UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

COLEGIO DE CIENCIA Y TECNOLOGÍA

LICENCIATURA EN INGENIERÍA EN SISTEMAS ELECTRÓNICOS Y DE TELECOMUNICACIONES

“Diseño e implementación de un sintetizador de voz concatenativo para interfaces hombre-máquina”

TRABAJO RECEPCIONAL
PARA OBTENER EL TÍTULO DE LICENCIADA EN
INGENÍERA EN SISTEMAS ELECTRÓNICOS Y DE TELECOMUNICACIONES

Presenta

Marisol Caballero López

Director del trabajo recepcional
Ing. Amaranto de Jesús Dávila Jáuregui

México, D.F. Mayo ,2015.

UACM
Universidad Autónoma de la Ciudad de México
Nada humano me es ajeno

CA Coordinación Académica

SISTEMA BIBLIOTECARIO DE INFORMACIÓN Y DOCUMENTACIÓN

UNIVERSIDAD AUTÓNOMA DE LA CIUDAD DE MÉXICO
COORDINACIÓN ACADÉMICA

RESTRICCIONES DE USO PARA LAS TESIS DIGITALES

DERECHOS RESERVADOS©

La presente obra y cada uno de sus elementos está protegido por la Ley Federal del Derecho de Autor; por la Ley de la Universidad Autónoma de la Ciudad de México, así como lo dispuesto por el Estatuto General Orgánico de la Universidad Autónoma de la Ciudad de México; del mismo modo por lo establecido en el Acuerdo por el cual se aprueba la Norma mediante la que se Modifican, Adicionan y Derogan Diversas Disposiciones del Estatuto Orgánico de la Universidad de la Ciudad de México, aprobado por el Consejo de Gobierno el 29 de enero de 2002, con el objeto de definir las atribuciones de las diferentes unidades que forman la estructura de la Universidad Autónoma de la Ciudad de México como organismo público autónomo y lo establecido en el Reglamento de Titulación de la Universidad Autónoma de la Ciudad de México.

Por lo que el uso de su contenido, así como cada una de las partes que lo integran y que están bajo la tutela de la Ley Federal de Derecho de Autor, obliga a quien haga uso de la presente obra a considerar que solo lo realizará si es para fines educativos, académicos, de investigación o informativos y se compromete a citar esta fuente, así como a su autor ó autores. Por lo tanto, queda prohibida su reproducción total o parcial y cualquier uso diferente a los ya mencionados, los cuales serán reclamados por el titular de los derechos y sancionados conforme a la legislación aplicable.

AGRADECIMIENTOS

Dedico este trabajo a mis padres (Caballero Herrera Pedro y Navarro López Soledad) por un testimonio de cariño y eterno agradecimiento por mi existencia, valores morales y formación profesional, ya que sin escatimar esfuerzo alguno, han sacrificado cada parte de su vida para formarme y porque nunca podré pagar todos sus desvelos, ni aun con las riquezas más grandes del mundo, gracias por su cariño, comprensión y los consejos que me dieron a lo largo de mi vida que me hicieron crecer. También les doy gracias por aguantarme en los años que me tarde en salir de la universidad, los amo.

A mi hermana (Caballero López Jaquelin Yenny) por su apoyo incondicional, amor, comprensión, esfuerzo y paciencia que me brindo cuando lo necesite, por compartir los momentos de tristeza, enojo, alegría, altas y bajas que tuve. Te quiero mucho hermanita.

Doy gracias especialmente para Antonio Ramírez por estar a mi lado ayudándome y motivándome a seguir adelante, gracias por estar en los momentos más difíciles de mi vida y me ayudaste a superar, gracias por escucharme y hacerme reír en los momentos tristes, te amo. También agradezco a su familia que siempre estuvieron a mi lado dándome ánimos, apoyo, cariño y motivándome a salir adelante, ya que sin ellos mi sueño no lo habría cumplido, gracias por esos momentos que me caí y me ayudaron a levantarme, los quiero mucho y los considero una familia muy linda.

Agradezco a todos mis profesores de la licenciatura de ingeniería en sistemas electrónicos y de telecomunicaciones, que día a día me compartían sus conocimientos en particular a Magali Cortes Vásquez que me enseño tanto de la profesión como de la vida, impulsándome a seguir adelante y por su amistad.

