"DISEÑO Y DESARROLLO DE PROTOTIPO DE SISTEMA DE TRADUCCIÓN INSTANTÁNEA DE HABLA Y TRANSMISIÓN EN TIEMPO REAL, SOBRE EL PROTOCOLO RTP UTILIZANDO TECNOLOGÍAS DE RECONOCIMIENTO DE VOZ"

RICHARD NOLBERTO ROJAS BELLO
Universidad Austral de Chile, Ingeniero Civil en Informática, rrojas1@inf.uach.cl

ERICK ARAYA ARAYA
Universidad Austral de Chile, Ingeniero Ejecución Electrónico, earaya@inf.uach.cl

LUIS HERNÁN VIDAL VIDAL
Universidad Austral de Chile, Ingeniero Civil en Informática, lvidal@inf.uach.cl

Resumen **- El presente documento, expone el desarrollo de un prototipo de sistema de traducción de habla como Proyecto de Tesis. Este consiste básicamente en la captura del flujo de voz del emisor integrando tecnologías de reconocimiento de voz avanzadas, traducción instantánea, y comunicación sobre el protocolo Internet RTP para enviar en tiempo real la información al receptor. Este prototipo no transmite imagen, sólo aborda la etapa de audio. Finalmente, el proyecto además de abarcar un problema de comunicaciones personales, pretende aportar al desarrollo de actividades relacionadas con el reconocimiento de voz, motivando nuevas investigaciones y avances en el área.**

Abstract **- The present document exposes the development of a prototype of speech translation system as a Thesis Project. It consists basically on the capture of a flow of voice from the emitter, integrating advanced technologies of voice recognition, instantaneous translation and communication over the internet protocol RTP to send information in real-time to the receiver. This prototype doesn't transmit image, it only boards the audio stage. Finally, the project besides embracing a problem of personal communications, tries to contribute to the development of activities related with the speech recognition, motivating new investigations and advances on the area.**

Palabras Claves **– habla, reconocimiento, síntesis, RTP, traducción, tiempo, VoIP, voz.**

1. INTRODUCCIÓN

Actualmente Internet brinda distintas y eficientes formas de comunicarnos casi instantáneamente y sin importar que tan lejanas se encuentren las personas. La tecnología presente nos provee acceso a correos electrónicos, servicios *news*, servicios de mensajería instantánea (por ejemplo *MSN Messenger*) y aplicaciones para video conferencia. No obstante, en el tema de video conferencia y específicamente en conversaciones por voz, todavía quedan obstáculos que dificultan una plena comunicación; uno de ellos es la diferencia de idiomas o lenguas. Este es el punto sobre el cual se enfoca la solución propuesta en la tesis; solución que aborda el problema integrando tecnologías de reconocimiento y síntesis de voz, junto a tecnologías de transmisión de voz sobre redes IP (VoIP).

El reconocimiento de voz ha evolucionado considerablemente y se presenta como una interfaz confiable y efectiva entre el usuario y un computador. Ya es posible encontrar en el mercado sofisticadas aplicaciones orientadas al uso en oficinas, comercio electrónico, medicina para rehabilitación y telefonía entre otras.

Por otra parte, la transmisión de voz sobre redes IP ya dejó de emplearse solamente en aplicaciones de videoconferencia; hoy en día es un elemento indispensable de comunicación en algunas empresas comerciales, y no sólo por su flexibilidad y confiabilidad, sino que también por los costos reducidos que conlleva. La telefonía IP es una prueba concreta.

2. FUNDAMENTOS SOBRE RECONOCIMIENTO DE VOZ

2.1. Antecedentes Generales

La capacidad auditiva de un ser humano se caracteriza por percibir audio en un rango de 16Hz a 16Khz [1], y por diferenciar y comprender fácilmente distintos tipos de fuentes sonoras. Sin embargo, para que las máquinas logren tener esta última habilidad, se está trabajando durante muchos años. Dichos esfuerzos, han culminado en resultados que ya están presentes en el mercado mundial, y que se integran poco a poco al diario vivir.

Básicamente el proceso de reconocimiento de voz se puede explicar en dos pasos:

1. Extracción de fonemas: Los fonemas son unidades lingüísticas, sonidos que al agruparlos forman palabras. Son la unidad fonológica más pequeña en que puede dividirse un conjunto fónico; por ejemplo la palabra /páso/ "paso", está

formada por una serie de cuatro fonemas, ya que el máximo de unidades mínimas en que puede ser dividida es /p/+/a/+/s/+/o/ [2].

Para extraer los fonemas de la voz de entrada, la señal se analiza espectralmente vía transformadas de Fourier. El espectro de la palabra "AUDIO" se ve en la figura 2.1-1:

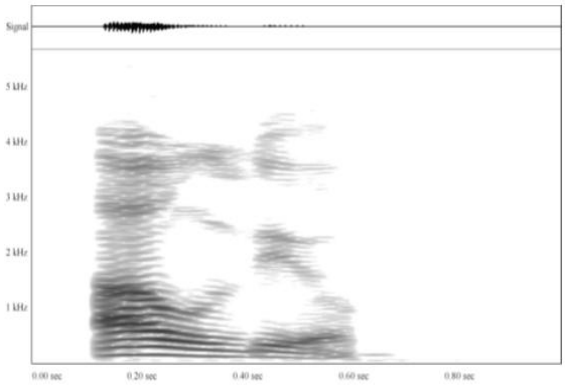

Figura 2.1-1: *Espectro de la palabra "AUDIO".*

2. Conversión de los fonemas en palabras identificables: Este proceso se puede realizar con ayuda de métodos topológicos, probabilísticos y de redes neuronales. Cada uno de ellos se detallarán en el punto 2.3 "Métodos de reconocimiento de voz".

2.2. Ramas del Reconocimiento de voz

Las técnicas de reconocimiento de voz se dividen en tres ramas principales [3]:

1. Reconocimiento de voz o Reconocimiento del habla: proceso que consiste en convertir un mensaje hablado en texto. Es la rama que más ha crecido en los últimos años.

