PILOTO DE SISTEMA DE COMUNICACIÓN CON VoWLAN CONTROLADO POR COMANDOS DE VOZ PARA UN ENTORNO HOSPITALARIO

KAROL VIVIANA MOSQUERA LÓPEZ
CLAUDIA XIMENA MUÑOZ RODRÍGUEZ

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE TELECOMUNICACIONES
GRUPO I+D NUEVAS TECNOLOGIAS EN TELECOMUNICACIONES
POPAYÁN
2007

PILOTO DE SISTEMA DE COMUNICACIÓN CON VOWLAN CONTROLADO POR COMANDOS DE VOZ PARA UN ENTORNO HOSPITALARIO

KAROL VIVIANA MOSQUERA LÓPEZ
CLAUDIA XIMENA MUÑOZ RODRÍGUEZ

Trabajo de grado presentado como requisito para obtener el título de Ingeniero en Electrónica y Telecomunicaciones

Director
I.E. GUEFRY LEIDER AGREDO MÉNDEZ

UNIVERSIDAD DEL CAUCA
FACULTAD DE INGENIERÍA ELECTRÓNICA Y TELECOMUNICACIONES
DEPARTAMENTO DE TELECOMUNICACIONES
GRUPO I+D NUEVAS TECNOLOGIAS EN TELECOMUNICACIONES
POPAYÁN
2007

Agradezco a Dios por iluminar mi camino

A mi madre por su gran esfuerzo e inmenso amor

A mi padre por su apoyo incondicional

A mi hermano por creer en mí y por ser mi ejemplo

A mis amigos por estar a mi lado siempre

A Karol Viviana por su esfuerzo

Claudia Ximena

Agradezco a Dios por darme la vida

A mis padres y hermana por su infinito amor y apoyo incondicional, por ser mi ejemplo y el mejor motivo para despertar todos los días

A mis amigos por su amistad sincera y por regalarme una sonrisa cada día

A J.G.

A Claudia Ximena por su dedicación

Karol Viviana

AGRADECIMIENTOS

Las autoras desean expresar sus agradecimientos a:

Ingeniero. Guefry Agredo Méndez, por su constante apoyo para la realización de este proyecto, por ser nuestro maestro y amigo.

Ingeniero Oscar Calderón, por su valiosa colaboración.

Lic. Ricardo García Jarquín, miembro del Grupo Linux México.

Compañeros Grupo de Nuevas Tecnologías en Telecomunicaciones

TABLA DE CONTENIDO

INDICE DE TABLAS

INDICE DE FIGURAS

LISTADO DE ACRONIMOS

ADPCM	Adaptive Differential Pulse Code Modulation (Modulación por Codificación de Impulsos Diferencial Adaptativa)
ANN	Artificial Neural Network (Redes Neuronales Artificiales)
AP	Access Point, (Punto de Acceso)
API	Aplication Programming Interface, (Interfaz de Programación e Aplicación)
ASR	Automatic Speech Recognition (Reconocimiento Automático del Habla)
ATA	Analog Telephony Adapter (Adaptador Telefónico Analógico)
CMU	Carnegie Mellon University (Universidad de Carnegie Mellon)
CS-ACELP	Conjugate Structure- Algebraic Code Excited Linear Prediction (Predicción Linear de Codigo Algebráico- Estructura Conjugada)
CSR	Continuous Speech Recognition (Reconocimiento de Habla Contínua)
CTI	Computer Telephony Integration (Integración Computador Telefonía)
DS	Distribution System (Sistema de Distribución)
DTMF	Dual Tone Multi Frecuency (Tono Dual Multifrecuencia)
DTW	Dynamic Time Warping (Tiempo Dinámico Distorsionado)
ESS	Extended Service Set (Grupo de Servicio Extendido)
FIET	Facultad de Ingeniería Electrónica y Telecomunicaciones
GNTT	Grupo Nuevas Tecnologías en Telecomunicaciones
GSM	Global System Mobile (Sistema Móvil Global)
HMI	Human Machine Interface (Interfaz Humano Máquina)
HMM	Hidden Markov Model (Modelo Oculto de Harkov)
IAX	Inter-Asterisk Exchange (Protocolo de Intercambio de Asterisk)
IEEE	Institute of Electrical and Electronics Engineers (Instituto de Ingenieros Eléctricos y Electrónicos)
iLBC	Internet Low Bitrate Codec (Codec de Internet de Baja Velocidad)
IP	Internet Protocol (Protocolo de Internet)
ISR	Isolated Speech Recognition (Reconocimiento de Habla Aislada)
ITU-T	International Telecommunication Union- Telecommunications. (Unión Internacional de Telecomunicaciones)
IVR	Interactive Voice Response (Respuesta de Voz interactiva)
LAN	Local Area Network (Red de Area Local)
MGCP	Media Gateway Control Protocol (Protocolo de Control de Pasarela de Medios)
MLP	Multi Layer Perceptron (Perceptrón MultiCapa)
MOS	Mean Opinion Score (Puntuación de Opinión Media)
MSS	Microsoft Speech Server (Servidor de Habla de Microsoft)

OSI	Open Systems Interconection (Interconexión de Sistemas Abiertos)
PBX	Private Branche Exchange, (Central de Intercambio Privado)
PCM	Pulse Code Modulation (Modulación por Impulsos Codificados)
PSCVoWLAN	Piloto de Sistema de Comunicación con Voz sobre WLAN
RDSI	Red Digital de Servicios Integrados
PSTN	Public Switching Telephonic Network
SALT	Speech Application Language Tags (Especificación de Etiquetas de Lenguaje para Aplicaciones de Voz)
SDSR	Speaker Dependent Speech Recognition (Reconocimiento de Habla Dependiente del Hablante)
SES	Speech Engine Services (Servicios del Motor de Habla)
SIP	Session Initiation Protocol (Protocolo de Iniciación de Sesión)
SISR	Speaker Independent Speech Recognition (Reconocimiento de Habla Independiente del Hablante)
TAS	Telephony Application Services (Servicios de Aplicación de Telefonía)
TTS	Text to Speech (Traducción de Texto a Voz)
VoIP	Voice over Internet Protocol (Voz sobre Protocolo de Internet)
VoWLAN	Voice over Wireless Local Area Network (Voz sobre Red Inalambrica de Área Local)
WiFi	Wireless-Fidelity (Fidelidad Inalámbrica)
WLAN	Wireless Local Area Network (Red Inalámbrica de Área Local)

RESUMEN

Las crecientes necesidades de comunicación en todo tipo de entornos, han sugerido la utilización e integración de tecnologías como el Reconocimiento Automático del Habla y las Redes Inalámbricas para cumplir los requerimientos de funcionalidad, flexibilidad y movilidad, con inversiones relativamente bajas que compensan su relación costo beneficio.

En el presente trabajo de grado se presenta el estudio teórico y técnico para la definición de criterios de diseño y finalmente la implementación un piloto en el que se utiliza el reconocimiento de la voz para el control de un sistema de comunicación PBX sobre una red inalámbrica de área local, definido con la colaboración de la “Clínica La Estancia S.A de la ciudad de Popayán”, que permite generar alarmas o establecer comunicaciones entre los miembros del cuerpo médico, o administrativo.

INTRODUCCION

Actualmente los sistemas de comunicación en muchas organizaciones se basan en redes telefónicas PSTN (Red Telefónica Publica Conmutada) o PBX (*Private Branche Exchange*), cuyo acceso básico se logra mediante teléfonos fijos o radioteléfonos de mediano alcance, con auriculares manos libres que utilizan señalización DTMF (*Dual Tone Multifrecuency*), obligando al personal a digitar números para acceder a un menú, a presionar teclas del teléfono o de su PC (en caso de que exista telefonía IP) para seleccionar posibles opciones o transcribir dictados, respectivamente y recibir mensajes de texto a través de beepers (sin poder generar una respuesta inmediata).

Lo anterior sin contar con las situaciones en las que debe hacerse uso de la telefonía celular, dada su indiscutible ventaja en cuanto a cobertura, pero con la desventaja de la amplia cantidad de operadores y diferencias en precios dentro del mercado, lo que la hace costosa y por tanto poco viable.

Ahora se toma como ejemplo la situación de un centro de atención médica, donde se solicita con frecuencia generar llamadas a dependencias o personal en turno, preferiblemente sin hacer uso de las manos y sin perder la atención visual mientras se esté diagnosticando o interviniendo a un paciente; y más aún donde el equipo que se manipula puede verse afectado por la interferencia que causan los sistemas de comunicación radiotelefónica o celular, lo cual restringe su uso.

Surge entonces la tecnología de Reconocimiento Automático del Habla (ASR-*Automated Speech Recognition*) por parte del PC, suscitando una gran revolución al ofrecer un nuevo método de acceso y control de la información en escenarios informáticos locales o remotos, por medio del lenguaje verbal.

Esta, ofrece interfaces sencillas para una amplia variedad de usuarios y aplicaciones que van desde *callcentres*[1], herramientas educativas y de oficina; tratamiento de enfermedades, discapacidades, seguridad, entre otras, desarrolladas y avaladas por el sector investigativo, académico y empresarial donde se destaca la labor de universidades y grupos de trabajo académico como es el caso del CMU Sphinx (*Carnegie Mellon University Sphinx*), el IEEE (*Institute of Electrical and Electronics Engineers*), el consorcio WWW (*World Wide Web*) y multinacionales como IBM, Microsoft, Dragon –por citar sólo algunas- que le han apostado a esta tecnología como nueva alternativa en la construcción de medios que faciliten la interacción hombre- máquina.

[1] Unidad funcional diseñada para manejar grandes volúmenes de llamadas telefónicas entrantes y salientes desde y hacia sus clientes, con el propósito de dar soporte a las operaciones cotidianas de una entidad

Por otra parte, aparece el concepto de VoWLAN (*Voice over Wíreless LAN*) que ha estado rondando desde hace algún tiempo, surgiendo tras un explosivo crecimiento de las tecnologías de Red de Área Local Inalámbrica y Telefonía IP, que han permitido crear sistemas de comunicación flexibles y móviles, gracias a lo cual por fin es posible crear escenarios en los cuales los usuarios puedan hacer y recibir llamadas telefónicas mientras recorren las instalaciones de sus empresas, sin que su disponibilidad se limite al instante en el que están en sus puestos de trabajo, teniendo en cuenta que no localizar un empleado para la toma de una decisión importante puede costarle mucho dinero a una organización, y en el escenario hospitalario, localizar a un médico o al personal de soporte es fundamental para prestar un servicio oportuno y con calidad humana a los pacientes.

Por lo anterior, si se integra el ASR con VoWLAN como soporte a la creación de sistemas de comunicación más completos e interactivos, se tendrá un nuevo sistema de comunicación con acceso manos libres (a través de comandos de voz) utilizando una misma infraestructura física de red para el transporte de datos y enrutamiento de llamadas de voz simultáneamente, en tiempo real y de forma inalámbrica, procurando alcanzar los objetivos a bajos costos sin sacrificar la eficiencia, precisión y capacidad de expansión.

El presente trabajo titulado “Piloto de Sistema de Comunicación con VoWLAN Controlado por Comandos de Voz para un Entorno Hospitalario- PSCVoWLAN” está enmarcado en la línea de Investigación del Departamento de Telecomunicaciones “Redes y Servicios Telemáticos” en el trabajo del Área de Sistemas Móviles e Inalámbricos del Grupo I+D Nuevas Tecnologías en Telecomunicaciones y establece un escenario en el que se utiliza el reconocimiento de la voz para el control de un sistema de comunicación PBX sobre una red inalámbrica de área local, que permita a los usuarios acceder a un servidor de Respuesta de Voz Interactiva (IVR- *Interactive Voice Response*) desde el cual se pueden generar alarmas o establecerse comunicaciones entre los usuarios[2]. De esta forma, se puede enrutar la solicitud de llamada y ubicar personal en casos de emergencia con sólo utilizar un conjunto limitado de comandos vocales a través de un dispositivo de captura de señales de voz (micrófono), mediante una interfaz de comunicación VoIP(*softphone*), sobre una infraestructura de red inalámbrica de área local, ofreciendo una solución de comunicación mediante un sistema que cuenta con las funcionalidades clásicas de una PBX convencional, con la ventaja que permite generar llamadas, establecer conferencias entre especialistas y solicitar un menú con información, a través de un comando de voz predeterminado para cada situación.

[2] Personal médico y administrativo de un centro de atención médica, en este caso, se contó con la colaboración de la Clínica La Estancia de la ciudad de Popayán.

El documento se divide en 5 capítulos, los cuales dirigirán al lector- capítulos 1 y 2- en el estudio de los aspectos más importantes y básicos de las tecnologías de reconocimiento, su integración con redes inalámbricas y la justificación técnica de la selección de las herramientas utilizadas para el desarrollo del Piloto.

Posteriormente, en los capítulos 3 al 5, se presenta la descripción del sistema, sus requerimientos, especificaciones, características, construcción, pruebas y resultados, para finalmente, presentar las conclusiones del desarrollo general del proyecto, recomendaciones y el sistema terminado completamente funcional.

1. FUNDAMENTOS DE LA TECNOLOGÍA DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

En este capítulo se consigna el fundamento teórico de la tecnología de Reconocimiento Automático del Habla, sus características, aplicaciones, sistemas operativos que las soportan y finalmente algunas de las herramientas que se han desarrollado, orientadas a la comunicación con equipos y creación de interfaces en aplicaciones individuales y Cliente- Servidor, con el fin de establecer los criterios de selección del motor de reconocimiento sobre el cual se soportará el desarrollo del proyecto.

1.1 INTRODUCCIÓN A LOS CONCEPTOS BÁSICOS EN LA TECNOLOGÍA ASR

Desde hace algún tiempo, se viene estudiando la posibilidad de desarrollar Interfaces Hombre- Máquina (HMI- *Human Machine Interface*) controlados por voz, para sustituir y/o complementar -en ciertas ocasiones- las interfaces tradicionales basadas en teclados, paneles, *mouses* y dispositivos similares, esto conllevó al desarrollo de las tecnologías de habla, dentro de la cual se destacan los procesos de *codificación, síntesis y reconocimiento* [1], entre otras, de las cuales se tratará en este proyecto el **reconocimiento**.

La utilización de la voz, y en este caso, del reconocimiento del habla, como posibilidad para comunicación con dispositivos informáticos, ofrece una gran cantidad de ventajas frente a los métodos tradicionales de interacción, tales como:

- Hace que las HMI sean más útiles para los usuarios, por ser el lenguaje hablado la manera más natural de comunicarse para el ser humano.
- Permite movilidad y acceso manos libres, lo que facilita la realización de otras tareas mientras se está en movimiento.
- Permite acceso remoto por medio de la integración con la telefonía. [2]

Las anteriores, entre otras ventajas, son las que han ayudado a incrementar el estudio multidisciplinario de esta tecnología, de la cual se destacan los siguientes factores determinantes, en la construcción y clasificación de aplicaciones basadas en el reconocimiento del habla, así como también los términos técnicos básicos necesarios para el entendimiento de la tecnología.

- **Expresiones**

  Cuando un usuario dice algo, ya sea una palabra, una exclamación o solo un balbuceo, esto se conoce como una expresión. Es la vocalización de una palabra simple, un conjunto de palabras, una oración o incluso múltiples oraciones (flujo de palabras entre dos periodos de silencio) que representan un significado simple para el computador. [3][4]

  El silencio en el reconocimiento del habla, es casi tan importante como lo que se está hablando, debido a que el silencio enmarca el inicio y fin de una expresión, ya que los motores de reconocimiento del habla están siempre "escuchando" o esperando hasta cuando haya una entrada de habla. Cuando el motor detecta una entrada de audio -en otras palabras, cuando detecta una ausencia de silencio- se señala el comienzo de una expresión, de igual manera, cuando se detecta un silencio muy largo seguido de una entrada de audio, marca el fin de la expresión. En este caso es importante resaltar que existen pequeños silencios entre las palabras de una misma frase, por lo que el usuario debe definir los tiempos de pausas entre palabras con el fin de no correr el riesgo de que la máquina corte las frases. [3][4]

- **Gramática**

  Define el dominio o contexto dentro del cual debe trabajar el motor de reconocimiento, el cual, compara la expresión que se ha introducido, con un conjunto de palabras y frases pertenecientes a la gramática definida. [3][4] La gramática hace uso de una sintaxis particular, o de un grupo de reglas para definir y limitar las palabras y/o frases que el motor puede reconocer dependiendo del número de combinaciones permitidas de las palabras del vocabulario. Si el usuario dice algo que no pertenece a este conjunto, el motor de reconocimiento no podrá descifrarlo correctamente y por lo tanto no se efectuará la acción a la cual estaba ligada dicha palabra o sentencia.

  En general la existencia de una gramática en un *reconocedor* ayuda a mejorar la tasa de reconocimiento al eliminar ambigüedades, disminuyendo los tiempos de respuesta, al limitar el número de palabras en una determinada fase del reconocimiento ("perplejidad" de la gramática).[3][4]

- **Dependencia del hablante**

  Esta característica describe el grado en el cual un sistema de reconocimiento requiere conocer las características individuales de la voz del usuario para llevar a cabo el proceso de reconocimiento.

Existen dos tipos, los sistemas dependientes del hablante y los sistemas independientes del hablante.[3]

Los *sistemas dependientes del hablante* se diseñan con alta precisión en torno a las características del habla de una persona o un hablante en particular, asumiendo la coherencia e invariabilidad de dichas características bajo previo entrenamiento; mientras, los *sistemas independientes del hablante*, se diseñan para una gran variedad de usuarios y no requieren entrenamiento previo por parte de los usuarios para que identifiquen sus rasgos personales de voz. [3]

- **Exactitud**

  La capacidad y el rendimiento de un motor de reconocimiento de voz se miden en su exactitud, es decir, qué tan bien reconoce las expresiones; esto no solo incluye la identificación correcta de las expresiones, sino también, identificar si la palabra hablada al motor de reconocimiento pertenece o no al vocabulario o *corpus*[3] y si se obtiene el resultado deseado ante determinada entrada. Esta medida se expresa como un porcentaje y representa el *número de expresiones reconocidas correctamente, sobre el total de expresiones habladas*. [3][4]

  A este concepto se hace referencia posteriormente en la sección 4.1 en la cual se definen los parámetros de prueba del piloto.

1.1.1 Definición

El reconocimiento del habla es el proceso que lleva a cabo una máquina, de analizar una señal de audio para determinar las palabras pronunciadas por un locutor a través de un teléfono o un micrófono, y transcribir o interpretar correctamente lo que se le ha dicho; esto aparentemente es una tarea sencilla, pero de hecho, ha representado en los últimos años uno de los retos más complejos para los investigadores y desarrolladores. [5][6]

La señal analógica de habla se captura por medio de un micrófono o un teléfono, posteriormente se digitaliza (proceso durante el cual, se elimina gran cantidad de información redundante e innecesaria para la identificación de habla o de sonidos y reduciendo al mismo tiempo la dimensión de los patrones para facilitar su clasificación), luego, teniendo en cuenta que el rango auditivo efectivo del ser humano está entre los 1000Hz y 6000Hz[4] se realiza un *pre-procesamiento*, el cual, toma muestras de la señal de habla y extrae únicamente los parámetros suficientes y necesarios para que haya exactitud en el reconocimiento del habla.

[3] El universo de palabras que existen para el reconocedor, será definido en la segunda parte de este capítulo.
[4] El rango auditivo teórico del ser humano es de 20Hz a 20000Hz.

Posteriormente, la señal se clasifica y se identifican los segmentos de voz procesados con símbolos fonéticos cuya longitud puede variar desde un fonema[5], una sílaba, hasta una palabra u oración completa dependiendo del tipo de sistema. Luego, el módulo de reconocimiento se encarga de buscar correspondencia entre dichos segmentos y los segmentos almacenados en la base de fonemas [3], por medio de diferentes técnicas conocidas actualmente.

1.1.2 Técnicas de reconocimiento

Poco a poco, las técnicas de reconocimiento de habla que están aplicándose actualmente, han alcanzado el grado de madurez necesario para ser incorporadas en el desarrollo de productos y servicios abiertos a todo tipo de usuarios, esto es debido a que su aplicación y variedad es tan amplia como las necesidades que surgen a diario en tecnología.

A continuación se presenta una breve descripción de las técnicas de reconocimiento del habla, ya que el objetivo de este proyecto no es ahondar en las características o propiedades de cada una de ellas, sino en los servicios o aplicaciones que sobre ellas se soportan.

1.1.2.1 Comparación de patrones

Esta técnica topológica, basada en el algoritmo de Tiempo Dinámico Distorsionado (DTW- *Dynamic Time Warping*) no requiere un conocimiento explícito del lenguaje, consiste en **parametrizar** la señal de voz a reconocer, para ello se divide en pequeñas ventanas de análisis de aproximadamente 20ms, y sobre cada una de esas ventanas se realiza un proceso de análisis por medio del cual se extraen un conjunto de parámetros (que pueden ser acústicos o coeficientes espectrales), los cuales al agruparse forman un *patrón acústico o plantilla*.

El sistema *reconocedor* dispone de un grupo de patrones de referencia que se calculan en la fase de **entrenamiento** y representan al conjunto de palabras del vocabulario que el sistema puede reconocer, de esta forma, una vez obtenida la plantilla de la palabra, la tarea del reconocedor consiste en compararla con todos los patrones de referencia que el sistema tiene almacenados, y dado que la pronunciación y la velocidad en el habla son variables, se calcula la "distancia" que las separa de las referencias para luego elegir como palabra reconocida aquella cuya plantilla de referencia tiene la menor distancia en la comparación, mediante una *regla de decisión*[6].

[5] Sonido individual distintivo en fonética
[6] La regla del vecino más cercano (NNR- Near Neighbor Rule) consiste en elegir el patrón con la menor distancia media (obtenida del algoritmo DTW) como el patrón reconocido.

Esta técnica de comparación de patrones incluye un alineamiento temporal no lineal y una medida de distancia o umbral de aceptación, con el fin de evitar que una entrada de ruido sea reconocido como habla, esto asegura que se ignoren las expresiones que no estén lo suficientemente cercanas a las plantillas establecidas en la base, esto es una gran ventaja ante otras técnicas, pero con la desventaja de que incrementa el uso de recursos computacionales. [1][2][7][8]

- **Reconocimiento acústico fonético**

Este es un tipo de reconocimiento basado en la teoría de la fonética acústica que postula: "existe un grupo de unidades fonéticas distintivas finitas en el lenguaje hablado, las cuales están ampliamente caracterizadas por un grupo de propiedades que pueden verse en la señal de habla, o en su espectro en el dominio del tiempo". Aún cuando las propiedades acústicas de las unidades fonéticas sean altamente variables, entre una persona y el resto de población, se asume que las reglas de la variabilidad son lo suficientemente fuertes y pueden aprenderse correctamente y aplicarse en situaciones prácticas.

Por ende, el primer paso en esta técnica consiste en una fase de *segmentación* y *etiquetamiento* que implica dividir la señal de habla en segmentos discretos (en el tiempo) donde las propiedades acústicas de la misma representan una de las diferentes unidades fonéticas o clases de unidades y luego, se coloca una o más etiquetas a cada región segmentada de acuerdo con las propiedades acústicas.

El segundo paso, pretende determinar una palabra válida (o una cadena de palabras) a partir de la secuencia de etiquetas fonéticas producidas en el primer paso, la cual al compararse debe coincidir con las almacenadas en el léxico del motor de reconocimiento y se selecciona la de mayor correspondencia. [9]

1.1.2.2 Métodos estadísticos o estocásticos

Este tipo de reconocimiento se fundamenta en la técnica de *Comparación de patrones*, basándose en el análisis estadístico y probabilístico para seleccionar la probabilidad más alta de una secuencia de muestras; la técnica más utilizada es la de Modelos Ocultos de Markov (HMM – *Hidden Markov Model*), debido a su capacidad inherente de representar eventos acústicos de duración variable y a la existencia de algoritmos eficaces para computar automáticamente los parámetros del modelo a partir de los datos de entrenamiento. Los HMM son una clase de modelos estadísticos útiles para el análisis de una serie de observaciones tales como un flujo de muestras acústicas extraídas de una señal de voz. Pueden explicarse como una máquina de estados finitos[7] en la que el siguiente estado depende únicamente del estado actual, donde se produce un vector de

[7] Que lleva asociados dos procesos, uno oculto no observable directamente, correspondiente a las transiciones entre estados y otro observable directamente asociado al primero que organiza los vectores de parámetros.

observaciones o parámetros, asociado a cada transición entre estados, lo que conforman la plantilla a reconocer.

En el caso aplicado al reconocimiento del habla[8], en términos técnicos, los HMM representan cada palabra del vocabulario del reconocedor con un modelo generativo que se calcula en la fase de entrenamiento y posteriormente, se calcula la probabilidad de que la palabra a reconocer haya sido producida por cada uno de los modelos de la base de datos del reconocedor. Para ello, se asume que durante la pronunciación de una palabra, el aparato fonador puede adoptar sólo un número finito de configuraciones articulatorias o estados, y que desde cada uno de esos estados se producen uno o varios vectores, que representan los ítems de la plantilla, cuyas características espectrales dependerán probabilísticamente del estado en el que se hayan generado, así las características espectrales de cada fragmento de señal dependen del estado activo en cada instante, y la evolución del espectro de la señal durante la pronunciación de una palabra depende de la ley de transición entre estados.

Lo anterior en términos más usuales significa que los HMM reconocen el habla mediante la estimación de la probabilidad de las unidades fonéticas elementales (normalmente fonemas contextuales), así como de las relaciones que se establecen entre dichas unidades para componer las palabras (transcripciones fonéticas) y entre las palabras para componer las frases (gramática); cada palabra se especifica en una lista de vocabulario, en términos de los fonemas que lo componen, luego, por medio de una búsqueda se determina la secuencia de fonemas con mayor probabilidad.

Por otra parte, la comprensión del habla utiliza adicionalmente el conocimiento semántico del dominio de la aplicación para captar el significado de la locución de entrada al sistema a partir de la cadena (o cadenas alternativas) de palabras que suministra el elemento reconocedor.

Actualmente, la mayoría de los sistemas de reconocimiento se basan en esta técnica estadística, ya que aunque sus prestaciones son similares a las de los sistemas basados en DTW, requieren menos memoria física y ofrecen un mejor tiempo de respuesta, aunque tienen como contrapartida una fase de entrenamiento mucho más lenta y costosa. [2][7][10]

1.1.2.3 Redes neuronales artificiales

Las ANN (*Artificial Neural Networks*), mediante un estilo de computación paralelo y adaptativo, son capaces de aprender a realizar determinadas tareas a partir de ejemplos de cómo realizarlas.

[8] Entendida el habla para este modelo, como una secuencia de diferentes sonidos producidos por un articulador de habla o hablante.

Las ANNs son sistemas físicos o simulados, que imitan de manera esquemática la estructura hardware (neuronal) del cerebro para tratar de reproducir algunas de sus capacidades. Así, una red neuronal artificial se compone de un conjunto de neuronas o nodos de procesadores interconectados y cuyas sinapsis o conexiones, que representan el flujo de datos, son modificadas mediante un proceso de entrenamiento consistente en la presentación de un conjunto de patrones-ejemplo, con la intención final de que el sistema aprenda a realizar por sí mismo determinada tarea ante determinado estímulo.

En el reconocimiento del habla, la red neuronal más utilizada debido a su desempeño, se llama MLP (*MultiLayer Perceptron*), que se desarrolla en dos etapas, una de **acceso** y otra de **entrenamiento**.

El objetivo de una red MLP es crear un modelo que une la capa de entrada con la salida utilizando el historial de datos, a través de capas ocultas intermedias conectadas, de tal modo que el sistema adquiere la capacidad de reconocer palabras o grupos de fonemas, cuando se conocen las palabras o fonemas que deberían estar a la salida.

En la figura 1 [13] se representa una red MLP.

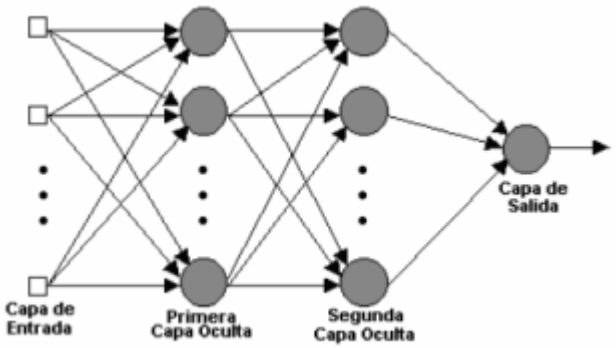

Figura 1. Diagrama en bloques de un MLP con dos capas ocultas.

En la **etapa de acceso**, las palabras ingresan a través de la capa de entrada y se multiplican hacia cada una de las sinapsis o conexiones hasta llegar a la primera capa oculta, luego se genera el mismo proceso de multiplicación a través de las sinapsis de capa en capa hasta alcanzar la capa de salida que produce la respuesta.

Luego, en la **fase de entrenamiento** se utiliza un algoritmo llamado *propagación hacia atrás*, en donde la palabra o fonema de entrada se ingresa repetidamente a la capa de entrada de la red neuronal y con cada ingreso, se compara la salida de la red neuronal con la salida deseada y se calcula un error que se realimenta

posteriormente a la entrada de cada iteración con el fin de disminuir ese error y producir una salida que se aproxime mas a la salida deseada.