A mis amigos que siempre me han acompañado en todo momento, especialmente a Estrella Alcántara que me ayudo y motivo en el término de mi tesis, alguna vez me dijiste una frase especial: “Si dios te pone ante una circunstancia también te dará la habilidad de sobrepasarla”

Agradezco a mi director de tesis Amaranto de Jesús Dávila Jáuregui, por su orientación, esfuerzo, dedicación y ayuda que me brindo para la realización de este trabajo, para concluir mis estudios con éxito.

A la profesora Diana Aurora por su apoyo y tiempo que dedico para este trabajo.

Gracias a mis sinodales por sus valiosas sugerencias y por todo su tiempo invertido en la revisión de la tesis.
Son muchas las personas que han formado parte de mi vida a las que me gustaría agradecerles su amistad, consejos apoyo, ánimo y compañía en los momentos más difíciles de mi vida. Gracias a todas las personas que día a día me dieron un consejo para ir creciendo como persona, y que con su ayuda he llegado a cumplir la primera de mis metas.

Agradezco a Dios y a la virgen María por sus bendiciones.

Se agradece al SECITI el apoyo para la realización de este trabajo, como parte del proyecto: Robot móvil de servicio para vigilancia y prevención del delito (PI2011 -1R), del convenio UACM – SECITI 60-2013.

RESUMEN

La voz es uno de los principales recursos con los que cuenta el ser humano para comunicarse, a partir del desarrollo de la tecnología y su uso intensivo en la vida cotidiana de las personas que trae como consecuencia la creciente interacción entre los seres humanos y las maquinas, surge el interés de desarrollar interfaces HMI (Human Machine Interface) que permiten al usuario ingresar y recibir información de una manera fácil, rápida y lo más cercano a una comunicación humana.

En este trabajo se implementó un sintetizador de voz por concatenación de unidades, en una computadora embebida Raspberry Pi B, que permite a una máquina sintetizar señales de voz con la finalidad de utilizarse como interfaz de comunicación de un robot móvil de servicio con una conexión remota HMI y tecnología Wi-Fi. Con la técnica de concatenación de unidades se obtiene la voz sintética con una mayor calidad e inteligibilidad, para el español de México.

Palabras clave: HMI, Wi-Fi, síntesis por concatenación de unidades, Raspberry Pi B.

INDICE

Capítulo 1

1.1 Introducción

Tradicionalmente las aplicaciones robóticas estaban orientadas a resolver las necesidades de los sectores industriales más desarrollados para la producción masiva, como el de la industria automotriz o farmacéutica. La introducción de los robots manipuladores y gracias a las grandes posibilidades que ofrecía su uso, impulsó el desarrollo de sistemas robóticos más rápidos, precisos y flexibles. Esto originó un gran adelanto en la automatización industrial, que permitió flexibilizar la producción, disminuyendo notablemente los costos de producción, incrementado la productividad y mejorando la calidad de los productos, permitiendo así el nacimiento de las celdas de manufactura robotizadas.

Sin embargo, el desarrollo de la robótica está tomando un giro muy importante, ya que se estima que en pocos años, otro tipo de robots, los robots móviles de servicio, superen el mercado de la robótica industrial (Huse, 2011). Los robots móviles son dispositivos de transporte automático, es decir, una plataforma mecánica dotada de un sistema de locomoción capaz de navegar a través de un determinado ambiente de trabajo, dotado de cierto nivel de autonomía para su desplazamiento.

Las aplicaciones de los robots de servicio son variadas, van desde aplicaciones sencillas como robots para jugar hasta aplicaciones que son riesgosas o nocivas para la salud humana, como ejemplos se tiene los robots para el traslado y acopio de materiales, tareas de mantenimiento a reactores nucleares, manipulación de materiales explosivos, exploración subterránea, entre otras. Actualmente la tendencia es aplicar estos robots a tareas de la vida cotidiana, se tienen robots que ayudan a personas discapacitadas, robots que asisten en hospitales para el traslado de medicamentos, realización de cirugías, robots limpiadores, podadores, sólo por mencionar algunas aplicaciones. Esto está provocando que cada vez más, los sistemas robóticos estén presentes en muchas de las actividades que realiza el ser humano por lo que es necesario que se desarrollen interfaces hombre – máquina HMI (*Human Machine Interface*) que permitan una interacción adecuada y confortable con las máquinas con las que interactúa.