2. Conversión texto-a-voz: generación de audio que emule la voz humana (síntesis de voz, TTS del inglés *Text-To-Speech*) a partir de información en formato texto digital.

3. Reconocimiento de Locutores: identificación o verificación de la persona que le habla a un sistema; su uso se proyecta como parte de medidas de seguridad.

La codificación de voz, también se postula como una rama del reconocimiento de voz; aunque pudiese considerarse un tema complementario al estar más relacionado con los canales de comunicación y el aprovechamiento del ancho de banda.

Como ya se mencionó anteriormente, otra área importante donde se aplica esta tecnología es el control Biométrico. Estudios descritos en [4] revelan el interés de crear repositorios de datos suficientemente robustos, que sirvan para ajustar sistemas de reconocimiento, incluyendo el de reconocimiento del habla. Detallan técnicas, métodos, dispositivos, resultados y conclusiones.

2.3. Métodos de Reconocimiento de voz

Tres son los métodos que han marcado la historia del reconocimiento de voz; ellos son: "Alineamiento temporal dinámico", "Modelos ocultos de Markov", y "Redes neuronales". Cada uno de estos métodos se aborda en los puntos siguientes.

2.3.1. Alineamiento temporal dinámico

El concepto de "Alineamiento temporal dinámico" (conocido como DTW, del inglés *Dynamic Time Warping)* se ha empleado para obtener la distorsión o diferencia entre dos palabras. Muchas veces una palabra puede no pronunciarse siempre a la misma velocidad o bajo las mismas condiciones del ambiente o del mismo locutor, es necesario entonces, ajustarla a un patrón para interpretar correctamente la información.

DTW está basado en la comparación todas las plantillas referencia - resultado de anteriores entrenamientos - contra plantillas compuestas de vectores de parámetros, calculados a partir de los distintos segmentos en que fue dividida la señal de entrada.

Para hacer la comparación se calcula la distancia mínima entre la referencia y la entrada, y finalmente se escoge la plantilla que entregue la menor distancia.

Los reconocedores de habla basados en DTW son fáciles de implementar y muy efectivos para vocabularios pequeños [5].

2.3.2. Modelos ocultos de Markov

Los sistemas basados en cadenas de Markov modelan procesos aleatorios, requiriendo menos memoria que los basados en DTW.

Un "Modelo oculto de Markov" (HMM, del inglés *Hidden Markov Models*) se puede ver como una máquina de estados finitos en la que el estado siguiente depende únicamente del estado actual, y asociado a cada transición entre estados se produce un vector de parámetros.

En reconocimiento de voz, las cadenas de Markov se encargan de ajustar los distintos fonemas captados a fonemas de palabras completas previamente establecidas, adquiridos por entrenamiento. Se asume que la cantidad de estados posibles para cada fonema es finita, por lo tanto el número de estados en la cadena también lo es.

Un tipo de HMM especialmente apropiado para reconocimiento de voz son los modelos "de izquierda a derecha" (Fig. 2.3.2-1); modelos en los que una vez abandonado un estado ya no se puede volver a él. Su plantilla se conforma de vectores que

se obtienen en cada uno de los nodos recorridos; cada nodo visitado genera un vector [6].

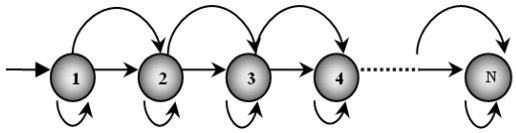

Figura 2.3.2-1: *Modelo de izquierda a derecha simplificado.*

Las cadenas de Markov no sólo son aplicables para extraer fonemas de la señal de voz, sino que también aplican para unir los fonemas y convertirlos en palabras, y luego tomar estas palabras y transformarlas finalmente en frases.

2.3.3. Redes neuronales

El empleo de redes neuronales en el reconocimiento de voz se justifica debido a que estas redes intentan emular complejos procesamientos cerebrales, y uno de ellos es precisamente el reconocimiento del habla. Además, su gran capacidad de resolver problemas que con otros métodos requieren mucha carga para los computadores, como son: el reconocimiento de patrones, evaluación de hipótesis y predicción.

Las redes neuronales organizan sus neuronas en capas (Fig. 2.3.3-1). Existe una capa de entrada y una de salida. La capa de entrada procesa directamente los vectores o plantillas, si el resultado de la operación de cada neurona supera un umbral predefinido la neurona realiza sinapsis con sus neuronas post-sinápticas. De esta forma, el resultado de la aplicación de una función de transformación no lineal a la combinación lineal de todos los puntos de la plantilla de entrada se traspasa a las neuronas siguientes.

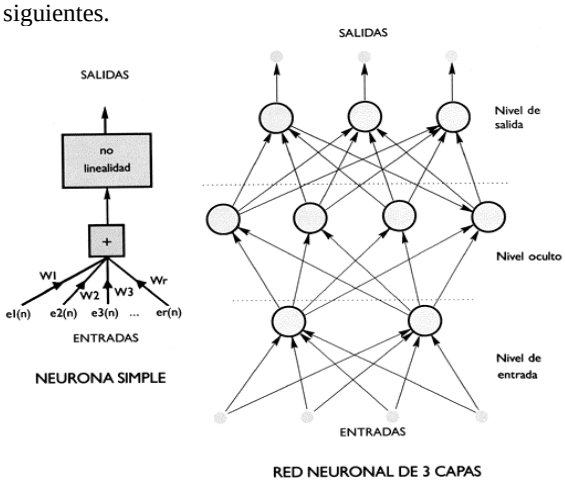

Figura 2.3.3-1: *Capas en una red neuronal* [6].

Así entonces, las redes neuronales pueden inferir sobre cuál es la palabra que viene en la señal de entrada. Sin embargo su eficiencia dependerá del entrenamiento al que haya sido sometida. Un excesivo entrenamiento de la red puede hacer que pierda su efectividad, llevándola a "exigir" demasiado a las plantillas de entrada para ajustarlas a sus modelos internos.