En la práctica, cuando la teoría probabilística debe hacerse “transparente” para los usuarios finales, para que un dispositivo “aprenda” a reconocer el habla es necesario que éste recuerde la manera como se dice cada palabra, debido a que cada persona habla con acentos e inflexión variantes, además, se debe tener en cuenta la pronunciación, el contexto y la frecuencia de su uso gramatical para finalmente poder predecir la palabra que se está pronunciando.

Esto se logra por medio de entrenamientos durante los cuales el sistema informático debe acostumbrarse a una voz y a un acento particular, que en el caso de este proyecto está predeterminado por la aplicación a utilizar, en la cual, lo más conveniente es que reconozca las palabras independientemente del locutor, pues son muchos los usuarios que van a hacer uso de ella. Este requerimiento y otros se mencionan posteriormente, cuando se establezcan los criterios de selección del motor de reconocimiento para el piloto. [9][11][12][13]

1.1.3 Tipos de reconocimiento de habla

El locutor, la forma de hablar, el vocabulario, la gramática y el entorno físico son las características más importantes que influyen en la clasificación de los sistemas de reconocimiento de habla, debido a su intervención directa en la capacidad del *reconocedor* para determinar quién es la persona que está hablando, dónde comienza y dónde termina una expresión, la precisión, no solo para reconocer cual es la palabra hablada por el usuario, sino, si la palabra existe en el léxico del motor de reconocimiento y más aun, si ejecuta correctamente la operación para la que se le ha programado.

De lo anterior depende el tipo de diseño y la técnica que debe aplicarse en la construcción de un motor de reconocimiento, lo cual, no es objeto de este proyecto, más, permitió en su momento seleccionar la herramienta más apropiada que se ajuste a las necesidades del mismo.

Los sistemas de reconocimiento del habla se clasifican según lo preliminar, de la siguiente manera:

1.1.3.1 Tipo de expresión

La forma de hablar es el segundo factor que determina la complejidad de un reconocedor de habla, ya que, el hombre pronuncia las palabras de una forma continua, y debido a la inercia de los órganos articulatorios, que no pueden moverse instantáneamente, se producen efectos coarticulatorios; ello, unido a las

variaciones introducidas por la prosodia[9], hace que una palabra al principio de una frase sea diferente cuando se dice en medio, o que sea diferente dependiendo de que es lo que le procede o le sigue.

Esta es una actividad bastante sencilla cuando se esta hablando a otra persona que ha desarrollado su capacidad de habla, aprendiendo desde el seno materno y posteriormente de la experiencia diaria en la interacción con otras personas de la manera más natural y fluida; pero para una máquina es una tarea bastante complicada, pues, de acuerdo con el diseño estará en capacidad de reconocer *palabras aisladas*, que no atienden a la naturalidad con la que se comunican los seres humanos, *frases conectadas*, que aunque un poco más largas implican la emisión de pausas que para el lenguaje humano son innaturales y para el oído completamente innecesarias; o en el mejor de los casos, *habla continua* que permite alcanzar un mejor nivel de naturalidad al hablar con la máquina, pero que le implica a esta última un muy alto nivel de refinamiento. [14][15][16]

Este tipo de reconocimiento se conoce también como la *Interfaz Oral*, y en resumen, indica la forma en la que se introducen las palabras en el sistema de reconocimiento, y de acuerdo con lo anterior, existen dos formas determinadas por este criterio.

- **Reconocimiento de palabras aisladas o reconocimiento discreto**

  Estos sistemas ISR (*Isolated Speech Recognition*) mantienen un modelo acústico separado para cada combinación de palabras o frases; las palabras son habladas individual y separadamente, no como una conversación, sino como un dictado.

  Por lo general, una máquina de reconocimiento de palabras aisladas requiere que cada expresión o palabra tenga un silencio antes y después, es decir a ambos lados de la ventana de muestreo, lo que no significa que acepte palabras simples, sino que requiere de la entrada de una sola expresión simple a la vez muy bien vocalizada, esto implica que el motor o sistema de reconocimiento maneje los estados: *escucha* o *no escucha*, para no tener que determinar inteligentemente donde comienza o donde termina cada grupo fonético o cada palabra, dentro de una oración.

  Los *reconocedores* de palabras aisladas son relativamente sencillos en su implementación y los más comerciales se basan en HMMs, aunque, también existe cierta variedad de implementaciones sobre todo académicas, basadas en redes neuronales[10], e híbridos entre ambos, pero su desarrollo es un poco menor (aunque creciente), debido a la complejidad de esta última técnica. No

[9] Parte de la gramática que enseña la correcta pronunciación de las palabras
[10] Para ampliar este tema, refiérase al documento “Redes Neuronales en reconocimiento de locutor”[17]

obstante, es amplio el número de aplicaciones de reconocedores de este tipo, siendo los IVR's una de las aplicaciones más frecuentes dado que las órdenes por lo general se reducen a reemplazar la marcación de números DTMF, por los mismos números solo que a través de comandos de voz. [17]

Esta característica hace al motor de reconocimiento menos pesado a la hora del procesamiento para el computador, consume menos recursos y facilita la implementación de sistemas basados en comandos como el que trata este proyecto, en la cual se requiere la utilización de equipos portátiles y de oficina que por lo general son de bajo o mediano perfil.

La aplicación más importante de un reconocedor de palabras aisladas es el reconocimiento de palabras conectadas, cuya entrada hablada es una secuencia de palabras de un vocabulario específico, y el reconocimiento se lleva a cabo basándose en la coincidencia de palabras de referencia aisladas. Ejemplos, son las cadenas de dígitos conectados donde el vocabulario es un conjunto de 10 dígitos, o el reconocimiento de letras conectadas, donde el vocabulario es el conjunto formado por el abecedario o comandos para comunicación en *callcentres* empresariales.

- **Reconocimiento de habla continua**

  Los sistemas de reconocimiento de habla continua CSR (*Continuous Speech Recognition*), son capaces de reconocer palabras habladas en discursos naturales, en los que como es natural para el ser humano, las expresiones habladas presentan adyacencia entre palabras y no existen pausas o aparentes divisiones entre ellas.

  Los *reconocedores* con características de reconocimiento continuo trabajan similarmente a los de habla discreta, con la diferencia de que el final de las palabras no se detecta por medio del silencio, por ello son más complicados de construir debido a que utilizan métodos especiales para determinar cuando "hipotéticamente" se alcanza el final de una palabra, produciendo otras miles de "hipótesis" que podrían preceder, para luego reducir sus posibilidades mediante un modelo de probabilidad de lenguaje.

  La variabilidad en la articulación, la tendencia a reducir el sonido de las consonantes o eliminar las pausas entre ellas, puede dar como resultado, que un motor de reconocimiento no pueda identificar las palabras, por ende, para incrementar la exactitud, los modelos para sistemas de habla continua incluyen información sobre combinaciones representativas de palabras y reglas de contexto haciendo de éste tipo de reconocimiento, el más natural, fácil y rápido para los usuarios, desde la perspectiva del número de palabras que pueden procesarse en determinado tiempo, mostrando con ello su

principal diferencia y ventaja con respecto a los sistemas de reconocimiento discreto o aislado. [18][19]

1.1.3.2 Tipo de aplicación

Este es un aspecto de gran importancia ya que la tecnología de habla se basa en la captura de las ondas sonoras producidas por la voz humana, que para cada persona, producen diferentes patrones.

Una persona no pronuncia siempre de la misma manera y otro individuo nunca poseerá idénticas características de habla que el primero, debido a situaciones físicas, psicológicas, lo que se conoce como *variabilidad intra-locutor*; culturales, regionales, edad, sexo, conocidas como *variabilidad inter-locutor*. Estas características, determinan el grado de dependencia que tenga un motor de reconocimiento respecto a un patrón específicamente, lo cual influye en la complejidad del sistema y en la selección de acuerdo a las aplicaciones para las cuales se va a emplear. [20]

- **Reconocimiento dependiente del hablante**

  Este tipo de reconocimiento obedece al grado de dependencia al cual un sistema de reconocimiento requiere conocimiento de las características individuales de la voz que se utilizan como parámetros en el proceso de reconocimiento.

  Todos los sistemas de reconocimiento dependientes del hablante SDSR (*Speaker Dependent Speech Recognition*) requieren algún tipo de entrenamiento para asegurar una alta tasa de exactitud, ello implica que pueden utilizarse por varias personas, pero requieren que cada uno de los usuarios entrene previamente el sistema para que éste pueda identificar su pronunciación, inflexiones y acentos. Por lo general, utilizan la técnica de *comparación de patrones*, para crear el modelo de habla, a través del entrenamiento del sistema, para reconocer si corresponde o no la forma de decir cada palabra del vocabulario. [3][18][19]

  El **Reconocimiento de voz**, es una tecnología derivada del reconocimiento de habla dependiente del locutor, llamada en muchos casos también *verificación de voz*. El objetivo de estos sistemas no es principalmente reconocer lo que el usuario dice, sino identificar quien está hablando basado en una serie de características incluidas en las señales de voz. [21]

  Actualmente este tipo de aplicaciones se utilizan frecuentemente en sistemas de seguridad computacional, comercial en donde se ha vuelto bastante común la utilización de técnicas de reconocimiento biométrico (la voz entre ellas) ya

que este tipo de parámetros representan características especiales y únicas en los seres humanos, convirtiéndolas en herramientas muy poderosas en dicho campo.

- **Reconocimiento independiente del hablante**

  Los sistemas de reconocimiento independientes del hablante SISR (*Speaker Independent Speech Recognition*) no requieren entrenamiento para usuarios específicos, sino que, utilizan modelos genéricos para reconocer el habla de cualquier usuario por medio de la combinación de plantillas o patrones existentes, provenientes de una gran variedad de usuarios, lo cual es ventajoso con respecto a los sistemas dependientes del usuario, pues no requiere entrenamiento individual y no limita el número de usuarios ni las aplicaciones posibles; aunque el tamaño de los vocabularios sea del orden de 3000 palabras y el grado de exactitud es menor debido a las múltiples diferencias inter-locutor e intra-locutor que se presentan en los sistemas que utilizan esta tecnología como base para sus aplicaciones. [3][18][19]

  A pesar de ello, los SISR son las solución más apropiada si se piensa en la implementación de sistemas de comunicación PBX y servicios basados en menús alfanuméricos, tales como *callcentres* y sistemas IVR, como es el caso del PSCVoWLAN cuyas opciones están limitadas, además, reducen el tiempo de respuesta de los usuarios frente a la selección de opciones, por tanto la longitud de las llamadas, y sus costos, traduciendo estas ventajas en eficiencia del sistema.

1.1.3.3 Tamaño del vocabulario

Por lo general, un sistema de reconocimiento de habla posee un vocabulario de 20 a 40.000 palabras, entre comandos, frases y palabras cotidiana (cuando se trata de sistemas de reconocimiento continuo).

Es una característica menos sobresaliente en cuanto a la clasificación de los sistemas de reconocimiento, pero tan importante como las anteriores. Está sujeta a las reglas que la gramática impone respecto a las secuencias de palabras permitidas de acuerdo con el conjunto de reglas del contexto y la estructura de las frases. Determina en un sistema la capacidad de almacenamiento del posible *corpus* de palabras que pueda reconocer, teniendo en cuenta que, grandes cantidades de vocabularios pueden disminuir la exactitud debido a la similitud en la pronunciación de muchas palabras y la ambigüedad en sus significados; y por otra parte, los sistemas con menor cantidad de palabras en su *corpus*, restringen su posibilidad de ampliación a otras aplicaciones.

1.2 HERRAMIENTAS DE RECONOCIMIENTO DE HABLA

Muchos factores entre ellos la disminución de costos, la amplia aceptación y creciente emergencia de las tecnologías del habla en la construcción de aplicaciones educativas, médicas y empresariales, han incrementado el desarrollo de una gran cantidad de herramientas de reconocimiento de habla para usuarios simples y/o aplicaciones cliente servidor, entre las cuales se destacan las siguientes, clasificadas así:

1.2.1 Software libre

Esta categoría corresponde a herramientas desarrolladas principalmente por proyectos avalados en grupos de investigación de universidades, los cuales en su mayoría son versiones que aun se encuentran en etapa de implementación. Entre estas se destacan las siguientes como las más conocidas y cuyo desarrollo continuo, les permite un lugar privilegiado a la altura de las necesidades del mercado.

1.2.1.1 CVoiceControl

Es un sistema de reconocimiento de palabras aisladas dependiente del hablante, que utiliza la técnica de comparación de patrones con DTW y permite a los usuarios -mediante entrenamiento previo- conectar comandos de habla con comandos de Unix. Se trata de un sistema muy básico, incluye una utilidad de configuración del nivel del micrófono, un editor o entrenador del vocabulario para ingresar nuevos comandos y expresiones (soporta hasta 1.000 expresiones), a través de una interfaz sin muchas opciones de configuración detallada, además, los requerimientos del equipo no son altos debido a que se trata de una aplicación de 2MB, con funcionalidades que no exigen máquinas con alto poder de procesamiento.

El reconocimiento inicia una vez el sistema detecta una entrada de voz a través del micrófono. En caso de que se haya ingresado correctamente alguna palabra definida previamente en el *corpus*, el sistema lleva a cabo una operación asociada a un comando de ejecución de determinada operación en Unix.

Esta herramienta es útil en aplicaciones domésticas de dictado y trascripción de texto monousuario y no soporta configuraciones cliente servidor, requiere el uso de micrófonos especiales con supresión de ruido, lo cual incrementaría los gastos, además de entrenamiento personal del usuario por lo que no es óptimo para los requerimientos de implementación del piloto que trata este proyecto. [22][23]

1.2.1.2 GVoice

No es una herramienta de reconocimiento como tal, sino que se trata de una librería de reconocimiento de Habla discreta dependiente del hablante que se instala sobre la versión para Linux del Via Voice de IBM desarrollada hasta el momento para controlar aplicaciones del GNOME.

Gvoice incluye librerías de inicialización, motor de reconocimiento, manipulación de vocabulario y configuración de opciones básicas para trascripción de texto y control de operaciones de telefonía básica para números limitados de usuarios en máquinas cuyas características sean equivalentes o superiores a un procesador Intel Pentium MMX 166MHz, 32MB de RAM, 70MB de disponibilidad en el disco duro, tarjeta de sonido compatible con Linux con entrada para micrófono, y sistema operativo RedHat 6.0, lo cual es por una parte ventajoso ya que no exige máquinas de altas prestaciones, haciéndolo económico, pero por otra parte limita las posibilidades de ampliación de los servicios que puede prestar haciéndolo ineficiente a largo plazo. [3]

1.2.1.3 CMU Sphinx

CMU[11] Sphinx es un motor de reconocimiento de habla de código abierto, desarrollado por la Universidad de Carnegie Mellon, basado en HMM, independiente del hablante adaptable a vocabularios cortos, mediano o largos (del idioma inglés y parcialmente en francés); es el más conocido sistema de reconocimiento de software libre en el mundo útil en variadas aplicaciones multiusuarios y simples, desarrollado inicialmente para Linux/UNIX y posteriormente para el sistema operativo Microsoft Windows NT o superior. [24]

Sphinx, más que un motor de reconocimiento de habla es una herramienta para el desarrollo de sistemas de reconocimiento adaptables a las necesidades propias de cada usuario, basado en los siguientes componentes:

- ***Modelo Acústico***, basado en HMM, representa estadísticamente un rango de posibles representaciones de audio para los fonemas o sonidos individuales del lenguaje. Se construye grabando un gran número de usuarios pronunciando listados de palabras (durante periodos de tiempo de 50 a 100 horas) que van a ser reconocidas posteriormente.

  A partir de la creación del modelo acústico es posible crear la base de datos de audio, que contiene todas las palabras que el sistema es capaz de reconocer dependiendo de la aplicación. En el caso específico del Sphinx la CMU ha puesto a disposición algunas bases de datos públicas grabadas a través de arreglos de micrófonos, con vocabulario limitado pero óptimo para

[11] Carnegie Mellon University

pruebas iniciales y desarrollo de sistemas de reconocimiento de mediano alcance, entre las cuales se tienen las presentadas en la tabla 1.

Tabla 1. *Bases de datos de audio disponibles para Sphinx*

Base de Datos	Características	
	Muestreo	**Descripción**
MicArray	16KHz - 16bit	La colección de datos se hace por medio de arreglos de micrófonos, ya sea de 15 u 8 elementos con espaciamiento determinado entre micrófonos, tomando muestras de habla a hombres y mujeres adultos acerca de nombres, números telefónicos, edades, fechas de cumpleaños, y expresiones de uso cotidiano, etc. (idioma inglés)
AN4	8KHz - 16bit	Utiliza un solo micrófono, maneja aproximadamente 130 expresiones tomadas de 948 entrenadores entre hombres y mujeres adultos, acerca de nombres, números telefónicos, edades, fechas de cumpleaños, etc. (idioma inglés)
Let's go	16KHz – 16bit	Recopila archivos de audio de aproximadamente 1464 llamadas telefónicas de personas adultas entre hombres y mujeres acerca de conversaciones y vocabulario de uso cotidiano. (idioma inglés)
CMU- SIN	Heredado del *Let's go*	Incluye una recopilación de 500 expresiones tomadas de llamadas telefónicas realizadas por un mismo locutor adulto masculino en inglés.

De las anteriores la más utilizada comúnmente es la AN4 debido a que ha sido entrenada por gran cantidad de voces lo cual ofrece mayor acercamiento a diversos locutores, esto pensando en que los usuarios finales no hacen parte del equipo de entrenamiento previo y es importante la variedad de acentos, tonos y timbres de voz. Además, es la más apropiada para aplicaciones de telefonía como el sistema de comunicación PBX sobre IP presentada en este proyecto, debido a la frecuencia de muestreo de 8KHz.

- ***Modelo del lenguaje***, que contiene información de probabilidades de palabras del idioma en cuestión, las cuales pueden ser individuales, o secuencias de dos o tres palabras. Este modelo permite utilizar y crear archivos de todas las palabras, oraciones o el *corpus* completo que se desearía que el decodificador reconociera, conocidos como Diccionarios, cada uno de los cuales consiste en un archivo que relaciona las palabras que van a ser reconocidas y su trascripción fonética, basada en la unidad fonética utilizada por el sistema, en este caso los **fonemas**. La CMU ha definido un diccionario con 39 fonemas con los cuales es posible pronunciar alrededor de 125.000 palabras de uso cotidiano en el idioma inglés. [25] Tales fonemas se relacionan en la tabla 2 [26], con un ejemplo de la forma en la que el Sphinx divide cada palabra y su pronunciación según la fonética del idioma inglés.

Tabla 2. *Fonemas del Diccionario de pronunciación del CMU para el idioma inglés*

Fonema	Palabra ejemplo	Pronunciación (Diccionario)
AA	odd	AA D
AE	at	AE T
AH	hut	HH AH T
AO	ought	AO T
AW	cow	K AW
AY	hide	HH AY D
B	be	B IY
CH	cheese	CH IY Z
D	dee	D IY
DH	thee	DH IY
EH	Ed	EH D
ER	hurt	HH ER T
EY	ate	EY T
F	fee	F IY
G	green	G R IY N
HH	he	HH IY
IH	it	IH T
IY	eat	IY T
JH	gee	JH IY
K	key	K IY
L	lee	L IY
M	me	M IY
N	knee	N IY
NG	ping	P IH NG
OW	oat	OW T
OY	toy	T OY
P	pee	P IY
R	read	R IY D
S	sea	S IY
SH	she	SH IY
T	tea	T IY
TH	theta	TH EY T AH
UH	hood	HH UH D
UW	two	T UW
V	vee	V IY
W	we	W IY
Y	yield	Y IY L D
Z	zee	Z IY
ZH	seizure	S IY ZH ER

- ***Decodificador***, del cual existen diferentes versiones adaptables a gran cantidad de aplicaciones, entre las que se encuentran *Sphinx 2*, *Sphinx 3*, *Sphinx 4* y *Sphinx Pocket*.

  Sphinx 2 (S2), es el sistema de reconocimiento de habla más rápido del CMU, consiste en un API que permite la captura de la señal de audio en a nivel de máquina sobre el dispositivo de adquisición de audio. Realiza la decodificación de la señal de audio en una cadena de texto, para lo cual utiliza un diccionario de datos que almacena el comando oral y su representación léxica, un modelo de lenguaje que debe ser cargado al iniciar el motor de reconocimiento de voz y que indica los parámetros del idioma en el que se está realizando el reconocimiento.

  Consta de un grupo de librerías escritas en lenguaje C, que incluyen funciones de reconocimiento de habla y pueden ser compiladas bajo plataformas Unix (Linux, DEC Alpha, Sun Sparc, HPs) y procesadores Pentium corriendo Windows XP/NT/95 y permite el desarrollo de aplicaciones ejecutables en tiempo real, la adición de nuevos modelos de lenguaje, nuevas palabras y archivos de audio, con la única desventaja hasta el momento, de manejar únicamente el idioma Inglés.

  Sphinx 2 utiliza la Aplicación de Voz de Perlbox-Voice como interfaz gráfica para evitar que los programadores tengan que configurar el motor de reconocimiento a través de línea de comandos.

  Este último consiste en un grupo de librerías diseñadas para habilitar el control por voz de aplicaciones para sistemas operativos basados en Unix, enlaza los *scripts* de reconocimiento del Sphinx, con los archivos de ejecución de dichas aplicaciones, por lo que requiere como principal requerimiento la instalación de Perlbox-Voice, sugiere mínimo un equipo con 128MB de RAM y 200MB de espacio libre en disco, con un procesador mínimo Pentium III o equivalente y preferiblemente su sucesor.

  La integración de estas dos herramientas funciona en términos generales de la siguiente manera: para que se ejecute una acción en Linux, ordenada a través de un comando vocal, el Perlbox-Voice ofrece en su interfaz la posibilidad de ingresar de manera escrita la palabra o comando que invocará la acción que se requiera y la acción a la cual debe ligarse, incluyendo la navegación a través del escritorio del sistema operativo. Además de ello, ofrece una especie de “protección” contra entornos ruidosos, pues permite la asignación de una palabra clave para el inicio de la aplicación sin correr el riesgo de que se inicie con cualquier expresión o palabra aleatoria. [27].

  Esto no es necesario en la implementación de este piloto debido a que los usuarios hablan normalmente el idioma español y para la utilización del

sistema lo harán en inglés, lo que minimiza dicho riego y mejora la eficiencia ya que es solo una palabra o comando lo que el sistema debe reconocer.

Sphinx 3 (S3), ofrece las mismas funcionalidades con un mayor número de librerías y en teoría incrementa la exactitud en el reconocimiento, pero con un significativo costo computacional, ya que, para alcanzar una baja tasa de errores del 27.4%, contra un 45.9% del Sphinx 2, incrementa en 120 veces la capacidad de captura en tiempo real, lo cual es ventajoso para el usuario final en cierto modo, pero a la vez, requiere mayor capacidad de procesamiento, lo que sacrifica su velocidad de ejecución.[12] [28][29][30][31]

Por su parte, **Sphinx 4 (S4)**, escrito completamente en JavaTM, consiste en un reconocedor de habla continua y discreta, con capacidad de reconocimiento de dígitos aislados y continuos, permite almacenar vocabularios pequeños, medianos y largos de 1000, 5000, y hasta 64000 palabras. Esta versión está disponible para el Ambiente Operativo SolarisTM, Sistemas Operativos Mac OS X, Linux y Windows de 32, y requiere la instalación de software adicional: Java 2 SDK Estándar Edición 1.5.0 o superior, Apache Ant 1.60[13] y Subversión SVN o Cygwin[14].

Un aspecto de gran importancia a tener en cuenta en el momento de la selección del decodificador de Sphinx, es que sea apto para la implementación de aplicaciones de telefonía, para lo cual se necesita manejar un modelo acústico de 8KHz, tal como lo hace el Sphinx 2, mientras que, Sphinx 3 y 4 se implementan mediante un modelo acústico PCM de 16bits a 16KHz. Hasta el momento, debido a que estas versiones, S3 y S4, son relativamente recientes y aún están en desarrollo, se utiliza una técnica de sobremuestreo conocida como interpolación para crear un modelo acústico y un modelo de lenguaje de 16KHz para Sphinx, pero esto disminuye la exactitud en el reconocimiento a un nivel comparable con la exactitud de S2, y teniendo en cuenta que S3 y S4 requieren equipos con mayores prestaciones, en ese caso, para la implementación del piloto es preferible utilizar S2 pues se obtendría el mismo nivel de exactitud con equipos menos exigentes. [32][33][34]

[12] Sugiere la utilización de equipos con procesador mínimo Pentium IV o superior, con 450MB de espacio libre en disco.

[13] Apache Ant es una herramienta independiente utilizada en programación para la realización de tareas mecánicas y repetitivas, normalmente durante la fase de compilación y construcción. Es independiente del sistema operativo, y se basa en archivos de configuración XML y clases Java para la realización de las distintas tareas, convirtiéndola en una solución multi-plataforma que permite gestionar el seguimiento de múltiples versiones en la misma unidad de información, esto es, mantener un historial de todos los avances en la implementación de un desarrollo software para plataformas Unix.

[14] Cygwin es una aplicación que consta de un conjunto de archivos que ofrece un comportamiento similar a los sistemas Unix en Windows, su objetivo es portar software que normalmente se ejecuta en Sistemas Operativos Portables basados en UNIX, con el fin de generalizar las interfaces de los sistemas operativos para que una misma aplicación pueda ejecutarse en distintas plataformas.

1.2.2 Software comercial

En esta categoría se destacan aplicaciones desarrolladas por empresas tales como IBM, Microsoft, Nuance, entre otras, que soportan sistemas operativos como Microsoft Windows y MAC OS X, entre las cuales se destacan:

1.2.2.1 IBM Via Voice

Es un desarrollo propietario de IBM basado en la técnica de Modelos Ocultos de Markov, consiste en una tecnología de reconocimiento de habla continua de vocabulario extenso, dependiente del hablante, que convierte voz en texto orientado a la navegación en el computador y aplicaciones de oficina basadas en Windows y MAC OS.

IBM ha desarrollado básicamente 6 versiones en las cuales se incluyen vocabularios en Ingles (Británico y Americano), Francés, Alemán, Portugués, Mandarin (China, Taiwan) y Español de Castilla, dichas versiones incorporan micrófonos de alta tecnología para dictado y procesamiento de texto, en programas como Microsoft Office para Windows XP Profesional y Home, 98SE, Me y 2000; ofrece la posibilidad de navegación por el sistema operativo y la creación de macros para controlar otras aplicaciones de procesamiento de texto que no están incluidas en el motor básico, además, ofrecen la capacidad de navegación por voz en Internet y por todo el sistema operativo a través de la creación de macros que soportan la mayoría de las utilidades predeterminadas de ambos sistemas operativos, Windows y Mac OS X. En la tabla 3 se muestran las características principales de cada una de estas versiones. [35]

1.2.2.2 Dragon Naturally Speaking™

Es un producto desarrollado por Dragon Systems basada en HMM, que ofrece una alta exactitud (99%) y rendimiento en el reconocimiento continuo de habla dependiente del usuario (requiere entrenamiento) orientado a la utilización de herramientas como Microsoft® Word y Excel®, Corel® WordPerfect®, y todas las aplicaciones basadas en Windows® 2000, XP Home y Professional, utilizado a través de dispositivos locales o manos libres Bluetooth y con la capacidad de adaptación en red. Implementado actualmente para el idioma inglés, permite el ingreso de nuevas palabras soportando un amplio vocabulario de aproximadamente 250.000 palabras de vocabulario de uso diario estándar o personalizadas por el usuario. Está enfocado hacia tres áreas básicas de funcionalidad: dictado, control y traducción de texto a habla TTS (Text to Speech) y ofrece una herramienta dirigida exclusivamente al sector médico, con vocabulario especializado y sistemas supresores de ruido en micrófonos especiales, con la desventaja de que no está orientado a aplicaciones cliente servidor ni aplicaciones IVR.

Tabla 3. *Comparación entre las versiones de ASR de IBM Via Voice*

Versión	Características generales		
	Sistema Operativo	**RAM**	**Procesador**
Pro Usb Edition	Microsoft® Windows 98SE y Windows Me	64 MB	Intel® Pentium® 300 MHz con 256K L2 en cache o AMD K6, o equivalente
	Microsoft® Windows 2000 Professional	96 MB	Intel® Pentium® III 600 MHz con 256K L2 en cache o AMD TM 600 MHz con 256K L2 en cache, o equivalente
	Microsoft® Windows XP Home & Professional SP2	192 MB	
	Puerto USB para entrada de Micrófono, 510MB de espacio libre en Disco Duro		
	Versión que permite la utilidad de dictado, edición y corrección de texto en el procesador de texto que maneja el sistema operativo por defecto o en aplicaciones de Microsoft® Office. Ofrece la posibilidad de crear macros para el control de otras aplicaciones propias del sistema operativo		
Advanced Edition	Microsoft® Windows 98SE y Windows Me	64 MB	Intel® Pentium® 300 MHz con 256K L2 en cache o AMD equivalente
	Microsoft® Windows 2000 Professional	192 MB	
	Microsoft® Windows XP Home & Professional SP2		
	Puerto USB para entrada de Micrófono, 510MB de espacio libre en Disco Duro		
	*Para Win98/Me, se requiere una tarjeta de sonido compatible de 16bits con salida de parlantes y entrada de micrófono.		
	Esta herramienta ofrece no solo la posibilidad de dictado y corrección en paquetes de oficina, sino que también permite navegación dinámica en Internet ya sea en el navegador propio de Windows o a través de la creación de macros para el control de otros navegadores y aplicaciones.		
Standard Edition	Microsoft® Windows 98SE, Windows Me	64 MB	Intel® Pentium® 266 MHz con MMX y 256K L en cache o equivalente, incluyendo AMD-K6® con 256K L2 en cache
	Microsoft® Windows XP Home & Professional SP2	192 MB	
	500MB de espacio disponible en el disco, Tarjeta de sonido compatible con Win 98/Me/Xp de 16 bits con entrada de micrófono		
	Esta diseñado para dictado directo en aplicaciones de Microsoft Office como Word 97, 2000, 2002 sin necesidad de crear macros para controlarlas, ofrece un vocabulario de alrededor de 300000 palabras con vocabulario personalizado que incluye direcciones, acrónimos y expresiones.		