En general HMI es el punto de acción en que un hombre entra en contacto con una máquina, el caso más simple es el de un interruptor, No se trata de un humano ni de una "máquina" (la lámpara), sino una interfaz

entre los dos que les permite de alguna forma interactuar entre sí. Para que una HMI sea útil y significativa para las personas, debe estar adaptada a sus requisitos y capacidades. Con el avance de la tecnología estas interfaces son cada vez más sofisticadas y permiten al hombre ir más allá del manejo de la máquina permitiéndole observar el estado del equipo y obtener la mayor información posible sobre la máquina que interviene en el proceso.

Los humanos por naturaleza están acostumbrados a interactuar y comunicarse utilizando su voz. La comunicación es un proceso en el cual se transmite información de una entidad a otra. Los procesos de comunicación son interacciones mediadas por signos o sonidos entre dos agentes o más, los cuales comparten un mismo repertorio y tienen reglas semióticas comunes. La comunicación se define como un intercambio de sentimientos, opiniones o cualquier otro tipo de información como escritura u otro tipo de señales por ejemplo: visuales o auditivas.

Ante el creciente uso de los sistemas robóticos en aplicaciones de servicio y su interacción con los humanos, surge la necesidad de un dispositivo de comunicación eficaz y se discute sobre la conveniencia de que consista en una interfaz natural en el idioma de la persona que lo use y que sea capaz de hacer que la máquina "*hable"* y "*entienda"* de una manera sencilla y eficaz.

Un sintetizador de voz es un dispositivo electrónico basado en una computadora o microcomputadora que convierten un texto escrito a una señal de voz que "dice" lo que el usuario ingresa en forma de texto. Estos sistemas son diferentes a los sistemas que reproducen mensajes que fueron pregrabados ya que el lenguaje de estos últimos está limitado a sólo la información grabada mientras que el sintetizador de voz es capaz de generar y *"decir"* cualquier palabra que se le ingrese, se puede decir que puede reproducir un número infinito de palabras.

El proceso de conversión de texto a voz se puede dividir en tres módulos básicos:

- Procesamiento del texto, aquí se realiza la normalización de texto, es decir expandir abreviaturas, convertir números y fechas en texto, entre otros.

- Generación de la prosodia, genera información prosódica para poder producir la voz, para esto se predicen frases entonativas, la entonación de la oración, duración y energía de los fonemas.

- Generación de voz sintética, produce la voz considerando la información provista por los módulos anteriores.

En este trabajo se propone el diseño inicial de un sistema de síntesis de voz basado en una computadora embebida que permita a una máquina sintetizar señales de voz con la finalidad de utilizarse como interfaz de comunicación de un robot móvil de servicio.

Se propone el uso de la técnica de síntesis de voz por concatenación de unidades, que además de ser una de las técnicas más sencillas para un sintetizador de voz, se obtienen mejores resultados en la calidad de voz.

Este sistema además de permitir una interacción natural entre una máquina y un humano, es posible aplicarlo en diferentes campos, por ejemplo en la ayuda a personas que sean incapaces de comunicarse a través de la voz, para difusión de información escrita en textos, entre otras.

1.2 Planteamiento del Problema

El desarrollo de la electrónica ha traído como consecuencia un incremento de las capacidades de redes de comunicación. Una de las mayores ventajas de tal desarrollo en la actualidad es el incremento de la comprensión de fenómenos de naturaleza física de altísima complejidad, como es la voz humana. Nuestra habilidad para manejar y analizar la enorme cantidad de información que implica cualquier lenguaje humano, se debe en gran medida a la “superación” de las complicaciones técnicas computacionales de capacidad de almacenaje y velocidad de procesamiento. Por ello, resulta ahora factible no sólo el análisis a mayor profundidad de los mecanismos que brindan naturalidad al lenguaje fonético humano, sino, también, el desarrollo de aplicaciones en las cuales podamos “imitar” tales mecanismos a nuestra conveniencia. Es por ello que construir interfaces que permitan una interacción natural entre las máquinas y las personas se hace necesario.