La señal de voz requiere de métodos con capacidad de proceso en dos dimensiones: espacio y tiempo. Las redes neuronales por sí solas sólo tienen capacidad de procesado espacial. Ello nos obliga a combinarlas con técnicas de Programación Dinámica como HMM; consiguiendo con ello modelar la variable tiempo, clasificaciones muy acertadas de las entradas de la red, y segmentación de la señal de entrada [7].

2.4. Métodos de Síntesis de voz

La síntesis de voz puede clasificarse dentro de tres tipos, según el modelo usado para generar la voz. Dichos modelos son: síntesis por formantes, síntesis articulatoria y síntesis por concatenación.

2.4.1. Sintetizadores por formantes

Las formantes son las resonancias características de cada articulador del tracto vocal (Fig. 2.4.1-1) [5]. Determinan el timbre particular de cada vocal y definen las características individuales de las voces. Cada palabra emitida puede definirse en términos de las frecuencias formantes propias de cada individuo.

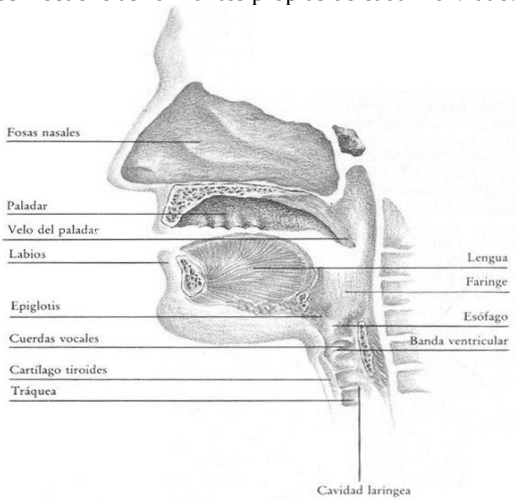

Figura 2.4.1-1: *Órganos de generación de sonido* [8]

Basándose en estos principios, los sintetizadores por formantes, modelan la resonancia del tracto vocal aplicando filtros para generar cada formante.

Los filtros son ajustables y poseen parámetros definibles mediante reglas. Éstas indican cómo modificar los parámetros entre un sonido y otro sin

perder la continuidad presente en los sistemas de generación de voz físicos [5].

Los sintetizadores por formantes involucran un procedimiento manipulable y flexible, son capaces de generar diversas voces modificando parámetros de sus filtros. Sin embargo, en la síntesis automática se necesita un número enorme de reglas, lo que requiere compiladores cada vez más sofisticados, capaces de integrar todo el conocimiento que se adquiere a base de experimentar con el sistema [3].

2.4.2 Sintetizadores articulatorios

Los sintetizadores articulatorios usan un modelo físico de la producción de la voz, simulando la propagación de las ondas acústicas [3]. Emplean parámetros - definidos por reglas, al igual que los sintetizadores por formantes - que modelan los movimientos mecánicos del aparato fonador, y de las distribuciones resultantes de volumen y presión de aire en pulmones, laringe, tracto vocal y nasal [5].

Los parámetros pueden obtenerse desde la voz real a través de rayos X y resonancias magnéticas, aunque posicionar los sensores en tracto vocal altera la forma en que se produce el habla e impide completamente sonidos naturales.

Esta técnica de síntesis no es capaz de generar voz con una calidad comparable a la síntesis por formantes y concatenación [5] e implica altos costos en investigación sobre el sistema humano de generación de voz.

2.4.3 Sintetizadores por concatenación

En la síntesis por concatenación, un segmento de voz se sintetiza simplemente reproduciendo la onda sonora con el fonema respectivo. Un discurso completo se sintetiza entonces concatenando varios fragmentos de voz [5].

Al contrario de los sintetizadores citados anteriormente, no necesita reglas ni ajustes manuales. Cada segmento es completamente natural, con lo que se puede esperar una salida del mismo tipo [5].

El problema que se presenta con los sintetizadores por concatenación es que si se unen dos segmentos de voz no adyacentes entre sí, pueden generarse discontinuidades espectrales o prosódicas. Las discontinuidades espectrales ocurren cuando las formantes no coinciden en el punto de concatenación; las discontinuidades prosódicas, cuando los tonos no coinciden en el punto de concatenación.

La discontinuidad puede llevar a clasificar un sistema de síntesis como deficiente, aunque esté basado en segmentos totalmente naturales.

Para enfrentar el problema de la discontinuidad se definieron “unidades”, que son representaciones abstractas de segmentos de voz. Las unidades pueden ir desde fonemas hasta sentencias completas. Mientras mayor tamaño tenga la unidad, mejor será la calidad de la síntesis, pero la cantidad de unidades a almacenar podría extenderse ilimitadamente. La figura 2.4.3-1 es un ejemplo que expresa el número de unidades necesarias para abarcar una cantidad N de los apellidos ingleses más frecuentes en EEUU.

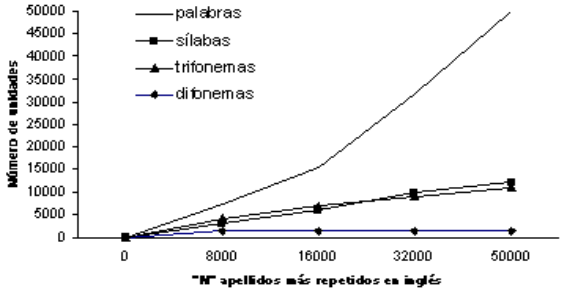

Figura 2.4.3-1: *Unidades requeridas, de diferentes tipos, para formar N apellidos* [5] (Adaptación).

Lo que se plantea actualmente es emplear híbridos que combinen unidades grandes (como palabras) con otras más pequeñas (fonemas), manteniendo el realismo y dando flexibilidad a la pronunciación.

3. FUNDAMENTOS SOBRE RTP

3.1. Antecedentes Generales

La tecnología VoIP consiste básicamente en la transmisión de voz sobre redes IP.

