14 March 2007.

[Gur05] Gurses, E. and Akan, O. B., "Multimedia Communication in Wireless Sensor Networks" Annals of Telecommunications, Volume 60, No. 7-8, Pages 799-827, July-August 2005.

[Dem06] Demirkol, Ilker et al., "MAC protocols for wireless sensor networks: a survey" Communications Magazine, Volume 44, Issue 4, Pages 115-121, April 2006.

[Mis07] Misra, Satyajayant et al., "A survey of multimedia streaming in wireless sensor networks", submitted for publication.

[Aky07] Akyildiz, Ian F. et al., "A survey on wireless multimedia sensor networks" Computer Networks, Volume 51, No. 4, Pages 921-960, 2007.

[Alk04] Al-Karaki, J.N. and Kamal, A.E., "Routing techniques in wireless sensor networks: a survey", Wireless Communications, IEEE [see also IEEE Personal Communications], Volume 11, Issue 6, Pages 6-28, Dec. 2004.

[Akk05] Akkaya, K. and Kamal, A.E., "A survey on routing protocols for wireless sensor networks", Ad-Hoc Networks, Volume 3, No. 3, Pages 325-349, May 2005.

[Soh00] Sohrabi, K. et al, "Protocols for Self-Organization of a Wireless Sensor Network," IEEE Personal Communications, Volume 7, Issue 5, Pages 16-27, 2000.

[Kul02] Kulik, Joanna et al, "Negotiation-based protocols for disseminating information in wireless sensor networks", Wireless Networks, Volume 8, No. 2/3, Pages 169-185, 2002.

[Int03] Intanagonwiwat, Chalermek et al, "Directed diffusion for wireless sensor networking", IEEE/ACM Transactions Networking, Volume 11, No. 1, Pages 2-16, 2003.

[Wan06] Wang S.Y., et al, "The protocol Developer Manual for the NCTUns 3.0 Network Simulator and Emulator", March 2006. Disponible en: http://nsl10.csie.nctu.edu.tw

[Dun04] Dunkels, J., et al, "Connecting wireless sensornets with TCP/IP Networks", In Proceedings of the Second International Conference on Wired/Wireless Internet Communications (WWIC'04), Frankfurt/Oder, Germany, 2004.

[May06] Mayer, K. and Fritsche, W., "IP-enabled wireless sensor networks and their integration into the internet", In Proceedings of the First International Conference on Integrated internet Ad Hoc and Sensor Networks (InterSense'06), Nice, France, 30-31 May 2006.

[Beu06] Beutel J., "Metrics for Sensor Network Platforms", In Proceeding of the ACM Workshop on Real-World Wireless Sensor Networks (REALWSN'06), ACM Press, New York, June 2006, Pages 26-30.

[Sal06] Saldivar Vicente, De la Cruz, Luis J., "Wireless Sensor Networks", Reporte interno RI-2006 Sal. , ENTEL 2006.

[Tya05] Tyan, Hung-ying, "J-Sim" [online], 2005. Available from: http://www.j-sim.org/

[Cru07] De la Cruz, Luis J., "Scalev Lite v2.3, Manual de usuario" [online], 2007. Available from: http://globus.upc.es/~ljcruz/

propósitos de sensado y trabajos de investigación que fueron llevados a cabo por diversos grupos obteniendo resultados sorprendentes, dando pie al nacimiento de una comunidad científica abocada a los sensores de fibra óptica.

Ciertamente, el campo de los sensores de fibra óptica se ha desarrollado enormemente desde sus primeros pasos y, en el presente, existen áreas de aplicación real donde compite con ventajas sobre los sensores tradicionales (principalmente electrónicos).

CARACTERÍSTICAS

Existe una serie de ventajas con esta tecnología, cuyos argumentos son conocidos y esto la hace ser bastante atractiva para su implementación. Entre las ventajas, se incluyen:

(1) Las fibras ópticas exhiben baja atenuación y gran banda ancha, haciendo posible grandes capacidades de datos a kilómetros de distancia. Consecuentemente, las fuentes ópticas y detectores pueden ser instalados alejadamente del ambiente en la cual la medición se lleva a cabo.

(2) Existe una amplia selección de fuentes luminosas, detectores ópticos, y en general, una gama importante de dispositivos ópticos adecuados para cada aplicación específica.

(3) Los sensores de fibra óptica están hechos de materiales dieléctricos que son químicamente inertes. Además, los sensores de fibra óptica son tanto inmunes a las interferencias electromagnéticas como a no generarlas. Es posible, por ejemplo, utilizar el mismo conducto donde pasa un cable eléctrico instalado, posibilitando la hibridización tanto de señales eléctricas como electrónicas con sensores de fibra óptica sin presentarse problemas de interferencia. Esto abre la opción para la aplicación de esta tecnología bajo circunstancias ásperas y hostiles donde los sensores eléctricos pudiesen fallar o requerir de una especial y costosa protección.

(4) Sus pequeñas dimensiones, geometrías simples y la naturaleza de transducción hacen posible la implementación de sistemas compactos y ligeros. Las consideraciones de dimensión son importantes cuando se trata de integrar dentro de unidades optoelectrónicas para alcanzar grandes unidades en escala.

(5) Las fibras ópticas son biocompatibles. Además, esta tecnología es aplicable en el desarrollo de instrumental biomédico.

(6) Los sensores de fibra óptica pueden fácilmente ser fijados dentro de materiales, debido a sus pequeñas dimensiones y cortes transversales uniformes.
(7) Los sensores de fibra óptica pueden ser empelados para monitorizar una amplia gama de parámetros incluyendo tensión, temperatura, presión, vibración, aceleración, emisión acústica, humedad, especies biomédicas, gas, distancia, posicionamiento, desplazamiento, nivel líquido, doblamiento, torsión, radiación, rotación, parámetros médicos, etc.
(8) Relativamente pueden resistir a elevadas temperaturas gracias al alto punto de fusión de la fibra óptica, y utilizando la adecuada protección pueden trabajar a muy altas temperaturas, como ejemplo de ellos se mencionan los sensores basados en la radiación de un cuerpo negro.
(9) A diferencia de sus contrapartes electrónicas, los sensores de fibra óptica pueden ser multiplexados, compartiendo la misma fuente y/o el detector.
(10) Existen varios esquemas de implementación, habilitándose la posibilidad de una configuración puntual, multiplexada, distribuida y cuasi-distribuida. Consecuentemente, un número de sensores similares como no similares pueden ser adjuntados a lo largo de una fibra óptica. Sin embargo, el sensado distribuido a lo largo de una fibra también puede llevarse a cabo. Esto es una de las ventajas más importantes sobre los sensores convencionales.
(11) Para finalizar, principalmente con las técnicas interferométricas, la sensibilidad, el rango dinámico y la resolución son características potencialmente mayores que las que presentan los sensores convencionales.

También los sensores de fibra óptica presentan algunas desventajas concernientes a los siguientes aspectos:

(1) El costo de los sensores de fibra óptica es superior comparado con los sensores convencionales. Esto es debido principalmente al costo de los componentes que están hechos en escasas cantidades. Algunas reducciones a los costos podrían derivarse si se fabricaran masivamente y se utilizasen las 3 ventanas de longitudes de onda para las telecomunicaciones (primera ventana ~850 nm; segunda ventana ~1310nm y tercera ventana ~1550nm).
(2) La sensibilidad es un concepto importante. La interpretación de los datos por parte de los sensores podría no ser del todo precisa ya que la salida se puede ver influenciada por otros parámetros exteriores que modifiquen la medición de interés, causando en el sistema sensor el arrojo

de datos erróneos. Un ejemplo de ello se encuentra en los procesos de medición tanto físicos como químicos, estos procesos se ven afectados por el parámetro de temperatura cuya influencia incide en la sensibilidad del sensor.

(3) El embalaje es un enorme desafío que desafortunadamente es compartido por todas estas tecnologías. Se sabe que en el dominio de las telecomunicaciones el embalaje de la optoelectrónica ejerce el mayor porcentaje de costos en componentes tecnológicos. La alineación precisa de la fibra con las fuentes y detectores ofrece un permanente reto a la industria.

CLASIFICACIÓN

Varios esquemas de clasificación han sido propuestos para los sensores de fibra óptica [5]. Existen diferentes puntos de vista basados, por ejemplo, en la cantidad física a ser medida, la determinación espacial de una variable, la naturaleza de transducción, las propiedades de radiación, sistemas de detección, tecnología sensora, técnicas de modulación. Los cuatro esquemas más representativos se muestran en la figura 2.

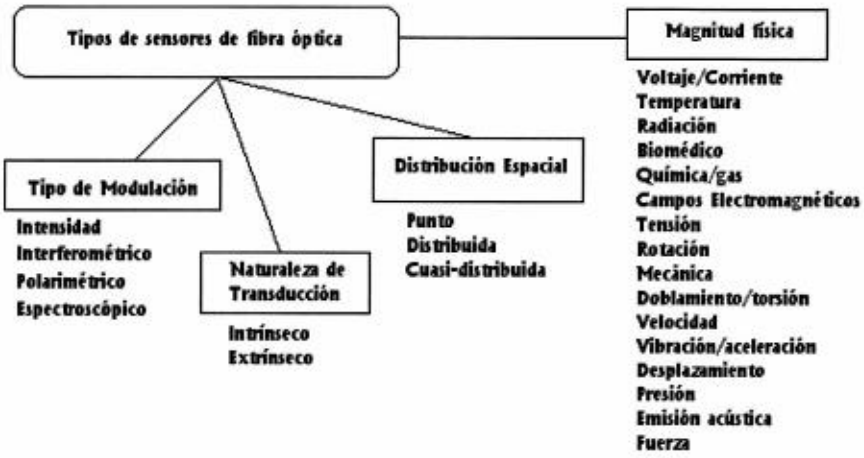


Figura 2. "*Tipos de sensores de fibra óptica categorizados de acuerdo al tipo de modulación, naturaleza de transducción, distribución espacial y magnitud física*"

De todas las clasificaciones existentes, la subdivisión más común para los sensores de fibra óptica por su naturaleza de transducción es la clasificatoria en extrínseco e intrínseco. Esta definición fue propuesta por Udd [6]. Los sensores

intrínsecos se caracterizan porque el sensado tiene lugar en la fibra misma. Esto implica la modificación ya sea de la transmisión o reflexión, dependiendo de la magnitud a ser medida. En este tipo de sensor, la radiación luminosa permanece total o parcialmente dentro de la fibra óptica siendo modulada por el efecto del parámetro a medir, por ejemplo el interferómetro Fabry-Perot ó Michelson y el sensado a través de redes de difracción Bragg. En los sensores extrínsecos la interacción entre la luz y la medición se produce fuera de la fibra misma. Un ejemplo de este tipo, son los sensores ópticos de apertura numérica basados en espejos flexibles al final de la fibra. Éstos se han usado para medir vibraciones y desplazamientos. Con aplicaciones en indicadores de cierre de puertas así como también para niveles de vibración en maquinaria.

Figura 3. *"Configuración general de la clasificación de sensores de fibra óptica, (a)Extrínseco, (b) Intrínseco"*

MECANISMOS DE SENSADO

Toda la gama de sensores presentan características que los agrupan en extrínsecos e intrínsecos, independientemente de ello, cada sensor emplea una técnica de modulación que hace variar la luz ya sea en amplitud, fase, polarización ó longitud de onda. Cada sensor utiliza un mecanismo de sensado de los cuales podemos desprender algunos ejemplos.

MANIPULACIÓN DE LA AMPLITUD

En este tipo de sensor, la magnitud a ser medida modula la intensidad óptica transmitida a través de la fibra. Algunos sensores basados en este rubro emplean un mecanismo de sensado cuyo principio es de la reflexión interna total. Un ejemplo

de este mecanismo de sensado es utilizado en los sensores para medición de nivel líquido.

Otra situación implica un sensor basado en la modulación de intesidad a través del mecanismo de sensado de apertura numérica también para medir nivel líquido e índice de refracción. Los sensores modulados en intensidad cuentan con una amplia historia y muchos de ellos ya han sido desarrollados. Aquí solo se han mencionado dos ejemplos de ellos basándose en mecanismos de sensado como la reflexión interna total y la apertura numérica. Sin embargo, se han reportado otros mecanismos de sensado en trabajos de investigación, de los cuales podemos mencionar los sensores basados en campo evanescente, sensores basados en micro y macro doblamiento, sensores de temperatura basados en la radiación del cuerpo negro.

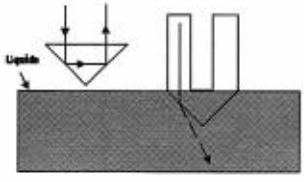


Fig. 4. "Sensor de nivel líquido basado en la reflexión interna total, detecta la presencia o ausencia de líquido con la presencia o ausencia de un retorno de señal"

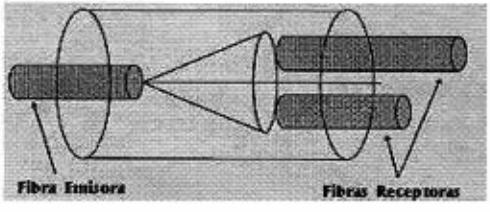


Fig. 5 "Sensor basado en la apertura numérica dirigido a la medición del índice de refracción. El medio varía en su índice de refracción con la presencia de líquido"

MANIPULACIÓN DE FASE

La variable a medir produce un cambio en la fase de la señal óptica (rotación, aceleración, emisión acústica, campo eléctrico, etc). Estos sensores basados en este tipo de modulación ofrecen una alta sensibilidad. Tradicionalmente, fuentes de alta coherencia, dispositivos de control con compleja polarización y unidades optoelectrónicas sofisticadas son requeridas. Recientemente, algunos sensores basados en interferometría, tal como la luz blanca y nanocavidades han estado introduciéndose arrojando resultados prometedores, por ejemplo, la utilización de interferómetros de luz blanca para mediciones de dispersión cromática ó mediciones en perfiles de superficie. Así como también el empleo de sensores basados en nanocavidades en fibra con propósitos interferométricos para la detección de

Peróxido de Hidrógeno [7]. El uso de interferómetros en fibras ópticas ha sido bien establecido desde hace algunas décadas. Su uso efectivo ha consistido en las técnicas de demodulación. Dentro de la operación normal de un sensor, veremos que los sensores interferométricos, el efecto que causa la medición, es modular la fase del campo eléctrico, donde la fase es convertida a un cambio de intensidad.

Existen algunos arreglos para convertir un cambio de fase a un cambio en la intensidad. Los esfuerzos sustanciales que se han emprendido se abocan a los interferómetros Sagnac, resonadores de anillo, interferómetros Mach-Zehnder y Michelson, interferómetros Fabry-Perot y Fizeau, polarímetros, fibras láser, etc. El sensor basado en la técnica de Fabry-Perot , que puede ser considerado como un interferómetro de haces múltiples, está compuesto por dos espejos formando una cavidad sensible, donde el efecto de transiciones múltiples de la señal hacia delante y atrás provoca un incremento sustancial en la sensibilidad. Entre mayor sea la reflectividad de los espejos, mayor será el propiedad fundamental "Finesse" obtenida a la salida, muy relacionada con parámetros como la reflectividad y sensibilidad. Esta propiedad es la relación entre la cantidad de energía almacenada en la cavidad con la energía que pasa a través de la cavidad misma.

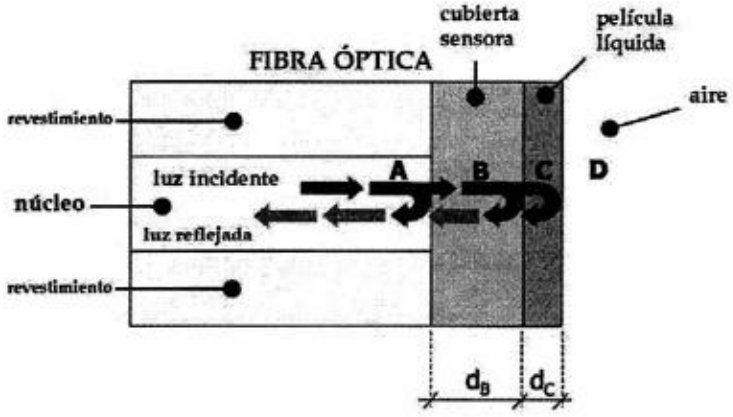


Figura 6. *"Interferómetro Fabry-Perot (FFP). Método de Autoensamblado iónico basado en una nanoestructura al final de la fibra"*

Recientemente, un novedoso arreglo cuyo interferómetro de Fabry- Perot está sobresaliendo por encima del resto, el representado en la figura 6. Existen distintos métodos de deposición convenientes para ser aplicadas en la fabricación de nanocavidades al final de la fibra. De entre esos métodos, el que mayor aplicación ha encontrado para el desarrollo de dispositivos y sensores optoelec-

trónicos basados en fibra es el método de autoensamblado iónico monocapa [8]. Actualmente, sensoras de fibra óptica basados en esta técnica son utilizados para la medición de temperatura, humedad, PH, acetona, diclorometano, etanol y amoniaco, entre otros, han sido fabricados de esta manera.