Personal Edition	Microsoft® Windows 98SE, Windows Me	64 MB	Intel® Pentium® 266 MHz con MMX y 256K L en cache o equivalente, incluyendo AMD-K6® con 256K L2 en cache
	Microsoft® Windows XP Home	192 MB	
	500MB de espacio disponible en el disco, Tarjeta de sonido compatible con Win 98/Me/XP de 16 bits con entrada de micrófono		
	Esta diseñado para dictado directo en aplicaciones de Microsoft Office como Word 97, 2000, 2002 sin necesidad de crear macros para controlarlas, ofrece un vocabulario de alrededor de 300000 palabras con vocabulario personalizado que incluye nombres, direcciones, coloquialismos y fechas. No ofrece la opción de creación de macros para control de otras aplicaciones.		
MAC OS X Edition	Mac OS X versión 10.1, 10.2 y 10.3	256 MB	Equipos G3 y G4 de 300 MHz
	600MB de espacio disponible en disco, Salida de audio convencional o USB		
	Esta herramienta permite dictar, corregir, editar y agregar formato a textos en el procesador de palabras del sistema operativo y además de ellos, controlar el navegador predeterminado. Por ahora esta versión no permite crear macros, controles o extensiones sobre otras aplicaciones		

1.2.2.3 Voice Xpress

Consiste en un reconocedor de habla continua (Tecnología de lenguaje natural) dependiente del hablante, entrenable, orientado al manejo -mediante comandos cortos o frases largas- de aplicaciones propias de Microsoft Windows 2000, XP Profesional y/o Home, tales como Microsoft Office e Internet Explorer, aunque también puede configurarse para controlar el entorno del sistema operativo en general, en inglés americano, británico, francés, holandés, español (castilla) y alemán, alcanzando una base de datos de más de 5000 palabras.

Debido a que es una aplicación netamente orientada para uso de oficina, no cuenta con la posibilidad de crear módulos de expansión o configuración con herramientas que no sean propias de Microsoft, aunque si asiente el ingreso de nuevas palabras, frases o comandos al vocabulario total.

1.2.2.4 Verbio

Verbio es un reconocedor conformado por un conjunto de librerías y utilidades que permiten incorporar herramientas de síntesis y reconocimiento de habla natural, maneja una extensa base de vocabulario o gramática independiente del hablante, disponible en francés, inglés español, catalán, gallego, portugués, brasileño y variantes del español hablado en Latinoamérica (argentino, chileno, mexicano, colombiano, venezolano) y EEUU, para sistemas operativos Windows 2000 y XP.

Está orientado principalmente para trabajar en entorno telefónico aunque ofrece amplia compatibilidad (bajo previa adaptación en recursos o modelos acústicos si se requiere), con distintos entornos de trabajo que van desde los sectores de *callcentres*, domótica, seguridad, portales de voz de servicios, aplicaciones de PC,

aplicaciones industriales, móviles, y en general, cualquier entorno que requiera o disponga de un sistema de manos libres. Está especialmente indicado para permitir la interacción hombre-máquina y comunicaciones personales mediante la voz en ámbitos como:

- Telefonía: *Callcentres*, IVR's, mensajería unificada, operadoras automáticas, portales de voz, etc.
- Multimedia: Realización de prototipos, CD's de información genérica y cambiante, temas de formación interactiva.
- Internet: Mediante las tecnologías asociadas a la VoIP (*Voice over Internet Protocol*) a través de aplicaciones con reconocimiento en la Red.
- Medicina: Aplicaciones de ayuda y soporte a distintos niveles de discapacidad, verificación de información o comandos.
- Industrial: Automatización industrial de procesos mediante la voz (logística, maquinaria, etc.)
- Terminales multimodales (móviles): Control del flujo de información y comandos en móviles a través de manos libres.

Los sistemas de habla desarrollados sobre Verbio siguen una estrategia de comunicación basada en la arquitectura cliente-servidor, de modo que en un mismo entorno de trabajo pueden coexistir varios servidores (todos ellos en máquinas distintas) y varios clientes (éstos sí pueden compartir máquina). Este escenario y la posibilidad de cada cliente (concretamente, de cada una de sus líneas por separado) de conectarse a un servidor distinto, permite distribuir la carga computacional entre todos los servidores presentes en el sistema y la característica de haber sido previamente entrenado con señales de audio procedentes de entornos telefónicos, tanto fijos como móviles, con el objetivo de obtener las mejores tasas de reconocimiento. [36]

En la tabla 4 [36] se indican las características y requerimientos más importantes a tener en cuenta en la implementación de un sistema basado en Verbio.

Verbio es una herramienta muy completa que combina funcionalidades de oficina como trascripción de texto y soporte a dispositivos especializados en aplicaciones de telefonía, con la desventaja de tratarse de un desarrollo propietario que exige una gran inversión, no solo en el software como tal, sino también en equipos que soporten sus requerimientos, lo que genera un criterio de selección de gran importancia teniendo en cuenta las características del entorno para el cual está siendo diseñado el escenario.

Tabla 4. *Características de un sistema de reconocimiento basado en el motor de Verbio*

Características	
Arquitecturas	Monousuario, Cliente-Servidor
Requerimientos de memoria	Motor de reconocimiento (Vox Server): >10 MB Configuración de reconocimiento monolingüe: 16 MB Configuración de reconocimiento bilingüe: 29 MB
Tasa de muestreo	8 Khz (ley A o ley Mu)
CPU	Req. Mínimos recomendables Pentium 4 - 3 Ghz >= 512 MB RAM
Plataforma	Windows NT, 2000, XP, 2003,Linux compatibles Red Hat, compatibles Debian, Pocket PC
Servidores de voz	Verbio ASR, Verbio TTS, y locutores TTS SAPI 4.0 y 5.x, Servidores MRCP*
Interfaces	Verbio API VoiceXML
Características destacadas	Gramáticas establecidas por el W3C[15], reconocimiento independiente del locutor, optimizado para entornos telefónicos (fijo y móvil) y ruidosos, cosibilidad de desarrollo de nuevos idiomas "on-demand", configuraciones de reconocimiento multilingües, múltiples hipótesis de reconocimiento, gramáticas básicas incorporadas, integración con módulo de verificación del locutor

1.2.2.5 Microsoft Speech Server MSS

Es una plataforma propietaria de Microsoft basada en el sistema operativo de Microsoft Windows®, el entorno de desarrollo Visual Studio® .NET, esquemas de integración Telefonía- Computación (CTI- *Computer Telephony Integration*) y la especificación de Etiquetas de Lenguaje para Aplicaciones de Voz (SALT- *Speech Application Language Tags*).

MSS cuenta con dos componentes principales, uno de Servicios de Aplicación de Telefonía (TAS- *Telephony Application Services*), que llevan a cabo el procesamiento de las aplicaciones Web basadas en telefonía y controladas por voz y un segundo componente o Servicio de Motor de Reconocimiento (SES- *Speech Engine Services*), que maneja las interaccione de voz, traducción de texto a voz, entre otras funciones de habla; permitiendo el desarrollo dos tipos de aplicaciones de habla: Sistemas IVR que soportan entradas de habla y por tonos DTMF (habitual), soluciones para dispositivos móviles basados en Web y habilitados por voz sobre redes WiFi o LAN (*Local Area Network*), la integración de centros de datos vía Web a través de *callcentres* y PBX, combinando el uso de la voz como objetivo principal, con entradas convencionales como teclado telefónico, alfanumérico y mouse. Incluye herramientas de desarrollo de software para la creación de nuevos módulos y aplicaciones para PCs de escritorio y dispositivos portátiles como Pocket PCs, Tablet PCs, Laptops y PDAs (Personal Digital Assistant). [37][38][39][40]

[15] World Wide Web Consortium

En la figura 2 [40] se muestra la Topología de un servidor de voz basado en MSS.

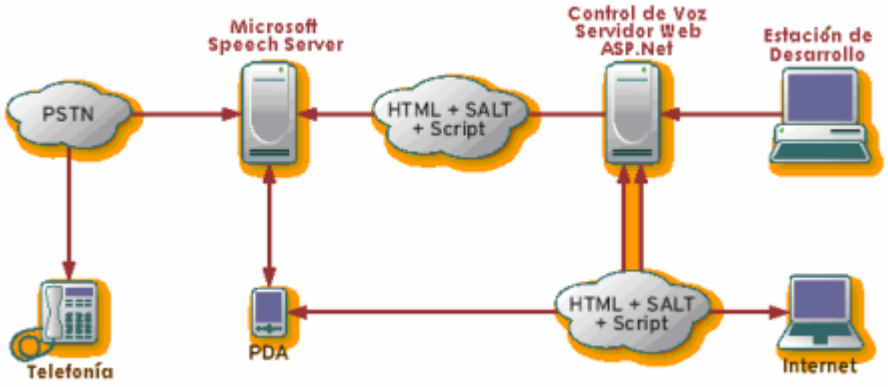

Figura 2. Topología de un servidor de voz basado en MSS

MSS representa una buena solución debido a que consiste en un paquete completo para el desarrollo de aplicaciones orientadas básicamente a IVR y *callcentres*, soportando gran cantidad de usuarios, pero su costo de implementación es elevado debido a que se trata de una plataforma propietaria de Microsoft, lo cual es un criterio importante a tener en cuenta en la selección de la herramienta, ya que, afecta la viabilidad y sostenibilidad económica, aunque no se descarta totalmente porque es una herramienta muy potente. No obstante, este no es un criterio único de selección, por lo que, más adelante en la sección 1.3 se realizará la comparación global que permite definir la herramienta adecuada para los propósitos perseguidos por el sistema.

Otro criterio importante de selección y que no puede pasarse por alto, consiste en la evaluación de los requerimientos mínimos del sistema para su instalación, los cuales se listan en la tabla 5 [41].

Tabla 5. *Requerimientos mínimos de un sistema para el soporte de MSS (en el servidor).*

Características	
Equipo	Procesador igual o superior a 2.5 GHz, 4GB de RAM o más, disco duro con formato NTFS con 1.7GB de espacio en disco disponible para la instalación y otros 10GB mínimo para otro software requerido, unidad de CD o DVD
Sistema Operativo	Microsoft Windows Server 2003 Standard Edition
Otro Software requerido	Internet Information Services (IIS) ASP.NET 1.1 con Service Pack 1 o superior Microsoft .NET 1.1 con Service Pack 1 o superior
Otros requerimientos	Tarjeta de video compatible con Windows 2003 con adaptador para una resolución mínima de 800x600, tarjeta de telefonía o de conexión VoIP

Como se puede observar en la tabla 5, el Microsoft Speech Server es una herramienta muy exigente en cuanto a las características del hardware que debe

adecuarse para su instalación en un esquema como el mostrado en la figura 3, y aunque es muy apropiada y potente en el campo de la telefonía, no es justificable económicamente para su implementación en el entorno sobre el cual se investigó en este proyecto, debido a que se trata de un entorno hospitalario mediano, cuya misión no es ofrecer servicios avanzados en comunicación, aunque deba hacer uso de ellos para su funcionamiento diario, pero en menor escala.

1.3 CRITERIOS DE SELECCIÓN DEL MOTOR DE RECONOCIMIENTO DE HABLA

Muchos factores están incrementando la aplicación de las tecnologías del habla en distintos entornos laborales cuyas exigencias se reducen a mantener una alta relación costo beneficio, teniendo en cuenta la aplicabilidad de los sistemas basados en ellas sobre sus escenarios de operación.

Pero además de ello, la selección de un ASR depende no solo de las características de la aplicación, sino también de las propiedades del ASR mismo, por ejemplo, del número de usuarios que puede soportar, la naturaleza de la expresión, es decir, si se trata de reconocimiento de palabras aisladas, reconocimiento continuo de habla o reconocimiento espontáneo, la complejidad y tamaño del lenguaje, lo que se refiere al tipo de palabras que han de utilizarse dependiendo del entorno de desempeño de la aplicación, las condiciones del entorno que puede degradar drásticamente el rendimiento del sistema, debido a fenómenos como el ruido, distorsiones, etc.; todo lo anterior sin contar con el aspecto económico que juega un papel fundamental, pues, debe definirse una herramienta que cumpla con los requerimientos técnicos que solicita el sistema, con una inversión adecuada, tanto en el momento de la implementación, así como a futuro cuando se requieran posteriores expansiones o actualizaciones.

Además de lo anterior, si se tiene en cuenta que el transporte de la voz se realiza en una infraestructura de red inalámbrica, mediante VoIP se debe evaluar si sus especificaciones soportan aplicaciones de telefonía, con respecto a lo cual, es preciso anotar que cada una de las herramientas presentadas como posible interfaz de acceso por voz, tienen su propios *codificadores*, que son de vital importancia en el envío de voz sobre redes de paquetes con VoIP, puesto que ellos inciden en la calidad y precisión del sistema. Sin embargo, como se utilizarán los *codificadores* propios que los reconocedores soporten, no se tomarán como un factor de selección de la herramienta como tal, sino que su selección dependerá de su comportamiento en la aplicación y de su efectividad para el reconocimiento del *Corpus* y la integración con el sistema de control (lo que implica que al momento de la selección del mejor *codec*, el motor de reconocimiento ya se ha seleccionado y probado).

Los criterios definidos para la selección, son los siguientes:

- **Soporte a aplicaciones telefónicas,** que cumpla las especificaciones de muestreo de voz, bajo parámetros adecuados para telefonía.

- **Soporte a aplicaciones de red,** ya que el PSCVoWLAN está basado en la construcción de un escenario cliente- servidor de telefonía, luego la herramienta debe funcionar no solo para un usuario aislado, sino en red.

- **Funcionalidad y Expansión.** Posibilidad de futura expansión, tanto para más usuarios, como para más aplicaciones. Esto se refiere a la posibilidad de configurar o crear módulos que faciliten la ejecución de la herramienta sobre diversas aplicaciones en el sistema operativo, es decir, que no se ofrezca únicamente de dictado y trascripción de texto, sino también control del entorno operativo.

- **Independencia del hablante.** Pues el sistema se debe implementar para el personal médico de la Clínica la Estancia S.A., el cual se estimó entre 20 y 60 profesionales de la salud, contando a médicos y enfermeras.

- **Licencia.** Tratándose de un sistema experimental se pretende que su implementación genere los resultados deseados, sin incrementar en grandes cantidades los presupuestos designados por la entidad para la parte de expansión tecnológica. Dentro de lo cual debe tenerse en cuenta si se trata de una herramienta de uso libre o si es necesario la compra de software y licencias legales.

- **Capacidad del *corpus*,** esto implica que además de las palabras que ya trae por defecto en sus diccionarios, la herramienta debe permitir la configuración de palabras externas a dicho diccionario.

- **Requerimientos computacionales.** Exigencias mínimas de los equipos para su instalación y ejecución, se medirá en la capacidad de adaptación a la infraestructura de red y a los recursos computacionales (hardware y software) existentes en la Clínica la Estancia S.A.

- **Idioma.** Significa el idioma para el cual funcionan los diccionarios del motor de reconocimiento.

- **Interfaz.** Lo que implica que la interfaz de interacción con el usuario final sea amigable y fácil de configurar y utilizar.

En la tabla 6 se listan los criterios de selección del motor de reconocimiento de habla y la calificación de 1-5 establecida para cada una de las diferentes herramientas consultadas, basadas en los requerimientos del PSCVoWLAN con el fin de concluir cual de las herramientas es la más apropiada para su

implementación. Entre más alta sea la calificación por cada criterio quiere decir que se adapta más a lo que se espera en este sistema particularmente.

Tabla 6. *Comparación de las herramientas de reconocimiento con base en los criterios de selección establecidos*

Criterio / Herramienta	Soporte a aplicaciones telefónicas	Soporte a aplicaciones de red	Funcionalidad y expansión	Licencia	Dependencia del hablante	Capacidad del *corpus*	Req. Computa-cionales	Idioma	Interfaz	Pro-medio
CVoice Control	No soporta aplicaciones de telefonía (1.0) ✗	(1.0) ✗	Trascripción de texto y ejecución de aplicaciones bajo Linux (5.0) ✓	Software Libre (5.0) ✓	Requiere entrenamiento para identificar la voz del hablante (1.0) ✗	Hasta 1.000 expresiones (3.0) ⚠	Bajos (5.0) ✓	Inglés (2.0) ✗	Con pocas opciones de configuración (2.0) ✗	2.8 ✗
Gvoice	Si soporta aplicaciones telefónicas (5.0) ✓	(2.0) ✗	Trascripción de texto y aplicaciones de telefonía (4.0) ✓	Software libre (5.0) ✓	Dependiente del hablante, requiere entrenamiento (1.0) ✗	No hay documentación especifica (1.0) ✗	Bajos (5.0) ✓	Inglés (2.0) ✗	Pocas opciones de configuración, herramienta de prueba (1.0) ✗	2.8 ✗
Sphinx 2	Si soporta aplicaciones telefónicas (5.0) ✓	(5.0) ✓	Trascripción de texto y ejecución de aplicaciones bajo Linux y Windows (5.0) ✓	Software libre (5.0) ✓	Independiente del hablante, aunque requiere entrenamiento para aprender a pronunciar las palabras según el modelo de lenguaje (3.0) ⚠	125.000 palabras (5.0) ✓	Bajos (4.0) ✓	Inglés y francés (2.5) ✗	Interfaz de configuración gráfica de Perlbox-Voice, de fácil manejo (4.0) ✓	4.2 ✓
Sphinx 3	No soporta aplicaciones de telefonía (1.0) ✗	(5.0) ✓	Trascripción de texto y ejecución de aplicaciones bajo Linux y Windows (5.0) ✓	Software libre (5.0) ✓	Independiente del hablante, requiere entrenamiento para aprender la pronunciación del léxico (3.0) ⚠	125.000 palabras (5.0) ✓	Superiores a los de Sphinx2 (3.5) ⚠	Inglés y francés (2.5) ✗	Interfaz programable en C++ acorde con la aplicación (3.0) ⚠	3.7 ⚠
Sphinx 4	No soporta aplicaciones de telefonía (1.0) ✗	(5.0) ✓	Trascripción de texto y ejecución de aplicaciones bajo Linux y Windows (5.0) ✓	Software libre (5.0) ✓	Independiente del hablante, requiere entrenamiento para aprender la pronunciación del léxico (3.0) ⚠	125.000 palabras (5.0) ✓	Medianos (3.0) ⚠	Inglés y francés (2.5) ✗	Interfaz programable en C++ acorde con la aplicación (3.0) ⚠	3.6 ⚠
IBM Via Voice	No orientado a telefonía	(1.0) ✗	Trascripción de texto en	Comercial (2.0) ✗	Dependiente del hablante, requiere	300.000 palabras	Bajos (4.0) ✓	Ingles Francés,	Interfaz gráfica con	2.9

Criterio Herramienta	Soporte a aplicaciones telefónicas	Soporte a aplicaciones de red	Funcionalidad y expansión	Licencia	Dependencia del hablante	Capacidad del *corpus*	Req. Computa-cionales	Idioma	Interfaz	Pro-medio
	aunque es configurable (3.0)		herramientas de oficina bajo Windows y Mac OS (2.0)		entrenamiento (1.0)	(5.0)		Alemán, Portugués, Mandarin y Español (3.5)	varias opciones, fácil de configurar (5.0)	
Dragon Naturally Speaking	No orientado a telefonía ni servicios IVR (1.0)	(1.0)	Trascripción de texto en herramientas de oficina Windows, orientado a aplicaciones médicas (3.5)	Comercial (2.0)	Independiente del hablante, aunque requiere entrenamiento para aprender a pronunciar las palabras según el modelo de lenguaje (3.0)	250.000 palabras de uso cotidiano y léxico médico (5.0)	Medianos (3.0)	Inglés (2.0)	Interfaz gráfica con varias opciones, fácil de configurar (5.0)	2.8
VoiceXpress	No orientado a telefonía ni servicios IVR (1.0)	(1.0)	Trascripción de texto en herramientas de oficina únicamente de la familia Microsoft (2.0)	Comercial (2.0)	Dependiente del hablante, requiere entrenamiento (1.0)	>5.000 palabras (5.0)	No hay documentación especifica (1.0)	inglés francés, holandés, español y alemán (3.5)	Interfaz gráfica configurable (5.0)	2.4
Verbio	Si soporta aplicaciones telefónicas (5.0)	(5.0)	Trascripción de texto y aplicaciones de telefonía para Windows 2000 y XP (3.5)	Comercial (2.0)	Independiente del hablante, no requiere entrenamiento (5.0)	>5.000 palabras (5.0)	Medianos (3.0)	Inglés (2.5)	Interfaz gráfica con varias opciones configurables (5.0)	4.0
Microsoft Speech Server	Si soporta aplicaciones telefónicas (5.0)	(5.0)	Trascripción de texto y aplicaciones de telefonía para Windows 2003 server (3.0)	Comercial (2.0)	Independiente del hablante, no requiere entrenamiento (5.0)	>5.000 palabras (5.0)	Altos (2.0)	Inglés,(otros idiomas en desarrollo) (3.5)	Interfaz gráfica con varias opciones configurables (5.0)	3.9

El PSCVoWLAN requiere una herramienta económica, independiente del hablante que soporte la implementación de aplicaciones telefónicas sobre redes IP, con posibilidades de expansión, por ello se seleccionó el CMU Sphinx 2 como la más conveniente para la implementación, pues, alcanza una alta calificación acode a los criterios establecidos y además de ello se prefiere por ser un desarrollo universitario que ha alcanzado grandes logros y que puede darse a conocer y al mismo tiempo mejorarse a partir de recomendaciones generadas sobre las conclusiones del proyecto.

1.4 *CORPUS* DE VOZ

Un *corpus* es una colección de grabaciones de voz con transcripciones de texto, los cuales se preparan y dividen para el desarrollo, prueba y entrenamiento de un sistema de reconocimiento de habla.

Los *corpus* de voz están diseñados para propósitos específicos que permiten determinar su contenido, es decir, que las características de los *corpus* cambian dependiendo del objetivo para el cual son creados, por ello, la creación de un *corpus* es una actividad fundamental en el desarrollo de un sistema con interfaz de habla, pues, estas palabras conforman el universo que éste puede reconocer y de allí depende su exactitud y precisión. [42][43]

Según las entrevistas realizadas a algunos integrantes del cuerpo médico de la Clínica La Estancia S.A. (Anexo A) acerca de las situaciones médicas más frecuentes, y el léxico comúnmente utilizado para cada una de estas circunstancias, se definió un *corpus* de 50 palabras clave con las cuales es posible configurar el Piloto para que ejecute las acciones correspondientes.

En este caso se trata de un *corpus* específico diseñado con palabras comúnmente utilizadas por el equipo médico, para el establecimiento llamadas y conferencias entre especialistas de diversas áreas y en determinadas circunstancias para la generación de la alarma de código azul y la solicitud de información en caso de no conocer las extensiones correspondientes.

Teniendo en cuenta que el motor de reconocimiento de habla estará basado en el Sphinx 2 de CMU, y que éste está definido para los idiomas Inglés y Francés, es necesario traducir el *corpus* de comandos del español al inglés para que éste sea adaptado al diccionario de fonemas correspondiente, resaltando que la pronunciación de cada palabra debe hacerse conforme a las reglas fonéticas del idioma inglés. Esto se puede observar en la tabla 7.

Tabla 7. *Corpus de Voz del personal médico de la Clínica La Estancia S.A.*

Instrucción en español	Comando en Inglés	División en unidades fonéticas según el diccionario de Sphinx
Enfermería – Área	Administration	AE D M IH N AX S T R EY SH AX N
Médico General – Área Administrativa	Administrative	AX D M IH N AX S T R EY DX IX V
Médico General – Admisiones	admisión	AE D M IH SH AX N
Enfermería – Admisiones	Admit	AX D M IH T
Enfermería – UCI Adultos	Adult	AX D AH L T(1) AE DX AX L T(2)
Médico General - Anestesiólogo	Anesthesiologist	AE N AX S TH IY Z IY AA L AX JH AX S T
Médico General – Área Hospitalaria	Area	EH R IY AX
Médico General – Auditor	Auditor	AO DX AX DX AXR
Enfermería – UCI neonatos	Baby	B EY B IY
Enfermería – Facturación	Billing	B IH L IX NG
Enfermería – Banco de Sangre	Bleed	B L Y ID
Médico General – Banco de Sangre	Blood	B L AH D
El Médico General da aviso de un Código Azul	Blue	B L UW
Médico General – UCI neonatos	Born	B AO R N
Médico General – Enfermera Jefe	Boss	B AO S
Enfermería – Terapia Respiratoria	Breathing	B R IY DH IX NG
Enfermería – Laboratorio clínico	Clinical	K L IH N AX K AX L
Conferencia (Medico general, traumatólogo, oncólogo)	Conference	K AA N F AXR AX N S K AA N F R AX N S
Enfermería – Banco de Datos	Data	D EY DX AX D AE DX AX
Médico General – Banco de datos	Database	D EY DX AX B EY S D AE DX AX B EY S
Enfermería – Imágenes diagnosticas	Diagnostic	D AY AX G N AA S T IX K
Auditor Interno – Auditor Externo	External	IX K S T ER N AX L
Médico General – Servicios Generales	General	JH EH N AXR AX L JH EH N R AX L
Médico General – Hematólogo	Hematologist	EH M AE T AA L AX JH AX S T
Enfermería – Área Hospitalaria	Hospital	HH AA S P IH DX AX L
Médico General – Imágenes Diagnosticas	Images	IH M AX JH AX Z
Médico General – IVR	Interactive	IH N T AXR AE K T IX V (1) IH N AXR AE K T IX V (2)
Auditor Externo – Auditor Interno	Intern	IH N T AXR N
Médico General – Facturación	Invoicing	IH N V OY S IX NG
Médico General – Laboratorio Clínico	Laboratory	L AE B R AX T AO R IY
Médico General – Enfermería	Nurse	N ER S
Médico General – Oncólogo	Oncologist	AA NG K AA L AX JH AX S T

Instrucción en español	Comando en Inglés	División en unidades fonéticas según el diccionario de Sphinx
Médico General – UCI adultos	One	W AH N (1) HH W AH N (2)
Médico General – Patólogo	Pathologist	P AX TH AA L AX JH AX S T
Médico General – UCI pediátrica	Pediatric	P IY DX IY AE T R IX K
Médico General – Pediatra	Pediatrician	P IY DX IY AX T R IH SH AX N
Médico General – Cirujano Plástico	Plastic	P L AE S T IX K
Enfermería – IVR	Response	R AX S P AA N S (1) R IY S P AA N S (2)
Enfermería – Servicios Generales	Services	S ER V AX S AX Z
Enfermería – Estadística	Statistic	S T AX T IH S T IX K
Médico General – Estadística	Statistical	S T AX T IH S T IX K AX L
Enfermería – Almacén	Store	S T AO R
Jefe de enfermería – Auditor Interno	Supervisor	S UW P AXR V AY Z AXR
Médico General – Cirujano Pediátrico	Surgeon	S ER JH AE N S ER JH AX N
Médico General – Cirugía	Surgery	S ER JH AXR IY
Médico General – Terapia Respiratoria	Therapy	TH EH R AX P IY
Médico General – Traumatólogo	Traumatologist	T R OW M AH T AA L AA JH AX S T
Enfermería – UCI pediátrica	Two	T UW
Médico General – Urólogo	Urologist	Y AXR AA L AX JH AX S T
Médico General – Almacén	Warehouse	W EH R HH AW S

Hasta este punto se han dado a conocer los aspectos teóricos más importantes de la tecnología ASR, y las herramientas a las que se puede acceder actualmente en el mercado, sus características y funcionalidades, que hicieron posible el establecimiento de los criterios económicos y técnicos necesarios para la selección del CMU Sphinx 2 como el motor adecuado para la implementación del módulo de reconocimiento de habla del piloto, teniendo en cuenta su integración con la PBX y la red de transporte sobre la cual se prestará el servicio, lo cual es objeto de estudio en el siguiente capítulo.

2. INTEGRACIÓN DE LA TECNOLOGÍA DE VOWLAN EN LA IMPLEMENTACIÓN DE SISTEMAS IVR BASADOS EN INTERFACES DE RECONOCIMIENTO DE HABLA

En este capítulo se presenta la justificación de la utilización de una Red Inalámbrica de Área Local (WLAN- *Wireless Local Area Network*) como infraestructura de transporte de voz en la implementación de un sistema de Respuesta de Voz Interactiva basado en una PBX controlada a través de la tecnología de reconocimiento del habla, incluyendo el estudio de la herramienta de implementación del sistema IVR y su integración con el motor de reconocimiento seleccionado en el capítulo anterior.