Se originó a partir de distintos factores que entre sí se potencian. El crecimiento de Internet y el desarrollo de métodos de compresión de voz, transmisión en tiempo real, y principalmente la necesidad de estar siempre comunicados son la base y antesala a la VoIP.

El creciente desempeño del protocolo IP y de las redes Ethernet, y la administración del ancho de banda, permiten aplicaciones como distribución automática de llamada, trabajo a distancia y mensajería instantánea; aplicaciones que se apoyan en estándares de constante evolución.

Dentro de los estándares más empleados para establecer sesiones multimedia se encuentran H.323 [9] y SIP [10]. Ambos marcan una fuerte presencia en Internet y emplean protocolos complementarios como RTP (*Real time Transport Protocol*) Y RTCP (*Real Time Control Protocol*).

3.2. Los Protocolos RTP Y RTCP

Para transmisión de datos en tiempo real como audio o video se introdujeron en SIP y H.323 mecanismos adicionales para garantizar una comunicación exitosa.

RTP al ser un protocolo de tiempo real, realiza sus operaciones manteniendo un comportamiento temporal estricto. Privilegia que las acciones se realicen intervalos de tiempo fijos, en lugar de ofrecer un desempeño a la velocidad más rápida.

RTP puede usarse para "media-on-demand" (servicio asíncrono de entrega de información) y para servicios interactivos. Su estructura básica está definida por dos partes: una relacionada con los datos a transmitir (sincronización, detección de pérdidas, seguridad e identificación de contenido) y otra dedicada a las funciones de control (identificación de fuente, soporte para gateways como puentes de audio y video, traductores de multicast a unicast).

RTP es independiente del protocolo de transporte; aunque fue desarrollado con el objetivo de residir sobre UDP, esfuerzos adicionales lograron usarlo sobre protocolos como IPX y CLNP, y experimentalmente sobre AAL5/ATM.

3.2.1 RTP

RTP se implementa sobre UDP, no aportando fiabilidad adicional ni reservas de recursos u otras garantías. Incluye información sobre los orígenes del tráfico y marcas de tiempo específicas para la sincronización de cada medio transportado [11].

La cabecera de un paquete RTP tiene la siguiente estructura:

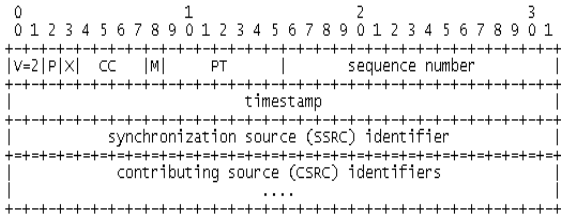

Figura 3.5.1-1: *Estructura de cabecera de un paquete RTP* [12]

Los doce primeros octetos están presentes en todos los paquetes RTP, mientras que la lista de los identificadores está presente solo cuando se inserta un "mezclador". El significado de cada campo se detalla en el ANEXO IV.

Entre emisores y receptores puede haber 2 tipos de nodos: mezcladores y receptores [12].

Mezclador, Recibe varios paquetes RTP, los combina, y envía otro nuevo. Puede utilizarse en casos donde algunos de los participantes pertenece a un área con bajo ancho de banda y los demás participantes son de un área con un ancho de banda privilegiado. En lugar de forzar a todos a usar una baja calidad en la comunicación se resincronizan los paquetes entrantes, mezcla el audio en un solo flujo, y recodifica el audio para enviarlo por bajos ancho de banda. Finalmente lo transmite a sus participantes en un paquete con un nuevo SSRC (*Synchronization Source*).

Traductor, Al igual que el mezclador, es un sistema intermedio. Ejemplos de traductor son:

- Conversores de codificación.
- Replicadores de multicast a unicast.
- Filtros a nivel de aplicación en cortafuegos.

El traductor reenvía paquetes tras modificarlos pero sin cambiar su identificador SSRC. Posibilita que los receptores identifiquen fuentes individualmente, aunque todos los paquetes pasen a través del mismo traductor y lleven la dirección fuente de la red del traductor.

Pero RTP solo se encarga de encapsular tráfico en tiempo real. El protocolo de reserva y garantía de calidad de servicio a determinados flujos se conoce como RTCP. Cada participante envía un paquete RTCP para que se sepa quienes están escuchando [13].

3.2.2. RTCP

RTCP se basa en la transmisión periódica de paquetes de control a todos los participantes de una sesión, y usando el mismo mecanismo de distribución que los paquetes de datos.

Se definen distintos tipos de paquetes RTCP que transportan variada información de control.

Los tipos pueden ser:

- SR (*Sender Report*): para estadísticas de transmisión y recepción de los participantes que son emisores activos.
- RR (*Receiver Report*): para estadísticas de recepción de los participantes que no son emisores activos.
- SDES (*Source Description Items*): ítems de descripción de fuente, incluye CNAME.
- BYE: Indica el fin de la participación en una sesión.
- APP: funciones específicas de la aplicación.

4. DISEÑO Y DESARROLLO DEL PROTOTIPO

Ya descritos los fundamentos tecnológicos sobre los cuales se basa el desarrollo expuesto se detalla a continuación su arquitectura, implementación y validación.

4.1. Arquitectura

La figura 4.1-1 representa el prototipo de sistema de reconocimiento, traducción y transporte de VoIP propuesto como solución. El diagrama muestra flujo enviado por un usuario A de lengua española hacia un usuario B de habla inglesa. El usuario B ejecuta una instancia del prototipo remotamente.

Figura 4.1-1: *Prototipo de sistema de traducción.*

La figura 4.1-2 representa un modelo simplificado del flujo de información dentro del prototipo.