MANIPULACIÓN DE POLARIZACIÓN

La operación de este tipo de sensor está basada en la modulación de la polarización de la luz, como ejemplo se encuentran los actuales sensores basados en el efecto Faraday para la medición de corriente. En principio, la idea de medir corriente con una fibra óptica (un dieléctrico) pudiese parecer equivocada, al final de cuentas, la corriente eléctrica no circulará por la fibra. Sin embargo, una medida de corriente se logra utilizando la fibra óptica afectándola por intermedio de un efecto denominado rotación de Faraday (fig.7). Una de las propiedades de la luz cuando esta se propaga en un medio cualquiera es la llamada polarización. Los tipos de polarización que la luz puede presentar son: lineal, circular y elíptica. El estado de polarización de la luz deniminado (SOP), se refiere al comportamiento del vector eléctrico en función de tiempo en una determinada posición en el espacio.

La polarización lineal puede expresarse como una superposición de dos polarizaciones circulares (mano-derecha y mano-izquierda). En consecuencia, un campo magnético inducido alrededor del elemento portador de corriente induce una birrefrigerancia circular dentro del embobinado de fibra. Como se sabe, la birrefrigerancia, es la característica que presenta un medio al poseer dos índices de refracción distintos [9]. Por lo tanto, después de pasar a través de la bobina, se genera una diferencia de fase relativa entre dos componentes circulares de la polarización, que da lugar a la rotación del ángulo lineal de la polarización en proporción con el actual y al número de las vueltas de la fibra.

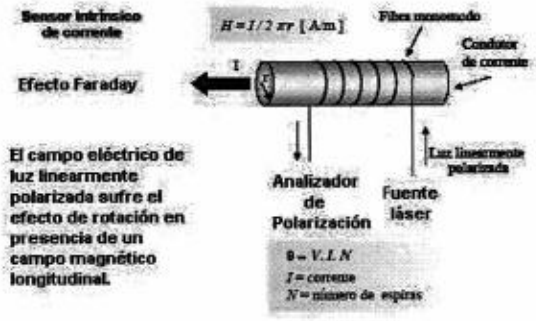


Figura 7. "*Sensor de corriente basado en el efecto Faraday.*

Los sensores de tensión utilizando fibra basan su principio de funcionamiento en el efecto electroóptico, o sea, el cambio de birrefrigerancia lineal debido a la acción de un campo eléctrico aplicado . Cuando un atraso de la fase es proporcional al campo eléctrico aplicado recibe el nombre de efecto Pockel [9], mientras que la variación cuadrática del campo eléctrico se denomina efecto Kerr [9]. En la figura 8 se presenta la configuración básica de un sensor de tensión extrínseco basado en el efecto Pockel, en este ejemplo se puede ver la luz proveniente de una entrada pasar por una lente y en seguida por un polarizador el cual produce una polarización lineal de 45 grados en relación al campo eléctrico aplicado. Una birrefrigerancia inducida por este campo aplicado causa desfasamiento entre los componentes del campo de luz incidente, haciendo que ésta presente una polarización elíptica. Tanto el polarizador de entrada como el de salida poseen la misma dirección de polarización, o sea, en ausencia de un campo aplicado, toda la luz es transmitida hacia el fotodetector, es decir, la salida de la fibra. Cuando hay un campo aplicado, una cantidad de potencia óptica detectada por el fotodetector dependerá del grado de desfasamiento entre las componentes de campo de luz incidente. Cuanto mayor sea este desfasamiento, mas inclinada se torna la elipse y menor será la cantidad de luz que incide en el fotodetector.

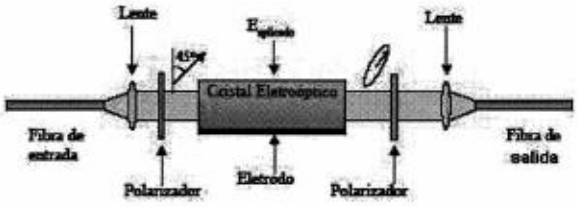


Figura 8. *"Sensor de tensión basado en el efecto Pockel"*

Aunque, el concepto es simple, para fines comerciales se han encontrado algunas dificultades que limitan la resolución de los sistemas de sensado. Unas de estas dificultades es la imperfección en la forma del núcleo de la fibra óptica, lo cual genera una birrefrigerancia lineal y por consecuencia mediciones erróneas. La birrefrigerancia puede reducirse considerablemente a través del proceso de templado. Los primeros reportes de manipulación de polarización se remontan a 25 años atrás. Recientemente, los diseños preliminares con sensores de temperatura y vibración transversal de sensibilad han dado apertura a un compacto y perfecto empaquetado, que ya es comercialmente disponible. La medición de campos electromagnéticos y radiación, ha abierto un frente importante en el sensado de fibra óptica, desde el punto de vista comercial como se indica en líneas anteriores.

MANIPULACIÓN DE LONGITUD DE ONDA

El objetivo es modular el espectro de radiación óptica por el parámetro en cuestión a medir (temperatura, tensión, etc). Los sensores en mención se basan en la absorción, luminiscencia (principalmente fluorescencia), y redes de difracción. Un componente que es ampliamente utilizado y tiene un enorme impacto tanto en las comunicaciones como en los sistemas de sensores son las redes de difracción en fibras. Estas rejillas o redes son simples elementos de sensado, las cuales son foto inscritas dentro del núcleo de la fibra de silicio (fig. 9). Una red de difracción es un dispositivo que refleja o refracta la luz por una cantidad parcial que varía según la longitud de onda. Por ejemplo, si la luz del sol cae en una red de difracción (en el ángulo correcto) entonces la luz del sol se separa en sus colores componentes para formar un arco iris. Esta función (difracción) es igual que la de un prisma.

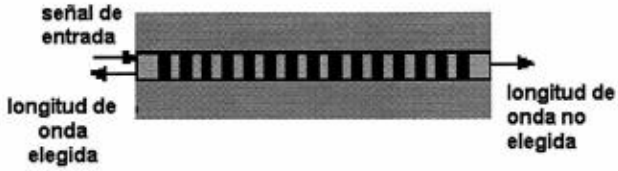


Fig. 9 "*Redes Bragg en fibra*"

Una red Bragg es justo un segmento de fibra monomodo ordinaria de algunos cuantos centímetros de largo. La rejilla o red es construida variando el índice de refracción del núcleo longitudinalmente a lo largo de la fibra. La luz de la longitud de onda especificada que viaja a lo largo de la fibra se refleja en la rejilla en la dirección de la cual procede. Las longitudes de onda que no son seleccionadas se pasan a través de la red con poco o nada de atenuación. Ésta es la característica más importante: las longitudes de onda resonantes son reflejadas hacia la fuente y las longitudes de onda no resonantes son transmitidas a través del dispositivo sin pérdida.

Un valor típico de sensibilidad de la red Bragg para temperatura es alrededor de 10 pm/C en la segunda ventana de telecomunicaciones. En la primera y en la tercera ventana, el valor es de 3pm/C, +/-, respectivamente. Todas las implementaciones de las redes Bragg han sido probadas en aplicaciones como, puentes, embalses, minas, laminas compuestas, control de tráfico, aviones, generadores, sistemas de vías férreas, aplicaciones médicas y en instalaciones nucleares, entre otras. Uno de sus métodos de fabricación es el denominado autoensamblado iónico monocapa, similar al utilizado en la fabricación de nanocavidades de Fabry-Perot. En la figura 10 se puede ver un número de capas depositadas, esto se comporta como un filtro de banda ancha.

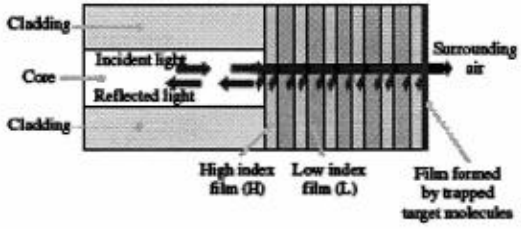


Fig.10 "*Esquema de Autoensamblado de la red en fibra y su mecanismo de sensado*" *(tomado de Encyplodia of sensors figure.12 pag.9)*

NANOTECNOLOGÍA - TÉCNICAS DE PREPARACIÓN

En el desarrollo de sensores de fibra óptica y dispositivos fotónicos en general, es preciso conocer perfectamente las técnicas y materiales susceptibles de ser utilizados, así como entender y aplicar las tecnologías necesarias para llevar a cabo cualquier desarrollo de este tipo. Los sensores ópticos basados en distintos materiales y productos químicos es un campo en el que existe un fuerte impulso investigador en el campo de nuevos materiales con diversas propiedades químicas que pueden ser susceptibles de ser aplicados en sensores ópticos. Para la realización de los sensores se hace uso de tales materiales como sustancias sensoras de los parámetros que se quieren medir.

Para fijar los materiales a la fibra óptica existen varias técnicas, siendo una de ellas la técnica sol-gel [11]. Las nuevas tendencias tecnológicas, especialmente en el campo de la electrónica molecular, requieren el desarrollo de técnicas que permitan la construcción de dispositivos cada vez más pequeños. Esta necesidad, ha propiciado que exista un gran interés por la preparación de las denominadas películas delgadas. Existen distintas técnicas para la realización de estas películas delgadas. Entre ellas se encuentran la de Langmuir-Blodgett (LB), el método de autoensamblado iónico monocapa (AIM-ESA). Existen otras, aunque a priori se ven limitadas a la construcción de estructuras con muy pocas monocapas, o se ven restringidas por el hecho de poder aplicarse solamente con determinadas reacciones químicas, como en el caso del autoensamblado químico monocapa.

MÉTODO SOL-GEL

Desde 1984 el proceso sol-gel se ha usado para la preparación de matrices (vidrios) sólidas porosas dopadas con moléculas orgánicas a temperaturas bajas (< 150 °C). Esos geles vitrificados (xerogeles) son ideales para albergar en su seno distintas moléculas. Este proceso implica la transición de un sistema líquido (sol) a una fase sólida (gel). El proceso de sol-gel permite la fabricación de materiales con una amplia variedad de propiedades: polvos ultra-finos, vidrios y cerámicos monolíticos, fibras plásticas, membranas inorgánicas, cubiertas de películas delgadas y aerogeles. Las aplicaciones derivadas de este proceso son numerosas, pero una de las áreas de mayor aplicación es en la de películas delgadas, las cuales pueden ser producidas sobre una pieza de sustrato ya sea por la técnica de 'spin-coating' ó 'dip-coating'.

Método de autoensamblado iónico monocapa (Electrostatic Self-Assembly)

Es una novedosa técnica de deposición de materiales mediante la fabricación de películas delgadas con una estructura de tipo multicapa, que ha sido ensayada sobre sustratos de diferente naturaleza como vidrio, silicio cristalino, polímeros e incluso metales [10]. El fundamento de la construcción de estructuras multicapa mediante el AIM es la atracción electrostática entre las cargas eléctricas de las moléculas que forman cada monocapa que se deposita. Utilizando esta nanotécnica es posible depositar una nanoestructura bastante estable al final de la fibra. Además a esto, una variedad de estructuras polímeras pueden ser seleccionadas. Las muestras moleculares de los componentes aniónicos y catiónicos y el espesor físico en el ordenamiento de las capas, contribuye a determinar las propiedades de la cubierta resultante.

Es importante hacer notar que los polianiones y policationes se superponen unos con otros a nivel molecular, y esto genera un material óptico homogéneo. En la figura 11 se observa que la primera capa, es decir, la capa individual del sustrato puede estar compuesta por metales, plásticos, cerámicos, y semiconductores. En cada conformación de capa el sustrato es sometido a agua purificada para la remoción de las partículas libres que no lograron enlazarse. El espesor de las películas pueden ser controladas por fuerza iónica ó pH.

Fig. 11 "*Proceso de Deposición de Autoensamblado iónico monocapa (ESA)*"

Método de Langmuir-Blodgett

Otra técnica interesante es la técnica Langmuir-Blodgett (LB), que permite la preparación de películas organizadas, donde es posible llevar un control en el orden y espesor a nivel molecular. La técnica LB, consiste fundamentalmente

en la preparación de multicapas de moléculas sobre un substrato sólido a partir de monocapas ordenadas previamente formadas en la superficie de una subfase líquida.

Estas películas son acumuladas a través de un proceso sucesivo de monocapas individuales sobre un sustrato sólido. La monocapa es formada por la separación de moléculas orgánicas sobre una subfase líquida. Las moléculas usualmente tienen partes hidrófilas e hidrófobas, así que cuando la película se forma, las moléculas se sustentan en sus partes hidrófilas. Las moléculas al ser separadas por primera vez en el agua se agrupan muy libremente formando una fase gaseosa. Esto significa que el área de agua disponible para cada molécula es bastante mayor, y la presión superfical baja. La presión superficial puede incrementarse mediante una o dos barreras deslizables. A cierto punto la presión superficial comienza a crecer más rápidamente indicando una transición hacia la fase líquida. Tanto como sea deslizada la barrera más allá del inicio de la fase sólida, se notará un crecimiento con comportamiento constante en la presión superficial (fig.12). Una o más monocapas pueden depositarse en el sustrato sólido bajándolo a este último y depositando sobre la monocapa ya existente. La monocapa es depositada uniformemente bajo el control de la presión superficial y el mecanismo de barreras deslizables (fig.13).

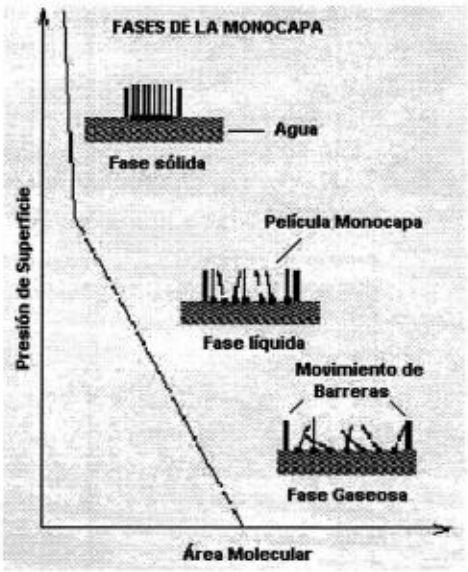


Fig. 12 "Presión superficial-Área isotérmica"

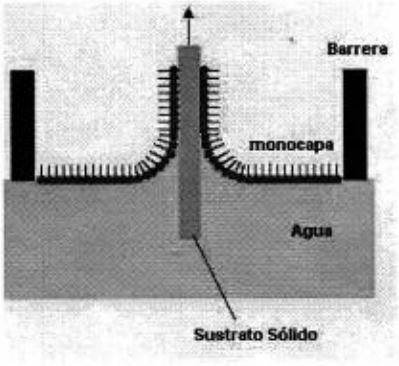


Fig.13 "Deposición de una monocapa sobre un sustrato sólido"

Introducción a los Biosensores de Fibra Óptica

La necesidad de llevar a cabo determinaciones analíticas de manera rápida, selectiva y con elevada sensibilidad ha dado lugar a la aparición y amplio desarrollo de los denominados biosensores. Estos dispositivos analíticos que incorporan un elemento biológico como fase sensorial asociado a un transductor ya que presentan un enorme potencial para la detección de numerosos analitos tanto en el ámbito del análisis clínico, industria alimenticia o medioambiental. Los biosensores pueden ser considerados como una alternativa importante a técnicas convencionales debido a su sensibilidad, selectividad, versatilidad, prontitud y capacidad en la supervisión de analitos múltiples. Un biosensor es un dispositivo autónomo integrado que proporciona la información cuantitativa o semicuantitativa analítica que incluye un elemento de sensado biológico en contacto directo con un transductor (fig. 14). Los elementos de reconocimiento biológico, también llamados receptores, juegan el papel clave en la sensibilidad y la selectividad del sensado. El receptor traduce la información biológica, como una concentración del analito referido, en una señal de salida con una sensibilidad definida. Esto provee del sensor de un alto grado de selectividad para el analito que es medido. El receptor detecta el analito por una interacción específica entre ambos, que generará una perturbación física o química que se pueda convertir en un efecto mensurable tal como una señal óptica o eléctrica, ésta señal es medible a través de un detector que permite observar los cambios de señal. Los puntos dominantes en el biosensado son el mecanismo del reconocimiento del analito y la conexión cercana entre el elemento de detección y el transductor.

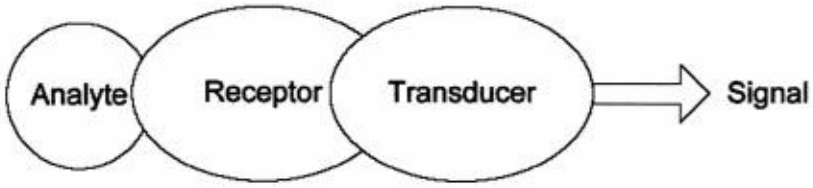


Fig. 14 *Estructura básica de un Biosensor. Yan Zhang "Miniature fiber-optic multicavity Fabry-Perot interferometric biosensor"*

CLASIFICACIÓN DE LOS BIOSENSORES

Los biosensors pueden ser clasificados según su mecanismo de reconocimiento biológico. Las propiedades de reconocimiento del receptor pueden ser clasificadas como catalíticas y por el principio de afinidad de pares. Los biosensors catalíticos monitorean la formación de un producto, la desaparición de un reactante o la inhibición de la reacción en esta misma biocatalizada. El biocatalizador puede representarse por enzimas, microorganismos ó tejidos. Los productos típicos en reacciones catalíticas son las moléculas de bajo peso que son preferidas para transductores electroquímicos. En esta clasificación, son necesarias las sustancias cromogénicas (aquellas que cambian de color y por tanto su índice refractivo) para determinar las señales ópticas que se produzcan por efecto de los enlaces. Los biosensores que emplean un mecanismo de reconocimiento biológico en base a la afinidad, son aquellos cuyos bioreceptores se enlazan a un analito (antígeno) por especificidad. Estos bioreceptores pueden ser anticuerpos, ácidos nucleicos, hormonas, etc. Para ambas clasificaciones, las técnicas ópticas para la detección de variaciones pueden ser a través de la reflectancia, absorbancia, resonancia de plasmones superficiales, onda evanescente, fluorescencia y bioluminiscencia.