2.1 JUSTIFICACIÓN DE LA UTILIZACIÓN DE VOWLAN EN LA IMPLEMENTACIÓN DEL PILOTO

Con el fin de lograr un acercamiento a la integración de servicios que involucren tecnología de habla dentro de entornos inalámbricos locales, a continuación se introducen los conceptos básicos de la tecnología VoWLAN, aclarando que no se profundizará en el estudio de esta tecnología, pues, esto se ha analizado suficiente en el trabajo de grado titulado “Prototipo Experimental de VoIP sobre WLAN para Entornos Empresariales”. [44]

Una red 802.11 constituye un sistema de comunicación de datos implementada como una extensión de una red local cableada dentro de un edificio o campus; se basa en una arquitectura donde el sistema se subdivide en células, cada una de las cuales está controlada por una estación base, llamada Punto de Acceso (AP- *Access Point*), conectados entre sí por un Sistema de Distribución (DS- *Distribution System*) o *backbone*.

Incluso, una WLAN podría consistir en una única célula, con un único AP y en determinados casos sin AP, todos estos componentes interconectados vistos en las capas superiores del modelo OSI como una simple red 802, conforman lo que se llama Grupo de Servicio Extendido (ESS- *Extended Service Set*), tal como se ilustra en la figura 3.[45]

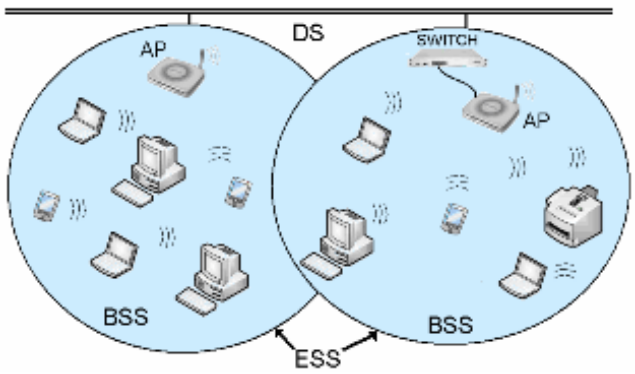

Figura 3. Esquema básico de una WLAN

802.11 divide el espectro en canales de tal forma que es posible instalar y configurar APs en diferentes canales para que pueden operar sin interferencia, por lo cual, define estándares identificados como "a," "b," "g" que especifican el manejo de varias frecuencias que se utilizan en las implementaciones prácticas actualmente. [46]

- 802.11a opera en la banda de 5GHz ofreciendo muchos más canales de radio, lo cual es ventajoso ya que no solo ayuda a evitar que se presente interferencia radio y microondas, sino que también permite aumentar el número de puntos de acceso en diferentes canales dentro de la misma área sin que interfieran uno con el otro, incrementando significativamente el rendimiento de la WLAN dentro de un área dada, permitiendo con mayor calidad el transporte de aplicaciones más exigentes como conferencia y salas de usuarios. Ofrece una velocidad de datos máxima teórica de 54Mbps y 12 canales no traslapados, 8 para red inalámbrica y 4 para conexiones punto a punto, en los cuales la velocidad real de transferencia es de 25Mbps aproximadamente. [45][47][48][49]

  Si bien, 802.11a es ventajosa puesto que la banda en la que opera se encuentra menos "contaminada", no obstante, la alta frecuencia en la que opera (comparado con 802.11g/b) limita su cobertura, lo que obliga a instalar más puntos de acceso para cubrir la misma área, además, incrementa las pérdidas y al mismo tiempo tiene mayor dificultad para penetrar paredes y superar otros obstáculos, debido a la reflexión, condicionando su movilidad.

  Por otra parte, la utilización de 802.11a es poco común debido a varios factores, principalmente que no es compatible con 802.11b/g, aunque puede coexistir con ellos sin afectar su desempeño. [45][47][48][49]

- 802.11g opera en la banda de 2.4GHz, y ofrece una velocidad de datos teórica de 54Mbps, aunque en la práctica la velocidad real de transferencia sea de 24.7Mbps aproximadamente, soportando mayor número de usuarios

simultáneos; utilizando 22MHz para transmitir cada señal, debido a que en esta tecnología sólo hay 3 canales sin traslape y por consiguiente, limita el número de puntos de acceso que puedan conectarse sin que ocurra traslape entre las celdas cubiertas por cada uno, haciendo difícil la asignación de canales cuando el área de cobertura es grande y la densidad de usuarios es alta, igual que con 802.11b. [45][47][48][49]

- 802.11b utiliza el espectro de 2.4GHz y cuenta con 11 canales cada uno de los cuales puede ofrecer una velocidad de datos máxima teórica de 11Mbps, aunque en la práctica es de 5.8 a 7Mbps. Es compatible con 802.11g debido a la frecuencia de operación y la técnica de modulación en la que se basan, lo cual facilita posteriores actualizaciones a este tipo de equipos y por consiguiente a sus capacidades. [45][47][48][49]

En la tabla 8 [50][51][52][53][54] se establece una comparación, entre los estándares de red inalámbrica, en donde se describen las principales especificaciones técnicas de cada uno, con base en lo cual, se concluye que el estándar más apropiado y por tanto, el seleccionado para el establecimiento del piloto es 802.11g, por permitir una gran variedad de ventajas técnicas entre las que se puede ver el aumento en el número de canales para evitar la interferencia, el rango de cobertura que puede alcanzar, la velocidad de datos y el número de usuarios que puede cubrir por celda, además también juegan a su favor el hecho de ser medianamente económico y más comúnmente utilizado, lo cual permite futuras expansiones, migraciones y aumento de capacidades.

Tabla 8. *Comparación de estándares inalámbricos*

Característica	Definición	802.11b	802.11g	802.11a
Canales RF disponibles	Cantidad de enlaces de comunicación	3 de 11 sin superposición (1,6 y 11)	3 de 11 sin superposición (1,6 y 11)	12 que no se traslapan
Velocidades teóricas de datos por canal (Mbps)	Valores teóricos configurables máximos por canal RF	1, 2, 5.5 y 11	6, 12, 18, 24, 36, 48, 54	6, 12, 18, 24, 36, 48, 54
Máxima velocidad real de datos por canal Mbps	Valor real máximo	5.8- 7	24.7	25
Banda de frecuencia	Intervalo de frecuencia de difusión	2,4 GHz ISM[16]	2,4 GHz ISM	5 GHz UNII[17]
Ancho de		22MHz	22MHz	20MHz

[16] Industrial Scientific and Medical, son bandas reservadas internacionalmente para uso no comercial de Radio Frecuencia electromagnética en áreas industrial, científica y médica.
[17] Unlicensed National Information Infrastructure

Característica	Definición	802.11b	802.11g	802.11a
banda de canal				
Alcance	Máxima distancia por velocidad*	3- 45m a 5.8- 7Mbps 50- 61m a 3.7- 5Mbps 65- 76m a 1.6- 3Mbps 80- 91m a 0.9- 2Mbps	3- 15m a 24.7Mbps 35m a 19.8Mbps 45m a 12.4Mbps 61m a 4.9Mbps 76m a 1.6Mbps 91m a 0.9Mbps	3- 15m a 24.7Mbps 30m a 19.8Mbps 45m a 12.4Mbps
Densidad de Usuarios	Número máximo de usuarios por AP (aprox)	32	64	64
Costos	Implementación y expansión	Menos costosa que g y a	b <g< a	Más costosa que b y g

*Tomada como velocidad máxima, la velocidad máxima real.

En la práctica, la complejidad de la configuración física de una WLAN puede ser variable, y se requiere identificar los requerimientos del sistema a implementar para lograr una estructura eficiente y que se adapte a las necesidades.

Una de las razones primordiales en la elección de una red inalámbrica es su transportabilidad, muy a menudo los equipos deben comunicarse con otros que pueden ser de cualquier tipo, bien sea portátiles, o lo más probable, computadores conectados a una LAN por cable. Estas estructuras se conocen como topologías, y en el caso de redes inalámbricas se conocen dos principalmente: redes AdHoc y las WLAN de infraestructura.

Las **redes AdHoc** corresponden a la configuración más simple frecuentemente llamada Conjunto Básico de Servicio Independiente (IBSS- *Independent Basic Service Set*), se trata de una red inalámbrica independiente que conecta un grupo de PCs por medio de sus adaptadores de red inalámbricos, sin necesidad de requerir servicios de una infraestructura cableada. [55]

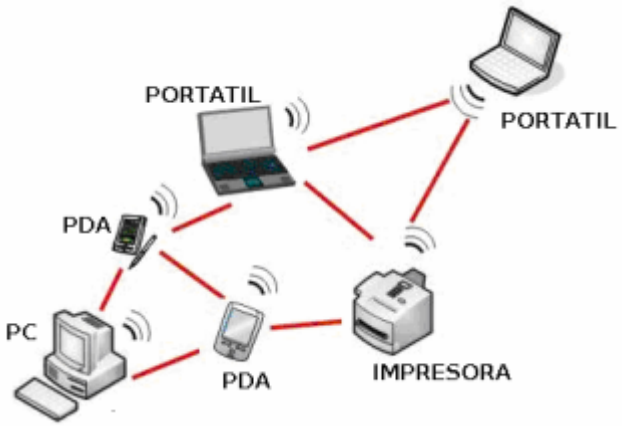

Figura 4. Red básica WLAN en configuración AdHoc

Como se observa en la figura 4, este tipo de redes no necesita un nodo central, sino que sus estaciones se comunican directamente entre si, y el único requisito deriva del rango de cobertura de la señal, ya que es necesario que los terminales móviles estén dentro de este rango para que la comunicación sea posible; esto varía dependiendo del estándar seleccionado. Por otro lado, estas configuraciones son muy sencillas de implementar y no es necesario ningún tipo de gestión administrativa de la red, de modo que comparten todos a la vez la posibilidad de ser clientes y servidores simultáneamente.

Por su parte, las **WLAN de infraestructura** se basan en concepto de celdas[18], utilizando APs, que funcionan como repetidores y por tanto son capaces de doblar el alcance de una red inalámbrica *AdHoc*, ya que la distancia máxima permitida no es entre estaciones, sino entre una estación y un punto de acceso (un AP puede funcionar en un rango de al menos treinta metros y hasta varios cientos de metros). [55] Tal como lo ilustra la figura 5.

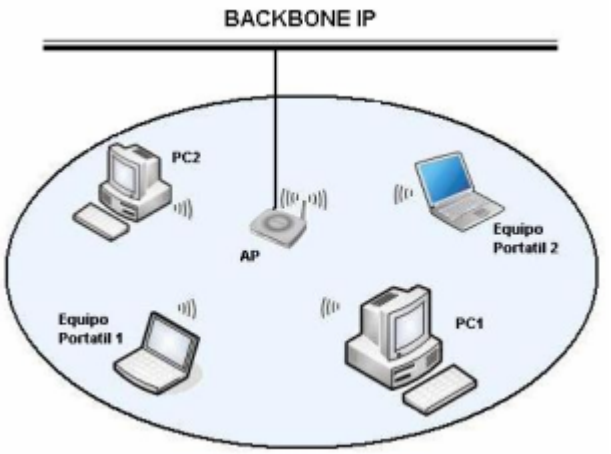

Figura 5. Red básica WLAN en configuración Infraestructura.

Esta topología además del evidente aumento del alcance de la red, permite el *roamming* gracias a la utilización de varios puntos de acceso que conforman las diferentes celdas que se entrelazan en algún punto de la red, como se indica en la figura 6.

Esto representa una de las características más interesantes de las redes inalámbricas ya que los terminales pueden moverse sin perder la cobertura y sin sufrir cortes en la comunicación mientras recorren las celdas.

[18] Área de cobertura en el que una señal radioeléctrica es efectiva

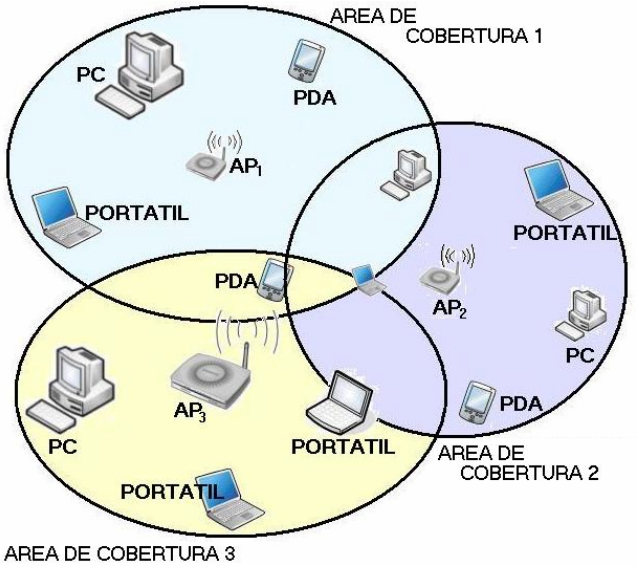

Figura 6. Roamming.

La topología en infraestructura es la más utilizada, pues ofrece la gran ventaja de facilitar la implementación de redes un poco más complejas y que combinan redes cableadas con redes inalámbricas, que se adapten perfectamente a gran cantidad de requerimientos de expansión y movilidad [55], como es el caso del PSCVoWLAN donde es importante contar con un tipo de cobertura como la que brindan los AP y que al mismo tiempo se puedan aprovechar las altas velocidades y la casi supresión de la interferencia para dejar al servidor en la parte de red cableada y que algunos clientes puedan pertenecer también a la red cableada. Esto se ilustra la figura 7.

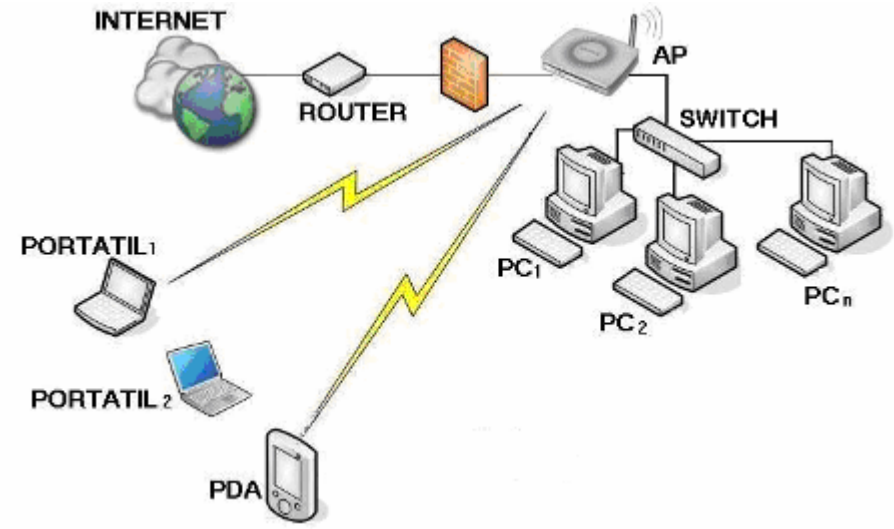

Figura 7. Combinación de WLAN y LAN

Voz sobre Protocolo de Internet, también llamado Voz sobre IP, VoIP, Telefonía IP, Telefonía por Internet, Telefonía *Broadband* y Voz sobre *Broadband* es el enrutamiento de conversaciones de voz sobre Internet o a través de alguna otra red basada en IP.

La implementación de WLAN para el transporte de voz consiste en transmitir Voz sobre IP, sobre una red Wi-Fi (red de paquetes), esto integra telefonía cableada e inalámbrica en la misma infraestructura IP, lo que generalmente es más simple y menos costoso que si se manejaran dos entidades separadas. Para ampliación de la información técnica respecto a la tecnología de VoWLAN, refiérase al documento "Prototipo Experimental de VoIP sobre WLAN para Entornos Empresariales. [44]

Básicamente, los sistemas VoWLAN trabajan en diferentes maneras, una de ellas consiste en enrutar llamadas de voz desde un teléfono a un AP y desde este a una gateway VoIP que lleve el tráfico a una red IP, para finalmente entregarlas a la PBX, en este caso las llamadas que se hagan desde teléfonos ubicados fuera del entorno de la red, se dirigirán a través de la PBX hacia la Public Switching Telephonic Network (PSTN).

Para el caso específico de la implementación de este piloto se utilizan *Softphones* (teléfonos basados en software) para generar las llamadas a través de la Intranet, de esta manera, el personal médico puede realizar llamadas desde cualquier lugar de la clínica ubicado dentro de los puntos de cobertura, utilizando sus PCs, equipos portátiles o PDAs en movimiento, aunque es destacable que estas implementaciones de VoIP están siendo cada vez más populares por medio de teléfonos especiales para la tecnología, y a pesar que aún no se ha descartado la telefonía fija convencional, la telefonía IP está emergiendo con gran fuerza y esta ocupando un importante lugar dentro de las cada vez mayores opciones de comunicación en hogares y empresas, precisamente por su integración natural con CTI.

2.1.1 Codificación/Decodificación

Para enviar audio a través de una red de datos, la forma de onda de audio analógica ha de ser codificada en bits de datos que puedan ser procesados por un PC. Se muestrea, se cuantifica y se comprime para que ocupe la mínima cantidad de ancho de banda; una vez el sonido llega a su destino, se invierte el proceso.

Este proceso se realiza mediante un *codec* de audio, también conocido como *codec* de habla, el cual, está diseñado para la compresión y descompresión de señales de sonido audible para el ser humano.

Estos dispositivos o aplicaciones comprimen las secuencias de datos, realizan cancelación de eco y aprovechan las pausas entre palabras, períodos de silencio y cambios predecibles en las amplitudes para reducir el consumo de ancho de banda para transmitir la voz humana.

Esto es especialmente interesante en los enlaces de poca capacidad y permite tener un mayor número de conexiones de VoIP simultáneamente.

De ellos existen fundamentalmente dos aplicaciones en las cuales son útiles los *codec* de audio y de lo cual depende su clasificación [56]:

- **Almacenamiento** útil para reproductores multimedia que pueden reproducir sonido almacenado, por ejemplo, en un disco duro, CD-ROM o tarjeta de memoria.
- **Transmisión** cuya aplicación permite principalmente implementar redes de videoconferencia, telefonía móvil e IP.

En el caso específico de desarrollo de este piloto se va a tratar únicamente los *codificadores* de audio orientados a transmisión, debido a su influencia directa en la implementación del mismo.

2.1.1.1 Caracterización de los *codificadores* de audio

Los *codificadores* de audio se caracterizan por los siguientes parámetros:

- **Número de canales.** Un flujo de datos codificado puede contener una o más señales de audio simultáneamente. De manera que puede tratarse de audiciones "mono" (un canal), "estéreo" (dos canales, lo más habitual) o multicanal cuya aplicación mas común es en sistemas de entretenimiento "cine en casa" ofreciendo seis (5.1) u ocho (7.1) canales. [56]

- **Frecuencia de muestreo.** Determina la calidad percibida a través de la máxima frecuencia que es capaz de codificar, que es precisamente la mitad de la de muestreo. Por tanto, cuanto mayor sea ésta, mayor será la fidelidad del sonido obtenido respecto a la señal de audio original. [56]

- **Pérdida.** Algunos *codificadores* pueden eliminar frecuencias de la señal original que, teóricamente, son inaudibles para el ser humano. De esta manera se puede reducir la frecuencia de muestreo, lo que se conoce como *codec* con pérdida. En caso contrario se dice que es un *codec* sin pérdida. [56]

- **Velocidad de datos o bit-rate.** Es el número de bits de información que se procesan por segundo, teniendo en cuenta la frecuencia de muestreo

resultante, la profundidad de la muestra en bits y el número de canales. A causa de la posibilidad de utilizar compresión (con o sin pérdidas), la tasa de bits no puede deducirse directamente de los parámetros anteriores. [56]

Existen dos métodos para probar la calidad de la voz, subjetiva y objetivamente, realizados por el ser humano y los computadores respectivamente.

Las medidas estándar de calidad objetiva, como la total distorsión armónica y relación señal a ruido no se corresponden muy bien con una percepción de calidad de voz humana; por lo tanto no respaldan el objetivo de las técnicas de compresión de voz que consiste en hacer cada vez más natural la voz digitalizada.

Existe la recomendación ITU-T P.861 [57], que trata de los mecanismos con los que se puede determinar objetivamente la calidad de voz utilizando la Medida de la Calidad de la Voz según la Percepción (PSQM- *perceptual speech Quality Measurement*), sin embargo, presenta muchos inconvenientes cuando se trabaja con codificadores de voz debido a que fue desarrollado para oír deterioros provocados por la compresión y descompresión y no por la perdida de paquetes o la fluctuación de fase. [44]

Por su parte, la Puntuación de Opinión Media (MOS- *Mean Opinion Score*) siendo un parámetro subjetivo, permite medir la calidad de la voz en una escala de 1 a 5 puntos, obtenida mediante el cálculo de la media de las calificaciones otorgadas por un variado grupo de oyentes a quienes se les da a escuchar muestras de voz obtenidas con determinado *codec*. [58][59] La escala de la MOS se observa en la tabla 9. [58]

Tabla 9. *Puntuación de Opinión Media*

MOS	Calificación	Descripción respecto al ruido
5	Excelente	Imperceptible
4	Bueno	Perceptible pero no molesto
3	Aceptable	Levemente molesto
2	Deficiente	Molesto
1	Malo	Muy molesto

2.1.1.2 *Codificadores* de audio

Existe una gran cantidad de *codificadores*, de los cuales sólo se describen los que están incluidos en los *softphones* del servidor y los clientes del piloto, SJPhone y X-Lite, respectivamente; con los cuales se realizaron pruebas documentadas posteriormente.

- **ITU G.711**

  G.711 es la estandarización de la ITU-T para la Modulación por Impulsos Codificados (PCM, *Pulse Code Modulation*) para representar señales de audio con frecuencias de la voz humana (con ancho de banda de 3.4KHz), mediante muestras comprimidas de una señal de audio digital con una frecuencia de muestreo de 8 kHz, proporcionando un flujo de datos de 56 o 64 Kbps.

  Para este estándar existen dos algoritmos principales de compresión, μ-law (ITU G.711 ley μ) y el A-law (ITU G.711 ley A) [60], las cuales son el requisito básico de la mayoría de los estándares de comunicación multimedia de la ITU, utilizando un método de compresión de amplitud logarítmica para alcanzar de 12 a 13 bits de calidad PCM lineal en 8 bits, pero se diferencian en detalles de compresión relativamente menores (la ley μ tiene una ligera ventaja en la capa baja y alto rendimiento en relación señal a ruido).

  G.711 es el método de codificación de señal de audio analógica más popular y es ampliamente utilizado en telefonía PSTN, PBX e IP, ofreciendo altos niveles de satisfacción en la calidad de la voz sobre WLAN, a pesar de no soportar compresión de ancho de banda, según estudios realizados por el Instituto de Investigación en Redes de Comunicaciones y el Instituto de Tecnología de Dublín. [61]

- **GSM**

  El *codec* GSM utilizado en telefonía móvil celular, es un *codec* libre, con una característica particular, poco consumo de procesador. Lo cual, es importante cuando se está trabajando a gran escala y sobre todo en terminales que tienen baja capacidad de procesamiento.

  Este *codec* utiliza una frecuencia de muestreo de 8KHz a una velocidad de 13 Kbits/s, las últimas versiones como la GSM 6.10 se ha perfeccionado para la reproducción del habla ofreciendo buena calidad de la voz. Se utiliza comúnmente en muchas aplicaciones de videoconferencia y telefonía IP, ya que, consigue una compresión elevada con una calidad aceptable de audio, por lo general voz humana. [62]

- **Speex**

  Optimizado para el habla y diseñado para la comunicación de baja latencia sobre una red de paquetes no confiable. No es tecnológicamente el *codec* más avanzado de habla disponible, sin embargo, Speex se adapta bien a Internet y posee características importantes que no están presentes en otros *codificadores*, como codificación de intensidad estéreo, múltiples frecuencias

de muestreo (8,16,32 KHz), es de fuente abierta, y permite comprimir voz a tasas de bits desde 2 Kbps a 44 Kbps.

No obstante, a pesar de sus grandes ventajas presenta la desventaja de que consume mayor cantidad de recursos computacionales que GSM, lo cual genera retardos en procesamiento y por tanto disminuye el rendimiento del sistema. Esto es un punto importante a tener en cuenta en la implementación de este piloto ya que sería demasiado exigente considerando los recursos computacionales de los que se dispone para las pruebas. [63]

- **iLBC (Internet Low Bitrate *Codec*)**

  Este *codec* diseñado principalmente para aplicaciones de voz en banda estrecha, es conveniente para las comunicaciones de voz sobre IP, ya que, en el caso de los *codificadores* comunes con una baja velocidad de bits aprovechan la dependencia entre las tramas de voz pero como resultado de esto se presentan errores de propagación cuando los paquetes se pierden o retardan. En contraste a lo anterior, iLBC codifica las tramas independientemente, lo que le da a iLBC mayor robustez en contra de los paquetes perdidos o con retardo. [64]

- **Broadvoice-32 y Broadvoice-32 FEC**

  Broadvoice es una familia de *codificadores* de audio desarrollados para aplicaciones de VoIP, logra una alta calidad en la voz con un bajo retardo en la codificación y una complejidad relativamente pequeña. Están disponibles dos versiones, una de Banda estrecha llamada Broadvoice-16 o BV16 y la de Banda ancha llamada Broadvoice-32 o BV32 que soportan telefonía estándar e IP respectivamente.

  BV16 muestrea a una frecuencia de 8KHz, con una velocidad de 16Kbps, lo cual es reducido comparado con los 64Kbps de G.711, esto permite incrementar el número de líneas que podría manejar un proveedor de servicio de telefonía IP.

  Por su parte, BV32, codifica la voz a una frecuencia de muestreo de 16 Khz con una velocidad de 32 Kbps, utilizando tramas de 5ms para minimizar el retardo de las comunicaciones bidireccionales en tiempo real. [65]

- **DVI4 y DVI4 Wideband**

  Trabaja a una frecuencia de muestreo variable (11,02 KHZ o 22.05 KHz), tomando muestras de 4 bits. La codificación se hace colocando en el

encabezado un valor predeterminado en lugar del valor de la muestra, ya que, solamente se tienen muestras de compresión de igual valor se utiliza el valor predeterminado para decodificar la primera muestra. [66]

En la tabla 10 [67] se presenta un resumen de los *codificadores* anteriormente mencionados, comparando sus características técnicas más importantes, resaltando que, algunos de ellos también están incluidos en el API traductor de *codificadores* de Asterisk, y se mencionan sin entrar en detalles, en la tabla 12 de la sección posterior cuando se trate del tema.

Tabla 10. *Codificadores de los softphones X-Lite y SJPhone*

Codec	**Velocidad de datos (Kbps)**	**Frecuencia de muestreo (KHz)**	**Tamaño de la muestra (ms)**	**Ancho de banda Ethernet (Kbps)**	**MOS**
G.711	64	8	10	87	4.2
GSM 6.10	13.2	8	22.5	31.2	3.5
Speex	8, 16, 32	2.15-24.6 4-44.2	30 34	17.63 – 59.63	2.92
iLBC	8	15.2, 13.3	25 30	30.83	3.95 3.88
DVI	32	11.02, 22.05	Variable	-	-
BV32/ BV16	16	8	-	-	-

En el anexo C se documentan las pruebas que se realizaron a los *codificadores* de los *softphones* del cliente y el servidor respectivamente con el fin de descartar las parejas que presentaran incompatibilidad, para, posteriormente hacer pruebas subjetivas basadas en MOS entre los *codificadores* compatibles con el objeto de comprobar la calidad de la voz, el reconocimiento y la movilidad, las cuales están documentadas en el capítulo 4.

2.2 SISTEMAS DE TELEFONÍA IP Y RESPUESTA DE VOZ INTERACTIVA

Los Sistemas de Respuesta de Voz Interactiva corresponden a la implementación más común de lo que actualmente se conoce como Integración Telefonía-Computación. Consiste en un sistema telefónico automático computarizado capaz de recibir llamadas e interactuar con el o los llamantes, a través de síntesis de voz o respuestas pregrabadas apropiadamente que conforman el cuerpo del menú.

Esta tecnología está orientada a entregar y/o capturar información automatizada a través del teléfono o dispositivo similar por medio de entradas DTMF o comandos de voz, permitiendo crear, acceder y gestionar múltiples servicios empresariales y domésticos, incluyendo identificación y enrutamiento de llamadas, solicitud de pedidos, reservas, servicios de información y operaciones autorizadas como el

acceso a bases de datos, acceso a correo electrónico, tele votación, entre otros. [68]

Un esquema de un sistema IVR se muestra en la figura 8 [69], en la que puede verse la integración de la telefonía IP con la PSTN, en la prestación de un servicio de respuesta de voz interactiva (basado en un servidor de procesamiento de llamadas, gestionable a través de un terminal de consola remoto o local) a usuarios de una red que incluye terminales IP tanto hardware como software, fijos y móviles, y teléfonos convencionales a través de una interfaz apropiada (gateway VoIP/PSTN).

En este esquema cualquiera de los usuarios a través de sus terminales puede comunicarse con el servidor de llamadas, el cual por medio de un menú de voz permite desplegar todas las posibles opciones, para de este modo enrutar llamadas entre los mismos usuarios, o establecer configuraciones de servicio a cada uno.

Por lo general, los menús de opciones presentados en la mayoría de los casos tienen niveles superiores a 1, esto implica que los usuarios del sistema deberán navegar por más opciones aparte de las desplegadas en primera instancia, lo cual genera un proceso largo y en algunos casos tedioso.