El acceso a *Google Language Tools* - a diferencia de *AltaVista BabelFish* – se puede obtener directamente de la página de traducción empleando la clase *C# WebWagon* (clase implementada por un desarrollador certificado de aplicaciones Microsoft, Jon Vote), la cual permite cargar el código completo de paginas HTML. Sin embargo, la sintaxis de *VC#*

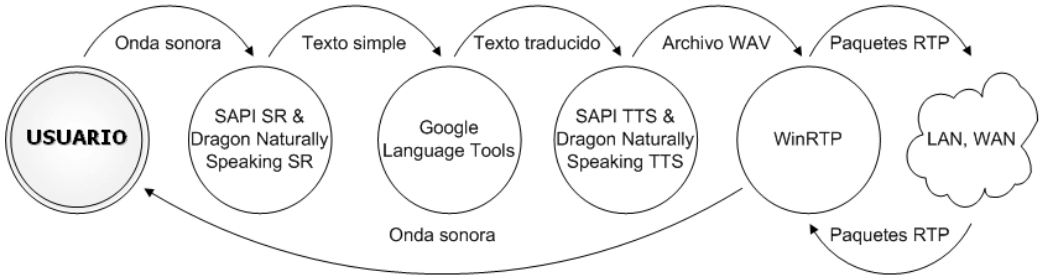

Figura 4.1-2: *Prototipo de sistema de traducción.*

Las bibliotecas empleadas para desarrollar la aplicación de reconocimiento y síntesis de voz fueron *Microsoft SAPI SDK* versión 4.0. Estas bibliotecas, añaden la ventaja de que la aplicación no solamente use los motores incluidos en SAPI, sino que motores de otros desarrolladores también pueden ser reconocidos por la aplicación si son compatibles con SAPI.

Para habilitar el reconocimiento de voz en español se empleó el motor *Dragon Spanish NaturallySpeaking*. El reconocimiento en inglés se logró de dos formas distintas, con *English Continuous* de *Microsoft SAPI SDK* (incluido en las bibliotecas) y con *Dragon English NaturallySpeaking.*

La síntesis en español se implementó haciendo uso de las voces masculina y femenina de *TTS3000* de Lernout & Hauspie; la síntesis en inglés con las voces de *Microsoft SAPI SDK* 4.0.

Se investigó sobre motores traductores de texto controlables mediante lenguajes de programación. Los resultados obtenidos arrojaron dos traductores en línea, *AltaVista BabelFish* y *Google Language Tools*.

Para acceder a *Altavista BabelFish* es necesario un cliente SOAP (*Simple Object Access Protocol*) que enlace el prototipo con el servicio de *Altavista*. En la práctica, al ejecutar su servicio de traducción, la conexión a *BabelFish* se comporta de manera inestable, y pierde reiteradamente la conexión con el servicio. Una causa posible, es una modificación de acceso al servicio que el cliente no es capaz de resolver.

no es compatible con *VC++* y por lo tanto no es directamente integrable al proyecto de tesis. Por esta razón, como solución a la incompatibilidad de la clase *WebWagon*, se exportó su código a una biblioteca TLB (biblioteca de tipos de datos, empleada para crear objetos COM en entornos de programación.). Así, puede invocarse desde otros desarrollos. Mayor detalle de este proceso se encuentra en el ANEXO VII.

Por otra parte, para realizar la transmisión de la voz se empleó el protocolo RTP. Como ya se describió en el capítulo 3, RTP es un protocolo ligero que proporciona soporte a aplicaciones de audio y/o video, o a aquellas que requieren transmisión en tiempo real. Además es un protocolo común a los estándares H.323 y SIP.

Las bibliotecas RTP que constituyen la base de la transmisión y recepción del audio son *WinRTP, las que están presentes en tecnologías comerciales como la* soluciones *Cisco AVVID*.

4.2 Implementación

El desarrollo del software fue por prototipado. Para implementarlo se buscó un lenguaje que además de ser compatible con las bibliotecas *SAPI* y *WinRTP*, fuese de un nivel lo más cercano posible al sistema operativo para obtener un mejor rendimiento.

Por estas razones, se optó por implementar el prototipo en Microsoft Visual C++ .NET accediendo directamente a las APIs de Windows. Cabe destacar que no se emplearon funciones propias de .NET tales como MFC (*Microsoft Foundation Classes*) y herramientas para servicios web.

La aplicación comienza con el cuadro de diálogo de la figura 4.2-1. El usuario debe ingresar un nombre y elegir el modo de reconocimiento de voz que necesite.

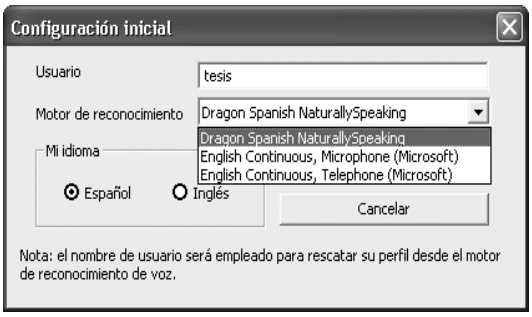

Figura 4.2-1: *Diálogo de configuración inicial.*

El nombre de usuario es necesario para asociar al locutor con el modo de reconocimiento escogido; de esta forma, mientras más sesiones de entrenamiento ejecute el locutor, mejor será el reconocimiento. Toda la información extraída acerca de su voz se guarda en un perfil creado con su nombre de usuario, que a la vez, es su identificador principal. Al hacer clic en "*Aceptar*" un mensaje informará si el perfil existe en el sistema o si ha creado uno.

El recuadro "*Mi idioma*" (fig. 4.2-2) es el menú de las dos opciones de traducción. Si por ejemplo el locutor escoge "Inglés", entonces todo el texto reconocido se traducirá de inglés a español.

Figura 4.2-2: *Selección de idioma del usuario.*

Luego, en el segundo cuadro de diálogo (Fig. 4.2-3) el usuario debe elegir una voz que lo represente. La voz escogida será la que escuchará el segundo participante de la sesión. Cada vez que el usuario seleccione un tipo de voz del menú, el personaje seleccionado se presentará al usuario, comunicándole que desde ése momento él será su voz sintetizada.