TÉCNICAS DE DETECCIÓN DEL GLUTEN EN ALIMENTOS (ESTADO DEL ARTE)

CELIAQUÍA

La enfermedad celíaca es un desorden intestinal producido por la intolerancia al gluten, proteína presente en el trigo y en otros cereales, cebada, centeno y avena, y alimentos derivados, que daña las vellosidades del intestino delgado. Es

una intolerancia permanente al gluten. Es la enfermedad crónica intestinal más frecuente en el mundo occidental [12]. Produce una atrofia de las vellosidades del intestino, lo que ocasiona una mala absorción de los nutrientes (proteína, grasas, hidratos de carbono, sales minerales y vitaminas). Dichas vellosidades son las que permiten absorber los nutrientes de los alimentos, por lo que al ingerir gluten las personas que sufren esta enfermedad desarrollan un síndrome de mala absorción que puede derivar en un proceso de desnutrición y sus consecuencias asociadas como cáncer, anemia, osteoporosis, epilepsia, infertilidad, e incluso la muerte. A ello se suman malestares físicos como dolores intestinales, diarreas crónicas y cansancio, entre otros síntomas.

Hoy en día no existe una manera de obtener información respecto a la composición de los alimentos y medicamentos que se venden a los consumidores, lo que dificulta mucho la posibilidad de llevar una dieta segura para los celíacos. El gluten, que es la proteína a la cual tienen intolerancia las personas que sufren este mal, se usa en muchos alimentos procesados como chocolates y cecinas por ejemplo, además de estar presente en el trigo, la cebada, el centeno y la avena. Ello implica un riesgo permanente de una ingesta involuntaria, con todas las complicaciones que ello genera, si no existe la información apropiada.

Se describen a continuación diferentes técnicas utilizadas para la detección de gluten en alimentos. Algunas de las técnicas descritas están muy implantadas en el control del gluten en los alimentos, como los ensayos ELISA, la técnica PCR, Western Blot, espectrometría de masas, cromatografía y tiras inmunocromatográficas. Por último, se describe la técnica propuesta en la detección de gluten, donde hace referencia a la utilización de un biosensor.

Ensayo Inmunoenzimático Elisa

La técnica ELISA (del inglés Enzyme Linked ImmunoSorbent Assay) consiste en un ensayo basado en el principio inmunológico del reconocimiento y unión de los anticuerpos a las moléculas que reconocen como extrañas (antígenos). Es un método inmunológico clásico, enormemente utilizado para una gran cantidad de aplicaciones, por ejemplo, en diagnóstico clínico, detección de virus, búsqueda de anticuerpos, etc.

En el caso de la detección de gluten se utilizan anticuerpos que reconocen fragmentos presentes en las proteínas del gluten (antígeno). En este ensayo se produce una unión del anticuerpo al antígeno sobre una superficie (generalmente el fondo del tubo de ensayo o similar) a la que previamente el anticuerpo o el antígeno se ha unido. Alguno de los componentes del ensayo (anticuerpo o antígeno)

se encuentra unido a una enzima que catalizará la formación de un producto coloreado, que podrá ser cuantificado mediante la medida de la luz absorbida por dicho compuesto (espectrofotometría). Existen distintos tipos de ensayos ELISA, siendo los más utilizados en la detección de gluten los ensayos tipo sandwich y los ensayos competitivos. En el ELISA tipo sándwich se utilizan dos anticuerpos, el anticuerpo primario y el anticuerpo secundario, unido a la enzima. En este ensayo se establece la unión directa del gluten a los dos anticuerpos, quedando el antígeno "atrapado" entre ambos (ver Figura 15). En el ELISA competitivo se incuba la muestra con el anticuerpo para después añadir esta preparación sobre una superficie recubierta de antígeno (por ejemplo, gliadinas de trigo) de tal forma que se une a la superficie el anticuerpo libre no unido al gluten de la muestra. Finalmente se detecta la cantidad de anticuerpo libre; cuanto más anticuerpo libre es detectado, menos cantidad de gluten contiene la muestra. Para la detección de gluten se han propuesto distintos anticuerpos y métodos de detección. A continuación se tratará el método de Skerrit y el método basado en el anticuerpo R5. El primero ha sido utilizado, en el pasado, a nivel internacional.

El método propuesto por Skerrit y Hill fue validado internacionalmente por la AOAC (Association of Official Analytical Chemists). La sensibilidad original de este método era de 160 ppm (Partes por millón. Miligramo de analito por kilogramo de muestra), aunque se ha distribuido comercialmente bajo distintas marcas, optimizándose hasta conseguir sensibilidades de 20 ppm. El método detecta prolaminas del trigo y centeno resistentes a alta temperatura, que pueden ser extraídas del alimento después de ser cocinadas sin perder su capacidad de unión al anticuerpo.

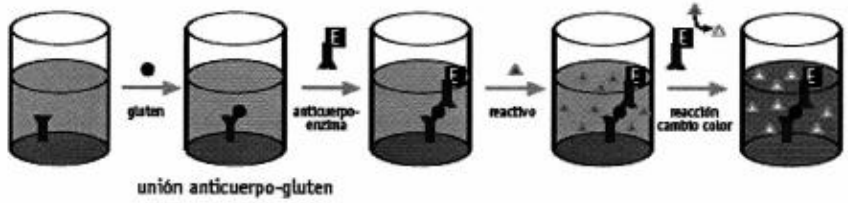


Fig. 15 *"Ensayo inmunoenzimático ELISA tipo sandwich. En este tipo de ensayos, el anticuerpo (en azul) se encuentra adherido al fondo del tubo de reacción. Cuando se añade la muestra, las moléculas de gluten (en negro) se unen al anticuerpo. A continuación, se añade nuevo anticuerpo unido a una enzima y al añadir el reactivo final, la enzima cataliza la reacción de formación de un producto coloreado en la mezcla"*

Ventajas

- Simple en su realización.
- Rapidez: tiempo medio del ensayo ELISA en placa de, aprox. 2 horas.
- Económico, versátil y robusto.
- Alta sensibilidad (3 ppm de gluten).
- La detección se realiza por medio de dispositivos ópticos.
- No produce reacciones cruzadas frente a prolaminas no tóxicas de maíz o arroz.

Desventajas

- Pueden producirse falsos negativos cuando se desnaturalizan las proteínas por

cambios de presión, temperatura o concentración de sales.

- Posibilidad de reacciones cruzadas entre proteínas estrechamente relacionadas.

Técnica de PCR

La tecnología de PCR ó reacción en cadena de la polimerasa consiste en la obtención de múltiples copias de un fragmento específico de ácido desoxirribonucleico (ADN) situado entre regiones de secuencia conocida a partir de una muestra compleja de ADN. La amplificación de ese fragmento elegido permite realizar su detección y estudio posterior. Esta tecnología fue desarrollada a principios de los años 80 por Kary Mullis y sus colaboradores y, desde su descubrimiento, se ha convertido en una herramienta de biología molecular indispensable, con aplicaciones tales como la búsqueda de mutaciones en enfermedades hereditarias, criminalística y ciencia forense, análisis genéticos, pruebas de paternidad, identificación de especies y búsqueda de microorganismos patógenos. Hoy en día, el proceso de amplificación de fragmentos de ADN se encuentra automatizado. Mediante PCR se puede detectar gluten de manera indirecta. Esta técnica detecta el ADN responsable de la síntesis de las proteínas del gluten, proceso conocido como expresión génica y que constituye la base del funcionamiento de todos los seres vivos. Si en el diseño del ensayo PCR se utiliza un fragmento presente en gliadinas, secalinas, hordeínas y aveninas puede detectarse el ADN de todos los cereales con prolaminas tóxicas.

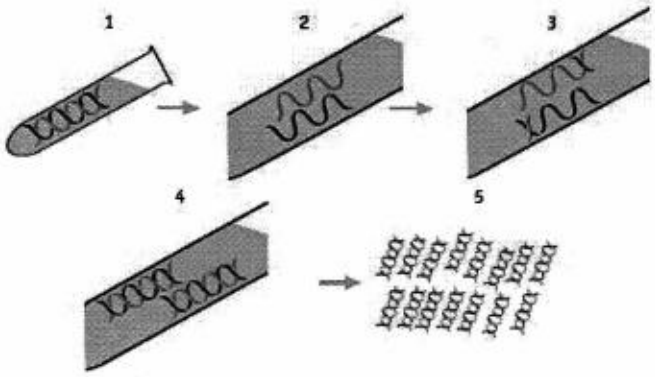


Fig. 16 *"Reacción en cadena de la polimerasa. 1. Obtención de moléculas de ADN. 2. Separación de las hebras de la doble hélice calentando a 90 ºC. 3. Unos fragmentos de ADN se unen a las hebras de ADN en la posición correcta. 4. La enzima ADN polimerasa, que se añade al medio, sintetiza dos nuevas copias de las hebras de ADN. 5. El proceso completo genera gran número de copias de la molécula de ADN original" http://nobelprize.org/nobel_prizes/chemistry/laureates/1993/illpres/pcr.html*

VENTAJAS

- Sensibilidad muy alta en la detección del ADN (5-50 picogramos de ADN).
- Permite identificar la especie de la que proviene el gluten presente, muy útil para identificar el origen de una contaminación cruzada.
- El ADN es menos susceptible de ser degradado durante el procesado de los alimentos.

DESVENTAJAS

- Se requiere tiempo y personal calificado en el análisis.
- El ADN puede fragmentarse durante el procesamiento de los alimentos.
- Técnica indirecta para detectar gluten (no cuantifica la presencia de gluten, sino la del ADN que codifica para el gluten).

Técnica Western Blot

La técnica Western Blot es, al igual que los ELISA, un inmunoensayo diseñado para detectar proteínas en muestras complejas. La detección final de las proteínas (el gluten) se realiza mediante la unión de anticuerpos específicos anti-gluten y la utilización de enzimas unidas a estos mismos anticuerpos gracias a las cuales puede registrarse su unión al gluten. El nombre de esta tecnología proviene de la transferencia de proteínas o blotting, característica principal de este tipo de ensayos, en los que se produce la inmovilización de las proteínas sobre membranas sintéticas, seguido de la detección.

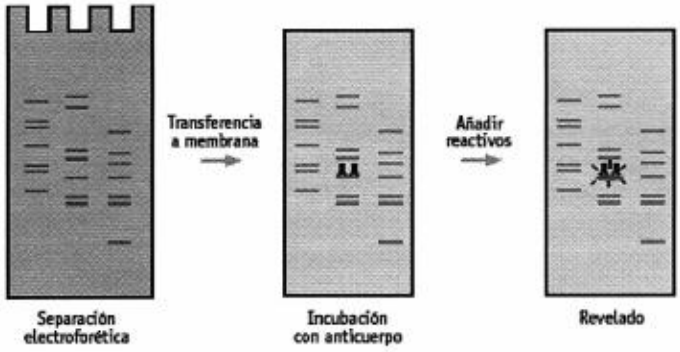


Fig. 17 *" Western Blot"*

Los procedimientos de Western Blot constan de varias etapas. En una primera etapa se separan las proteínas, en función de su tamaño, mediante un proceso denominado electroforesis en gel que utiliza una corriente eléctrica para la separación de las proteínas. A continuación, con el objetivo de hacer accesibles las proteínas para su detección por el anticuerpo, se transfieren a membranas sintéticas donde se localizan sobre la superficie de la membrana para su detección. Por último, se produce la unión específica de la proteína a los anticuerpos y éstos son detectados gracias a la enzima unida a ellos. Existen distintos métodos de detección, por ejemplo el quimioluminiscente, basado en una reacción química para la formación de un producto cuya energía se emite en forma de luz. La técnica Western Blot es utilizada como una prueba confirmatoria de los resultados de un ELISA cuando se detecta presencia de gluten.

VENTAJAS

- Útil para identificar si el gluten de la muestra proviene de trigo, cebada ó centeno.
- Método altamente específico.
- Sensibilidad de 5-10 ppm de gluten en función del anticuerpo utilizado.
- Valor confirmatorio al acoplar el proceso de caracterización del tamaño del gluten con la unión específica de los anticuerpos.
- Eficaz en la detección de proteínas insolubles.
- Método muy adecuado para la detección del contenido en gluten de alimentos crudos y procesados.

DESVENTAJAS

- Método lento (se necesitan 48 horas para realizar un análisis completo).
- Se requiere formación y especialización adecuada de los analistas.

ESPECTROMETRÍA DE MASAS

La espectrometría de masas es una técnica analítica utilizada para medir la masa molecular de compuestos químicos o biológicos, deducir datos estructurales e identificarlos. Para el análisis de gluten se ha empleado la técnica de espectrometría de masas MALDI-TOF (desorción/ionización mediante láser asistida por matriz acoplada a un detector de tiempo de vuelo). En la espectrometría de masas las muestras son ionizadas, es decir dotadas de carga, porque los iones son más fáciles de manipular que las moléculas neutras. En la ionización MALDI los analitos se mezclan con una matriz orgánica y se convierten en iones mediante la acción de un láser. Esta técnica de ionización de la muestra es muy apropiada para moléculas no volátiles de alta masa molecular, como las proteínas. En el analizador de tiempo de vuelo (TOF) los iones se separan en función de su masa y carga tras ser acelerados en el vacío por un campo eléctrico. El tiempo que tardan en recorrer el analizador depende de estos dos parámetros y permite determinar la masa.

Esta técnica puede aplicarse a la detección de gluten en extractos de cereales. El análisis del gluten de trigo, centeno, cebada y avena proporciona el patrón

de masas característico de las prolaminas de estos cereales (gliadinas, secalinas, hordeínas y aveninas, respectivamente) y permite la compleja identificación de las prolaminas del gluten.

Ventajas

- Rapidez del análisis (pocos minutos).
- Manipulación de la muestra sencilla.
- Reproducibilidad.
- Precisión en la determinación de la masa de las prolaminas.
- Puede indicar la especie de la que proviene el gluten.
- La interpretación rutinaria de los espectros es relativamente sencilla.

Desventajas

- Instrumentación compleja.
- Equipamiento costoso.
- No es una técnica cuantitativa.
- El equipo requiere instalaciones amplias.
- Complejo proceso de elaboración de librerías de perfiles de espectros.
- Compleja calibración del equipo.

Técnicas Cromatográficas

La cromatografía es un método físico de separación, en el que los componentes se distribuyen en dos fases, la fase móvil y la estacionaria. El análisis de muestras alimentarias puede realizarse mediante cromatografía líquida, en la que la fase móvil es un líquido, por tanto los componentes de la muestra deben ser solubles en ese líquido. Las especies separadas se pueden caracterizar mediante los detectores apropiados. Mediante cromatografía se puede detectar la presencia en alimentos de péptidos y proteínas en general y de gluten en particular y se puede también cuantificar su concentración. La separación puede realizarse en función de la carga, el tamaño, la hidrofobicidad.

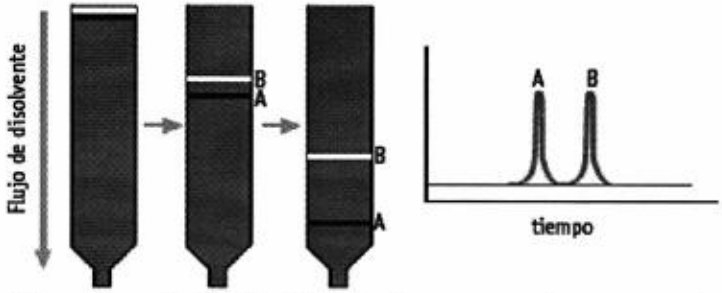


Fig. 18 *"Cromatografía líquida. Se aplica la muestra en la parte superior de una columna de cromatografía. El disolvente fluye por la columna y los componentes de la muestra se van separando según viajan por el lecho de la columna. Finalmente, se detecta la elución (salida) de cada compuesto en función del tiempo"*

VENTAJAS

- Gran capacidad para la separación de distintos péptidos.

DESVENTAJAS

- Se requiere mucho tiempo para el análisis.
- Difícil de automatizar para muchas muestras.

TIRAS INMUNOCROMATOGRÁFICAS

La detección de gluten mediante tiras inmunocromatográficas es un método muy sencillo y rápido, similar a los test de embarazo. El gluten debe extraerse de la muestra de alimento utilizando una disolución de extracción. Después de la extracción, la muestra se aplica sobre la tira, donde se encuentran los anticuerpos que reconocen el gluten unidos a esferas coloreadas de látex. El gluten unido a estos anticuerpos se desplaza a través de la tira en un proceso cromatográfico de separación. Finalmente, el gluten y el anticuerpo unido se inmovilizan en una región de la tira, donde se puede detectar la presencia del gluten como una banda coloreada.