En el caso particular de la construcción de este Piloto, es importante destacar que su escenario de aplicación es hospitalario y que no se realizará la implementación de un menú para las comunicaciones entre el personal médico o de enfermería, dado que no es conveniente, pues, en casos de emergencia no tienen demasiado tiempo para esperar que el sistema lea todas las opciones ni navegar a través de ellas.

No obstante, con el objetivo de mostrar la funcionalidad de la tecnología IVR y su posible aplicación en entornos hospitalarios, donde sea necesaria, se realizó la programación de un servicio en el cual los usuarios pueden invocar las opciones de un servidor de respuesta interactiva que despliega un menú que ofrece información sobre las extensiones telefónicas y la posibilidad de establecer llamadas entre si, a través de marcación DTMF.

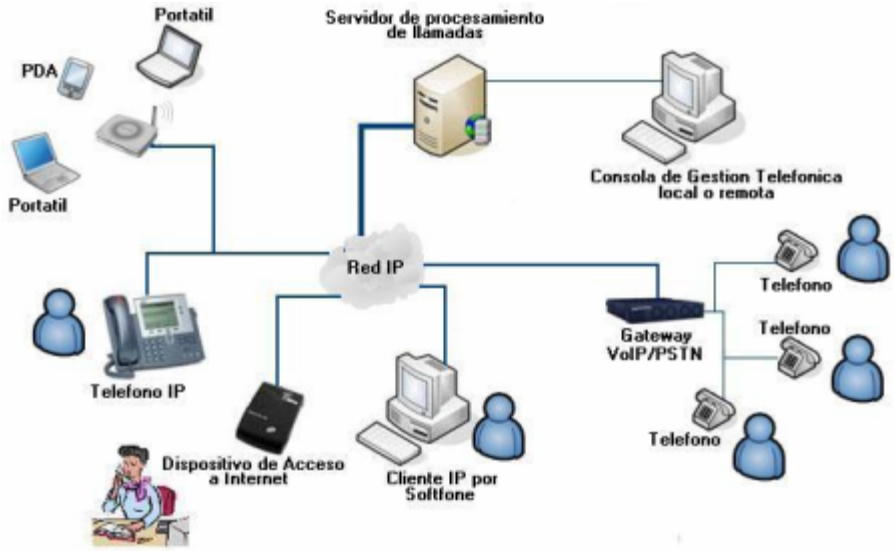

Figura 8. Esquema general de un sistema IVR

Actualmente, existen muchas opciones posibles en el mercado, para la implementación de sistemas de comunicación que combine telefonía IP, telefonía convencional PSTN y que permitan la implementación de aplicaciones de respuesta de voz interactiva, sin embargo, teniendo en cuenta que para la construcción de este piloto se ha adoptado la alternativa de trabajar con software libre, se analizaron las PBX de VoIP más completas, cuya comparación se observa en la tabla 11[44].

Tabla 11. *Comparación entre servidores de telefonía IP*

	Asterisk	**OpenPBX**	**Bayonne**
Ventajas	Muy Completo y confiable, software libre, tiene muchas opciones de personalización, diseñado para funcionalidad completa PBX Compatible con Pasarelas VoIP-PSTN, muy bien documentado, muchos servicios, y posibilidad de agregar nuevos, posibilidad de integración con Sphinx	Es software libre Interfaz amigable Fácil configuración	Muy Confiable Es software libre Compatible con Pasarelas VoIP-PSTN
Desventajas	Soporte	Solo para implementaciones pequeñas Muy limitado Soporte	Configuración compleja, no tiene muchas opciones de personalización, no es diseñado para funcionalidad completa PBX, no hay mucha documentación ni soporte

En general un servidor de telefonía IP no necesita hardware adicional para el establecimiento de llamadas de VoIP; mientras que, para interconexión con equipo telefónico analógico y digital, se debe configurar una tarjeta Hardware que soporte T1 o E1. Sin embargo, en el caso particular de este piloto el plan de marcado sólo incluye comunicaciones entre los terminales de la red interna, por tanto, no se realizaron pruebas que involucren conexión con la PSTN para llamadas externas a la red y no se tendrá en cuenta su utilización dentro del plan de pruebas. [44]

Comparando algunas de las ventajas y desventajas de estos servidores de telefonía IP, se observa claramente que Asterisk es el más completo de todos, por tal motivo es la herramienta seleccionada para la construcción del sistema telefónico, principalmente por que ofrece documentación detallada sobre su configuración e integración con el motor de reconocimiento CMU Sphinx 2, elegido previamente. Además es de libre utilización y sus capacidades ya han sido probadas en proyectos de laboratorio que confirman su alto desempeño y potencia. Sin embargo se deja libre esta elección para futuras implementaciones, en las que pueda ser más conveniente otra de las opciones.

En la siguiente sección se presentan las características fundamentales de Asterisk y su integración con Sphinx 2.

2.2.1 Asterisk

Asterisk es una plataforma PBX basada en software de código abierto desarrollado originalmente para sistemas operativos Linux, actúa como middleware conectando tecnologías de telefonía, tales como servicios VoIP (MGCP[19], SIP[20], IAX[21], H.323) y telefonía tradicional (T1, RDSI PRI y BRI, PSTN); con aplicaciones de telefonía, con la gran ventaja que los usuarios pueden crear nuevas funcionalidades escribiendo los programas en el lenguaje de script[22] de Asterisk o añadiendo módulos escritos en lenguaje C o en cualquier otro lenguaje de programación soportado por Linux. [70]

Entre las funcionalidades de llamada (tipo central PBX) más sencillas que Asterisk ofrece se tienen:

- Transferencia
- Música en espera
- Registro de llamadas en BD
- Llamada en espera
- Salas de Conferencia

[19] Media Gateway Control Protocol
[20] Session Initiation Protocol
[21] Inter-Asterisk eXchange
[22] Los lenguajes de script (o lenguajes interpretados) forman un subconjunto de los lenguajes de programación, que incluye a aquellos lenguajes cuyos programas son habitualmente ejecutados en un intérprete en vez de compilados.

- Contestador automático y/o desvío de llamadas
- Identificación de llamante
- Buzón de Voz personal
- Bloqueo de llamante
- Colas de llamada
- Desvío si no responde
- Timbres distintivos
- Colas con prioridad

Otras funcionalidades, más avanzadas que las anteriores son las siguientes:

- **IVR,** gestión de llamadas con menús interactivos.
- **LCR (Least Cost Routing),** encaminamiento de llamadas por el proveedor VoIP más económico.
- **AGI (Asterisk Gateway Interface),** integración con todo tipo de aplicaciones externas.
- **AMI (Asterisk Management Interface),** gestión y control remoto de Asterisk.

Dado que la elección de la herramienta para el servicio de telefonía es clave, debe tenerse en cuenta la estabilidad y el soporte que ofrece a los componentes de la aplicación. Asterisk puede instalarse en plataformas GNU/Linux 2.X, MacOSX 10.x y Microsoft Windows por medio de Cooperative Linux, que es una aplicación Open Source para correr de forma nativa Linux sobre una plataforma de Microsoft Windows, ya sea Windows 2000, Windows XP Home o Professional, aunque esta última opción no es recomendada puesto que no ofrece todo el soporte y las funcionalidades que ofrece Linux original, por tal motivo se escogió una distribución GNU/Linux muy conocida y apropiada como Debian para la implementación del piloto.

En la figura 9 [71] se muestra un escenario sencillo de la implementación de una central telefónica empresarial gracias a una PBX desarrollada con Asterisk en un equipo servidor dentro de una LAN que transporta voz y datos simultáneamente a través de protocolos VoIP, en la que todos los usuarios acceden al servicio telefónico a través de terminales IP software o hardware.

Los teléfonos analógicos convencionales pueden conectarse a la red a través de adaptadores (ATA - *Analog Telephony Adapter*) que se encargan de transformar tanto la señalización como la información de voz.

El servidor de Asterisk, dispone de una o más tarjetas que le permiten conectar directamente una serie de teléfonos analógicos y también líneas RDSI, así como

un proveedor externo de VoIP, de forma que cada usuario de la oficina podrá comunicarse con otro usuario de la red VoIP de ese proveedor.

Además, la red local en la que se encuentra la central está conectada a Internet a través de un router, lo que quiere decir que utilizando exclusivamente VoIP, cualquier usuario de la oficina puede comunicarse telefónicamente con cualquier otro usuario VoIP a nivel mundial. Esta comunicación podrá realizarse directamente si ese otro usuario está registrado en la centralita, por ejemplo, si la compañía tiene varias oficinas repartidas por todo el mundo puede tener registrados todas las extensiones de sus trabajadores.

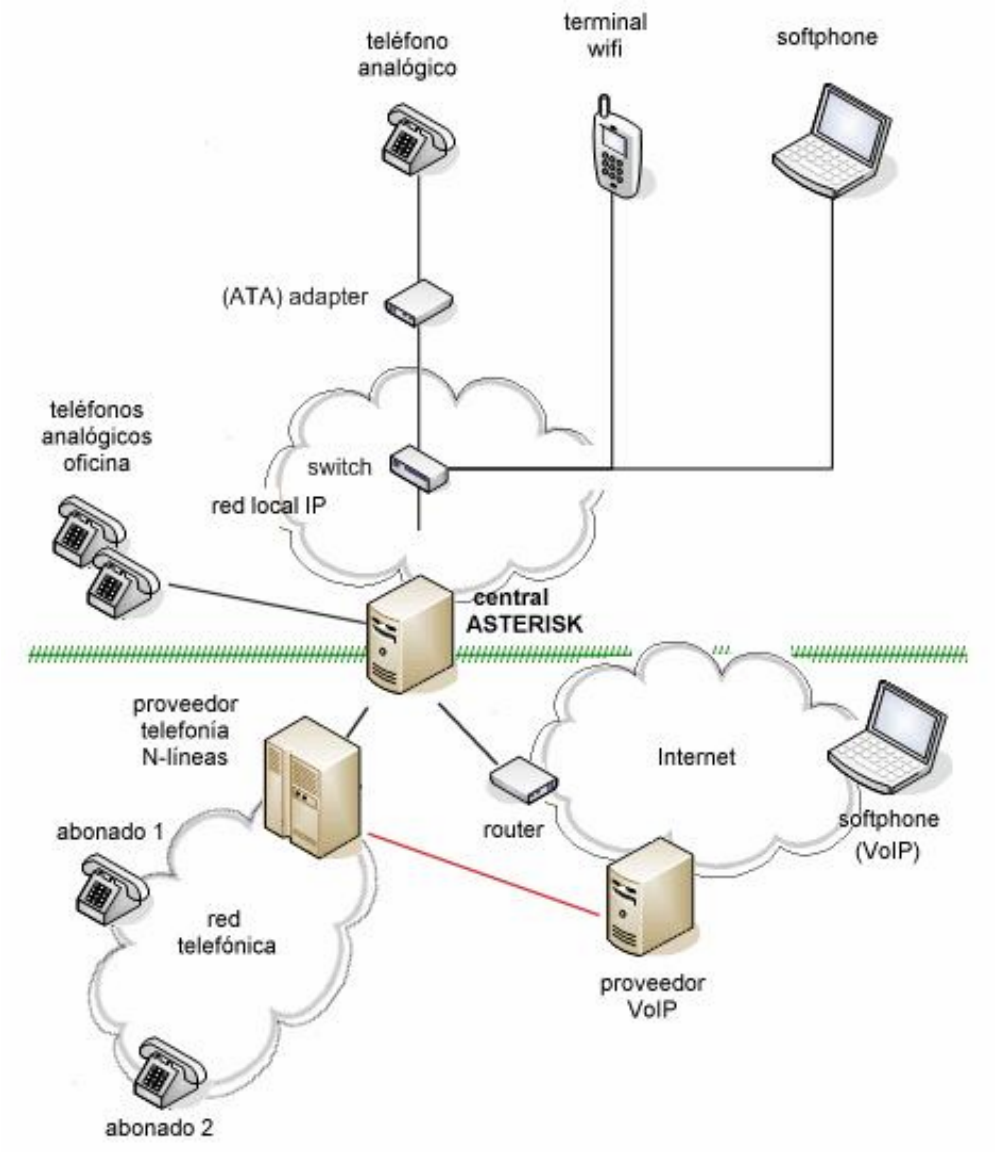

Figura 9. Esquema general de un sistema implementado con Asterisk

2.2.1.1 Arquitectura

La arquitectura tal y como se observa en la figura 10 [72], está conformada principalmente por 4 APIs:

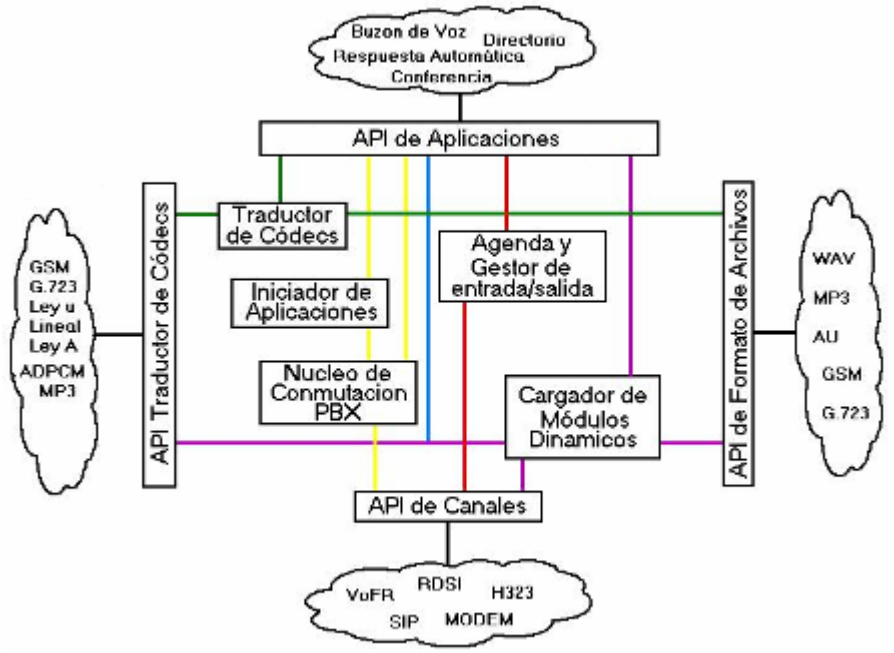

Figura 10. Arquitectura de Asterisk

- **API de Aplicaciones:** permite correr diversos módulos de aplicaciones al mismo tiempo, para realizar varias funciones a la vez. Entre estos módulos se encuentran por ejemplo, el de buzón de voz, conferencia, directorio telefónico, etc.

- **API de Interfaces y Canales:** Todas las llamadas que se realizan, sean entrantes o salientes pasan a través de una interfaz en un canal diferente, estos permiten manejar el tipo de conexión entre el llamante y el servidor, ya sea VoIP, RDSI, PSTN o cualquiera, a través del núcleo de conmutación, según los protocolos y las interfaces correspondientes.

- **API de formato de Archivos:** manipula las actividades de lectura y escritura de varios formatos de archivos de audio para el almacenamiento de los datos en los archivos del sistema y su futura reproducción, entre ellos se encuentra WAV y MP3.

- **API Traductor de *Codificadores*:** carga *codificadores* para soportar varios formatos de codificación y decodificación de audio. Estos *codificadores* determinan la velocidad de datos que requiere cada canal. Asterisk soporta

los mencionados en la sección 2.1.1.2 además de los mencionados en la tabla 12 [73] [74].

Tabla 12. *Codificadores incluidos en Asterisk*

Códec	Tasa de bits	Tasa de muestreo	Descripción
G.726	16, 24 o 32 Kbps	8 KHz	(Reemplazo al G.721) Es también conocido como ADPCM[23], ofrece calidad idéntica al G.711utilizando la mitad del ancho de banda. Esto es posible ya que envía únicamente la información suficiente para describir la diferencia entre la muestra actual y la anterior.
G.723.1	5.3 o 6.3 Kbps	8 KHz	Este *codec* de cadencia dual (reemplaza al G.723) esta diseñado para habla de baja velocidad de datos es uno de los más utilizados con el protocolo H.323. Este *codec* solicita licencia para decodificar las llamadas.
G.729a	8 Kbps	8 KHz	Es un *codec* que entrega una excelente calidad de audio con bajo ancho de banda, debido a que utiliza CS-ACELP[24], pero utiliza gran cantidad de recursos por parte del PC, lo cual no es conveniente para la máquina que está soportando Asterisk.

De los anteriores *codificadores* que Asterisk soporta, se escogieron para la implementación y pruebas con el piloto, los que cumplían criterios importantes dependientes de la relación entre la calidad que se espera del sistema y los recursos económicos y de infraestructura disponible. Además de ello se escogieron los que están en común con los de los *softphones* para asegurar su compatibilidad (Ver Anexo C). Tales *codificadores* son:

- GSM
- G.711
- Speex
- iLBC

La integración de las APIs, se realiza mediante 4 módulos para hacer que Asterisk en conjunto funcione de la siguiente manera: Cuando apenas se inicia Asterisk, el Cargador de Módulos Dinámicos inicializa cada uno de los *drivers* que ofrecen control de canales, formatos de archivo, registro detallado de llamadas, *codificadores* y aplicaciones, entre otras, enlazándolas con la correspondiente API interna.

Luego, el Núcleo de Conmutación comienza a aceptar llamadas desde las interfaces y las gestiona de acuerdo con el plan de marcado utilizando el Iniciador de Aplicaciones para generar el timbre de llamada, conectar al buzón de voz o conectar con troncales externas. Este núcleo incluye una Agenda y Gestor de

[23] Adaptive Differential Pulse-Code Modulation
[24] Conjugate-Structure Algebraic-Code-Excited Linear Prediction

entradas y salidas disponible para los canales y aplicaciones, que realiza tareas de registro de actividades para gestión del sistema. El Traductor de *codificadores*, por su parte, permite a los canales que están utilizando diferentes *codificadores* comunicarse entre si. [74]

El **plan de marcado** se trata de la configuración de la central Asterisk que indica el itinerario que sigue una llamada desde que entra o sale del sistema hasta que llega a su punto final. Se trata en líneas generales del comportamiento lógico de la central, el cual se puede programar mediante una serie de instrucciones y funciones pre-configuradas que permiten hacer casi cualquier cosa, como por ejemplo autocontestar, redireccionar la llamada, desplegar un menú, etc, y además, se pueden extender utilizando lenguajes de programación estándar como C o Perl. [74]

2.2.1.2 Requerimientos del sistema

Linux y Asterisk por lo general son un sistema operativo y una herramienta, respectivamente, que consumen una moderada cantidad de recursos computacionales, más aún, teniendo en cuenta su integración con el motor de reconocimiento de habla, por tanto es importante contar con un equipo servidor que permita correr las aplicaciones simultáneamente sin que se generen retardos en el procesamiento de las órdenes o bloqueos continuos y largos.

Además, la aplicación fundamental del Asterisk radica en la creación de pequeñas centrales telefónicas, y siendo ese el propósito, la selección de los complementos tanto hardware como software debe ser cuidadosamente pensada no solo para el sistema diseñado actualmente y también para futuras expansiones.

En la tabla 13 [75] se detallan las características mínimas tanto HW como SW que deben cumplir los equipos clientes y/o servidores para la implementación apropiada de un sistema telefónico bajo Linux/Asterisk. Asimismo, se encuentra una comparación de las características de los equipos configurados para el Piloto diseñado para el entorno hospitalario de la Clínica La Estancia S.A., donde las pruebas se realizan con un equipo servidor de telefonía en el cual esta instalado el Asterisk, y clientes mediante PCs de escritorio y portátiles con *softphones* mediante micrófonos.

Tabla 13. *Requerimientos del sistema*

	Ideales mínimas		**Reales para PSCVoWLAN**	
Características	**Cliente**	**Servidor**	**Cliente**	**Servidor**
Procesador	Pentium III, 1GHz o equivalente	Pentium IV, 2 GHz o equivalente	Pentium III, 866 MHz	Pentium III, 866 MHz
Memoria	128MB	512MB	128MB	256MB

Disco Duro	5GB	40GB	10GB	10GB
Otros	Adaptador Ethernet, o tarjeta de red 802.11g Tarjeta de sonido, micrófono, audífonos o parlantes	Adaptador Ethernet, , o tarjeta de red 802.11g, Slots para tarjetas de interfaz telefónica	Tarjeta de red Ethernet 3Com 3C918, tarjeta de red 802.11b Tarjeta de sonido	Tarjeta de red Ethernet 3Com 3C918, tarjeta de red 802.11b, tarjeta de sonido
Sistema Operativo	Windows, Linux	Linux Kernel minimo 2.6 RedHat, Fedora, Debian, Mandrake o Gentoo	Windows	Debian
Programas Instalados	*Softphone*	Todos los paquetes de Asterisk	X-Lite	SJPhone Perlbox-Voice

2.3 INTEGRACIÓN DE ASTERISK CON SPHINX

Perlbox-Voice permite ejecutar *scripts* de acuerdo a la palabra reconocida por el Sphinx, para el control de las acciones del IVR implementado con Asterisk.

Teniendo en cuenta que la plataforma seleccionada para la implementación, es Linux/Debian, para que se ejecute una acción ordenada a través de un comando vocal, el Perlbox-Voice ofrece en su interfaz la posibilidad de ingresar de manera escrita la palabra o comando que invocará la acción que se requiera y la acción a la cual debe ligarse, lo que corresponde en el caso de Asterisk, al plan de marcado que determina el flujo de operaciones que se deben realizar, esto es, asignar a cada comando de voz una correspondiente extensión telefónica y una acción a ejecutar para dicha extensión, desde establecer la llamada, hasta contestar automáticamente, generar conferencia, etc.

Durante la integración de Sphinx con Asterisk es importante tener en cuenta que el motor de reconocimiento de Sphinx 2 trabaja con un modelo acústico de 8KHz, por tanto se requiere configurar los *codificadores* que trabajen a dicha frecuencia de muestreo.

El proceso de instalación e integración con sus respectivos archivos de configuración se describen más adelante en el capítulo 3 y se complementa en el anexo B.

En este capítulo se describió la base teórica que permitió definir a 802.11g como la infraestructura física de red ideal para la implementación del piloto. Además, se establecieron los criterios técnicos y económicos que condujeron a la selección de Asterisk para la construcción del sistema de comunicación, cuyo proceso se detalla en el siguiente capítulo.

3. DISEÑO, CONSTRUCCIÓN, OPERACIÓN Y LÓGICA DE FUNCIONAMIENTO DEL MÓDULO DE IVR INTEGRADO CON LA HERRAMIENTA ASR SOBRE UNA RED VOWLAN

En este capítulo se describe el diseño configuración y prueba del piloto de manera modular, comenzando por el motor de reconocimiento de habla, luego la central PBX y finalmente la integración de éstos dos sobre la red inalámbrica.

3.1 DISEÑO Y CONSTRUCCIÓN

El diseño de la estructura general del piloto, se comienza definiendo cada una de sus etapas de manera general:

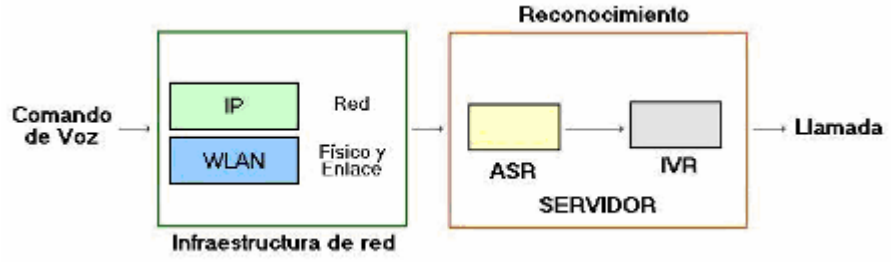

Figura 11. Diagrama en bloques del Piloto

La primera etapa según se observa en el diagrama en bloques de la figura 11, consta de una combinación de redes cableadas e inalámbricas basadas en IP y corresponde a la **infraestructura de red** VoWLAN sobre la cual se transportan los paquetes desde que se digitaliza la entrada de voz en el terminal del usuario, hasta su destino final en el equipo servidor.

La segunda etapa corresponde a la de **reconocimiento,** está conformada por dos módulos internos, el ASR encargado del reconocimiento de cada palabra y de dar la orden para que se ejecute la acción asociada a ella y el IVR, que funciona como una PBX y se encarga de ejecutar y enrutar las llamadas de acuerdo con la palabra reconocida en el módulo previo. Es aquí donde se programan los diferentes servicios que puede prestar el sistema, es decir el *plan de marcado*.

3.1.1 Etapa 1: Infraestructura de red

La red para la cual está diseñado el piloto, como se mencionó previamente, está basada en una solución IP sobre una infraestructura de red Wi-Fi. Esto en términos del modelo de referencia de Interconexión de Sistemas Abiertos (OSI-Open Systems Interconection) corresponde a los niveles de *Red*, *Enlace* y *Físico* respectivamente.

A **nivel físico**, según los estudios realizados en el capítulo 2 de este documento y teniendo en cuenta criterios técnicos y económicos, el diseño ideal de la red está basado en equipos con tecnología 802.11g en configuración de infraestructura utilizando canales sin traslape con el fin de que puedan agregarse más APs, por ejemplo, en el caso en que los usuarios de la red ya hayan sobrepasado la capacidad de un AP, es posible la introducción de un segundo AP omnidireccional en la misma área y configurado en un canal que no presente traslape con el que ya está operando duplicando la capacidad del sistema e incrementando el rango de cobertura. Sin embargo, como se observa en las consideraciones técnicas de la sección 3.1.1.1, además del diseño ideal con 802.11g se realiza un diseño de pruebas para este piloto basado en 802.11b, cuya justificación se encuentra en dicha sección.

3.1.1.1 Elementos del escenario

En la definición del escenario hospitalario se realiza una aproximación espacial posible, en el cual se muestra el número de empleados, dimensiones de las instalaciones, utilización de las redes de voz y datos; así como la definición de utilización y distribución del tiempo y los recursos de la red.[25]

- **Consideración de los usuarios**

  En condiciones reales, un entorno hospitalario mediano por lo general, se trata de un contexto de aproximadamente 500 personas, de las cuales el 20% tienen un equipo de cómputo[26] como herramienta fundamental de trabajo diario (100 PCs), 30% comparten 1 equipo por cada 5 personas (es decir 30 PCs), para un total de 130 PCs; y el 50% restante no utilizan el computador en sus labores diarias. Lo anterior es un factor determinante y más si se tiene en cuenta que el tráfico corresponde en su mayoría a transferencia de archivos, impresiones en red, entre otros, que no representan mayor carga para la red.

  Por otra parte, es importante destacar que los sistemas de comunicación actualmente utilizados están basados en telefonía convencional o en centrales PBX distribuidas aproximadamente 1 terminal telefónico por cada 5 personas. Lo que implica un número total de 100 terminales telefónicos dentro del edificio.

  Pero es necesario aclarar que las condiciones de construcción del piloto, según las encuestas realizadas al personal de la Clínica la Estancia S.A, están dadas para realizar las comunicaciones a través de comandos de voz entre los usuarios predeterminados en las tablas 14 y 15 de la sección 3.2.2, de las cuales se obtiene que el número de terminales para el sistema de

[25] Este diseño está basado en los resultados obtenidos en el "Prototipo Experimental de VoIP sobre redes WLAN para Entornos Empresariales. [44]
[26] Equipo de cómputo o PC ya sea de escritorio, portátil o un dispositivo como una PDA

comunicación por voz es de 33, sin embargo, también deben considerarse como usuarios de la red de datos, a los 97 equipos restantes y a los 100 teléfonos, aclarando que éstos pueden pertenecer a la PBX y que es posible generar llamadas entre ellos, pero no están incluidos dentro del plan demarcado por voz, sino a través de la marcación su propia extensión por medio de DTMF.

Terminales controlados por comando de voz:	33
Terminales pertenecientes a la red de datos:	97
Terminales telefónicos:	100
Total terminales de la red:	**230**

Esta consideración es importante ya que con la construcción de este piloto se intenta integrar los servicios de telefonía con los de transferencia de datos a través de la red IP.

- **Escenario Espacial**

  En condiciones ideales, teniendo en cuenta que un entorno hospitalario puede compararse con un escenario empresarial debido a la cantidad de personal y a la extensión de la estructura, se calcula aproximadamente el área total de cobertura en unos 6400m^2 *indoor*, suponiendo que se trata de una única sede de 1 piso, distribuida aproximadamente de la siguiente manera:

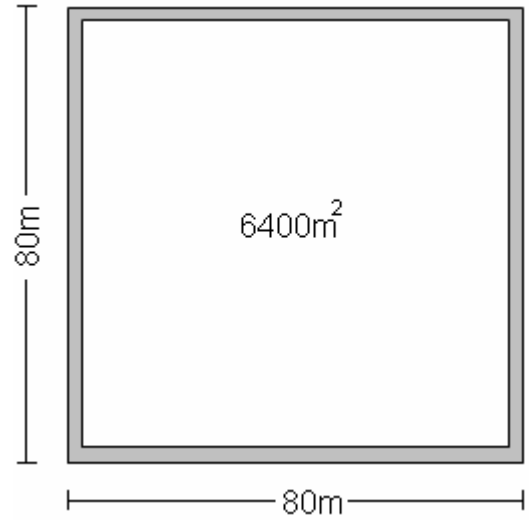

Figura 12. Área de un entorno hospitalario para el diseño del piloto

- **Consideraciones técnicas**

  En este punto se realizan dos modelos completamente funcionales, uno ideal y el modelo de pruebas, basados en equipos 802.11g y 802.11b, respectivamente. Esto se debe a que teórica y técnicamente la mejor opción

es la primera, pero en el piloto de prueba se trabajó con 802.11b, ya que, se quería probar el funcionamiento en el caso más crítico de tal forma que si es viable con éste, muy seguramente lo es con 802.11g que da más velocidad. Por otra parte, la mayoría de las PDAs lo traen integrado y para la otra se requiere de tarjetas, teniendo en cuenta que estos son los dispositivos que tienen la posibilidad más alta de ser utilizados por el usuario.