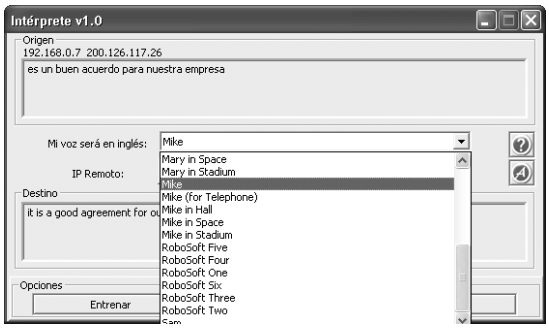

Figura 4.2-3: *Diálogo principal de la aplicación.*

Es necesario para el locutor, conocer el número IP del computador de la segunda persona que ejecuta una instancia de la aplicación (fig. 4.2-4). La comunicación para este prototipo es unicast y se inicia al presionar el botón "Transmitir". Este botón tiene un doble propósito: iniciar, y terminar la transmisión y recepción.

Figura 4.2-4: *Transmisión hacia IP remoto.*

Dos cuadros de texto muestran por separado el texto reconocido (fig. 4.2-5) y la traducción obtenida desde *Google Language Tools* (fig. 4.2-6).

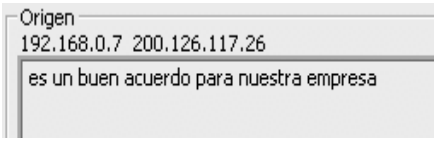

Figura 4.2-5: *Cuadro de texto para voz reconocida.*

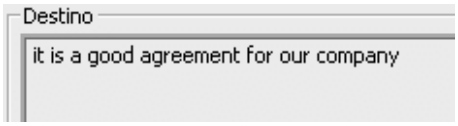

Figura 4.2-6: *Cuadro de texto para traducción recibida.*

En la parte inferior del dialogo de la figura 4.2-3 se observan tres botones, "Entrenar", "Sensibilidad" y "Audio". Sus funciones son:

- Entrenar: asistente para entrenar el motor de reconocimiento de voz (fig. 4.2-7). Este asistente es propio de los motores Microsoft SAPI, no obstante, es probable que motores ajenos a Microsoft – como *Dragon NaturallySpeaking* - no permitan el entrenamiento por esta vía; en tal caso, dicho motor debe ser entrenado directamente con el software con que fue adquirido.

Figura 4.2-7: *Asistente para entrenamiento de Microsoft SAPI.*

- Sensibilidad: configuración de precisión del reconocimiento, tiempo de respuesta, rechazos y reconocimientos exitosos (fig. 4.2-8 y fig. 4.2-9). Las opciones varían entre distintos motores.

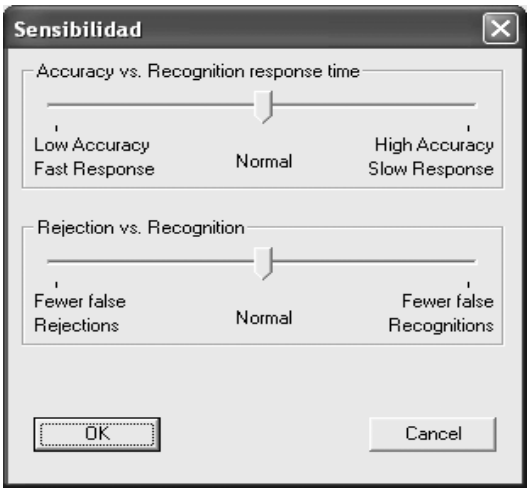

Figura 4.2-8: *Configuración de sensibilidad en Microsoft SAPI.*

Figura 4.2-9: *Configuración de sensibilidad en Dragon NaturallySpeaking*

- Audio: asistente para configuración de los niveles de audio en micrófono y parlantes del PC (fig. 4.2-10).

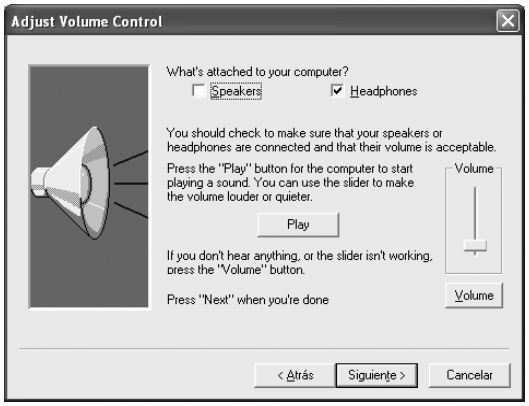

Figura 4.2-10: *Configuración de micrófono y audífonos.*

Para asistir al usuario durante la ejecución del prototipo, el software incluye un archivo de ayuda en formato HLP. Informa sobre los nombres de usuarios, motores, idiomas y ejecución de sesiones. Se abre al hacer clic en el botón  del cuadro de diálogo principal.

4. VALIDACIÓN DEL PROTOTIPO

El mejor rendimiento obtenido por el prototipo fue usando *Dragon NaturallySpeaking Spanish*. Alcanzó una precisión superior a un 96% con sólo 5 sesiones de entrenamiento.

La incorporación de *Google Language Tools* se ajusta a las necesidades de traducción del prototipo; considerando la estabilidad del servicio, rapidez y la adecuada semántica de las traducciones.

La variedad de voces disponibles para síntesis de voz pone a disposición del usuario distintas formas de identificarse. El prototipo cuenta con voces nítidas masculinas, femeninas, con variados tonos y en ambientes diversos.