Las tiras permiten detectar el gluten sin necesitar complejo material de laboratorio, de tal manera que las empresas que disponen de laboratorio con pequeñas infraestructuras para tratamiento de muestras pueden realizar el análisis fácil-

mente. La detección ofrece un resultado positivo o negativo para el gluten pero no permite conocer su concentración.

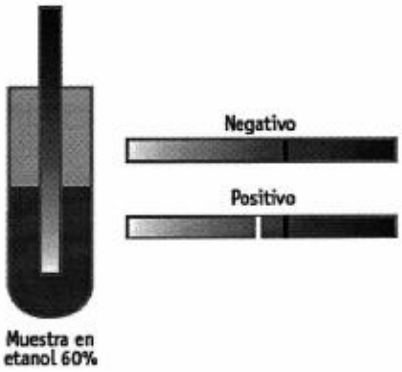


Fig. 19 *"Ensayo de una tira inmunocromatográfica".*

Ventajas

- Es el método más sencillo de todos.
- Rápido.
- Interpretación visual.

Desventajas

- No permite conocer la concentración de gluten en la muestra.

Diseño Preliminar de un Biosensor de Fibra Óptica para la Detección de Gluten

Inmunoensayo

Se propone una técnica de análisis que permitirá detectar alimentos aptos para los celíacos utilizando un biosensor de fibra óptica, bajo el principio de afinidad de pares, es decir, propiciar la inmovilización de anticuerpos (inmunosensor) fijados con la técnica de autoensamblado iónico monocapa (ESA) a un sustrato (extremo de fibra óptica), para enlazarse por especificidad a las proteínas a ser analizadas (fig. 20).

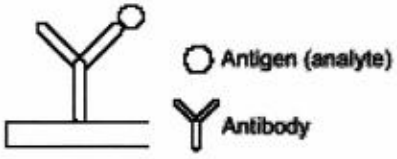


Fig. 20 " Formato de ensayo directo para inmunosensores"

ANTÍGENOS Y ANTICUERPOS

Un antígeno (Ag) es cualquier cosa que hace que el sistema inmune responda generando anticuerpos (Ab). Los antígenos pueden vivir o permanecer fuera de los organismos, tales como virus, bacterias, y también podían ser proteínas, polisacáridos, lípidos, o aún el polvo. Los anticuerpos son las proteínas que atan o enlazan al antígeno específico con alta afinidad (fig.21). Combinando la especificidad inherente de las reacciones del antígeno-anticuerpo (Ag-Ab) con la alta sensibilidad de varios transductores físicos, los biosensores inmunoensayo son superiores en sensibilidad y selectividad.

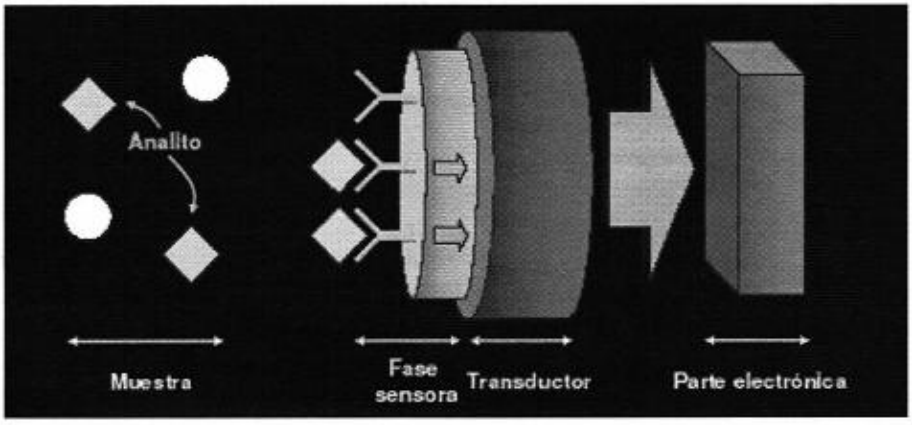


Fig. 21 " Diseño de un inmunosensor de fibra óptica, esquema tomado de Ciencia Viva "Sensores Electroquímicos"

Las fuerzas de enlace entre antígenos y anticuerpos se pueden considerar todas como interacciones no covalentes. En las reacciones químicas ordinarias, las moléculas se unen por la creación de enlaces covalentes sólidos e irreversibles. Pero en los sistemas biológicos, tales enlaces no serían bastante flexibles o adaptables para los fines de biosensado. En cambio, los enlaces no covalentes

son la manera rápida y reversible de formar complejos, y permiten reutilizar las moléculas de anticuerpos en una manera que sería imposible si se establecieran enlaces covalentes. Los enlaces no covalentes suelen formarse a distancias intermoleculares relativamente pequeñas, exhibiendo una estrecha correspondencia por especificidad, análogamente se pudiera decir de una llave con su cerradura. De las fuerzas de enlace que unen a los antígenos con los anticuerpos podemos mencionar 4 grupos: Interacción Electrostática, Puentes de Hidrógeno, Interacción Hidrófoba y Fuerzas de Van der Waals. A la fuerza de los enlaces entres Ag's y Ab's se le llama "afinidad" y se puede calcular recurriendo a la ley de acción de masas, pues la interacción entre antígenos y anticuerpos es reversible.

La ley de acción de masas establece que la velocidad de reacción es proporcional a las concentraciones de las partes reaccionantes;

$$[Ab]+[Ag]\underset{k_d}{\overset{k_a}{\rightleftharpoons}}[Ab\cdot Ag],$$

donde y son las constantes de velocidad, mientras que es la concentración del complejo formado por el anticuerpo y antígeno.

Esquema de Implementación

La inclusión de anticuerpos se propone realizar por el método de deposición de autoensamblado iónico monocapa (ESA) visto en la figura 11 como fijador al sustrato a través de polianiones y policationes. La configuración propuesta es empelando la técnica de Nanocavidades Interferométricas de fibra Fabry-Perot (Nano-FFP), (fig. 6). Esta nanocavidad funciona usando la variación de la distancia que existe entre sus espejos para producir un cambio de fase y por tanto un cambio en la intensidad óptica reflejada. Sin embargo, para aplicaciones de biosensado, las moléculas que quedan enlazadas en la superficie de la cubierta sensora, cambian la reflectividad óptica. Consecuentemente, la longitud inicial de la nanocavidad, d_B, se incrementa hacia un valor final $d_B + d_C$.

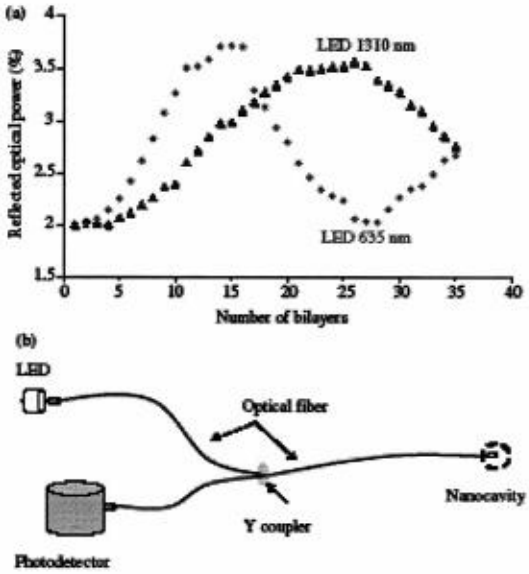


Fig.22 (a) Energía óptica reflejada obtenida durante el ensamblado de una cubierta en el extremo de una fibra, utilizando dos diferente fuentes lumínicas a 635 nm y 1310 nm. (b)Esquema experimental para monitorizar el ensamble de las bicapas. R. O. Claus, I. R. Matias, and F. J. Arregui, "Optical Fiber Sensors" in "Encyclopedia of Sensors" Vol. X:Page 7, 2006.

La figura 22 ilustra el proceso de ensamblado bicapa tras bicapa de la nanocavidad, exhibiendo las curvas experimentales típicas de la la naturaleza interferométrica de la cavidad. Como es evidente se puede observar el comportamiento de la cavidad así como la similitud de las curvas aunque con diferente periodicidad debido al valor de los LED`s utilizados en distintas ventanas.

CONCLUSIONES Y MEJORAS

Este reporte presenta un panorama general introductorio en los sistemas de sensado basados en fibra óptica, en lo general. Se mencionaron las ventajas y desventajas que posee frente a sus contrapartes, así como también las clasificaciones en que se subdivide para propósitos de estudio. Se ha mencionado con ligero grado de detalle los mecanismos de sensado en los que se basa cada sistema en particular, con su respectiva técnica de modulación. Trasladándose hacia lo particular para el enfoque de los biosensores de fibra óptica. Desde mi punto de vista, este campo de investigación que se encuentra en desarrollo por sus características de manejo

sencillo, rápido y con precisión responderá a la demanda creciente de las empresas alimenticias para integrarse a los procesos de producción. Consecuentemente, detonará la expansión del mercado global ya que por sus ventajas de manipulación con técnicas a escalas nanométricas atraerá otras ramas de la ciencia como el control de procesos industriales, instrumental médico, robótica, áreas medioambientales, entre otras. Los esfuerzos se deben centrar en mejorar aspectos como incrementar las respuestas rápidas, que sea independiente a la temperatura, fiable, exacto, miniaturizable, reutilizable y, bajo costo de producción.

Metodología

La elaboración del presente trabajo requirió la realización de distintas actividades en el que se ha desarrollado a partir de análisis diversos que consistieron en observar aspectos bioquímicos y biológicos en ciertas proteínas de algunos cereales que implicaban afectaciones en pacientes con predisposición genética y, cuyos cuadros clínicos ocasionaban confusión en su diagnóstico, según las asociaciones de celiacos de Ibero América, principalmente España. Consecuentemente, se presupuso por asociación de ideas (método deductivo) un diseño preliminar novedoso en base a las técnicas actuales que se vienen empleando, complementadas con el uso de tecnologías de óptica de fibras. Esto derivó en la conformación de un modelo (diseño del sistema) que sintetice el análisis y a la hipótesis que nos diera respuesta. Las estrategias de búsqueda de información en la que se estructura este trabajo, es soportado por consultas a través de bases de datos de publicaciones científicas que se auspician en servidores especializados en el tema y cuyos reportes son emitidos por grupos de investigación.

Bibliografía

[1] J. M. López-Higuera, Ed., "Advanced Photonic Topics." Universidad de Cantabria, Santander, Spain, 1997.
[2] R. O. Claus, I. R. Matias, and F. J. Arregui, "Optical Fiber Sensors" in "Encyclopedia of Sensors" Volume X:Page 2, 2006.
[3] J. C. Simon and E. Spitz, Communications à la Societé Française de Physique 24, 1449 (1963).
[4] C. K. Kao and G. Hockman, Proceedings of the IEE 113, 1151(1966).
[5] B. Lee, Optical Fiber Technology 9, 57 (2003).
[6] Eric Udd, Review of Scientific Instruments 66 4015 (1995).

[7] ESA-based in-fiber nanocavity for hydrogen-peroxide detection .Del Villar, I. Matias, I.R. Arregui, F.J. Claus, R.O. Electr. & Electron. Eng. Dept., Public Univ. of Navarra, Pamplona, Spain.

[8] F. J. Arregui, R. O. Claus, I. R. Matias, M. Olza, E. Luquin, and K. L. Cooper, Science and Engineering of Composite Materials 10, 19 (2002).

[9] E. Udd, Fiber Optic Sensors – An Introduction for Engineers and Scientists, John Wiley, New York, 1991.

[10] Francisco. J. Arregui, Yanjing Liu, Kristie M. Lenahan, Ignacio R. Matías and Richard O. Claus "Optical fiber nanometer-scale Fabry-Perot interferometer formed by the Ionic Self-Assembled Monolayer Process" Optics Letters, vol. 24 (9); pp. 596-598. Mayo 1999.

[11] M.L. Rodríguez Méndez, "Langmuir-Blodgett Films of Rare-Earth Lanthanide Bisphthalocyanines. Applications as Sensors of Gases and Volatile Organic Compounds", Comments Inorg. Chem. Vol.22, no. 3-4, 227-239, (2000).

[12] www.celiacosmadrid.org

12. Estudio de Factibilidad para el Desarrollo de un Nuevo Sensor de Corrosión Electroquímica.

Carlos E. Martínez del Ángel, Julio Laria Menchaca

Resumen

En el presente trabajo se hace un estudio de los sensores de corrosión que más se utilizan en la actualidad y como estos dependen en gran medida del tipo de técnica de monitoreo que lo utiliza, también realizamos un análisis de estas. A partir del estudio de las ventajas y desventajas de cada una de las técnicas de monitoreo así como de su evolución, seleccionamos algunas características que nos permitieron realizar un estudio sobre la factibilidad de desarrollar un nuevo sensor de corrosión electroquímica. En específico realizamos un análisis que nos dio como resultado que es factible el desarrollo de un sensor de corrosión conjugando algunas de las ventajas de la técnica de espectroscopia acústica y la técnica de vibraciones. Este tipo de sensor complementará los ya existentes con ese fin y tendrá algunas características particulares que pudiesen encontrar alguna aplicación más específica.

1. Introduccion

1.1. Corrosión

La ISO Standard 8044 [3] define la corrosión como la reacción química entre un metal y su medio ambiente, el cual conduce hacia un cambio de las características del metal y el cual puede conducir a un sustancial deterioro de las funciones del metal. La corrosión frecuentemente ocurre en presencia de fluidos conductivos, humedad atmosférica o altas temperaturas. La corrosión también se encuentra frecuentemente en la presencia de gases como por ejemplo CO2 [4].

En los materiales metálicos se generan formas típicas de corrosión que afectan a una mínima parte de la superficie metálica, dando lugar a penetraciones considerables sin apenas pérdidas de material, característica que las hace extremadamente peligrosas, pues solo se detectan cuando se ha producido el daño y ya no son posibles las medidas preventivas. Estos son los fenómenos de corrosión por picaduras, en resquicios, ínter granular y el agrietamiento por corrosión-fatiga y por corrosión bajo tensiones.

En la mayoría de los casos existe una interrelación entre las formas de corrosión, aunque algunas veces es posible observar una forma única de corrosión. De acuerdo a la forma en que se manifiesta la corrosión sobre un metal, es conveniente establecer una clasificación particular. La mayoría de los autores se inclina por ocho formas comunes de corrosión sin tener un orden particular de importancia:

1. Corrosión galvánica.- Corrosión totalmente electroquímica, consiste en la unión de dos metales diferentes y se hace necesario la presencia de una pequeñísima parte de electrolito que funcione como conductor.
2. Corrosión por separación.- El termino separación es mas usado para indicar un tipo de corrosión selectivo, para un elemento en particular que se separa de una aleación.
3. Corrosión por erosión.- Podría compararse a un barrido de los átomos que están en la superficie de un metal. La velocidad de destrucción del metal se ve acelerada por el movimiento del líquido que rosa su superficie.
4. Corrosión uniforme.- Es la forma más benigna. Consiste en un ataque homogéneo en toda la superficie. Existe igual penetración en todos los puntos. Se puede calcular la vida útil de los materiales expuestos.
5. Corrosión en placas.- Caso intermedio entre uniforme y localizada. Ocurre un ataque general pero más extenso en algunas zonas.
6. Corrosión por picado.- Es una forma peligrosa. El ataque no es proporcional a la magnitud de los daños. El ataque se localiza en puntos aislados de superficies metálicas pasivas y se propaga al interior del metal. En ocasiones por túneles microscópicos. Provoca la perforación de cañerías o tanques.
7. Corrosión ínter granular.- Se propaga a lo largo de los límites de grano. Se extiende hasta inutilizar el material afectado.
8. Corrosión bajo tensión.- Ocurre cuando el metal es sometido simultáneamente a un medio corrosivo y a tensión mecánica de tracción. Aparecen fisuras que se propagan al interior del metal hasta que se relajan

o el metal se fractura. La velocidad de propagación varía entre 1 y 10 mm/h.

Existen otros tipos de corrosión, pero, aparentemente caen en alguno de los enunciados anteriormente, entre otros se puede mencionar: Ataque en línea, ruptura, cavitación, rozamiento, fatiga, gradientes de temperatura, ataque de hidrogeno, ataque microbiano, catastrófica, etcétera. Además se considera que todos esos tipos de corrosión o son físicos, químicos o electroquímicos.

- Física: Golpes sobre el metal, esfuerzos, agotamiento del metal.
- Química: El metal reacciona con un medio no iónico, por ejemplo la oxidación de un metal en aire a altas temperaturas.
- Electroquímica: Ocurre transporte simultáneo de electricidad a través de un electrolito. Ejemplos: corrosión en soluciones salinas, agua de mar, atmósfera, suelos, etcétera.

En la actualidad se considera que la corrosión electroquímica es la que ocurre en mayor proporción.