Modelo 1: Partiendo del hecho que los APs que hacen parte de la implementación del piloto están basados en **802.11g**, para la selección del número óptimo es necesario tener en cuenta el alcance de los mismos trabajando en canales sin traslape en términos reales según los datos consignados en la tabla 8.

Dado que estos equipos no ofrecen mayor número de canales y por consiguiente hay mayor posibilidad de *que* se presente interferencia co-canal, para solventar este inconveniente y con el fin de incrementar la capacidad del canal se va reducir intencionalmente el rango de cobertura de cada AP (potencia de salida) y así incrementar a la vez la capacidad con el número de APs.

En este caso, se supone la instalación de 4 APs trabajando a una velocidad de datos real máxima de 19.8Mbps, y un cubrimiento límite por AP[27] de 75 usuarios dentro de un radio de 35m, en los canales 1,6,11 y 1 tal como lo ilustra la figura 13a.

Terminales controlados por comando de voz por celda:	9
Terminales pertenecientes a la red de datos por celda:	25
Terminales telefónicos por celda:	25
Total terminales por celda:	**59**
Total terminales en la red:	**59*4= 236**

Dando con ello cubrimiento a un terminal más por celda, lo que suma 6 terminales más en la red, de lo esperado y que se encuentran dentro del límite establecido para esta tecnología.

[27] Tomado de la tabla 8

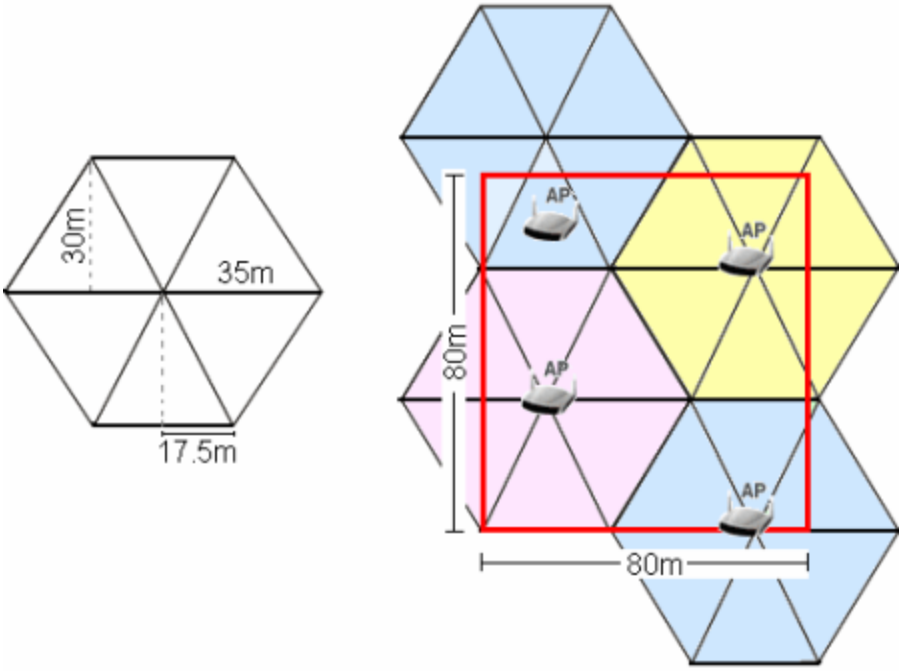

Figura 13a. Diseño para red basada en APs 802.11g

Modelo 2: Para el segundo diseño, partiendo del hecho que los APs que hacen parte de la implementación del piloto están basados en **802.11b** y según las mismas consideraciones que para el primero, adaptadas a su caso propio, se tiene lo siguiente:

Instalación de 7 APs a una velocidad de datos real máxima de 7Mbps, y un cubrimiento límite por AP[28] de 45 usuarios dentro de un radio de 25m seleccionando un esquema basado en grupos de 7 celdas, trabajando en los canales 1,6 y 11, como lo muestra la figura 13b.

Terminales controlados por comando de voz por celda:	5
Terminales pertenecientes a la red de datos por celda:	14
Terminales telefónicos por celda :	15
Total terminales por celda:	**34**
Total terminales en la red:	**34*7= 238**

Dando con ello cubrimiento a un terminal más por celda, lo que suma 8 terminales más en la red, de lo esperado y que se encuentran dentro del límite establecido para esta tecnología.

[28] Tomado de la tabla 8

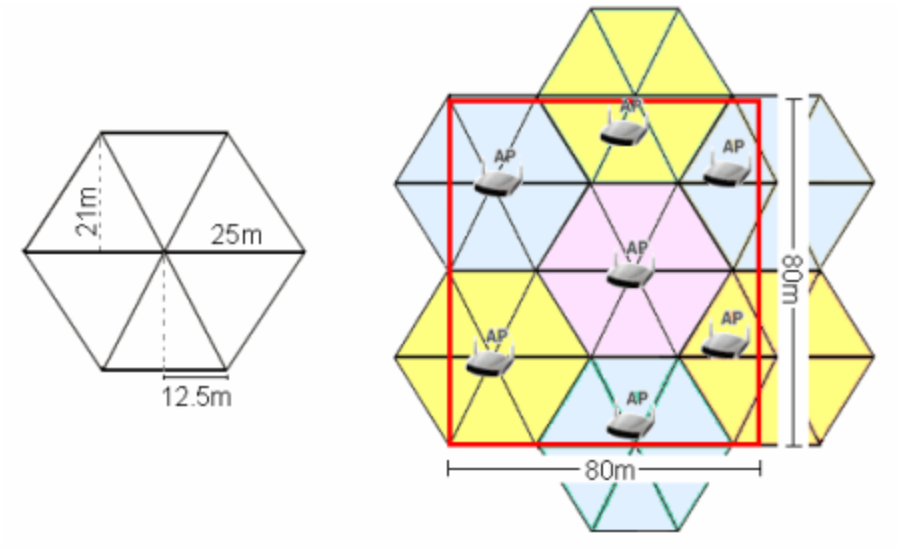

Figura 13b. Diseño para red basada en APs 802.11b

Por su puesto, es importante destacar que debido al patrón hexagonal de cobertura celular que se presenta en ambos modelos, cabe la posibilidad de que si se disminuye el número de APs y se aumenta su cobertura individual para atender toda el área, algunas zonas queden sin cobertura dentro del edificio, y si se incrementa el número habrá algunas zonas en las que la señal podrá extenderse incluso fuera del edificio, como se observa en las figuras 13 a y b respectivamente.

En cuyo caso es preferible tener más puntos de acceso y disminuir la potencia de salida de los APs que están en las fronteras para que la señal no interfiera con posibles sistemas externos, además, con esto se consigue que en futuras expansiones se pueda incrementar la densidad de usuarios por celda.

Cabe resaltar también, que las pruebas de conectividad y funcionalidad van a realizarse sobre un segmento de la red comprendido por el grupo de usuarios dentro de una celda, en los pasillos del segundo piso del edificio de la FIET y no en las instalaciones de la Clínica La Estancia S.A., debido a que no es un espacio apropiado para realizar este tipo de procedimientos, conveniente evitar molestias a los pacientes y por otra parte entorpecer la labor del personal de turno. Posteriormente en el capítulo 4 se detallará un plano del lugar de las pruebas y la ubicación de los AP y equipos de la red.

3.1.2 Etapa 2: Reconocimiento

- **Motor de reconocimiento de habla**

La construcción del piloto inicia con la instalación y configuración del motor de reconocimiento de habla Sphinx 2, seleccionada con base en los criterios definidos en el capítulo 1.

La instalación del Sphinx 2 se llevo a cabo siguiendo los pasos recomendados en el documento del grupo CMU Sphinx [25]. Esto es un proceso largo y que requiere gran atención, ya que es necesario hacer todos y cada uno de los pasos allí indicados para que se instalen todos los componentes debidos.

Además de lo anterior, para facilitar la gestión de Sphinx 2 se utiliza la interfaz gráfica del Perlbox-Voice, vista en la figura 14, ésta permite configurar y verificar que las palabras estén siendo reconocidas. Lo cual, se hace con el fin de evitar que la configuración y el control del motor se hagan a través de la ventana de comandos. El proceso de instalación se explica, más en detalle en el Anexo B.

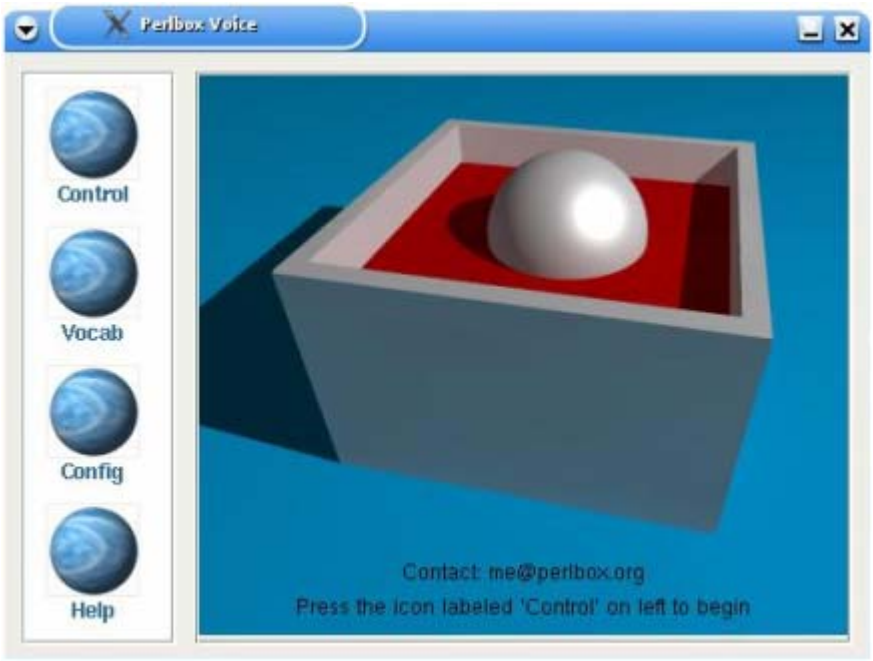

Figura 14. Interfaz gráfica del Perlbox-Voice

La instalación del Sphinx puede verificarse mediante algunas pruebas utilizando las palabras que trae por defecto el Perlbox-Voice. Por ejemplo, cuando se pronuncia la palabra *home*, y el programa reconoce esta palabra, en la pantalla de ejecución del Perlbox-Voice aparece "*I executed home*" e inmediatamente se debe abrir la ventana del navegador Konqueror. Las figuras 15 y 16 ilustran el ejemplo.

Figura 15. Configuración básica de Perlbox-Voice para un ejemplo de prueba

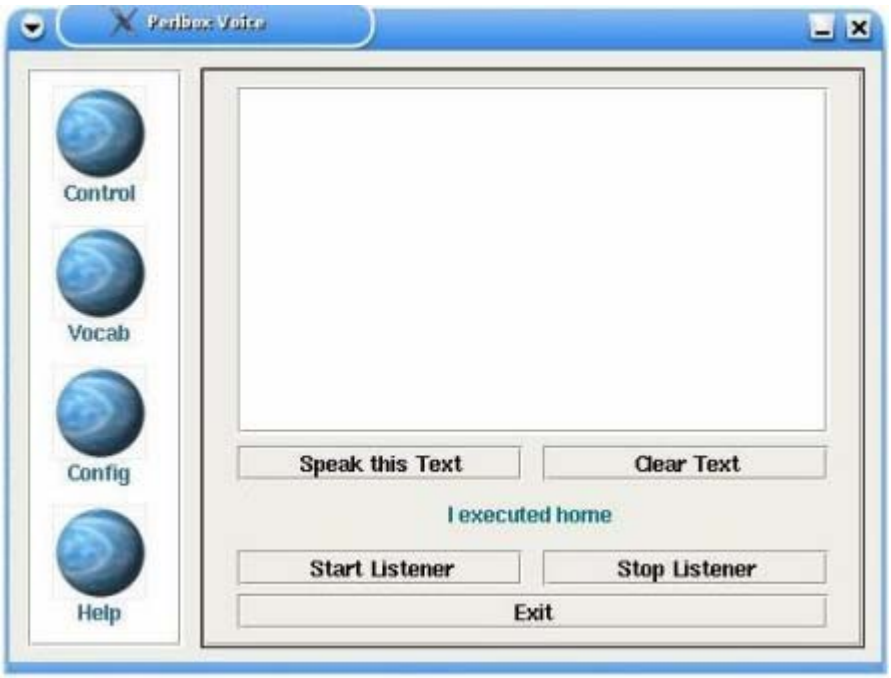

Figura 16. Confirmación de la palabra reconocida

Después de probar el programa de reconocimiento de voz, se instala y configura el plan de llamadas de Asterisk. Estos procedimientos se encuentran detallados en el anexo B.

3.1.3 Integración del motor de reconocimiento y la central PBX

Para la integración de los módulos de IVR y ASR (Sphinx2, Perlbox-Voice, Asterisk), el primer paso consiste en buscar la forma en la cual el Perlbox-Voice pudiera hacer que el Asterisk ejecute la acción correspondiente al comando de voz una vez este llegue al servidor.

Este fue el proceso más difícil de realizar y en el cual el proyecto logró un aporte significativo, ya que, inicialmente la única posibilidad de hacerlo era mediante un código lo cual resultaba muy complejo. Entonces al investigar mas sobre sphinx se encontró al Perlbox-Voice como controlador y se descubrió que este interactúa directamente con el PC, o sea, que decirle un comando de voz para que lo reconociera era lo mismo que escribirle en consola el comando, luego, así surgió la idea de la ejecución de los *scripts*.

Esto se logra haciendo que el Perlbox-Voice mueva el `archivo.call` al directorio `/var/spool/Asterisk/outgoing` para generar la llamada de manera automática.

El `archivo.call` define a que usuario se va a dirigir la llamada y el contexto que va a manejar la misma, más adelante se explicara en detalle la estructura de este archivo.

La forma en la cual el Perlbox-Voice genera el movimiento es mediante un *script*, el cual también debe realizar una copia de los archivos debido a que cada vez que se genera el movimiento para ejecutar la llamada, el `archivo.call` era eliminado de la carpeta donde se encontraba, así entonces se crea una carpeta temporal que contiene a todos los archivos de tal manera que después de realizado el movimiento de cada uno siempre haya una copia.

En el servidor se programa un usuario de Asterisk al cual los médicos deben llamar y posteriormente decir la palabra que desean que el sistema reconozca. Para mayor facilidad se instala un *softphone*[29] que tenga la opción de contestar automáticamente ya que en el servidor no hay quien pueda contestar la llamada, es importante aclarar que el Perlbox-Voice solo escucha por medio del micrófono entonces para que capte la señal de voz desde el *softphone* se utilizó un cable **miniplug** como el mostrado en la figura 17 entre la salida de los parlantes y la entrada del micrófono.

[29] La instalación y configuración de este softphone se encuentra en el Anexo B.

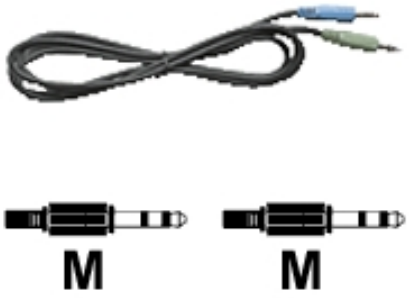

Figura 17. Cable miniplug

La figura 18 muestra la integración de todas las etapas y sus correspondientes módulos dentro del piloto y la arquitectura general del sistema.

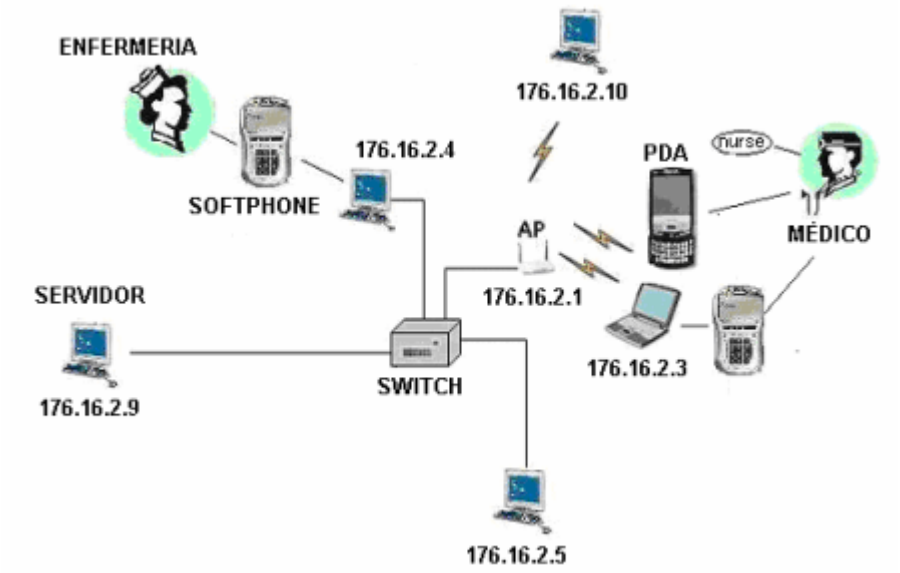

Figura 18. Esquema de funcionamiento del piloto

3.2 OPERACIÓN Y LÓGICA DE FUNCIONAMIENTO

3.2.1 Lógica de funcionamiento y servicios

En busca de un escenario real de aplicación como lo es un entorno hospitalario, la lógica de funcionamiento de esta implementación está basada en las entrevistas realizadas al personal de la Clínica La Estancia S.A. de Popayán.

A partir de éstas se observa que el *médico general* es quien más requiere el servicio de comunicación, ya que, durante su turno se encuentra haciendo rondas en toda el área, por esta razón se plantea este piloto de modo que le permita por medio de un dispositivo portátil, tener acceso a la PBX y así comunicarse con

algún especialista o con la estación de enfermería en caso de requerir instrumental, mientras se moviliza por todo el entorno hospitalario.

También, gracias a los datos obtenidos con las entrevistas puede verse que durante el proceso de generación de una llamada en casos de emergencia no es conveniente que se despliegue un menú de posibilidades, ya que, tratándose de dicha situación no hay tiempo de esperar hasta que el IVR lea todas las opciones, por tal motivo el enrutamiento de las llamadas se realiza directamente, es decir, una vez el médico solicite la comunicación con determinada extensión a través del comando vocal asociado.

No obstante, con el objetivo de mostrar más claramente la funcionalidad de la tecnología IVR en el presente piloto, se configuró un servicio en el cual los usuarios pueden interactuar con otras extensiones a través de un comando de voz que invoca las opciones de un servidor de respuesta interactiva, el cual despliega un menú que ofrece información sobre las extensiones telefónicas y la posibilidad de establecer llamadas entre si, a través de marcación DTMF.

Además del médico general, la enfermera jefe y las auxiliares de turno tienen su vocabulario para generar una orden, claro está que a parte del comando de voz también pueden comunicarse entre si marcando la extensión de la persona que soliciten ya que todos los usuarios están conectados a una red IP que a través de Asterisk trabaja como una PBX.

En la lógica del piloto se definen palabras que el médico general y el personal que se encuentre en la estación de enfermería deben manejar de acuerdo al servicio que requieran, se aclara que las palabras para el médico, la enfermera y los auditores son diferentes porque cada una está configurada de manera distinta, por ejemplo, si el médico es quien requiere la comunicación con la enfermera se debe generar la llamada a ella y cuando conteste se redirecciona al médico nuevamente, mientras que si es la enfermera la que solicita el servicio es a ella a quien debe ser devuelta la llamada, si se utilizaran las mismas palabras la llamada siempre se devolvería a la misma persona.

En la tabla 14 se definen el directorio de servicios con los cuales puede comunicarse el médico general, el *corpus* que deberá utilizar el medico para cada uno de ellos y la pronunciación correspondiente a cada palabra según el diccionario fonético de Sphinx.

Tabla 14. *Servicios y palabras que utiliza el Médico General*

Médico general	***Corpus* del médico**	**Pronunciación (lenguaje fonético)**
Médico General – Área Administrativa	administrative	AX D M IH N AX S T R EY DX IX V
Médico General – Admisiones	admission	AE D M IH SH AX N
Médico General – UCI adultos	One	W AH N (1) HH W AH N (2)
Médico General - Anestesiólogo	anesthesiologist	AE N AX S TH IY Z IY AA L AX JH AX S T
Médico General – Área Hospitalaria	Area	EH R IY AX
Médico General – Auditor	Auditor	AO DX AX DX AXR
Médico General – Banco de Sangre	Blood	B L AH D
El Médico General da aviso de un Código Azul	Blue	B L UW
Médico General – UCI neonatos	Born	B AO R N
Médico General – Enfermera Jefe	Boss	B AO S
Conferencia (Medico general, traumatólogo, oncólogo)	Conference	K AA N F AXR AX N S K AA N F R AX N S
Médico General – Banco de datos	Database	D EY DX AX B EY S D AE DX AX B EY S
Médico General – Servicios Generales	General	JH EH N AXR AX L JH EH N R AX L
Médico General – Hematólogo	hematologist	EH M AE T AA L AX JH AX S T
Médico General – Imágenes Diagnosticas	Images	IH M AX JH AX Z
Médico General – IVR	interactive	IH N T AXR AE K T IX V (1) IH N AXR AE K T IX V (2)
Médico General – Facturación	Invoicing	IH N V OY S IX NG
Médico General – Laboratorio Clínico	laboratory	L AE B R AX T AO R IY
Médico General – Enfermería	Nurse	N ER S
Médico General – Oncólogo	oncologist	AA NG K AA L AX JH AX S T
Médico General – Patólogo	pathologist	P AX TH AA L AX JH AX S T
Médico General – UCI pediátrica	Pediatric	P IY DX IY AE T R IX K
Médico General – Pediatra	Pediatrician	P IY DX IY AX T R IH SH AX N
Médico General – Cirujano Plástico	Plastic	P L AE S T IX K
Médico General – Estadística	Statistical	S T AX T IH S T IX K AX L
Médico General – Cirujano Pediátrico	Surgeon	S ER JH AE N S ER JH AX N
Médico General – Cirugía	Surgery	S ER JH AXR IY
Médico General – Terapia	Therapy	TH EH R AX P IY

Médico general	*Corpus* del médico	Pronunciación (lenguaje fonético)
Respiratoria		
Médico General – Traumatólogo	traumatologist	T R OW M AH T AA L AA JH AX S T
Médico General – Urólogo	Urologist	Y AXR AA L AX JH AX S T
Médico General – Almacén	Warehouse	W EH R HH AW S

En la tabla 15 se encuentra la definición de los servicios, el *corpus* y la pronunciación que deben utilizar las enfermeras y los auditores.

Tabla 15. *Servicios y palabras que utilizan los auditores y las enfermeras*

Comunicaciones Enfermería	*Corpus* de la enfermera jefe, auxiliares y auditores	Pronunciación (lenguaje fonético)
Auditor Externo – Auditor Interno	Intern	IH N T AXR N
Auditor Interno – Auditor Externo	External	IX K S T ER N AX L
Enfermería – IVR	Response	R AX S P AA N S (1) R IY S P AA N S (2)
Enfermería – Admisiones	Admit	AX D M IH T
Enfermería – Almacén	Store	S T AO R
Enfermería – Área Administrativa	administration	AE D M IH N AX S T R EY SH AX N
Enfermería – Área Hospitalaria	Hospital	HH AA S P IH DX AX L
Enfermería – Banco de Datos	Data	D EY DX AX D AE DX AX
Enfermería – Banco de Sangre	Bleed	B L Y ID
Enfermería – Estadística	Statistic	S T AX T IH S T IX K
Enfermería – Facturación	Billing	B IH L IX NG
Enfermería – Imágenes diagnosticas	diagnostic	D AY AX G N AA S T IX K
Enfermería – Laboratorio clínico	Clinical	K L IH N AX K AX L
Enfermería – Servicios Generales	Services	S ER V AX S AX Z
Enfermería – Terapia Respiratoria	Breathing	B R IY DH IX NG
Enfermería – UCI Adultos	Adult	AX D AH L T(1) AE DX AX L T(2)
Enfermería – UCI neonatos	Baby	B EY B IY
Enfermería – UCI pediátrica	Two	T UW
Jefe de enfermería – Auditor Interno	supervisor	S UW P AXR V AY Z AXR

Un servicio adicional que presenta este piloto es la conferencia la cual se invoca con el comando "*conference*" y genera una comunicación entre tres usuarios diferentes. En este caso se propuso teniendo en cuenta los servicios de telefonía actual y un caso especial de comunicación, que es, una conferencia entre el médico general, el pediatra y el traumatólogo. Cabe aclarar que en los entornos hospitalarios y en particular en la Clínica La Estancia S.A. este servicio no es utilizado, ya que, no se cuenta con un sistema de comunicación que lo permita.

Otro servicio que se ofrece es el de anuncio de Código Azul, el cual, permite generar una llamada a enfermería mediante el comando "*blue*" y cuando se conteste se reproducirá una voz diciendo "esto es un código azul por favor encienda la alarma", este servicio es utilizado para que el equipo médico pueda tener sus manos libres para atender mas rápidamente la situación de emergencia y evitar buscar la alarma y luego encenderla.

3.2.2 Operación del piloto

La operación del piloto inicia cuando un usuario dentro de la red genera una llamada al servidor, éste contesta por medio del *softphone* y recibe el comando de voz (palabra) que es escuchado por el Perlbox-Voice quien lo reconoce e inmediatamente ejecuta un *script* para que Asterisk genere el evento asociado al comando introducido.

3.2.2.1 Programación del *Script* de interacción con Asterisk

El *script* que permite esta operación esta escrito en lenguaje Bash y tiene la siguiente configuración:

```
#!/bin/bash
#! enfermera_script
mv scriptsperlbox/archivoscall/enfermera.call var/spool/Asterisk/outgoing
cp scriptsperlbox/temporales/enfermera.call scriptsperlbox/archivoscall
```

Por ejemplo, este *script* es el que se ejecuta después que se reconoce la palabra "nurse", por esta razón su nombre es `enfermera_script`, y en su caso propio, se utiliza para generar la llamada a la enfermería por parte del médico general.

La primera línea muestra que está escrito en lenguaje Bash y que se ejecuta en el directorio `/bin/`, esto quiere decir que es en éste donde se deben guardar todos los *scripts* que el Perlbox-Voice va a ejecutar.

```
#!/bin/bash
La segunda línea es el nombre del script.
#!enfermera_script
```

La tercera línea es la más importante pues es la que hace que se genere la llamada de manera automática,

```
mv scriptsperlbox/archivoscall/enfermera.call var/spool/Asterisk/outgoing
```

La cuarta línea es la que hace una copia de cada archivo.call después de ser utilizado para que de esta manera no desaparezca cuando se haga el movimiento.

```
cp scriptsperlbox/temporales/enfermera.call scriptsperlbox/archivoscall
```

El directorio /`scriptsperlbox/archivoscall` es el que contiene a todos los `archivos.call` que se moverán para generar las llamadas y el directorio /`scriptsperlbox/temporales` es el que contiene copias de cada uno de ellos, y de esta manera generar llamadas automáticamente cuantas veces se requiera.

3.2.2.2 Configuración del *Script* para generar el plan de marcado

Para que una llamada se genere correctamente, los `archivos.call` deben tener una configuración determinada por los siguientes parámetros:

```
Channel: SIP/2151
MaxRetries: 1
RetryTime: 60
WaitTime: 30
Context: enfermera
Extension: s
Priority: 1
```

- **Parámetros antes de generar la llamada**

Channel: El canal que se va a utilizar para la salida de la llamada.
MaxRetries: número de re-intentos antes de que falle la llamada (si se coloca 0 de todas maneras se hace un intento).
RetryTime: Segundos entre re-intentos.
WaitTime: Segundos para esperar una respuesta.
CallerID: se utiliza para que se pueda saber quien genera la llamada automática, puede no funcionar si no se respeta el formato:

```
CallerID: cualquier nombre<1234>
Account: código de cuenta para usar
```

- **Especificaciones si la llamada es contestada**

Context: es el contexto que se especifica en el archivo de configuración de Asterisk `extensions.conf`, el cual describe el marcado automático a la extensión que estableció todo el proceso.
Extension: extensión definida en `extensions.conf` para la ejecución de la aplicación, en este caso DIAL.
Priority: prioridad definida en `extensions.conf` para cuando empiece la llamada.
Set: grupo de variables para usar en la lógica de la extensión (ejemplo: file1=/tmp/to), en Asterisk 1.0.x se debe usar 'SetVar' y no 'Set'.
Aplicacion: Aplicación de Asterisk para ejecutar (se usa en lugar de un contexto específico, extensión y prioridad).

Data: Las opciones para ser pasadas a la aplicación.