Se analizó también el comportamiento del ancho de banda (AB) al ejecutar una instancia del prototipo. Las pruebas se hicieron en dos conexiones distintas, y se observó el tráfico de entrada/salida en intervalos de 1 segundo en los nodos de menor velocidad. El software empleado para observar el tráfico fue *PRTG Traffic Grapher V4.3.0.470 Freeware Edition.*

Para una conexión entre un nodo de 320/128 Kbps y otro de 512/256 Kbps, se obtuvo el siguiente gráfico:

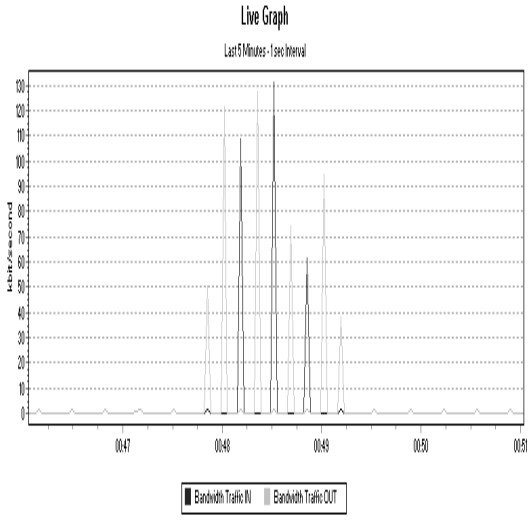

Figura 5-1: *Consumo de ancho de banda en conexión de 320/128 Kbps.*

Del gráfico se desprende la tabla 1, que refleja que el consumo de AB de salida en el nodo 320/128 Kbps no superó los 128 Kbps, y el flujo de entrada

alcanzó como máximo los 131 Kbps. Las frases transmitidas se recibieron con un retardo aproximado de 1 seg. y sin jitter perceptible.

Frase sintetizada	Kbps
buenos días	50
me gustaría acordar una fecha para la reunión	122
we will travel Monday of the next week	109
entonces los esperamos en el aeropuerto	128
our airplane will arrive at the 11 hours	131
perfecto ahí estaremos	74
we see Monday	62
buen viaje y nos vemos pronto	95
Adiós	39

Tabla 1: *Consumo de ancho de banda en conexión de 56/48 Kbps.*

En una sesión entre un nodo con Internet por RTC (*Red Telefónica Conmutada*) 56/48 Kbps y un nodo ADSL 320/128 Kbps, el resultado se ve en la figura 5-2.

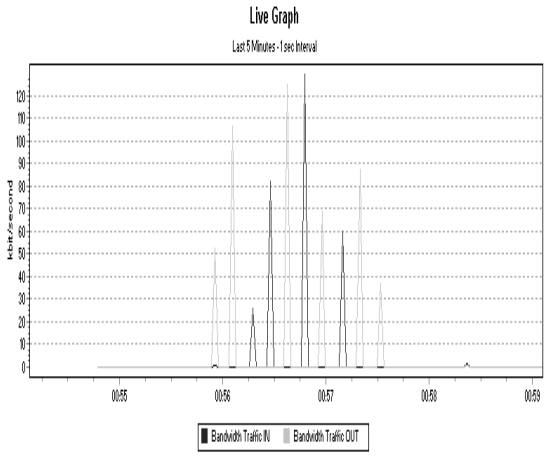

Figura 5-2: *Consumo de ancho de banda en conexión de 56/48 Kbps.*

Como se ve en la tabla 2, en esta conexión el consumo de AB de salida y entrada sigue el mismo comportamiento de consumo que en la prueba anterior. No obstante, las frases traducidas se escuchan interrumpidas por silencios, lo que significa una disminución en la calidad del audio. Además, el retardo aumentó a 3 segs. aproximadamente.

Para comprobar cómo se comporta el consumo de AB con frases más extensas, desde el nodo 320/128 se realizó la transmisión de frases con distintos tamaños. En la tabla 3 se observan los tamaños (en palabras y caracteres) de 11 frases. Cada frase tiene asociado el tiempo necesario para sintetizarla, y el consumo máximo de AB de salida medido en intervalos de 1 seg.

Frase sintetizada	Kbps
buenos días	52
me gustaría acordar una fecha para la reunión	108
we will travel Monday of the next week	82
entonces los esperamos en el aeropuerto	125
our airplane will arrive at the 11 hours	128
Perfecto ahí estaremos	70
we see Monday	60
buen viaje y nos vemos pronto	88
Adiós	38

Tabla 2: *Consumo de ancho de banda en conexión de 56/48 Kbps.*

Frase	Caracteres	Palabras	Kbps	Segundos
1	500	77	400	35
2	450	66	400	31
3	400	60	400	27
4	350	50	400	25
5	300	46	385	21
6	250	38	380	18
7	200	32	370	14
8	150	26	350	10
9	100	19	230	7
10	50	9	140	4
11	25	5	75	2

Tabla 3: *Rendimiento por caracteres en conexión 320/128 Kbps.*

Al observar los datos de la tabla en el gráfico de la figura 5-3, se distingue un consumo ascendente de AB para frases de hasta 150 caracteres. Luego, el consumo de AB se estabilizó en los 400 Kbps.

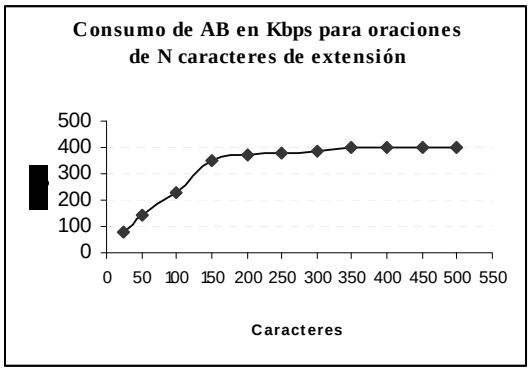

Figura 5-3: *Consumo de ancho de banda en Kbps para frases de N caracteres de extensión.*

La calidad del audio recibido en el nodo 512/256 Kbps, no se percibió con jitter hasta la recepción de frases de 100 caracteres. La utilización de AB registró 230 Kbps en el nodo 320/128.

El tiempo necesario para generar el buffer de voz sintetizada se comportó de forma lineal (fig. 5-4). La extensión de la frase es proporcional al tiempo empleado para sintetizarla.