1.3 Objetivo General

Diseñar e implementar un sistema electrónico de síntesis de voz por concatenación de unidades en una computadora embebida, que genere una señal de voz suficientemente entendible y pueda ser aplicado como HMI.

1.4 Objetivos particulares

- Construir una base de datos que contenga las palabras más comunes y todos los difonemas existentes en el español de México.
- Diseñar un programa para el sintetizador de voz en lenguaje de programación C.
- Implantar el código del sintetizador de voz en la tarjeta Raspberry Pi B.
- Dotar la tecnología de Wi-Fi en la tarjeta Raspberry Pi B y generar una conexión remota para interactuar con el sintetizador.

1.5 Justificación

Como se mencionó anteriormente el uso creciente de sistemas robóticos en aplicaciones cotidianas y su inherente interacción con los humanos, genera la necesidad de dispositivos de comunicación eficaz entre el humano y las máquinas, por lo que resulta conveniente el desarrollo de interfaces naturales al ser humano como lo es la voz, con una conexión rápida y de bajo costo. Hoy en día existen sintetizadores de voz como Vozme que es una aplicación que permite el cambio de texto a voz en múltiples idiomas, entre ellos el español y el catalán. Vozme genera un archivo MP3 con el texto traducido a voz y se puede descargar gratuitamente. Otro ejemplo es Text to voice que es un addon para Firefox que permite seleccionar cualquier texto de un sitio web y escucharlo. Sin embargo, estas aplicaciones no están diseñadas específicamente para reproducir audio en español de México ni tampoco para aplicaciones en tiempo real.

1.6 Metodología

Se realizará un estudio sobre la naturaleza y características de la señal de voz, con la finalidad de entender su comportamiento y conocer sus parámetros característicos. Un estudio sobre las técnicas de síntesis de voz concatenativas permitirá conocer su operación e identificar sus principios de diseño que servirán de base para el diseño del sintetizador propuesto en este proyecto.

Se obtendrá una base de datos de audio de la cual se extraerán las unidades fonéticas para el sintetizador por medio de segmentación de unidades. Se creará una base de datos con los audios de las unidades fonéticas segmentadas.

Tendiendo la base de datos de audio se diseñarán programas que permitan un análisis y procesamiento de texto básico, realice la selección de unidades y la concatenación para formar la señal de voz requerida.

El sistema se implementará en una computadora embebida con sistema operativo Linux, se diseñará una interfaz remota que permita enviar texto al sistema vía internet.

1.7 Alcances y Limitaciones.

El sistema propuesto en este proyecto es un primer diseño de un sintetizador de voz que permita una interacción hombre máquina de manera natural e inteligible. El sistema será portátil al estar implementado en una computadora embebida con sistema operativo Linux y podrá ser utilizado en diversas aplicaciones que requieran de una interfaz de voz. Para este primer diseño de sintetizador de voz no se trabajara con la prosodia.

1.8 Organización

En el capítulo 2 se estudiaran las diferentes formas en que las personas se comunican, el lenguaje y cómo la comunicación varía dependiendo de la situación o entorno en el que se encuentra. En el capítulo 3 se explicará la síntesis de voz, su arquitectura y las diferentes técnicas para los sintetizadores de voz. En el capítulo 4 se construirá el corpus de voz para el sintetizador de voz. En el capítulo 5 se diseñara las etapas del programa para el sintetizador de voz. En el capítulo 6 se implementará el programa en la tarjeta Raspberry Pi B. En el capítulo 7 se explicará la tecnología de Wi-Fi y la conexión remota para interactuar con el sintetizador.

Capítulo 2

En este capítulo se estudiarán las diferentes formas en que las personas se comunican y cómo la comunicación varía dependiendo de la situación o entorno en el que se encuentran.

2.1 Lenguaje y Comunicación

La comunicación y el lenguaje son dos términos conceptualmente diferentes. La comunicación es el proceso de transmitir información de un emisor a un receptor a través de un sistema de señales: olfativas, visuales, entre otras y signos muy distintos desarrollados específicamente para comunicarse: vocalizaciones, palabras, gestos.