Es muy importante la definición del contexto en `extensions.conf` dado que la llamada se establece de una manera inicial y luego se transfiere a su destino final de la siguiente manera: Suponiendo la llamada desde el Usuario A al Usuario B, el usuario que genera la llamada debe marcar el número de la extensión del usuario B, cuando esto sucede, la llamada se establece entre el equipo servidor y el usuario B, luego esta llamada es devuelta al usuario A y contestada automáticamente para dar la apariencia de que se estableció directamente, así los dos usuarios pueden establecer su comunicación.

Esto puede observarse en las figuras 19 y 20 en las que se muestra el funcionamiento del piloto a nivel de los usuarios y a nivel de programas respectivamente.

Las configuraciones de los *softphones* de los clientes y el servidor, la programación de los *scripts* de cada palabra se encuentran en el Anexo B.

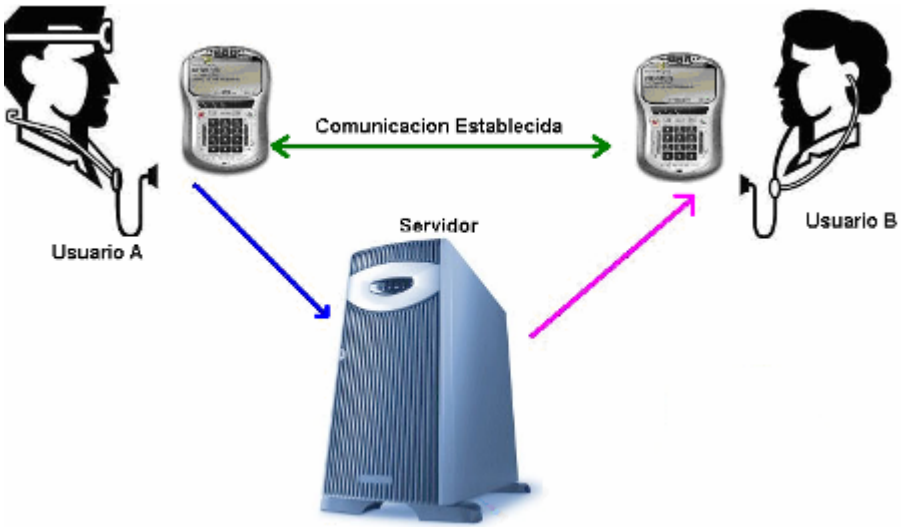

Figura 19. Esquema general del establecimiento de la llamada

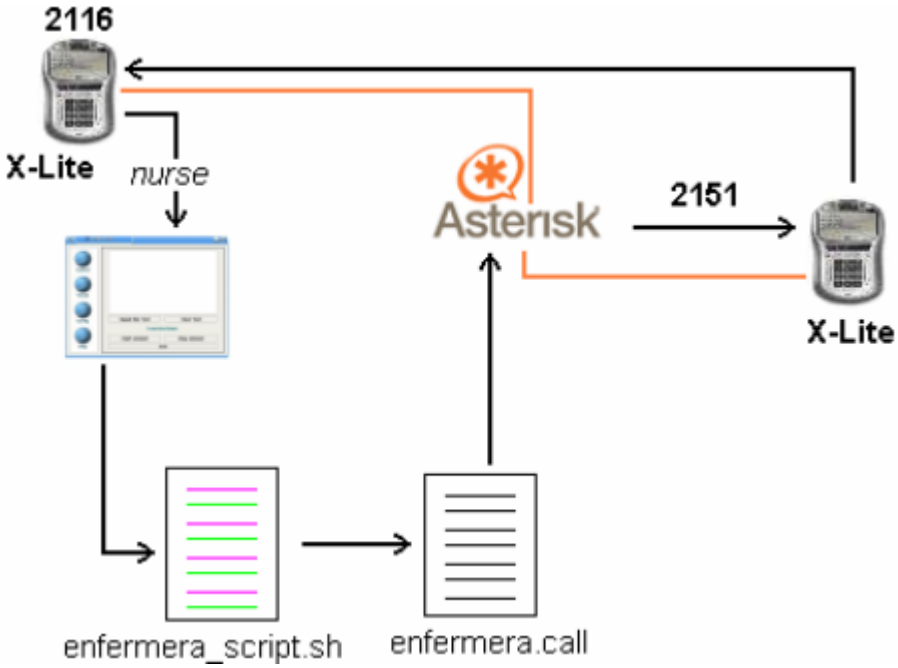

Figura 20. Diagrama de implantación

En el servidor deben estar corriendo siempre los tres programas principales que son: Asterisk, Perlbox-Voice, y el *softphone* que se maneje de acuerdo al sistema operativo, que no necesariamente es el mismo en los clientes.

Hasta este punto se ha definido el diseño y la estructura de los componentes que hacen parte del piloto, explicando su configuración, funcionamiento y servicios que se ofrecen de acuerdo a los requerimientos extraídos del estudio de las encuestas realizadas al personal de la Clínica La Estancia S.A., con lo cual se concreta que la construcción de este piloto está orientada a la prestación del servicio de comunicación interna a través de una PBX con marcación por voz, desarrollada sobre Asterisk a través del motor de reconocimiento Sphinx 2.

Estas herramientas ofrecen gran variedad de posibles configuraciones, pero en este caso específico se enfocará este proceso en el enrutamiento de las llamadas, ya que es el servicio mas requerido por parte del personal, no sin incluir una funcionalidad que permita la posibilidad de comprobar las aplicaciones de los sistemas IVR en la construcción de aplicaciones a nivel empresarial.

Por otra parte, es importante aclarar que teóricamente se define el diseño general y funcional que en la práctica es conveniente para la construcción de este piloto en un entorno hospitalario real; además existe el diseño y la implementación con la cual se realizaron las pruebas de funcionamiento y conectividad consignadas en el siguiente capítulo.

4. PRUEBAS Y RESULTADOS

El plan de pruebas establecido para la funcionalidad del Piloto está basado en un segmento de la red total diseñada en el capítulo 3 debido a que las pruebas se realizaron en el segundo piso del edificio de la FIET.

Este segmento de red de prueba, corresponde al de la figura 21 y está conformado por los siguientes elementos:

- 4 PCs de escritorio
- 1 equipo portátil
- 1 AP 802.11b Cisco AIRONET serie 340
- 1 SWITCH Ethernet

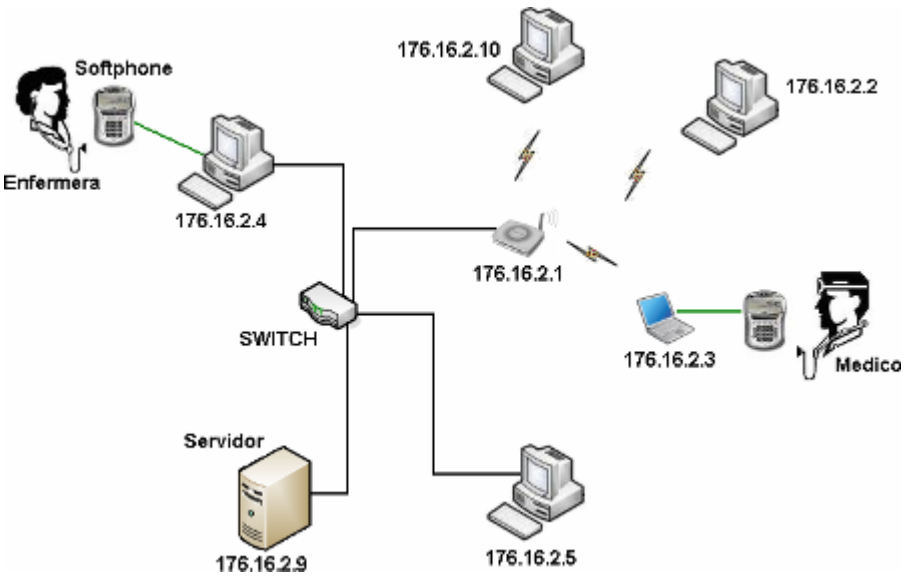

Figura 21. Montaje del piloto utilizado para realizar las pruebas

4.1 DEFINICIÓN DE LOS PARÁMETROS DE PRUEBA

Teniendo en cuenta que ya están dadas todas las condiciones de implementación del piloto y que las pruebas de funcionalidad hasta este punto han sido satisfactorias, a continuación se definen y describen individualmente los parámetros técnicos que se probaron con el fin de validar tecnológicamente su construcción.

- **Calidad del reconocimiento:** Este parámetro es el más importante en la construcción del piloto, ya que, se evalúa conforme a la capacidad del motor

para identificar la palabra y ejecutar correctamente la acción asociada a la misma, lo cual se conoce como exactitud y se describe en la sección 1.1.

- **Calidad de la voz:** Tal como se mencionó en el capítulo 2, acerca de las formas de probar la calidad de la voz después de pasar por un proceso de codificación/ decodificación, y para el caso específico de este piloto se evaluó siguiendo la metodología MOS, con base en la percepción y calificación de 5 personas diferentes. Este parámetro evalúa la calidad de la voz mientras la llamada se encuentra establecida, es decir, durante la conversación.

- **Cobertura:** Determina la distancia en metros que es capaz de cubrir un AP (lo que establece el tamaño de cada celda), sin que ocurran interrupciones en la transmisión.

- **Movilidad:** Definido para este caso como la posibilidad de los usuarios de realizar llamadas sin interrupciones ya sea desde un terminal fijo en cualquier lugar del edificio o desde un terminal portátil en movimiento dentro de la zona total de cobertura.

- **Congestión:** Determina en este caso la capacidad del sistema de funcionar exitosamente, desde el marcado hasta el establecimiento de la llamada cuando hay gran cantidad de trafico de datos en la red.

Estos parámetros fueron evaluados y los resultados se establecen con base en una escala cuantitativa entre 1 a 5 y una descripción cualitativa definida especialmente para este piloto acorde a la percepción de 5 usuarios diferentes, como se indica en la tabla 16.

Tabla 16. *Tabla de puntuación para la evaluación del piloto*

Cuantificación	Calificación	Representación
1.0-2.9	Malo	
3.0-3.9	Regular	
4.0-4.9	Bueno	
5.0	Excelente	

4.2 DESCRIPCIÓN DE LAS PRUEBAS

Se realizaron pruebas para evaluar los parámetros definidos anteriormente, algunos por separado y posteriormente combinándolos con variaciones entre ellos para que los resultados tengan más sentido en un escenario real.

- **Prueba 1**

Para comprobar la **calidad del reconocimiento** en primer lugar se prueba la compatibilidad entre los *codificadores* y posteriormente se pronuncia el comando de prueba *nurse* con el fin de verificar el parámetro más importante del piloto: la exactitud[30] del reconocedor para que se ejecute la acción correspondiente, que consiste en establecer una comunicación con la enfermera. Esto se hace variando los *codificadores* configurados en los *softphones* tanto de los clientes como del servidor y verificando el número de intentos antes de obtener una respuesta del sistema. Los resultados de esta experiencia se describen en la tabla 17.

- **Prueba 2**

Para comprobar la **calidad de la voz** se parte del hecho de que ya se ha establecido una comunicación entre dos terminales (independientemente de cuantas veces se marco para establecerla), para esta prueba se varían los *codificadores* configurados en los *softphones* tanto de los clientes como del servidor y se miden características subjetivas de la calidad de la conversación, tales como retardo, distorsiones e intensidad, que alteren la percepción de la voz o el mensaje original, tal como se observa en la tabla 18.

- **Prueba 3**

La prueba de **cobertura** se realiza cuando el usuario pronuncia el comando para establecer la comunicación desde un punto dentro de la zona de cobertura del AP y continúa alejándose del mismo en intervalos de 10m hasta definir la distancia límite, variando los *codificadores* tanto del cliente como del servidor. Esto se traduce en la evaluación del número de intentos que se debe hacer para que el sistema reconozca la palabra correctamente cuando se varía la ubicación del terminal portátil con respecto al AP. Por ejemplo, si el número intentos aumenta cuando crece la distancia entre el cliente y el servidor o si por el contrario no se ve afectado de ninguna manera. Los resultados se consignaron en la tabla 19.

- **Prueba 4**

En la prueba de capacidad de **movilidad** de los usuarios se experimenta la misma situación de la prueba 1, para comprobar la exactitud, exceptuando que en este caso, el usuario va caminando dentro del área de cobertura mientras pronuncia la palabra para establecer la llamada. Los resultados se encuentran en la tabla 20.

[30] El porcentaje más alto de reconocimiento que se obtenga

- **Prueba 5**

En la prueba de **congestión** del sistema, se inyecta tráfico de datos desde un cliente ubicado dentro de la red, al servidor, al mismo tiempo que éste atiende una llamada y realiza el reconocimiento de un comando desde otro cliente inalámbrico. Luego, utilizando el Ethereal se puede observar la congestión generada hacia el servidor y si existe pérdida de paquetes de voz cuando se inyecta tráfico en la red. El registro de dichas pruebas se encuentra en las figuras 22, 23 y 24.

Nota: las pruebas de movilidad y cobertura se llevaron a cabo de acuerdo al plano de la figura 25. Los puntos en rojo señalan los sitios específicos en los cuales se ubico el equipo portátil y desde los que se realizaron las pruebas.

Tabla 17. *Pruebas de calidad de reconocimiento*

		Clientes				
		Codec	**G.711**	**GSM**	**iLBC**	**Speex***
Servidor	**Compatibilidad**	**G.711**	Compatible (5.0)	Compatible (5.0)	-	Compatible (5.0)
		GSM	Compatible (5.0)	Compatible (5.0)	Compatible (5.0)	-
		iLBC	-	Compatible (5.0)	Compatible (5.0)	-
		Speex*	Compatible (5.0)	-	-	Compatible (5.0)
	Exactitud	**G.711**	Se logro el reconocimiento inmediatamente y la exactitud es del 90% (4.5)	Reconocimiento después de 2 intentos, con exactitud del 50% (2.0)	-	Reconocimiento inmediato y su exactitud fue del 70% (4.0)
		GSM	Reconocimiento después de 2 intentos y su exactitud fue del 50% (2.0)	Se logro el reconocimiento inmediatamente y la exactitud es del 70% (4.0)	Reconocimiento después de 4 intentos y su exactitud fue del 60% (2.0)	-
		iLBC	-	Reconocimiento después de 4 intentos y su exactitud fue del 60% (2.0)	Se logro el reconocimiento inmediatamente y la exactitud es del 90% (4.0)	-
		Speex*	Reconocimiento inmediato y su exactitud fue del 70% (4.0)	-	-	Se logro el reconocimiento inmediatamente y la exactitud es del 80% (4.3)

* Corresponde al Speex de 15.2KHz, 40ms.
**Esta calificación numérica está basada en la puntuación establecida en la tabla 16.
*** Los porcentajes expresados se toman con base en un total de 10 intentos realizados.
****Los campos sin llenar corresponden a los que no tuvieron compatibilidad, así *que* no se pudo hacer ninguna otra prueba.

Tabla 18. *Prueba de calidad de la voz*

		Clientes				
		Codec	**G.711**	**GSM**	**iLBC**	**Speex**
Servidor	***Retardo**	**G.711**	4.5	3.5	---	3.5
		GSM	3.5	4.5	3.5	---
		iLBC	---	3.5	4.5	---
		Speex	3.5	---	---	4.5
	***Variacion**	**G.711**	4.0	2.8	---	3.8
		GSM	2.8	4.0	3.5	---
		iLBC	---	3.5	4.0	---
		Speex*	3.8	---	---	4.0
	***Intensidad**	**G.711**	4.5	3.8	---	3.8
		GSM	3.5	4.5	3.8	---
		iLBC	---	3.5	4.5	---
		Speex*	3.8	---	---	4.5

* Estas calificaciones están asignadas con base en la tabla 16.

Tabla 19. *Prueba de cobertura para calidad de reconocimiento*

	Clientes				
	Codec	**G.711**	**GSM**	**iLBC**	**Speex**
Servidor	**G.711**	Entre 0-40m, el reconocimiento se obtiene inmediatamente con un nivel de voz normal	Entre 0-40m, el reconocimiento se obtiene después de 2 intentos a un nivel de voz normal	---	Entre 0-40m, el reconocimiento se obtiene en el tercer intento con un nivel de voz muy elevado
	GSM	Entre 0-40m, el reconocimiento se obtiene después de 3 intentos a un nivel de voz muy elevado	Entre 0-40m, el reconocimiento se logro después de 3 intentos con un nivel de voz mas elevado de lo normal	Entre 0-40m, el reconocimiento se logro después de 6 intentos con un nivel de voz muy elevado	---
	iLBC	---	Entre 0-40m, el reconocimiento se logro después de 6 intentos con un nivel de voz muy elevado	Entre 0-40m, el reconocimiento se logra de inmediato con un nivel de voz normal	---
	Speex	Entre 0-40m, el reconocimiento se obtiene en el segundo intento con un nivel de voz mas elevado de lo normal	---	---	Entre 0-40m, el reconocimiento se obtiene inmediatamente con un nivel de voz normal

*Esta es la distancia ±1m entre el terminal portátil y el AP de la celda.

Tabla 20. *Prueba de movilidad para calidad de reconocimiento*

		Clientes			
	Codec	**G.711**	**GSM**	**iLBC**	**Speex**
Servidor	**G.711**	Reconocimiento después de 2 intentos con un nivel normal de voz y buena vocalización	Reconocimiento después de 4 intentos hablando fuerte y vocalizando despacio	---	Reconocimiento al tercer intento vocalizando despacio
	GSM	Reconocimiento al cuarto intento	Reconocimiento después de 3 intentos con un nivel de voz muy elevado	Se logro reconocimiento después de 6 intentos con un nivel de voz muy elevado y vocalización cuidadosa	---
	iLBC	---	Reconocimiento después de 6 intentos con un nivel de voz muy elevado	Reconocimiento se logra de inmediato con un nivel de voz normal y pronunciación cuidadosa	---
	Speex	Se obtiene reconocimiento al tercer intento con nivel de voz elevado y vocalización cuidadosa	---	---	Se obtiene reconocimiento al segundo intento con un nivel de voz mas elevado de lo normal y buena vocalización.

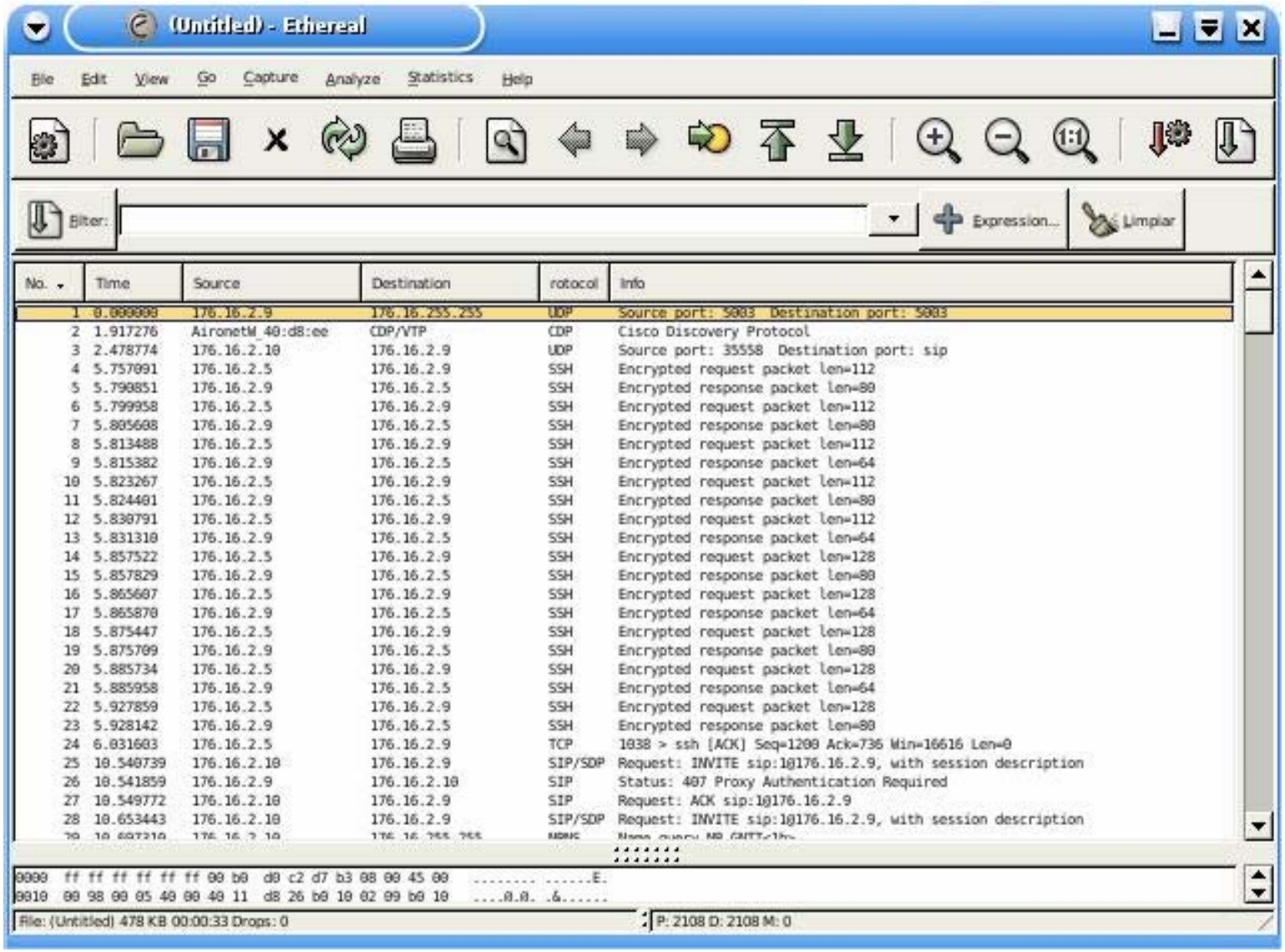

Figura 22. Prueba de congestión analizada con Ethereal

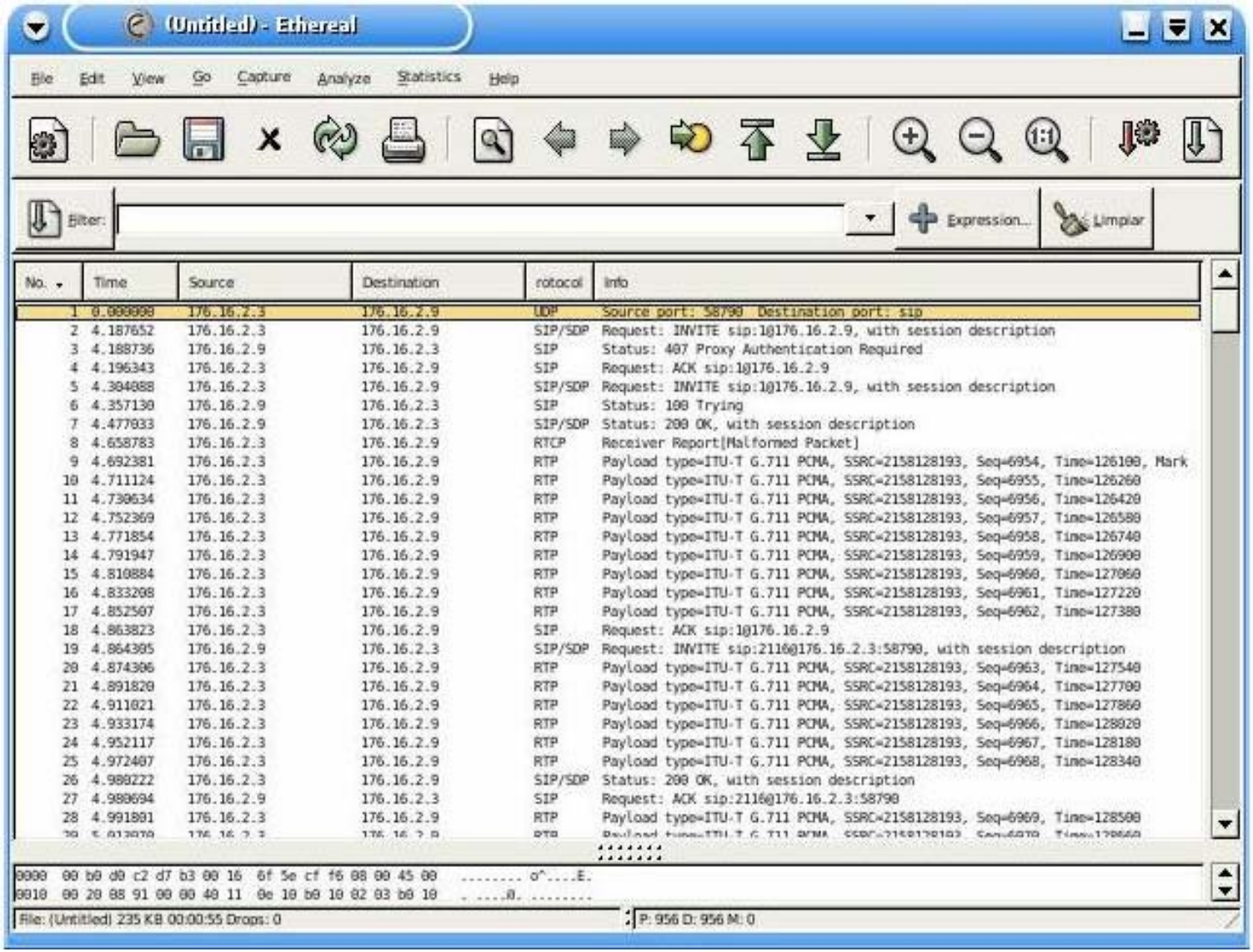

Figura 23. Prueba de congestión y movilidad analizada con Ethereal

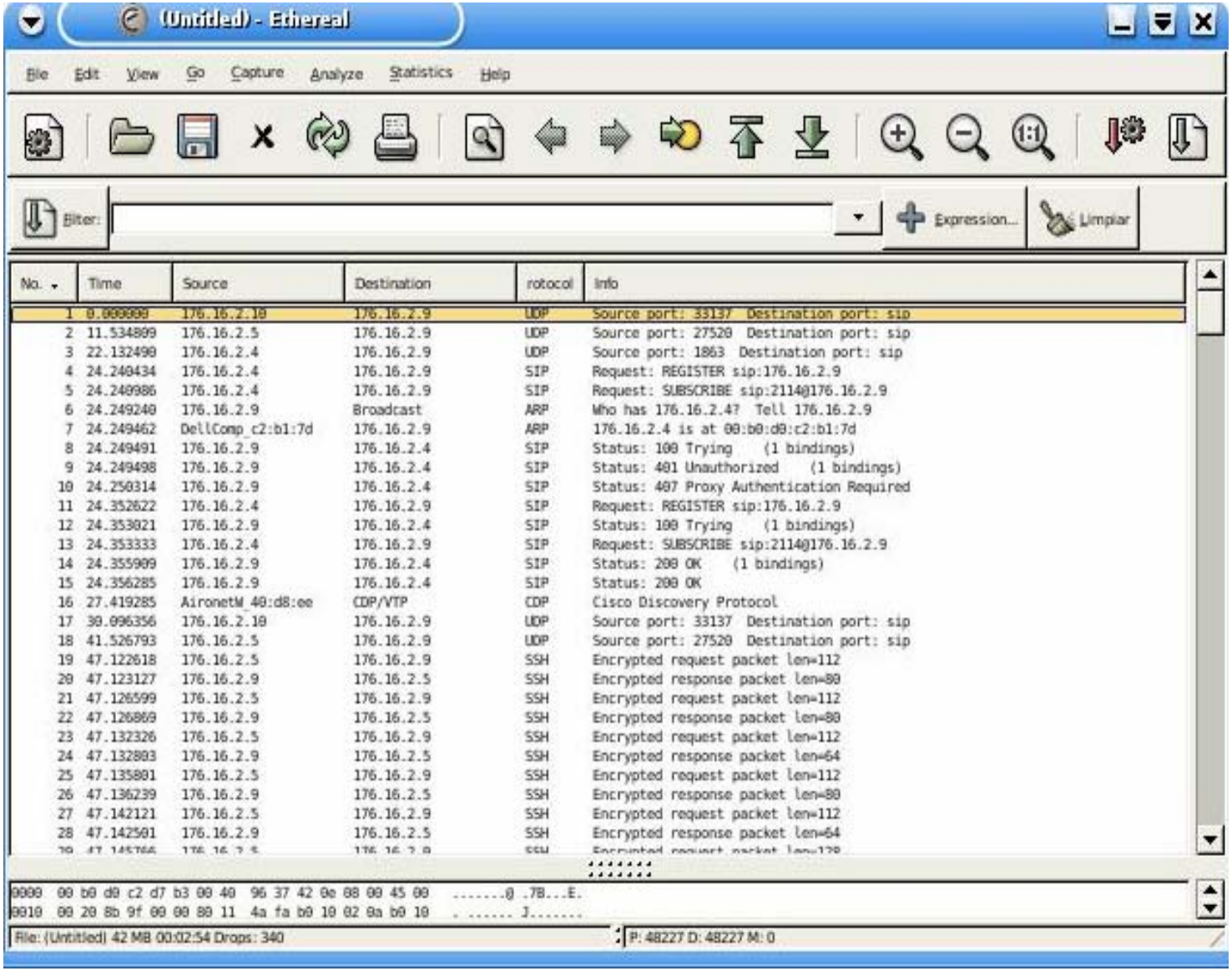

Figura 24. Tráfico de datos y voz enviado por los clientes al servidor.