1.2. Técnicas electroquímicas para el monitoreo de la corrosión.

En la literatura se pueden encontrar varias técnicas que se utilizan para el monitoreo de los procesos de corrosión pero entre estas las más importantes y de mayor amplio uso son:

1-Probetas de prueba (Cuppons)
2-Resistencia a la Polarización Lineal (LPR)
3-Espectroscopia de Impedancia Eléctrica (EIS)
4-Ruido Electroquímico (EN)
5-Espectroscopia Acústica (AE)

El empleo de probetas es una de las técnicas más antiguas y que aún en la actualidad se utilizan. Esta consiste en introducir en el medio corrosivo unas probetas del material que se desea analizar. Existe una norma para la confección de dichas probetas así como para el cálculo de la velocidad de corrosión del material el cual se basa en la diferencia de peso debido a la pérdida de este como consecuencia de la corrosión. La desventaja fundamental de esta técnica es que los re-

sultados se obtienen ya cuando existen altos niveles de corrosión que conlleven a pérdidas del material de la probeta. Además esta técnica solo permite determinar velocidad de corrosión y no el tipo de corrosión que esta teniendo lugar.

1.2.1. RESISTENCIA A LA POLARIZACIÓN LINEAL (RPL)

La resistencia a la polarización lineal (RPL) es una de las técnicas electroquímicas que ha sido más utilizadas en los últimos cincuenta años. Con el paso del tiempo se han desarrollado otras herramientas experimentales que son más complejas que la RPL, aportan información mecanística e implican el uso de instrumentación cara y sin embargo, no han conseguido desplazar a esta técnica de un lugar importante en el ámbito de la ingeniería de corrosión. El conocimiento de las limitaciones de la RPL y de sus ventajas y bondades, resultará en un uso correcto y una clara interpretación de los resultados que se obtengan al aplicar esta técnica en sistemas simples y complejos.

Algunas de las ventajas que podemos citar sobre esta técnica son: (a) que es una técnica no destructiva debido a que emplea bajas perturbaciones, (b) proporciona velocidades de corrosión instantánea, (c) no hace falta instrumentación muy sofisticada pues solamente necesita un potenciostato y, en consecuencia, es una técnica económica, y (d) para aplicar la metodología no hace falta personal altamente especializado. Y sobre sus limitaciones podemos mencionar que: (a) es necesario que el potencial sea estable, (b) caída óhmica en sistemas altamente resistivos, lo que la hace poco recomendable, y (c) es necesario seleccionar una velocidad de barrido adecuado [7].

1.2.2. ESPECTROSCOPIA DE IMPEDANCIA ELECTROQUÍMICA (EIS)

Esta técnica consiste en medir la respuesta eléctrica de un material en función de la frecuencia. Utilizando la EIS podemos hacer el estudio de las propiedades físicas y químicas del material y el movimiento de cargas dentro de su estructura lo que permite realizar un estudio sistemático de los procesos de corrosión uniforme y localizada como lo son el de picadura y agrietamiento además brinda información sobre los procesos que ocurren en la interfase metal-electrolito. La EIS ha sido establecida como una poderosa técnica para el estudio del comportamiento de la corrosión en metales recubiertos con películas anticorrosivos o protectoras.

Entre las desventajas de estas técnicas es que debido a que utilizan una señal de perturbación externa es difícil su aplicación en línea y la electrónica para la medición es compleja [7].

1.2.3. Ruido Electroquímico (EN)

Esta técnica se ha ido desarrollando en los últimos 20 años paralelamente con el resto de las técnicas electroquímicas y es posiblemente el método electroquímico más factible para el monitoreo de la corrosión localizada del tipo de picadura y agrietamiento [4-8] así como la corrosión uniforme. Lo atractivo de esta técnica a diferencia de las técnicas electroquímicas convencionales antes expuestas es su habilidad de poder identificar el tipo de corrosión que esta ocurriendo en el material y sin la imposición de una señal de perturbación externa. Por otra parte, los picos de las señales de corriente y voltaje de ruido asociados a los procesos de corrosión son detectados en tiempo real y mucho antes de que sean evidentes los daños en la interfase material/medio lo que permite realizar acciones preventivas para minimizar la corrosión.

De esta manera el ruido electroquímico se manifiesta como fluctuaciones espontáneas de potencial y corriente de baja frecuencia (<10Hz) y de baja amplitud (1mV) [8,9]. Al respecto podemos mencionar que algunas de las ventajas y desventajas sobre esta técnica son respectivamente: (a) que es una técnica que no necesita señal de perturbación externa, (b) da información si el metal esta activo o pasivo, (c) brinda información sobre la morfología del tipo de corrosión, y (d) mecanismo del tipo de corrosión. Y una de sus limitaciones es generalmente se usan tiempos de medición relativamente largos.

1.2.4. Espectroscopia Acústica (AE)

De acuerdo con la definición de la ASTM, las emisiones acústicas (AE) son un efecto dinámico transitorio debido a la propagación de la onda elástica de esfuerzo, generada por la rápida liberación de energía por micro-fracturas en el material.

De acuerdo con lo anterior, una inspección por emisiones acústicas requiere la aplicación de carga externa en la pieza o elemento que se inspecciona, así como de sensores que se utilizan para "escuchar", y de un sistema de análisis que correlacione las señales recibidas por los sensores, e identifique la fuente del sonido.

En realidad existen varias fuentes que producen las emisiones acústicas entre las cuales destacan las siguientes según su orden de importancia:

1. Formación y crecimiento de grietas,
2. Realineamiento molecular, o crecimiento de dominios magnéticos por procesos magneto-mecánicos (efecto Barkhausen),

3. Cambios micro estructurales como el movimiento de dislocaciones o cambios de fase,
4. Fractura de inclusiones frágiles o películas,
5. Fractura de fibras y delaminación en materiales compuestos,
6. Actividad química como la corrosión,
7. Sismos (fenómeno de emisiones acústicas a gran escala)

Las emisiones acústicas son una técnica pasiva de evaluación no destructiva, ya que la señal acústica es generada por el material mismo y no por una fuente externa. Por lo anterior, para que esta técnica se pueda aplicar, se necesita un estímulo mecánico, térmico o magnético que induzca esfuerzo en el material para provocar la emisión acústica, por tanto, sin la aplicación del estímulo no hay emisión sonora.

Es necesario el análisis de las señales recibidas por los sensores para ubicar la fuente del sonido; para ello hay que sincronizar en el tiempo las señales, y mediante triangularización localizar la fuente en función de las diferencias en tiempo entre los diversos sensores.

Casi todos los materiales tienen propiedades o condiciones para emitir emisiones acústicas que favorecen la inspección no destructiva. Entre las que incrementan el nivel de emisión acústica se encuentran las anisotropías; las impurezas; la fragilidad; los granos de gran tamaño; las transformaciones de fase; las grietas que se propagan, y las secciones laminares. Las condiciones que ayudan al incremento de las emisiones, son los niveles altos de esfuerzo y las bajas temperaturas.

Entre las ventajas de esta técnica tenemos:

- Localizan defectos o grietas que se propagan o crecen
- A pesar de que se debe aplicar una carga al objeto de prueba, no es necesario generar un pulso externo para obtener la medición
- Los equipos de medición son relativamente sencillos y de bajo costo
- Las señales pueden ser almacenadas para un procesamiento y análisis posterior
- La técnica se puede aplicar en forma remota y continua
- Se pueden detectar defectos ocultos
- Todo un sistema o estructura puede ser monitoreada en una sola prueba
- Es una técnica adecuada para monitoreo a largo plazo y en ambientes Hostiles

Pero esta técnica también tiene las siguientes desventajas:

- Se debe aplicar una carga, y no detecta grietas o defectos estabilizados que no se propagan
- Algunos defectos no emiten señales acústicas; por tanto, no son detectables
- No se puede aplicar a todos los materiales y ni tampoco durante el tratamiento térmico de una pieza
- A pesar de que se localizan los defectos, no se puede estimar su tamaño.
- Como cada condición de carga y crecimiento de grieta es único, la técnica es no repetible
- No existen procedimientos normalizados de prueba o interpretación de datos
- Estructuras complejas dificultan la localización de las fuentes sonoras
- La sobrestimación de la técnica ha generado falsas expectativas; de ahí; el descrédito del uso de las emisiones acústicas

Sin embargo, las emisiones acústicas se han convertido en la técnica de evaluación no destructiva más importante para los materiales compuestos y, en algunos casos, es la única.

1.3. Sensores de corrosión electroquímica

Los sensores de corrosión se desarrollan en dependencia de la técnica que se va a utilizar para el monitoreo de la corrosión y que se relacionaron anteriormente. Por tanto, veremos el principio de diseño y funcionamiento de los sensores según esas técnicas.

1.3.1. Sensores que se utilizan con la técnica de RPL y EIS

En estas técnicas como elemento sensor se utiliza por lo general una celda electroquímica. La celda electroquímica es un dispositivo en el cual se puede convertir la energía química en energía eléctrica, o viceversa, al ocurrir un proceso redox. Típicamente, una celda electroquímica esta constituida por tres electrodos sumergidos en una solución electrolítica. De los tres electrodos uno de ellos el de trabajo (WE, Working Electrode) lo constituye el material de estudio. Los otros dos electrodos son el de referencia (RE, Reference Electrode) y el auxiliar (CE, Counter Electrode). Estas técnicas pueden utilizar también cuatro electrodos

siendo el cuarto electrodo otro electrodo auxiliar (CE), como en la que se muestra en la figura 1.

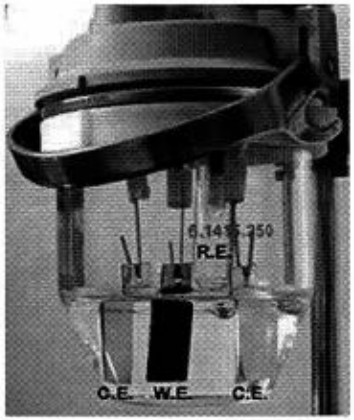


Figura 1. Celda electroquímica de cuatro electrodos.

1.3.2. SENSORES QUE SE EMPLEAN EN LA TÉCNICA DE RUIDO ELECTROQUÍMICO

En la bibliografía es posible encontrar diversos esquemas y dispositivos empleados para realizar las medidas mediante la técnica de EN (ver figuras 2 y 3). Entre ellos, el más extendido es denominado sistema de tres electrodos (ver figura 2). Este consiste en medir de manera simultánea las fluctuaciones espontáneas de la corriente entre dos electrodos de trabajo (WE, Working Electrode) idénticos que se corroen libremente y las fluctuaciones espontáneas del voltaje entre uno de esos dos electrodos de trabajo y un tercer electrodo que puede ser del mismo material o un electrodo de referencia [9-11]. Las fluctuaciones de potencial que tienen lugar durante un proceso de corrosión se conocen como ruido electroquímico de potencial (EPN). Por su parte, las fluctuaciones en la intensidad de corriente observadas cuando un metal se corroe libremente constituyen lo que se denomina ruido electroquímico de corriente (ECN) [5].

b) Análisis estadístico en el dominio temporal. En el análisis de señales de intensidad y voltaje han sido utilizados diferentes parámetros estadísticos como los valores medio, máximo y mínimo, los coeficientes de sesgo y de curtosis y la desviación estándar. Estos parámetros constituyen una forma sencilla de analizar el ruido electroquímico, especialmente con propósitos de vigilancia industrial. Además, es frecuente el uso de la denominada resistencia de ruido, Rn, que es un parámetro derivado de parámetros estadísticos. A partir de Rn es posible evaluar la velocidad de corrosión de la misma forma que se haría mediante la resistencia de polarización, Rp, obtenida mediante métodos de corriente continua [17].

c) Análisis en el dominio de frecuencias. Una magnitud ampliamente utilizada en bibliografía es la densidad de potencia espectral (PSD, Power Spectral Density) de las señales de ruido de corriente y potencial. La PSD se relaciona con la amplitud de las oscilaciones de una señal en función de las frecuencias de las mismas. Para ello, es necesario transformar los datos del dominio temporal, en el que son registrados, al de frecuencias. Esta transformación se realiza normalmente por medio de la Transformada de Fourier (FFT, Fast Fourier Transform) o el método de la Máxima Entropía (MEM, Maximum Entropy Method). Estos espectros son utilizados para extraer información relacionada con la cinética y el tipo de proceso de corrosión estudiado. Por otro lado, se denomina impedancia del ruido electroquímico, Zn, al cociente entre las PSDs de voltaje y de corriente. Esta magnitud esta relacionada con el modulo de la impedancia obtenida mediante Espectroscopia de Impedancia Electroquímica (EIS, Electrochemical Impedance Spectroscopy) [5].

d) Análisis basado en la Teoría del Caos. Estos métodos tratan de obtener información sobre el mecanismo de corrosión a través de un estudio del orden y la correlación de los datos. Normalmente, la información obtenida es de tipo cualitativa, ya que se evalúa la estructura de la señal y no la magnitud de la misma. Mediante este tipo de análisis ha sido posible establecer, por ejemplo, que la corrosión localizada es un proceso caótico relativamente simple, mientras que la corrosión uniforme es un proceso aleatorio [18].

e) Análisis basado en la Transformada de Wavelets. La Transformada de Wavelets es una extensión de la Transformada de Fourier que se adapta

mejor al estudio de señales no estacionarias como suele ser el caso en el ruido electroquímico. Recientemente, se han desarrollado diversas aplicaciones, basadas en la Transformada de Wavelets, que van desde el reconocimiento de patrones en señales de ruido electroquímico para distinguir distintos mecanismos de corrosión a la detección y caracterización automática de tránsitos [5].

2.3. METODOLOGÍA PARA EL ANÁLISIS DE LAS EMISIONES ACÚSTICAS

Las señales de emisión acústica cubren un amplio rango de frecuencias y niveles de energía, sin embargo, existen dos tipos básicos de señales: los estallidos cortos (burst), y las señales continuas de sonido (figura 6). Los estallidos sonoros son de corta duración que ocurren por eventos particulares; en cambio, las señales continuas se originan por eventos rápidos que permanecen durante un tiempo. Ambos tipos se propagan en una combinación de modos (longitudinal, transversal u ondas superficiales).

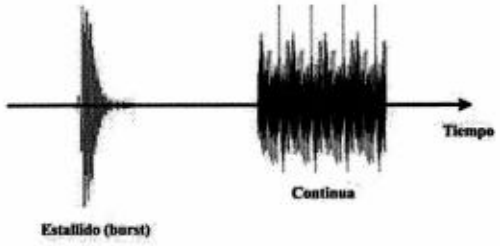


Figura 6. Señales típicas de emisiones acústicas

Las señales pueden variar en amplitud y frecuencia, pero de manera general se puede decir que las señales detectables están debajo de los 50 MHz; y el rango típico, entre los 20 y 1,200 kHz.

Cuando una emisión acústica es emitida por una carga inicial, la siguiente emisión no se producirá por una recarga, sino hasta que su valor exceda el de la carga inicial. Este fenómeno se denomina efecto Kaiser (curva BCB, figura 7). El factor de recarga, FR, es la razón del valor de la recarga al momento que se reinicia la emisión acústica entre el valor de carga antes de que sea liberada (FR=carga F/carga D figura 7). En general, valores de FR mayores o iguales a 1,0 indica que no hay daño antes de la carga; y valores menores de 1,0 denotan la presencia de daño.

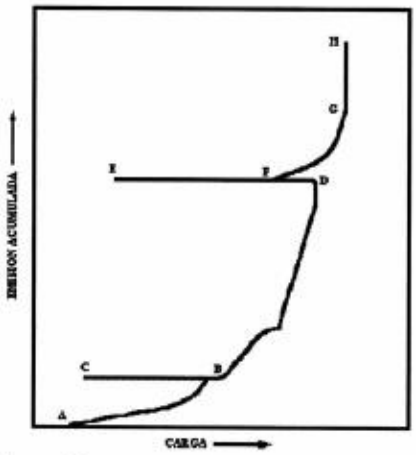


Figura 7.Historial de emisiones acústicas acumuladas con respecto a la carga. Nótense los efectos Kaiser BCB y de recarga FR

Existen cinco parámetros que se deben calificar en las señales de emisiones acústicas. En la figura 8 se muestra el intervalo de tiempo de formación, R, y el tiempo de duración de la señal, D; los cuales se determinan en función de un determinado valor del umbral, que generalmente se define para filtrar el nivel de ruido y eliminar las señales de baja intensidad. El valor pico de la amplitud, A, es el valor máximo de la señal, y determina el nivel de detección de una señal. El número de puntos de cruce N, es el total de veces que la señal cruza la línea del umbral. El área bajo la curva de la envolvente de la señal se identifica por el parámetro E, y se denomina, MARSE, el cual es una medida de la señal acústica. El MARSE es la forma mas aceptada para medir la una señal respecto al valor del número de cruces, N, ya que es sensible a la duración y amplitud de la señal, y depende menos del valor del umbral. Para medir el MARSE, el sistema de detección requiere una electrónica más compleja que lo calcule a partir de la señal obtenida.