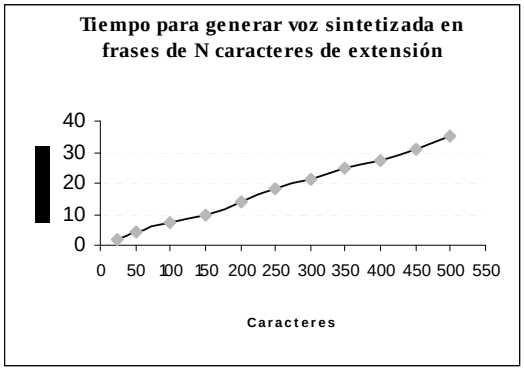

Figura 5-4: *Tiempo para generar voz sintetizada en frases de N caracteres de extensión.*

5. CONCLUSIONES

Esta tesis se enfocó como un trabajo de ingeniería que busca solución a una limitante social y de comunicaciones, la diferencia de idiomas. Se enfrentaron procesos de apertura de paquetes tecnológicos e integración de reconocimiento de voz, síntesis de voz, VoIP y las funciones de un servicio web.

El diseño de solución propuesto, otorga al proyecto un potencial equiparable con complejas propuestas universitarias. Un ejemplo es el sistema JANUS (de Carnegie Mellon University y Universität Karlsruhe), que ha sido empleado para tomar decisiones entre personas que no comparten el mismo idioma, y que necesitan una traducción inmediata.

Se logró la implementación de un prototipo que hace converger plenamente las tres tecnologías estudiadas: reconocimiento de voz, síntesis de voz y VoIP.

El software diseñado sintoniza con las necesidades del usuario. La elección de bibliotecas compatibles con soluciones ajenas a Microsoft, hizo que el usuario final tenga en sus manos la posibilidad de adquirir motores (de reconocimiento y/o síntesis), que estén a su alcance económico y se ajusten a sus requerimientos específicos.

El módulo de traducción implementado cumple con la funcionalidad requerida para el prototipo. El enlace a *Google Language Tools* proporcionó frases traducidas exitosamente en ambos sentidos (español/inglés, inglés/español) y con una correcta sintaxis y semántica.

El diseño del prototipo propuesto posee una conectividad extensible. Incorporar un módulo SIP para abordar la telefonía IP y tradicional, no representaría un cambio radical en la arquitectura propuesta. Además, SIP emplea RTP para transmitir información en tiempo real.

Las conexiones ADSL soportan el flujo de entrada/salida generado por las traducciones, mejorando la comprensión del mensaje hablado. Por lo tanto, basándose en los resultados del gráfico G5-3 - donde el consumo de AB se estabiliza en 400Kbps - se recomienda este tipo de conexiones para un mejor desempeño.

La información de la figura 5-4, indica que el óptimo sería sintetizar traducciones menores de 50 caracteres (10 palabras aprox.). Una extensión de 50 caracteres no supera los 4 segundos en sintetizar. La implementación de un algoritmo que divida frases extensas, en nuevas de menor tamaño, podría reducir aún más el tiempo de síntesis.

La utilización de motores de reconocimiento con porcentajes altos de precisión - cómo es el caso de *Dragon NaturallySpeaking* - reduce la ocurrencia de reconocimientos erróneos. Esto beneficia directamente la eficiencia del módulo de reconocimiento y del sistema en general.

REFERENCIAS

[1] M. Möser, J. Barros, *Ingeniería Acústica Teoría y Aplicaciones*. Universidad Austral de Chile, p. 1.

[2] A. Quilis, J. Fernández, *Curso de fonética y fonología españolas*. Instituto Miguel de Cervantes, 1968, p. 9.

[3] L. Hernández, F. J. Caminero. (2001, Sep). *Estado del arte en Tecnología del Habla.*. Universidad Politécnica de Madrid, Telefónica investigación y desarrollo. [Online]. Disponible: http://www.tid.es/presencia/publicaciones/docs_comtid/numero10.pdf

[4] C. Vivaracho, I. Moro. (2002, Oct.). *Creación de una base de datos para reconocimiento de personas mediante multimodalidad biométrica.* [en línea]. Universidad de Valladolid, Universidad del País Vasco. [Online]. Disponible: http://www.infor.uva.es/biometria/Documentos/Articulos/Biometria_SA.pdf

[5] X. Huang, A. Acero, *Spoken Language Processing*. Prentice Hall, 2001, pp. 4-5, 27, 793, 796, 803-805, 807, 809.

[6] M. Poza, L. Villarrubia. (2001, Nov.). *Teoría y aplicaciones del reconocimiento automático del habla*. Telefónica Investigación y Desarrollo. [Online]. Disponible: www.tid.es/presencia/publicaciones/docs_comtid/numero3.pdf

[7] J. Colás. (2004, Dic.). *Estrategias de incorporación de conocimiento sintáctico y semántico en sistemas de comprensión de habla continua en español.* Escuela Técnica Superior de Ingenieros de Telecomunicación- [Online]. Disponible: http://elies.rediris.es/elies12/

[8] Enciclopedia Británica Publisher, Inc. *Enciclopedia Hispánica Macropedia*, vol 6. Editorial Ercilla Galicia 1996.

[9] *H.323 : Sistemas de comunicación multimedios basados en paquetes*. International Telecommunication Union. (2003, Jul.). Disponible: http://www.itu.int/rec/recommendation.asp?lang=s&type=folders&parent=T-REC-H.323

[10] *Request for Comments: 3261 "SIP: Session Initiation Protocol"*. (2002, Jun.). Network Working Group. [Online]. Disponible: ftp://ftp.rfc-editor.org/in-notes/rfc3261.txt

[11] Grupo de Sistemas y Comunicaciones. *Protocolos de transporte con entrega en tiempo real*. (2003, Abr.). Universidad Rey Juan Carlos. [Online]. Disponible: http://gsyc.escet.urjc.es/docencia/cursos/fse-mbone/transpas/node9.html

[12] *Request for Comments: 1889 "RTP: A Transport Protocol for Real-Time Applications*". (1996, Ene.). Network Working Group. [Online]. Disponible: ftp://ftp.rfc-editor.org/in-notes/rfc1889.txt

[13] J. Salvachúa. *Realtime Transport Protocol RTP*. (2002, Nov.) Departamento de Ingeniería de Sistemas Telemáticos. Universidad Politécnica de Madrid. [Online]. Disponible: http://www.lab.dit.upm.es/~labscom/almacen/sld/rtp.pdf