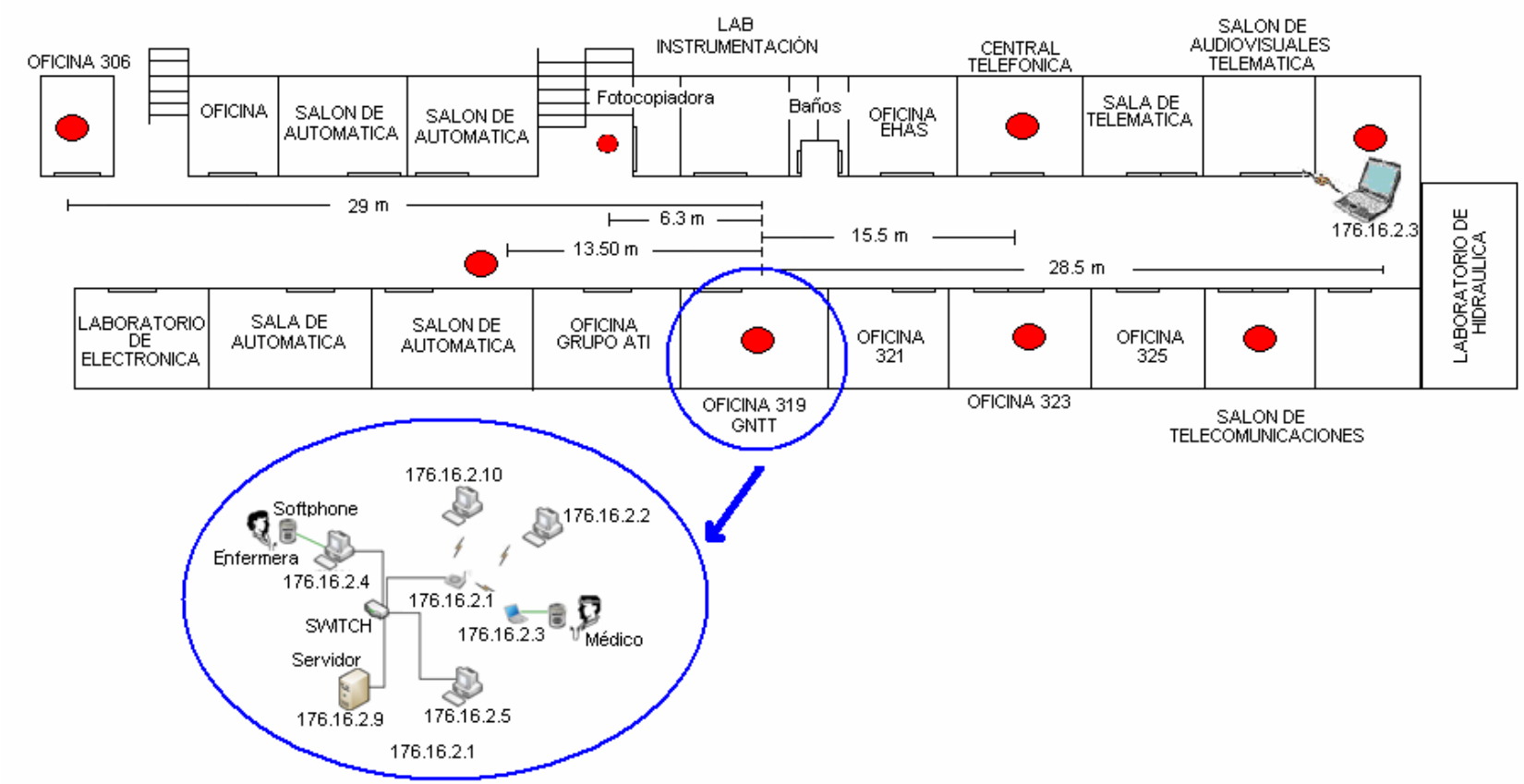

Figura 25. Plano del sitio de pruebas

El AP estaba ubicado a 1.75 metros del servidor en la oficina 319 del Grupo I+D Nuevas Tecnologías en Telecomunicaciones, la máxima distancia de cobertura fue hasta la oficina 306 de la Rama estudiantil IEEE que se encuentra a una distancia del AP de 35,84m y hacia el otro extremo, el área de cobertura incluía el salón de audiovisuales, que se encuentra a una distancia de 40,50m del mismo.

4.3 INTERPRETACIÓN DE RESULTADOS

Partiendo de los resultados de las pruebas anteriores se pueden concluir importantes aspectos para consolidar la construcción de futuros sistemas.

- Cuando en el cliente y el servidor se configura el mismo *codec*, a pesar de no tratarse del mismo sistema operativo ni del mismo *softphone*, existe completa compatibilidad y el reconocimiento se logra inmediatamente sin hacer muchos intentos, con alta exactitud.
- Cuando se utilizan *codificadores* combinados, es decir un *codec* A en el servidor y uno B en el cliente, los resultados de compatibilidad, exactitud en el reconocimiento y calidad de la voz, son los mismos que cuando se configuran el *codec* A en el cliente y el B en el servidor. Y dado el caso de que se presente alguna variación, ésta es mínima y no afecta sustancialmente el funcionamiento del sistema. No obstante, estas características se degradan notablemente comparado con la situación en la que se configura el mismo *codec*, aparecen fenómenos como el eco y la interferencia.
- El *codec* más apropiado para una implementación de este tipo es el G.711 y se recomienda que se configure en el cliente y el servidor a la vez, debido a su alta calidad para trasmitir la voz tanto para reconocer la palabra correctamente y generar la llamada, como también durante la misma.
- Cuando en el cliente y el servidor se configura el mismo codec, a pesar que internamente se presentan retardos o disminución de la calidad de la voz, debido a la capacidad de predicción del cerebro dentro de una conversación, estos cambios son imperceptibles al oído humano, de tal manera que la voz parece natural y con muy buena y apropiada intensidad.
- En cualquier situación, sea que los codificadores se configuren igual o no, en el cliente y el servidor, es importante la vocalización, y pronunciación correcta de las palabras, dado que el motor de reconocimiento es en inglés y con mayor razón existe la tendencia de que el servidor entienda la palabra equivocada o en su defecto que no entienda.
- En cuanto a la cobertura y movilidad, se encontró que la capacidad de reconocimiento es independiente de la distancia a la que se encuentre el terminal de llamada del AP y por el contrario, depende de la combinación de codificadores que se haya definido para el cliente y servidor. Además, también se encontró que cuando se pierde la conexión al salir de la zona de cobertura, al retornar, se recupera la comunicación sin necesidad de remarcación.
- Un experimento que no estaba especificado dentro del plan de pruebas con este piloto, es la capacidad de reconocimiento en entornos ruidosos, pero debido al escenario universitario en donde fue probado el sistema, se demostró que ésta condición no representa un inconveniente mayúsculo, pues

una vocalización adecuada de la palabra, a una intensidad normal permiten, después de 2 o 3 intentos establecer la comunicación adecuadamente. Además existen micrófonos especiales con capacidad de supresión de ruido que pueden mejorar ese tiempo de respuesta desde el primer intento.

- Como resultado de las pruebas de congestión, en la figura 21 se muestra cómo el cliente identificado en la red con la IP 176.16.2.5 envía tráfico de datos al servidor con IP 176.16.2.9 y como el cliente inalámbrico con IP 176.16.2.10 genera una llamada hacia el servidor y cuando esta se establece el reconocimiento se hace sin el menor inconveniente. Se puede ver entonces que el tráfico de datos y el de voz son atendidos al mismo tiempo y no se genera congestión. Por su parte la figura 22 confirma los resultados de la figura 21 bajo la condición de movilidad.
- Finalmente, la figura 24 muestra todo el tráfico de datos y voz generado cuando todos los dispositivos están generando llamadas al servidor o enviando datos al mismo tiempo, sin que se presente congestión.

Con base en los resultados obtenidos a lo largo del estudio teórico de las tecnologías de reconocimiento de habla, redes inalámbricas, sistemas de respuesta de voz interactiva; y la experimentación con las herramientas tanto software como hardware, fue posible construir un piloto de sistema de comunicación con VoIP sobre una red inalámbrica de área local, controlado por un *corpus* de voz definido para un entorno hospitalario.

5. CONCLUSIONES Y RECOMENDACIONES

CONCLUSIONES

- Gracias al estudio del fundamento teórico de la tecnología de Reconocimiento Automático del Habla, fue posible establecer los criterios económicos y técnicos fundamentales para la selección del CMU Sphinx 2 como motor de reconocimiento apropiado para la implementación del piloto, debido a sus considerables ventajas como por ejemplo la independencia del hablante, el soporte a aplicaciones telefónicas y de red y principalmente por tratarse de una herramienta libre. Esta combinación de criterios hacen de dicho motor una herramienta especialmente poderosa para este proyecto, comparada con otras herramientas actualmente disponibles en el mercado.
- Con este proyecto se genera un aporte significativo al software libre, en lo que respecta a la integración del Sphinx, Perlbox-Voice y Asterisk y se establece una base teórica y práctica para la socialización y futuro estudio de las tecnologías de habla tanto en el ambiente académico, así como también corporativo.
- Una característica fundamental en la implementación del sistema es que el motor de reconocimiento sea independiente del hablante, ya que para una aplicación como la desarrollada en este proyecto debe poder adaptarse a las diferentes formas prosódicas en las que pueden hablar los integrantes del cuerpo médico, y así, poder ofrecer un reconocimiento exitoso.
- La definición del *corpus* de un sistema de este tipo, es una labor importante y que requiere tiempo y dedicación, pues su generación parte de una investigación de las condiciones sociales y culturales de la población para la cual se va a aplicar.
- Es importante la definición de un *corpus* apropiado para cada aplicación, debido a que los motores de reconocimiento, por lo general, contienen un léxico predeterminado de miles de palabras, que pueden utilizarse indistintamente, pero que podrían no ser memorizables para el personal por no tener correspondencia con el entorno y la terminología normalmente utilizada por los profesionales de cada escenario, en este caso hospitalario, correspondiente a la Clínica La Estancia S.A.
- Una WLAN representa una alternativa muy conveniente en la implementación de sistemas de comunicación que combinen datos y telefonía IP, ya que su ejecución implica costos relativamente bajos comparados con todas las ventajas que ofrecen a los usuarios, tales como movilidad, capacidad de expansión y convergencia de servicios.

- En la construcción de un sistema como el PSCVoWLAN la selección del *codec* es importante ya que el envío y recepción de la voz, es el eje de la aplicación en general, de ello depende la exactitud con que el motor de reconocimiento capte las palabras y el tiempo de respuesta del sistema ante determinado comando.
- Los sistemas de Respuesta de Voz interactiva representan soluciones actualmente convenientes tanto económica como tecnológicamente para entornos en donde las comunicaciones internas y con los clientes son fundamentales para cumplir los objetivos de una empresa. No obstante, con la investigación realizada en un entorno médico de la ciudad de Popayán como lo es la Clínica La Estancia, con el objeto de acercar la construcción de este piloto a los requerimientos del mismo, fue posible concluir que la implementación de un sistema IVR para las comunicaciones internas entre el personal médico, no es conveniente desde el punto de vista operativo, pues en situaciones de emergencia no hay tiempo suficiente para escuchar todas las opciones del menú. Sin embargo, para comprobar que es posible la integración de un servicio de este tipo, con la tecnología de reconocimiento de voz, sobre WLAN se implemento un módulo que, bajo solicitud expresa, permite el despliegue de un menú de posibilidades para la comunicación entre el personal.
- La utilización de herramientas libres se ha convertido en una gran ventaja en la construcción de todo tipo de aplicaciones en cualquier entorno, por representar una alternativa económica y flexible para adaptar y mejorar de acuerdo a los requerimientos deseados. En el caso particular de la construcción de este piloto, se escogieron herramientas libres para la construcción de la PBX y el motor de reconocimiento como Asterisk y CMU Sphinx 2, respectivamente, con las cuales se consiguió el objetivo deseado y mediante su exploración se pudo observar que estas herramientas permiten la implementación de sistemas robustos y con amplias posibilidades de futura expansión y la viabilidad que permite a un entorno hospitalario contar con un sistema de comunicación que se adapte a sus necesidades y sin realizar elevadas inversiones en software ni equipos.

RECOMENDACIONES

- Debido a la flexibilidad que ofrecen las herramientas libres, y en particular el motor de reconocimiento del CMU Sphinx 2, se recomienda para trabajos futuros su estudio y aplicación para el idioma español, ya que el utilizado en este proyecto está implementado para el idioma inglés. No obstante ello no representó un verdadero inconveniente a la hora de realizar las pruebas ni de sugerir su utilización con el personal médico de la Clínica La Estancia S.A, pues, se encontró que el personal maneja el idioma inglés en un nivel intermedio, lo cual le facilita el aprendizaje y pronunciación de los comandos

del *corpus*. Además, el diccionario utilizado por Sphinx 2 ofrece la segmentación en fonemas de las palabras, facilitando mucho más su utilización.

Sin embargo, es de gran importancia generar diccionarios en español para dicho motor de reconocimiento y su aceptación permitiría la expansión de éste y otro tipo de servicios no sólo para entornos donde el requisito sea manejar un idioma a parte del materno, sino también para aplicaciones más sencillas enfocadas a todo tipo de público.

- Se recomienda configurar el mismo *codec* de voz tanto en el servidor como en los clientes, ya que, como se pudo observar en los resultados obtenidos en el capítulo 4, esto permite mayor exactitud en el reconocimiento y mejor calidad de la voz durante la conversación. Sin embargo, este parámetro no es un elemento clave cuando se trata de la cobertura del sistema, pues, no influye directamente en ella. Esta aclaración debe hacerse debido a que, para este sistema la exactitud del reconocimiento es crucial en el funcionamiento del sistema y es la primera prioridad a suplir, pero puede darse el caso en otras implementaciones en las cuales la calidad de la voz no sea tan crucial como algún otro criterio y por ello se decida utilizar combinaciones de *codificadores*. Estos son aspectos a tener en cuenta y a definir de acuerdo a los requerimientos propios.
- Es importante tener en cuenta la selección de los micrófonos a utilizar, especialmente de acuerdo a la aplicación que se va a desarrollar. En el caso particular de este piloto, en donde la movilidad es uno de los principales objetivos del sistema, es recomendable que se utilicen diademas inalámbricas con supresión de ruido y eco. Actualmente se encuentran en el mercado variadas opciones, algunas de las cuales pueden alcanzar distancias de cobertura de hasta 150m, con interfaces USB y amplificadores de cobertura (solo en caso de superar los 150m de distancia).
- A pesar que el motor de reconocimiento CMU Sphinx 2 es independiente del hablante, lo cual implica que no necesita entrenamiento, es importante que los usuarios del PSCVoWLAN realicen un entrenamiento previo, no para que el sistema aprenda sus voces, sino para que los mismos usuarios aprendan cómo deben pronunciar las palabras adecuadamente, ello se debe a que para este piloto se utilizó un motor con diccionario en inglés.

REFERENCIAS

[1] Laboratorio de Fonética, Universidad de los Andes, Mérida Venezuela, "Tutorial de Fonética". Disponible en: http://ceidis.ula.ve/cursos/humanidades/fonetica/tutorial_de_linguistica/recono.html

[2] M. J. Poza Lara, L. Villarrubia Grande, J A. Siles Sánchez, Telefónica I+D, "Teoría y aplicaciones del reconocimiento automático del habla". disponible: http://www.tid.es/presencia/publicaciones/comsid/esp/articulos/vol23/habla/habla.html

[3] Stephen Cook. "Speech Recognition How To". Disponible en: www.ibiblio.org/pub/linux/docs/HOWTO/other-formats/pdf/Speech-Recognition-HOWTO.pdf

[4] Kimberlee A. Kemble, Program Manager, Voice Systems Middleware Education, IBM Corporation, An Introduction to Speech Recognition, Disponible en: http://www-900.ibm.com/cn/software/websphere/products/download/whitepapers/Introduction_to_Speech_Recognition.pdf

[5] P. Coxhead. Natural Language Processing & Applications, Speech Synthesis and Recognition, disponible en: http://www.cs.bham.ac.uk/~pxc/nlpa/2006/NLPA-Phon2.pdf

[6] D. Oran Network Working Group. Cisco Systems, Inc. RFC4313. Requirements for Distributed Control of Automatic Speech Recognition (ASR), Speaker Identification/Speaker Verification (SI/SV), and Text-to-Speech (TTS) Resources. disponible en: http://www.ietf.org/rfc/rfc4313.txt

[7]Pasamontes Colás José. Escuela Técnica Superior de Ingenieros de Telecomunicación, Madrid- España. Estrategias de incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en español, disponible en: http://elies.rediris.es/elies12/index.html#indice

[8] Fifth Generation Computer Corporation. Speaker Independent Connected Speech Recognition, disponible en: http://www.fifthgen.com/speaker-independent-connected-s-r.htm

[9] Pawar R.V, Kajave P.P, Mali S.N, "Speaker Identification using Neural Networks", disponible en: http://www.enformatika.org/data/v7/v7-86.pdf

[10] Rouchka Eric C. Department of Computer Science, Washington University. "Pattern Matching Techniques and Their Applications to Computational Molecular Biology - A Review", disponible en: http://kbrin.a-bldg.louisville.edu/brg/papers/WUCS-99-09.pdf

[11] Veera Ala-Keturi, Helsinki University of Technology, Speech Recognition Based on Artificial Neural Networks, disponible en: www.cis.hut.fi/Opinnot/T-61.6040/pellom-2004/project-reports/project_07.pdf

[12] J. Orozco García, Carlos A. Reyes García, Luis Enrique Erro, Instituto Nacional de Astrofisica Optica y Electronica, Tonantzintla, Puebla, México, Clasificación de Llanto del Bebé Utilizando una Red Neural de Gradiente Conjugado Escalado Disponible en: http://ccc.inaoep.mx/~llanto-de-bebe/pages/llantoMicai02.pdf

[13] NeuroSolutions, ¿What is a Neural Network ?, disponible en: http://www.nd.com/welcome/whatisnn.htm

[14] Ron Mains, Tim Meier, Scott Nainis, Henry M. James, "White Paper on Speech Recognition in the SESA *Call center*", disponible en: www.itsc.state.md.us/PDF/C-5_Speech_Recognition_SESA_CC_White%20Paper.pdf

[15] An Introduction to Speech Technology in Language Learning, disponible en: http://www.ilo.uva.nl/Ontwikkeling/imictll/docs/ICT3c_English.pdf

[16] Hugo L. Rufiner, Diego H. Milone, "Sistema de Reconocimiento Automático del Habla", disponible en: www.bioingenieria.edu.ar/grupos/cibernetica/milone/pubs/rah_RCDT2004draft.pdf

[17] Carlos Vivaracho Pascual, Luis Alonso Romero, Departamento de Informática. Universidad de Valladolid, Departamento de Informática y Automática. Universidad de Salamanca, España. "Redes Neuronales en reconocimiento de locutor", disponible en: http://lisisu02.usal.es/~airene/capit2.pdf#search=%22CAPIT2.PDF%22

[18] Deroo Olivier, "A short Introduction to Speech Recognition", disponible en: http://www.babeltech.com/download/SpeechRecoIntro.pdf

[19] Michael A. Grasso, "Automated Speech Recognition in Medical Applications", disponible en: www.csee.umbc.edu/~mikeg/papers/asrmed.pdf

[20] Eugenio Lázaro Cañedo-Argüelles, Manuel Domínguez Somonte. Universidad Pontificia Comillas, Escuela Técnica Superior de Ingeniería, Universidad Nacional de Educación a Distancia, Escuela Técnica Superior de Ingenieros Industriales. "Interfaz para el manejo de aplicaciones de diseño asistido por ordenador mediante dispositivos captadores de voz", disponible en: http://www.egrafica.unizar.es/%20ingegraf/pdf/Comunicacion17098.pdf

[21] Jim Baumann, "Voice Recognition". disponible en: http://www.hitl.washington.edu/scivw/EVE/I.D.2.d.VoiceRecognition.html

[22] Daniel Kiecza, "CVoiceControl - command and control for Linux!" disponible en: http://www.kiecza.net/daniel/linux/

[23] Coca Bedoya, Oscar Julian y Ramírez Rendón, Carlos Alberto, "Gnome Environment Recognition Voice", disponible en: https://glec.umanizales.edu.co/index.php/corporate/content/download/171/676/file/ArticuloGERvoice.pdf

[24] Department of Electrical & Computer Engineering, Carnegie Mellon University, CMU Sphinx Group, "Robust group's Open Source Tutorial Learning to use the CMU SPHINX Automatic Speech Recognition system" disponible en: http://www.speech.cs.cmu.edu/sphinx/tutorial.html

[25] Department of Electrical & Computer Engineering, Carnegie Mellon University, CMU Sphinx Group, "The CMU Sphinx Group Open Source Speech Recognition Engines", disponible en: http://cmusphinx.sourceforge.net/html/cmusphinx.php

[26] Department of Electrical & Computer Engineering, Carnegie Mellon University, CMU Sphinx Group, "The CMU Pronouncing Dictionary", disponible en: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[27] Ravishankar Mosul, CMU Sphinx Group, "Sphinx-II User Guide", disponible en: http://cmusphinx.sourceforge.net/sphinx2/sphinx2.html

[28] Department of Electrical & Computer Engineering, Carnegie Mellon University, CMU Sphinx Group, "CMU Sphinx's Development Page" disponible en: http://www.cs.cmu.edu/~archan/

[29] Arthur Chan, "Do we have a true open source dictation machine?", disponible: http://www.cs.cmu.edu/~archan/personal/whyNoOpenSourceDictationDraft4.html

[30] P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer, Carnegie Mellon University, "The 1996 hub-4 sphinx-3 system", disponible en:

http://www.cs.cmu.edu/~pwp/papers/h496_system/H496CMU.HTM

[31] A. G. Hauptmann, R. E. Jones, K. Seymore, S. T. Slattery, M. J. Witbrock*, and M. A. Siegler, School of Computer Science, Carnegie Mellon University, "Experiments in Information Retrieval from spoken documents", disponible en: http://www.nist.gov/speech/publications/darpa98/html/sdr20/sdr20.htm

[32] http://www.taug.ca/list/index.cgi?0::3354

[33] http://turnkey-solution.com/Asterisk-sphinx.html

[34] http://www.voip-info.org/wiki/view/Sphinx

[35] http://www-306.ibm.com/software/voice/viavoice/

[36] http://www.verbio.com/webverbio/html/productes.php?id=2

]37] Microsoft®, 2005, White Paper "Microsoft Speech Server 2004, Estimating Capacity for Speech-enabled Interactive Voice Response Solutions", disponible en: http://download.microsoft.com/download/6/6/2/6620D5B5-AA10-406A-AE9D-DBACE672D6E1/EstimatingCapacity.doc

[38] Microsoft®, 2004, White Paper "Log Analysis and Tuning with Microsoft® Speech Server 2004", disponible en: http://download.microsoft.com/download/A/6/1/A6102822-C703-4C86-9CA8-90618C7E6A12/Log%20Analysis%20and%20Tuning%20with%20Microsoft%20Speech%20Server%202004_MR.doc.

[39] Microsoft®, 2004, White Paper "Microsoft Speech Server 2004 Evaluation Guide", disponible en: http://download.microsoft.com/download/6/d/d/6dd08f10-3379-4eb4-9411-659286c10c7b/MicrosoftSpeechServer2004EvaluationGuide.doc

[40] "Microsoft Speech Server 2004 Lets You Speak to IT Look Who's Talking", disponible en: http://www.networkcomputing.com/showitem.jhtml?docid=1509sp1

[41] Microsoft®, 2004, White Paper, "Microsoft Speech Server System Requirements", disponible en: http://www.microsoft.com/speech/evaluation/requirements/default.mspx

[42] López Moreno, J. Tesis Licenciatura. Ingeniería en Sistemas Computacionales. Departamento de Ingeniería en Sistemas Computacionales,

Escuela de Ingeniería, Universidad de las Américas, Puebla. “Desarrollo de un reconocedor de dígitos con distinción de énfasis”. Disponible en: http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/lopez_m_j/capitulo_3.html

[43] Torruella J, Llisterri J. Departamento de filología Española, Universidad Autónoma de Barcelona. “Diseño de Corpus Textuales y Orales”, disponible en: http://liceu.uab.es/~joaquim/publicacions/Torruella_Llisterri_99.pdf

[44] Lombana Monica y Gonzáles Juan. Director: Guefry Agredo. Facultad de Ingeniería Electrónica y Telecomunicaciones, Universidad del Cauca. Popayán, 2005. Proyecto de Grado, "Prototipo Experimental de VoIP sobre WLAN para Entornos Empresariales".

[45] Brenner Pablo. Breezecom Wireless Communications. “A Technical Tutorial on the IEEE 802.11 Protocol”, disponible en: www.sss-mag.com/pdf/802_11tut.pdf

[46] ¿What is 802.11?. disponible en: http://searchmobilecomputing.techtarget.com/sDefinition/0,,sid40_gci341007,00.html

[47] Phd. Nedeltchev Plamen, 2001, "Wireless Local Area Networks and the 802.11 Standard", disponible en: www.cisco.com/warp/public/784/packet/jul01/pdfs/whitepaper.pdf

[48] García Rico F.J. "Redes Inalambricas, Sistemas de Transporte de datos", disponible en: http://www.unap.cl/davidcontreras/Wireless/libro%20de%20sistema%20de%20transporte%20de%20datos%20redes%20inalambricas.pdf

[49] Hernández Rodríguez P. A, Facultad de Contaduría y Administración, Universidad Veracruzana, Jalapa, México, 2005. "WLAN un complemento para LAN", disponible en: http://www.ilustrados.com/documentos/wlan.pdf

[50]l Khan Afzal R. "Comparison of IEEE 802.11a and IEEE 802.11b",disponible en: http://www.codeproject.com/useritems/IEEE_WLAN_Standards.asp

[51] MOBILEINFO.com,"Wireless LANs", disponible en: http://www.mobileinfo.com/Wireless_LANs/802.11a_802.11b.htm

[52] Hewlett-Packard Company, 2004. "Pásese al mundo inalámbrico-Descubra nuevas posibilidades de trabajar y jugar", disponible en : http://h41320.www4.hp.com/gomobile/es/es/learning_centre/pdf/wireless_brochure10_ene.pdf

[53] ©Intel Corporation. "Redes inalámbricas para la pequeña empresa" disponible en : http://www.intel.com/espanol/smallbusiness/wireless/wlan_choice.htm

[54] Broadcom®. White Paper "IEEE 802.11g, The New Mainstream Wireless LAN Standard", disponible en: http://www.54g.org/pdf/802.11g-WP104-RDS1.pdf

[55] Ponce E.M, Molina E, Mompó V, "Redes Inalámbricas 802.11", disponible en: www.todo-linux.com/manual.todo-linux.com/redes/Manual%20redes%20inalambricas.pdf

[56] "Codec de audio", disponible en: http://es.wikipedia.org/wiki/C%C3%B3dec_de_audio

[57] http://www.itu.int/rec/T-REC-P/en

[58] "Mean Opinion Score". disponible en: http://en.wikipedia.org/wiki/Mean_Opinion_Score

[59] "Mean Opinion Score" disponible en: http://searchnetworking.techtarget.com/sDefinition/0,,sid7_gci786677,00.html

[60] Pulse code modulation (PCM) of voice frequencies, disponible en: http://www.itu.int/rec/T-REC-G.711-198811-I/en

[61] "Telefonía IP", disponible en: http://www.euskalnet.net/apetxebari/nu_tecs/tele_ip.htm#CODECS

[62] "GSM Codec'', VoIP- Info, disponible en : http://www.voip-info.org/wiki-GSM+Codec

[63] "Speex: a free codec for free speech", disponible en: http://ww.speex.org

[64] Sitio oficial iLBC, disponible en : http://www.ilbcfreeware.org

[65] "Broadcom", disponible en: http://en.wikipedia.org/wiki/Broadcom

[66] RFC 1890, Network Working Group, "RTP Profile for Audio and Video Conferences with Minimal Control", disponible en: http://www.freesoft.org/CIE/RFC/1890/10.htm

[67] "Codecs", disponible en : http://www.voipforo.com/codec/codecs.php#g711

[68] “Interactive Voice Response”, disponible en:
http://en.wikipedia.org/wiki/Interactive_voice_response

[69] IVR Technologies, disponible en :
http://images.voip-news.com/voip_news/sp/ivr/images/diag2.jpg

[70] R. P. Rodgers C, The Lister Hill National Center For Biomedical Communications, U.S. National Library of Medicine. "A Report to the Board of Scientific Counselors", disponible en:
http://lhncbc.nlm.nih.gov/lhc/docs/reports/2004/tr2004006.pdf

[71] AEXIT, Asociación Extremeña del Ingenieros de Telecomunicación. Artículo “Asterisk PBX - Centralitas telefónicas de altas prestaciones y bajo coste”, disponible en:
http://www.aexit.es/aexit/ap/aexit.php/doc/Asterisk-PBX---Centralitas-telefonicas-de-altas-prestaciones-82.htm?sesion=54c6ec623d1af989e7ed322f5129e7ee

[72] Spencer M, Allison M, Rhodes C, The Asterisk Documentation Team. "The Asterisk Handbook Version 2", disponible en: www.digium.com/handbook-draft.pdf

[73] Van Meggelen J, Smith J, Madsen L. "Asterisk The Future of Telephony", editorial O'Reilly, 2005.

[74] Mahler P.S, "VoIP Telephony with Asterisk, A technical overview of the open source PBX"

[75] Asterisk@ Home Handbook, disponible en:
http://asteriskathome.sourceforge.net/handbook/