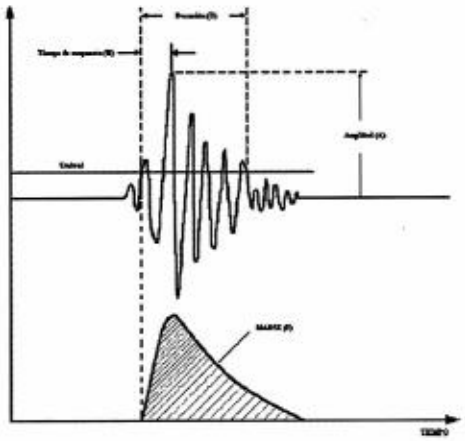


Figura 8. Parámetros típicos que describen una señal de emisiones acústicas

Existen varios formatos para presentar o desplegar la información procesada de las señales acústicas. En la figura 9 se muestran tres de los formatos más utilizados; el primero muestra el número de cuentas o energía acumulada en función del tiempo (figura 9a), y es útil para valorar la energía acústica total. El segundo corresponde al número de emisiones acústicas, o la energía acústica en función del tiempo (figura 9b), lo cual indica todos los cambios durante el tiempo de medición. En la figura 9c se muestra el tercer formato típico con una comparación de dos señales acumuladas de energía acústica en función de la carga; en este último ejemplo se identifica el caso de un buen material respecto a uno malo, y es el formato más empleado ya que relaciona las emisiones con el nivel de carga. Otras gráficas del nivel de energía acumulada como función de la carga corresponden a las figuras 10a y 10b

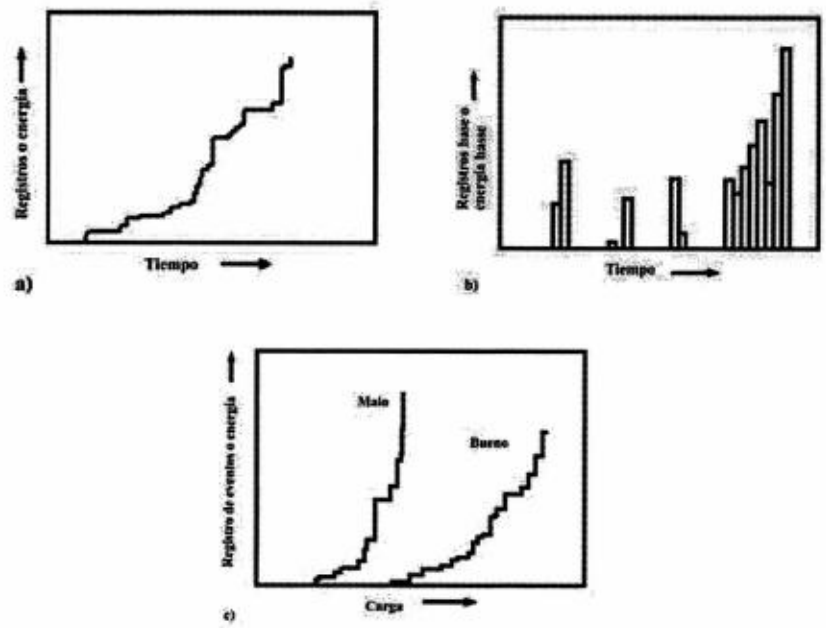


Figura 9. Formatos típicos para la presentación de resultados por AE
a) Energía acumulada vs tiempo
b) Razón de emisión vs tiempo
c) Energía acumulada vs carga

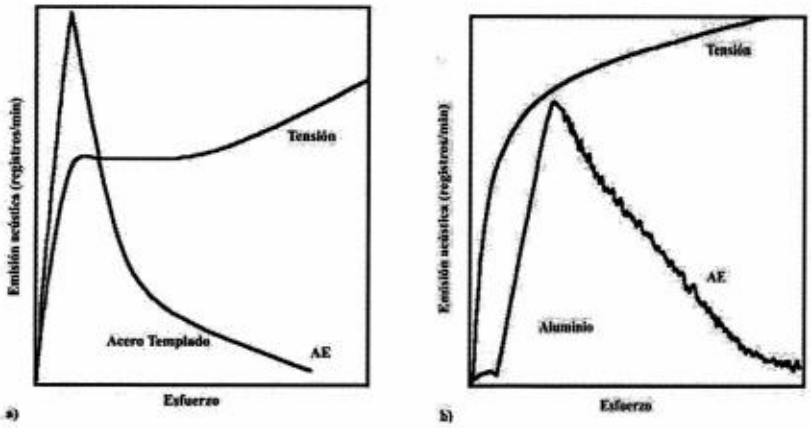


Figura 10. Gráficas de la energía de las emisiones acústicas vs esfuerzo
a) acero , b) aluminio.

La atenuación de las señales acústicas por la estructura del material mismo es un parámetro importante, ya que influye en el número y localización de los transductores, y se afecta por los mecanismos de dispersión y absorción de las ondas sonoras, y por factores geométricos (esquinas y discontinuidades). En términos generales, la dispersión en metales ocurre por pequeñas discontinuidades como precipitados o fronteras de grano; los mecanismos de absorción se presentan en dislocaciones, por efectos magnéticos y por amortiguamiento termoelástico; finalmente, los factores geométricos producen efectos de difracción, acoplamiento de ondas y atenuación por reflexión.

3. Resultados y Discusión

3.1. Selección de las técnicas a emplear en el desarrollo del sensor de corrosión

En general las técnicas electroquímicas para el monitoreo de la corrosión (LPR, EIS y EN) utilizan gran variedad de sensores. Se tienen reportes del uso de esas técnicas también para el monitoreo de la corrosión atmosférica. De todas ellas la que más se esta empleando en los últimos años es la técnica de EN debido fundamentalmente a que no se necesita señal de perturbación externa y permite detectar tanto la corrosión generalizada como la localizada desde prácticamente el inicio de su presencia. Sin embargo, esta técnica aún presenta limitaciones para su uso en el monitoreo de la corrosión atmosférica debido fundamentalmente a que los sensores, en este caso lo más utilizados que son los de multi-elementos en forma de sándwich, se deterioran con los cambios atmosféricos y en específico con la lluvia y la humedad. No obstante se continúan utilizando [ ].

En la literatura existen varios trabajos donde se emplea el uso de la AE para el monitoreo de la corrosión. Entre estos, en el trabajo [51] se utilizó la técnica de AE para la identificación de las señales generadas durante la propagación de fisuras por corrosión bajo tensión (intergranular y transgranular) en α-latón, expuesto en solución de NaNO2 1M y en solución de Mattsson. Para ello fue importante la selección de un umbral de trabajo adecuado que permitiera separar las señales de AE generadas por la propagación de fisuras por corrosión bajo tensión de aquellas provenientes de otros procesos que pueden ocurrir sobre la superficie del material, tales como: deformación elástica y plástica, corrosión generalizada, desarrollo de picaduras, evolución de burbujas, ruptura de películas superficiales y ruptura dúctil [52, 53].

La actividad de EA registrada durante la propagación de fisuras por corrosión bajo tensión transgranular es más de un orden de magnitud más alta que la actividad de AE registrada durante la propagación de fisuras por corrosión bajo tensión intergranular. El método utilizado para lograr la medición de señales de AE debidas exclusivamente a la fisuración por corrosión bajo tensión, fue seleccionar un umbral de trabajo para el cual la contribución a la actividad acústica de todos los otros procesos fuera despreciable. Para seleccionar dicho umbral, se expusieron durante 1 hora muestras de latón-α en condiciones estáticas a los mismos medios corrosivos y a los mismos potenciales utilizados en los ensayos de corrosión bajo tensión. Simultáneamente, se midieron las señales de AE, generadas durante la exposición utilizando diferentes umbrales de detección de la señal. La Tabla 1 y la Figura 11 muestran los resultados obtenidos.

Tabla 1. Velocidad de eventos [n° de eventos cada 1000 s] para cada umbral de detección y medio de corrosión.

Umbral	125 mV	150 mV	175 mV	200 mV	250 mV
$NaNO_2$	2334,3	337,8	24,5	17,2	9,6
Mattsson	1376,3	367,5	24,4	4,0	1,6

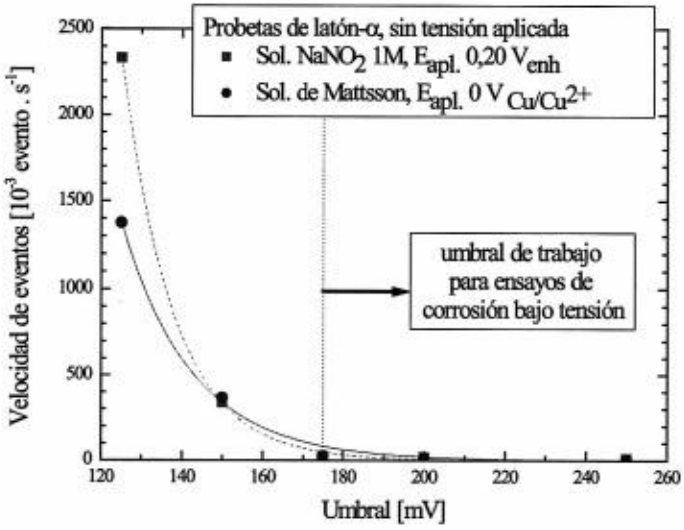


Figura 11. Velocidad de generación de eventos de emisión acústica en función del umbral de detección para los ensayos estáticos.

Se observa que cuando se trabaja a umbrales superiores a 250 mV la velocidad de eventos de EA es inferior a 0,01 eventos por segundo, lo que indica que para esos valores de umbral la probabilidad de adquirir señal de AE proveniente de procesos diferentes a corrosión bajo tensión es muy baja. Sin embargo, también debe tenerse en cuenta que al trabajar a umbrales cada vez más altos, se puede perder información sobre los eventos generados por la propagación de fisuras de corrosión bajo tensión. Por este motivo, para la medición de las señales generadas durante los ensayos de corrosión bajo tensión se seleccionó un umbral de detección de 200 mV.

Es importante destacar que por debajo del umbral seleccionado en ese trabajo que fue de 200mV es donde se encuentra la información sobre los otros procesos de corrosión tales como la corrosión generalizada y la localizada (picaduras, grietas, etc). Al igual que para el latón- α, los demás materiales tienen en el caso de la AE sus umbrales de trabajo. Sin embargo, estos umbrales están seleccionados a partir de la amplitud en mV de la señales de respuestas de los sensores. Aunque en ese trabajo no se presenta, la respuesta de los sensores también se puede analizar desde el punto de vista de distribución en frecuencias utilizando técnicas como la Transforma de de Fourier o Wavelets. En este sentido se puede utilizar sensores más sensibles como los que se utilizan en la técnica de análisis de vibraciones mecánicas los cuales permiten detectar con una alta sensibilidad las vibraciones que ocurren dentro del material.

4. CONCLUSIONES Y RECOMENDACIONES

El estudio presentado en el presente trabajo nos permite entender el estado actual del desarrollo de los sensores de corrosión y las técnicas que se utilizan así como las ventajas y desventajas de cada una de ellas. Basado en ello, vemos la necesidad de continuar trabajando en este sentido fundamentalmente en el desarrollo de nuevos sensores más confiables, sensibles y que puedan trabajar en condiciones de fuertes cambios climáticos. Por otra parte, las potencialidades que ofrecen actualmente las técnicas de AE y análisis de vibraciones mecánicas pueden ser la base para el desarrollo del sensor.

Por tanto, como recomendación proponemos que se continúe trabajando en el desarrollo de un sensor de corrosión en base a resultados que se obtengan tras un análisis de frecuencias a las señales de AE y de las vibraciones que se monitoreen en diferentes materiales bajo diferentes tipos de procesos de corrosión.

BIBLIOGRAFIA

[1] Control de la corrosión: Estudio y medidas por Técnicas Electroquímicas; José Antonio Fernández González. Consejo superior de investigaciones científicas, Centro nacional de investigaciones metalúrgicas, Madrid, 1989.

[2] Técnicas Electroquímicas para el estudio de la corrosión; Juan Genescá LLongueras, José M. Malo Tamayo, Yunny Meas Vong, Jorge Uruchurtu Chavarín; Primera jornada sobre técnicas electroquímicas para el control y estudio de la corrosión; XVI Congreso Sociedad Mexicana Electroquímica, Querétaro 2001.

[3] D. A. Eden and A.N. Rothwell, Electrochemical Noise data: Analysis, Interpretation and Presentation, Corrosion 92, Paper 292, NACE, Houston (1992).

[4] Jeffery. R. Kearns, David. A. Eden, Max. R. Yaffe, Jefferson. V. Fahey, David. L. Reichert, and David C. Silverman, ASTM Standardization of Electrochemical Noise Measurement, Electrochemical Noise Measurement for corrosion Applications, ASTM STP 1277, pp 446-470 (1996).

[5] Álvaro Aballe Villero, Francisco Javier Botana Pedemonte Y Mariano Marcos Barcena; Ruido Electroquímico. Métodos de Análisis; Septem Ediciones, Oviedo, 2002.

[6] Maria Luisa Cerón, Andrés Soto - Lubert; Elementos de Electroquímica; Julio 2004.

[7] J.L. Dawson, Electrochemical Noise Measurement for Corrosion Applications, Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 3-35.

[8] G.C. Barker, Journal of Electroanalytical Chemistry 21 (1969) 127.

[9] U. Bertocci y J. Krugger, Surface Science 101 (1980) 608.

[10] S.P. Mattin y G.T. Burstein, Proc. 10th European Corrosion Congress, Progress in the Understanding and prevention of corrosion Vol.2, London, EE.UU., 1993, Institute of Material, p. 1109.

[11] I. Al-Zanki. J.S. Gill y J.L. Dawson, Mater. Sci. Forum 8 (1986) 463.

[12] D.A. Eden, A.N. Rothwell y J.L. Dawson, Proc. Corrosion/91, paper 444, Houston, EE.UU., 1991, National Association of Engineers.

[13] D.A. Eden, A.N. Rothwell, Proc. Corrosion/92, paper 292, Houston, EE.UU., 1992, NACE.

[14] J. Gollner, I. Garz y K. Meyer, Korrosion 17 (1986) 244.

[15] K. Nachstedt y K.E. Heysler, Electrochim. Acta 33 (1988) 311.

[16] D.E. Williams, Proc. Electrochemical Corrosion Testing Ferrara, Italy, 1985, Dechema.

[17] J. Chen y W.F. Bogaerts, Corros. Sci. 37 (1995) 1839.

[18] P.R. Roberge, Electrochemical Noise Measurement for Corrosion Applications, ED. ASTM., West Conshohocken, EE.UU., 1996, pp. 142-156.

[19] Mas allá de la herrumbre Javier Ávila /Joan Genezcá. Primera edición, 1987. Segunda reimpresión, 1996. La ciencia para todos es proyecto y propiedad del Fondo de Cultura Económica, al que pertenecen también sus derechos. Se publica con los auspicios de la Subsecretaría de Educación Superior e Investigación Científica de la SEP y del Consejo Nacional de Ciencia y Tecnología. D.R. © 1986, FONDO DE CULTURA ECONÓMICA, S. A. DE C. V. D. R. © 1995, FONDO DE CULTURA ECONÓMICA Carretera Picacho-Ajusco 227; 14200 México, D.F.
ISBN 968-16-2396-7 Impreso en México

[20] National Instruments,http://www.ni.com. Recursos Web (ni.com). Zona de Desarrollo NI (zone.ni.com). Notas de Aplicaciones. Grupo de noticias labview (www.info-labview.org/). Instrument Driver Library (www.ni.com/idnet)

[21] ISO, Corrosion of metals and alloys. Basics terms and definitions (ISO 8044:1999), 1999.

[22] G. Schmitt and S. Feinen "Effect of anions and cations on the pit initiation in CO2 corrosion of iron and steel", in proceedings of corrosion 2000, National Association of Corrosion Engineers, 2000.

[23] Ives, J. D. and Janz, G. J., References Electrodes, Theory and Practice, Academic Press, New York, 1961.

[24] Bendat, J.S. and Piersol, A.G., Random Data: Analysis and Measurement Procedures, John Wiley & Sons, New York, 1986.

[25] Box, G.E.P. and Jenkins, G.M., in Time Series Analysis: Forecasting and Control, G.M. Jenkins and E. Parzen, Eds., Holden-Day Series in Time Series Analysis, Holden-Day, Inc., san Francisco, CA, 1970.

[26] Otnes, R.K. and Enochson, L., Applied Time Series Analysis: Basic Techniques, J. Wiley & Sons, New York, 1978, Vol 1.

[27] Andersen, N., "On the Calculation of Filter Coefficients for Maximum Entropy Spectral Analysis," Geophysics, Vol. 39, No. 2, 1974, pp. 69-72.

[28] Sabino Menolasina. Fundamentos y aplicaciones de electroquímica., Universidad de los Andes Consejo de Publicaciones Mérida – Venezuela 2004.

[29] R.A. Cottis, S. Turgoose y J. Mendoza-Flores, Electrochemical Noise Measurement for Corrosion Applications. Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 93-100.

[30] U.Bertocci, C. Gabrielli, F. Huet, y M. Keddam, J.lectrochem. Soc. 144 (1997) 31.

[31] D. A. Eden, Proc. Corrosion/98, paper 386, Houston, EE.UU., 1999, NACE.

[32] J.R. Kearns, D.A. Eden, M.R. Yaffe, J.V. Fahey, D.L. Reichert y D.C. Silverman, Electrochemical Noise Measurement for Corrosion Applications, Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 446-470.

[33] M.S. Al Ansari y R.A. Cottis, Proc. The 13th International Corrosion Congress, Vol. 2, Clayton, Australia, 1996, Australian Corrosion Assoc., pp 212/1-212/6.
[34] F. Mansfeld y H. Xiao, Electrochemical Noise Measurement for Corrosion Applications, Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 59.
[35] C.C. Lee, C.C. y F. Mansfeld, Corr. Sci. 40 (1998) 959.
[36] F. Mansfeld y H. Xiao, J. Electrochem. Soc. 140 (1990) 2205.
[37] P.C. Pistorius, Electrochemical Noise Measurement for Corrosion Applications, Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 343-358.
[38] R.G. Hardon, P. Lambert y C.L. Page, British Corr. Jour. 23(4) (1988) 225
[39] D.L. Reichert, Electrochemical Noise Meausrement for Corrosion Applications, Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 79-89.
[40] A. Bautista, L. Mariaca, P. Rodríguez y J.A. González, Proc. 50 Congreso Iberoamericano de Corrosión y Protección, II-oral-01, Tenerife, España, 1995, Universidad de La Laguna.
[41] F. Huet, U. Bertocci, C. Gabrielli y M. Keddam, Proc. Corrosion/97, Advanced Monitoring and Analytical Tachniques, New Orleans, EE.UU., 1997, NACE, p.11.
[42] D.E. Williams, Proc. Electrochemical Corrosion Testing Ferrara, Italy, 1985, DECHEMA.
[43] J. Chen y W.F. Bogaerts, Corrs. Sci. 37 (1995) 1839.
[44] D.A. Eden, D.G. Johm y J.L. Dawson, Proc. Corrosion/86, paper 274, Houston, EE.UU., 1986, NACE.
[45] U. Bertocci y F. Huet, Corrosion 51 (1995) 131.
[46] C. Gabrielli, F. Huet y M. Keddam, Electrochim. Acta 31 (1986) 1025.
[47] R.G. Nelly, M.E. Inman y J.L. Hudson, Electrochemical Noise Measurement for Corrosion Applications, Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 101-113.
[48] Y.J. Tan, S. Bailey y B. Kinsella, Corrosion 55(5) (1999) 469.
[49] J. de Damborenea y B. Fernandez, Electrochemical Noise Measurement for Corrosion Applications, Ed. ASTM., West Conshohocken, EE.UU., 1996, pp. 348-410.
[50] Alexander M. Lowe, Estimation of Electrochemical Noise Impedance and Corrosion Rates form Electrochemical Noise Measurements. January 2002, pp.
[51] P. A. Lapitz1,3, J. Ruzzante2 y M. G. Alvarez1. Análisis de la Amplitud de las Señales de Emisión Acústica Generadas por Corrosión Bajo Tensión en Latón-α.
[52] REBAK, R., CARRANZA, R., GALVELE, J., "The SCC Mechanism of α-brass in NaNO2 Solutions", Corrosion Science, v. 28, n. 11, pp. 1089¬1106, 1988.

[53] LAPITZ, P., ALVAREZ, M.G., FERNANDEZ, S., GALVELE, J.R, "Passivity Breakdown and Stress Corrosion Cracking of α-brass in Sodium Nitrite Solutions", Corrosion Science, v. 47, pp. 1643¬1652, 2005.

13. Biosensor de glucosa con fibra óptica

Arturo Medina Puente

Resumen

Biosensor de fibra óptica para la detección de concentración de glucosa se ha diseñado, basado en el método electrostática mismo-asamblea. Al final de la fibra óptica multimodo es cubierta con poly (allylamine hydrochloride) en combinación con prussian blue y la enzima glucosa oxidase (GOx). La concentración de glucosa puede ser medida entre 0.06 y 2 milímetros. La velocidad inicial del cambio de la señal del biosensor de salida se ha encontrado para ser lineal en la concentración de la glucosa. El sensor se recupera después de la inmersión en ácido ascórbico. Los valores de pH analizados entre el rango de 5 y 7. La sensibilidad del dispositivo ha sido aumentada ajustando el número de bilayers y agregando algunos bilayers que capsulaban en el extremo de la estructura.

1. Introducción

La detección de glucosa es probablemente uno de los campos en los cuales incluye un gran número de métodos, dependiendo de la aplicación. Hay dos grupos principales para la concentración de glucosa en sangre humana: los métodos invasivos y noinvasivos. [1,2] El Método noinvasivo se basa en permitir que la luz infrarroja penetre una región del tejido fino del cuerpo humano para excitar las moléculas de la sangre y usar el método de calibración para procesar el espectro de absorción. Este método no es apropiado para otras aplicaciones diferentes para la detección de glucosa en sangre a través de la piel humana y este trabajo se enfoca en el otro grupo .

De esta manera, dependiendo del rango de la concentración de glucosa estas son detectadas y que la necesidad del sensor debe ser inmune a interferencias electromagnéticas y que además el sensor también podrá se utilizados para diferentes campos como son: en análisis clínico, alimentaría y farmacéutico.

Los métodos invasivos para la detección de glucosa pueden ser subdivididos en dos grandes grupos que son: la detección electroquímica [3-6] y espectroscopia [7-12]. Entre estos métodos se ha seleccionado el método espectroscopia porque evita el uso de voltaje y el uso de membranas selectivas. El método espectroscopia también puede ser subdividido en quimioluminesencia [7], espectrofluorometria [8,9] y espectrofotometría [10,11]. Este último método se basa en los cambios de absorción experimentados por el sistema redox-sensible. El Prussian White (PW) puede ser oxidad por Prussian Blue (PB) por el peroxido de hidrogeno (H2O2). Por lo tanto, si la glucosa oxidasa (GOx) es también depositada en la estructura, esto generara (H2O2) y glucolactone en la presencia de glucosa y oxigeno [10]. El producto de la reaccion anterior H2O2, oxidara PW a PB. Este efecto se ha utilizado con éxito en estos sensores, y los cambios en la absorbencia son importantes en un amplio rango de longitud de onda, por lo menos entre 600 y 1000 nm. Esto incluye la primera ventana de comunicación óptica, donde son útiles los detectores y las fuentes por sus bajos costos. Para reducir el PB a PW se aplica un agente reductor tal como el acido ascórbico. El proceso puede ser repetitivo y los resultados pueden ser productivos.

Con respecto a la técnica de deposición, electrostatic layer-by-layer (LBL) se ha seleccionado el método de self-assembly. Este es una método basado en la construcción de multicapas moleculares por la atracción electrostática entre cargas opuestas de poliectrolitos en cada deposito monocapa, e involucra varios pasos [13]. Se utiliza para acumular una gran variedad de capas como substratos tales como cerámica, metales y plomimeros de diversas formas, incluyendo substratos planos, prismas los lentes cóncavos y convexas. En nuestro caso, la fibra óptica sirve como un substrato, puesto que presenta cualidades importantes tales como inmunidad contra interferencias electromagnéticas y la posibilidad de multiplexor sensores, y de que son de tamaño pequeño. Hasta ahora algunos ejemplos de sensores de fibra óptica son fabricados con esta técnica [14-16].

Otros sensores de glucosa se basan en electropolymerizacion o la técnica del sol-gel [5,7]. Sin embargo esta técnica presentada en este trabajo ofrece algunas ventajas. El sistema sensible puede ser utilizado en diversas estructuras de polímero. Esto debe permitir al seleccionar una estructura de polímero para evitar la sensibilidad cruzada causada por la otra especie molecular. Además, el pequeño grosor del nanocavity permite usar un LED simple como fuente de luz para la detección del circuito, cual ahorrara costos.

Fue probado en la referencia [17] que la solución tope influencia la liberación del tinte en una estructura depositada por electrolytic salf-assemble (ESA). Dos estrategias se pueden tomar en cuenta, el primero es un tratamiento térmico. En la Ref. [11] probamos que aunque a temperaturas muy altas de 90°C la enzima

irreversible se desnaturaliza y destruye su actividad, un sensor fue probado con el tratamiento térmico a 90°C y exhibe una sensibilidad de glucosa que es similar a la del sensor untreated pero se muestra un buen comportamiento después de la inmersión en la solución. La segunda opción, que se considera en este trabajo, es agregando bicapas de otros materiales que encapsulan al sensor [17,18]. En la referencia. [18] se sugiere para construir una heteroestructura integrada por el suplemento [PAH+/PAA-] y [PAH+/PSS-] bicapas.

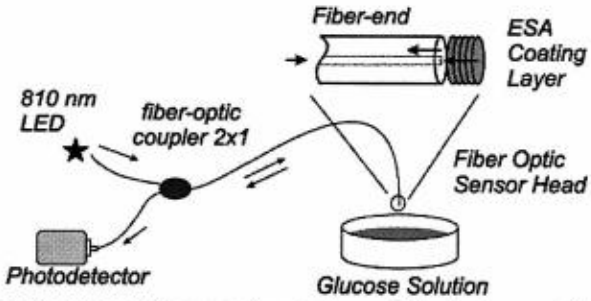


Fig. 1 Configuración experimental para detectar la energía óptica reflejada por la cavidad de nano Fabry-Perot.

De esta manera, el poly (el ácido de acrílico) (PAA) es un polyanion débil que atrapa a los nanoparticles de plata cargados positivamente de AG. Sin embargo, el poly(sulfonate de estireno) (PSS) es un polyanion fuerte y está viene limitado totalmente al poly (allylamine clorhidrato) (PAH), de modo que no pueda atrapar a los nanoparticles de AG. Si esta teoría se extrapola al atascamiento del PB cargado negativamente, un polycation fuerte debe ser el recurso para que las capas encapsulen. Esta segunda opción tiene la ventaja que no influencia la parte biológica del sensor y puede ser aplicada a otras enzimas con un rango de temperatura baja estable que la oxidasa de glucosa. Por otra parte, hay una mejora importante de sensibilidad, tal como sucedido con el mismo procedimiento aplicado a un sensor de peróxido de hidrógeno [19,20].

Las concentraciones de glucosa pueden ser monitoreadas estas entre 0.06 y 2 mM. Sin embargo, si se desea una aplicación específica, este rango puede ser extendida, dependiendo del número de bilayers depositadas y de las técnicas usadas para mejorar la sensibilidad. El pHs analizado entre el rango de 5 y 7, y la respuesta del sensor es lineal en la concentración de la glucosa y en todos estos valores de pH. Otros valores de pH más extremos podrían ser analizados, pero es confirmado que en especificaciones de glucosa oxidasa en el rango de actividad de esta enzima está entre pH 4 y 7, con un grado óptimo en pH 5.5.

2. EXPERIMENTALES

2.1 REACTIVOS

El polycation PAH y el polyanion enzima glucosa oxidasa GOx de Aspergillius Niger de tipo II (21) se obtuvo de Aldrich. Soluble (PB)(C6Fe2KN6 * xH2O), fue obtenido de Riedel-de Haën AG. El phosphate y el acetato soluciones tope eran usados para fabricar soluciones tope en Phs específicos. Ácido hidroclórico (HCI) e hidróxido de sodio (NaOH) fue utilizado para ajustar el pH de las soluciones de polycation y del polyanion. Toda el agua fue filtrada a través de un sistema de purificación Millipore Q plus 185.

2.2 INSTRUMENTACIÓN

Las medidas de la energía fueron hechas usando un esquema de reflexión donde estaba una fuente de luz Hewlett-Packard 9537 HFBR 1424 LED en 810 nm y un Photodetector de Ophir LaserStar detecta las medidas reflejadas de la energía en la primera ventana de las comunicaciones ópticas, según lo demostrado en fig. 1. El pH de las diferentes soluciones fue ajustado con un Crison GLP22 pH metros. Para obtener el espectro de PB depositado en la estructura del polímero, una fuente blanca de halógeno y un espectrómetro de Avantes AVS S2000 substituyen al LED y al fotodetector, respectivamente, en el esquema de la reflexión de Fig. 1.

2.3 EL PROCEDIMIENTO DE LA INMOVILIZACIÓN

Los materiales usados para la deposición de capas catiónicas y aniónicas es PAH y la enzima GOx, respectivamente. El método usado para incluir el PB entre la estructura polimérica es premezclado con el polycation, (15) que consiste en mezclar un tinte y un polyion con carga opuesta. Todas las soluciones se ajustan a pH 5. El propósito de seleccionar este valor de pH es evitar la degradación de la enzima GOx. Después de concluir en la construcción de la estructura LBL es agregado un pigtail al final de la fibra que es un procedimiento importante asociado a la durabilidad del sensor en soluciones agresivas. Sin embargo, GOx es una enzima, que implica que se degradada a temperaturas altas. Por lo tanto, no se realizó ningún endurecido.

2.4 EL MÉTODO DE DETECCIÓN

Para cada medida se repite el mismo procedimiento: el sensor se sumerge en un agente de reducción, tal como la solución ácida ascórbico, para reducir el PB a PW. Después de esto, el sensor se sumerge con un tapón en la solución. Cuando se estabiliza la señal, la glucosa se inyecta en el almacenador intermediario, que en la presencia de GOx y del oxígeno genera el peróxido de hidrógeno (H2O2) como producto de la reacción. Entonces el H2O2 oxida el PW a PB.

Entonces en la energía óptica reflejada se produce un cambio. Este proceso se puede repetir para otras medidas, que indica que es un sensor de multiples usos.

El tiempo de reacción del sensor en función de la concentración de la glucosa es exponencial. Evitar este problema, una mejor opción es medir la velocidad inicial de la reacción enzimática, que es proporcional a la concentración de la glucosa. Se propone la medida siguiente: la diferencia de la energía entre el 90% y el 10% del cambio completo del nivel de la señal, dividido por el momento en que es transcurrido entre estos dos valores. Aunque otros métodos de proceso de datos más complicados pueden dar resultados más exactos, el cálculo de la cuesta muestra buena linealidad como función de la concentración. Henceforward se refiere, por simplicidad, como la cuesta del cambio óptico de la energía reflejada detectada en el esquema de la reflexión.

3. RESULTADOS Y DISCUSIONES

La selección del PB se basa en el uso de sus características redox para la detección de H2O2 en análisis electroquímico y spectroscopy. [10,19,20,22,23] por otra parte, GOx es ampliamente usado de enzima debido a su alta estabilidad y rango de operación del pH: 4 a 7 con un grado óptimo en 5.5. Además, su costo es competitivo con muchas otras enzimas.

En relación con la absorción del PB, en Refs. 10 y 20 se afirma que la máxima longitud de onda está situada en 720 nm. Esto conviene con el máximo de este tinte sea cubierta en la solución. Sin embargo, si un indicador se inmoviliza en una matrix, un cambio puede ocurrir en el máximo de absorción, como se ha divulgado en la referencia. 24.

En este caso, en vez de (PAH+/PAA-) de la matriz en la cual el PB fue introducido en Refs. 19 y 20, la matriz se compone de (PAH+/GOx-).

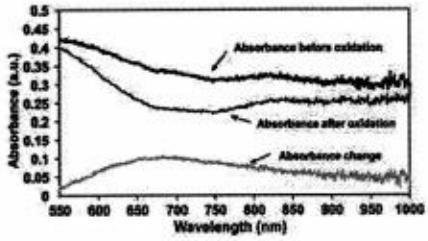


Fig. 2 absorbencia del espectro oxidación antes y después del PAH (c+pb+/GOx) estructura con glucosa. La diferencia entre los dos diagramas e incluye también.

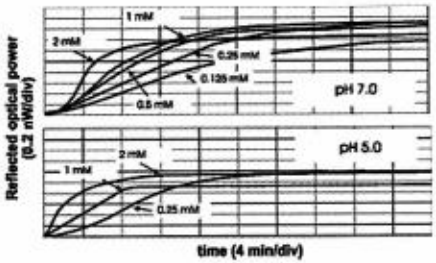


Fig. 4 La energía óptica reflejada para diversas concentraciones de glucosa en los pHs constantes de 5 y 7. Estructura del sensor: Pah (c+pb+/GOx) 23.

Por esta razón, en fig. 2 la absorción del espectro se demuestra para el estado reducido y oxidado del sistema de PW-PB. La diferencia de absorción entre los dos estados también se incluye. En este caso se ha experimentado un cambio, en contraste con los resultados de la referencia. 20. El máximo de absorción esta situado en 680 nm y no en 720 nm. De hecho, hay una amplio rango entre 600 y 1000 nm donde no se considera ninguna variación importante de la absorción. Esto permite que el uso de cualquier dipositivo óptico emitido en esta región. Los fotodetectores de 810 nm son una opción adecuada, que reduce el mínimo de costos. Ésta es la razón de seleccionar esta fuente del LED en nuestros experimentos.

Las medidas podrían ser mejoradas usando un espectro entero de la absorción obtenido con un analizador de espectro de sistema óptico, pero los costos del dispositivo aumentarían obviamente.

Se analizan dos sensores diferentes. El Primero se compone de 23 PAH+ PB+/GOx- bi capas. En la Fig. 3 el cambio de energía-reflejada producida por la disposición de la estructura LBL en los extremos de las dos fibras se muestra un pigtail.

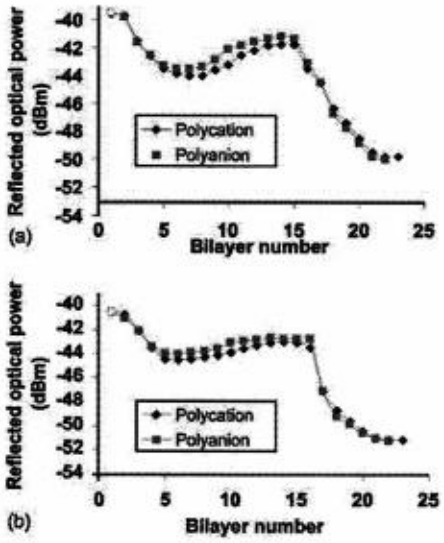


Fig. 3 reflejó energía en dos sensores construidos en paralelo en función del número de bilayers depositado. Se representan los diagramas positivos y negativos del polyion

Los sensores han sido construidos en paralelo sumergiéndolos en la misma solucion positiva y negativa. Los dos diagramas son muy similares el uno al otro. Por lo tanto, puede ser afirmado que la estructura está construido correctamente si se depositan 23 bi capas. Más bi capas podrían ser agregados, mejorarían la sensibilidad del sensor debido a la cantidad alta de GOx y PB depositada. Sin embargo, podría conducir a una estructura construida menos estable al final de la fibra. (19) Así que una compensación entre la sensibilidad y la robustez tenía que ser alcanzada, que condujeron a la selección de 23 bi capas.

En la fig. 4 la respuesta del sensor es presentada por cinco diferentes concentraciones de glucosa en pH 7 y otras tres en pH 5. Las concentraciones analizadas entre el rango 0.125 y 2 mM. Para realizar cada medida, el procedimiento explicado en el Sec. 2.4 fueron repetidos. Cuando una concentración de glucosa se introduce en la solución tope, la energía óptica reflejada se incremento gradualmente hasta que se estabilizo. La estabilización es debido a la disminución de la cantidad de disponible de PW para la reacción. Después de esta estabilización el sensor es tratado con un agente reductor (ácido ascórbico) para reducir el PB. Después de que, el sensor es sumergido en la solución tope y recupera esta energía original óptica reflejada antes de que una nueva inyección de glucosa

se realice. La cuesta del cambio de la reflejar-óptico-energía detectada en el esquema de reflexión (Sec. 2.4) se utiliza para calibrar el sensor de la glucosa. Esta cuesta conduce a una buena linealidad en función de la concentración de glucosa. El procedimiento explicado para medir la concentración de glucosa se ha repetido para otros ejemplos mostrados en la fig. 4, que indica que es un sensor de múltiple uso. Se ve que la cuesta del cambio en la energía óptica reflejada aumenta con la concentración de glucosa. La acumulación del producto de la reacción GOx es lineal en tiempo, y, en consecuencia, como la concentración aumenta, la acumulación del producto se convierte más rápida y la cuesta se incrementa.

En fig. 5 la linealidad de la cuesta para dos pHs diferentes son analizados. Los valores del coeficiente correlación R2 son indicados, y están cerca de uno. Además, también se prueba que hay una dependencia de pH, al igual que en el caso con el sensor H2O2 de la referencia. 20.

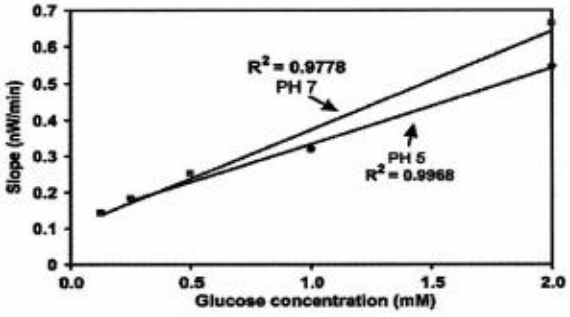


Fig. 5 Slope of the change in the reflected power for different concentrations of glucose at constant pHs of 5 and 7. Sensor structure: _PAH+PB+ /GOx−_23

Para probar que la sensibilidad del sensor no disminuye después de almacenar el sensor a una temperatura de aproximadamente 25° C en contacto con el aire, otro sensor es analizado en la fig. 6 con las mismas características de los sensor presentados en la fig. 5. De hecho, los dos sensores fueron fabricados en paralelo. Las medidas en pH 7 fueron tomadas tres semanas después de su construcción. La cuesta del cambio de energía óptico reflejada en función de la concentración es analizada, y en este caso el múltiple coeficiente correlación no está tan cerca de 1. Esto puede ser debido a las inexactitudes en la inyección de glucosa y del hecho de que la señal no era tan estable como en los sensores de la fig. 5 cuando las medidas fueron tomadas. Sin embargo, almacenar en condiciones ambientales el aire no causa importante variación en los resultados, como es pro-

bado comparando los dos sensores construidos en paralelo (con y sin almacenar) en las fig. 5 y 6.

Para reducir la influencia de la solución tope se libera el tinte en una estructura depositada con el método ESA, las bi capas cubiertas son de otro material y fueron agregados. El polycation poly (diallydimethyl ammonium chloride) (PDDA) fuerte fue seleccionado como el polycation y el GOx como el polyanion para las bi capas encapsuladas. Resultados con un sensor de 27 bi capas de [PAH + PB+/GOx-] y encapsulación en el extremo de la estructura de 2 bi capas de [PDDA+/de GOx-] presenta la sensibilidad mejorada, con valores debajo de 0.1 mM. Si únicamente se deposita una bi capa , hay una pérdida más rápida de sensibilidad, y cuatro o seis bi capas no permiten detectar la glucosa de una manera tan sensible. En la fig. 7 la respuesta de este sensor en pH 7 se presenta para diferentes concentraciones de glucosa. El rango analizado (0.06 a 2 mM) mejora en el sensor sin cubiertas de bi-capas. En la esquina derecha superior de fig. 7 la forma lineal de la cuesta del cambio de la energía reflejada en función de la concentración de glucosa se prueba otra vez. La línea de la tendencia es mucho más precipitada que las de fig. 5 y de la fig. 6, donde no se agregó ninguna cubierta de capas. Después de estas medidas, el sensor fue probado otra vez en pH 7. Once días después de su construcción y después de dos experimentos, puede ser visto en fig. 8 que el sensor todavía puede medir concentraciones de glucosa, pero con una sensibilidad más baja debido a la lixiviación del PB. En conclusión, la sensibilidad se mejora con la cubierta de bi capas, pero todavía hay una pérdida de sensibilidad después de muchos experimentos.

En adición a esto, para reducir la sensibilidad óptica cruzada del sensor son analizadas otras sustancias bien conocidas que pueden interferir con el sensor de glucosa. Las concentraciones 2.5-mM de cloruro, de bromuro, de yoduro, de oxalate, de thiocyanate , de urea, de sulfato, y el fosfato eran analizados, y ningún cambio fue detectado. Solamente ácido ascórbico como reductor del PB, y H_2O_2 es una interfase oxidante.

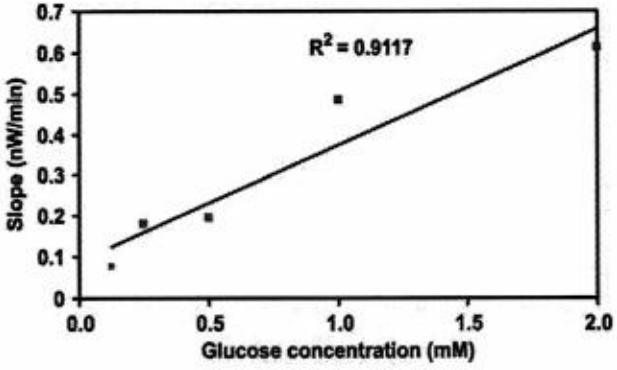


Fig. 6 cuesta del cambio en la energía reflejada para diversos centrations conde la glucosa en pH constante 7. El sensor 1 fue medido tres semanas después de su construcción, y el sensor 2 el día después de su construcción. Ambos sensores tienen la estructura (PAH + PB+/GOx-)[23].

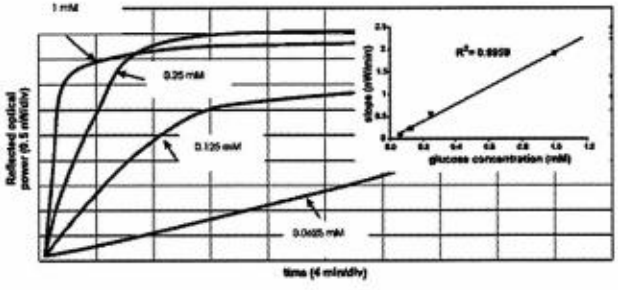


Fig. 7 reflejó la energía óptica para diversas concentraciones de glucosa en pH constante 7. Se han agregado dos bilayers que capsulaban. Estructura del sensor: (PAH+PB+/GOx-)[27] + (PDDA+/GOx-)[2].

Finalmente, es importante comentar una característica que sigue habiendo una confusión. En teoría, la detección de glucosa viene como resultado de su oxidación debido a la presencia de GOx en la extremidad de la fibra óptica, y de la generación consiguiente de H2O2 que oxide el PB. Considerar que muchas enzimas tienen una vida corta de almacenamiento a temperatura ambiente, el hecho de sensor almacenado a la temperatura ambiente todavía

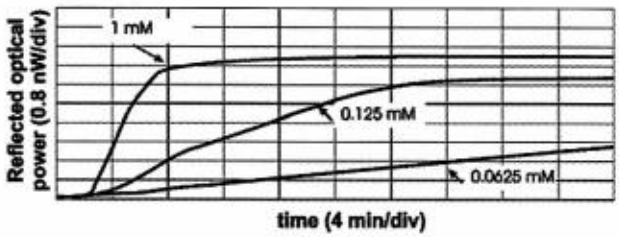


Fig. 8 reflejó la energía óptica para diversas concentraciones de glucosa en pH constante 7. Se han agregado dos bilayers que capsulaban. Medidas tomadas después de dos experimentos y de 11 días después de la construcción del sensor. Estructura del sensor: (Pah + PB+/GOx 27 + PDDA+/GOx) 2. Higo. Energía óptica reflejada 8 para diversas concentraciones de glucosa en pH constante 7. Se han agregado dos bilayers que capsulaban. Medidas tomadas después de dos experimentos y de 11 días después de la construcción del sensor. Estructura del sensor: (Pah + PB+/GOx [27] + PDDA+/GOx) [2].

detecta glucosa después de tres semanas es absolutamente sorprendido. Además de esto, algunos experimentos se han realizado en los sensores con bi capas que capsulaban de la referencia. 20, donde se ha probado que hay un cruce-sensibilidad con glucosa. Esto no fue visto en sensores sin bi capas que capsulaban. Como resultado, puede ser concluido que es posible que la detección de glucosa sea debido a la actividad combinada con la enzima oxidasa de glucosa y del azul prusiano en los valores de pH analizados.

4. CONCLUSIONES

Para concluir, un sensor nuevo de fiber-optic de glucosa ha sido diseñado aplicando el método de uno mismo-ensambla capa-por-capa. El polycation se compone de una premezcla del PAH y PB, y el polyanion contiene la enzima de GOx. La medida de las concentraciones de glucosa se basa en la cuesta del cambio de la energía reflejada mientras que este producto se inyecta en la solución. Esto permite que uno obtenga los resultados similares para una variedad de pHs entre 5 y 7. Esta rango cubre el rango predicho de operación de glucosa oxidasa y el aprovechamiento del pH de la sangre. Además, el sensor es sensitivo sobre un rango de concentraciones entre 0.06 y 2 mM. La recuperación del PB oxidado ha sido probada satisfactoriamente después de la inmersión en un agente de reductor tal como ácido ascórbico. Además de esto, el sensor es inmune a interferencias por una gran cantidad de productos a excepción del ácido ascórbico, que es el

agente de reductor de nuestro sensor, y de H2O2. Finalmente, para mejorar la sensibilidad del sensor, algunos bi capas se han agregado a la cubierta al extremo de la estructura. Esto reduce la lixiviación de PB y aumenta grandemente la sensibilidad.

5. REFERENCES

1. T. Koschinsky and L. Heinermann, "Sensors for glucose monitoring: technical and clinical aspects," Diabetes Metab. Res. Rev. 17, 113 _2001_.
2. T. W. King, G. L. Cote, R. McNichols, and M. J. Goetz, "Multispectral polarimetric glucose detection using a single Pockels cell," Opt. Eng. 33_08_, 2746–2753 _1994_.
3. J. Perdomo, H. Hinkers, C. Sundermeier, W. Seifert, O. Martínez Morell, and M. Knoll, "Miniaturized real time monitoring system for L-lactate and glucose using microfabricated multi-enzyme sensors," Biosens. Bioelectron. 15, 515 _2000_.
4. Y. Mishima, J. Motonaka, K. Maruyama, I. Nakabayashi, and S.Ikeda, "Glucose sensor based on titanium dioxide electrode modified with potassium hexacyanoferrate_III_," Sens. Actuators B 65, 343–345 _2000_.
5. T. D. Chung, R. A. Jeong, S. K. Kang, and H. C. Kim, "Reproduciblefabrication of miniaturized glucose sensors: preparation of sensing membranes for continuous monitoring," Biosens. Bioelectron. 16, 1079–1087 _2001_.
6. P. Se-Ik, B. J. Sang, J. Beom, P. Sejin, C. K. Hee, and S. J. Kim, "Application of a new Cl-plasma-treated Ag/AgCl reference electrode to micromachined glucose sensor," IEEE Sens. J. 3, 267–273 _2003_.
7. L. Qingwen, L. Guoan, W. Yiming, and Z. Xingrong, "Immobilization of glucose oxidase in sol-gel matrix and its application to fabricate chemiluminescent glucose sensor," Mater. Sci. Eng. 11, 67–70 _2000_.
8. V. Sanz, J. Galbán, S. de Marcos, and J. Castillo, "Fluorometric sensors based on chemically modified enzymes. Glucose determination in drinks," Talanta 60, 415–423 _2003_.
9. Z. Rosenzweig and R. Kopelman, "Analytical properties of miniaturized miniaturized oxygen and glucose fiber optic sensors," Sens. Actuators B 36, 475–483 _1996_.
10. T. Lenarczuk, D. Wencel, S. Glab, and R. Konscki, "Prussian bluebased optical glucose biosensor in flow-injection analysis," Anal. Chim. Acta 447, 23–32 _2001_.
11. I. Del Villar, I. R. Matias, and F. J. Arregui, "Nanosensor for detection of glucose," Proc. SPIE 5502, 259–262 _2004_.
12. Z. Gryczynski, I. Gryczynski, and J. R. Lakowicz, "Simple apparatus for polarization sensing of analytes," Opt. Eng. 39_09_, 2351–2358 _2000_.

13. G. Decher, "Fuzzy nanoassemblies: toward layered polymeric multicomposites," Science 277, 1232 _1997_.
14. F. J. Arregui, Y. Liu, I. R. Matias, and R. O. Claus, "Optical fiber humidity sensor using a nano Fabry-Perot cavity formed by the ionic self-assembly method," Sens. Actuators B 59, 54–59 _1999_.15. P. S. Grant and M. J. McShane, "Development of multilayer fluorescent thin film chemical sensors using electrostatic self-assembly," IEEE Sens. J. 3, 139–146 _2003_.
16. F. J. Arregui, I. Latasa, I. R. Matias, and R. O. Claus, "An optical fiber pH sensor based on the electrostatic self-assembly method," in Proc. 2nd IEEE Int. Conf. on Sensors _2003_.
17. A. J. Chung and M. F. Rubner, "Methods of loading and releasing low molecular weight cationic molecules in weak polyelectrolyte multilayer films," Langmuir 18, 1176 _2002_.
18. T. C. Wang, "Polyelectrolyte multilayers as nanostructured templates for inorganic synthesis," PhD Thesis, Dept. of Chemical Engineering, Massachusetts Inst. of Technology _2002_.
19. I. Del Villar, I. R. Matias, F. J. Arregui, and R. O. Claus, "ESA based in-fiber nanocavity for hydrogen peroxide detection," IEEE Trans. Nanotechnol. 4, 187–193 _2005_.
20. I. Del Villar, I. R. Matias, F. J. Arregui, J. Echeverría, and R. O.Claus, "Strategies for fabrication of hydrogen peroxide sensors based on electrostatic self-assembly _ESA_ method," Sens. Actuators B 108, 751–757 _2005_.
21. E. J. Calvo, E. Forzani, and M. Otero, "Gravimetric and viscoelastic changes during the oxidation-reduction of layer-by-layer self assembled enzyme multilayers wired by an Os-containing poly_allylamine _ polymer," J. Electroanal. Chem. 538–539, 231–241 _2002_.
22. A. Karyakin, E. E. Karyakina, and L. Gorton, "The electrocatalytic activity of Prussian blue in hydrogen peroxide reduction studied using a wall-jacket electrode with continuous flow," J. Electroanal. Chem. 456, 97 _1998_.
23. L. V. Lukachova, E. A. Kotel'nikova, D. D'Ottavi, E. A. Shkerin, E.Karyakina, D. Moscone, G. Palleschi, C. Antonella, and A. A. Karyakin, "Nonconducting polymers on Prussian blue modified electrodes: improvement of selectivity and stability of the advanced H2O2 transducer," IEEE Sens. J. 3, 326 _2003_.
24. A. Lobnik and M. Cajlakovic, "Sol-gel based optical sensor for continuous determination of dissolved hydrogen peroxide," Sens. Actuators B 74, 194 _2001_

Esta obra se terminó de imprimir en agosto del 2010
en los talleres de Ultradigital Press, S.A. de C.V.
Centeno 162 - 3, Col. Granjas Esmeralda
CP 09810, México, D.F.