El lenguaje es la capacidad de comunicación o transmisión de información mediante signos arbitrarios, sonidos verbales o gestos manuales, que tienen una forma convencional y un significado, se combinan siguiendo unas reglas determinadas. El lenguaje es la capacidad específica humana, que se conforma en el conocimiento y uso de las diversas lenguas construidas a lo largo de la historia.

Algunos de los factores para la comunicación son:

- Capacidad de crear información
- Capacidad de transmitir la información
- Capacidad de percibir información creada por otro

2.2 Comunicación

La comunicación es un acto de relación entre dos o más sujetos que tienen un propósito en común, ésta nos lleva a un proceso humano interindividual: el hablante impulsado por estímulos interiores o externos, en un contexto emite varios signos lingüísticos portadores del mensaje al receptor y éste interpreta el mensaje (ver figura 2.1).

Comunicare, en latín, significa poner en común, estar en relación. Fue en el siglo XVI cuando la palabra comunicar adquiere la significación de transmitir. En la actualidad se define como un sistema de conducta integrado, que tiene por efecto ajustar, calibrar y hacer posibles las relaciones humanas, es por tanto el núcleo de la interacción persona (Ancinas, 2004).

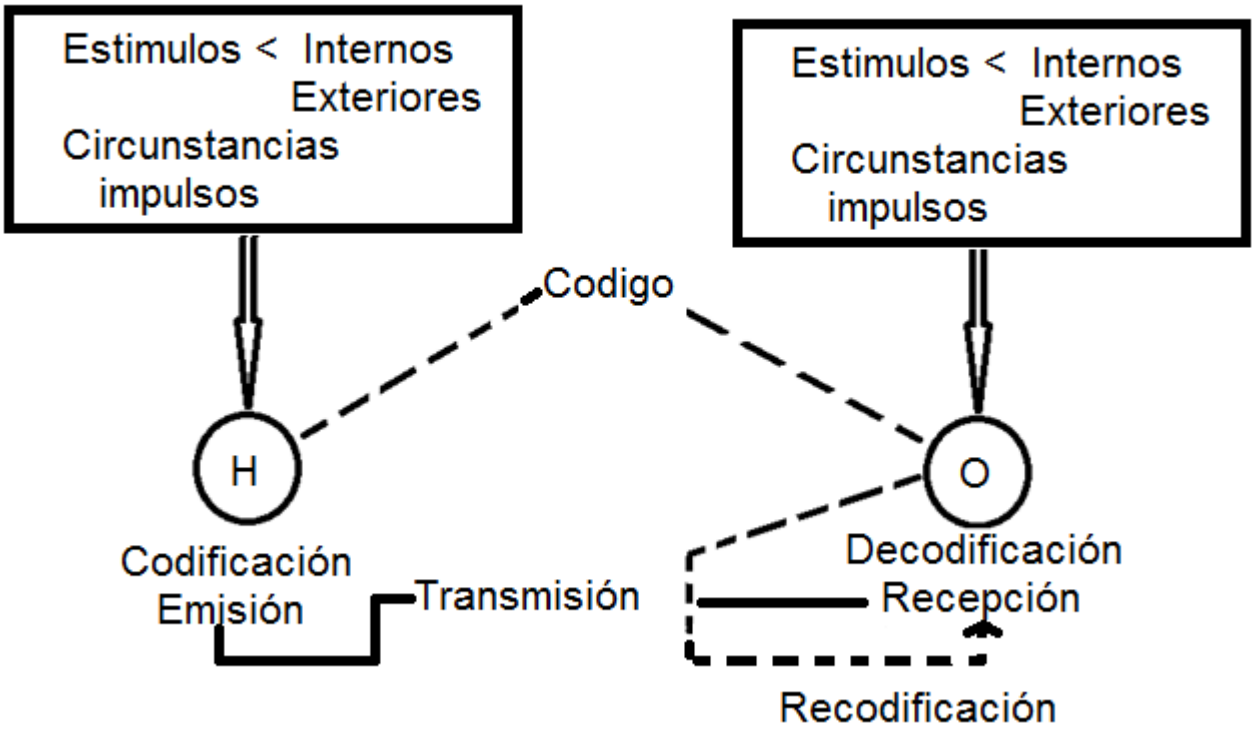


Figura 2.1. Proceso de comunicación (Quilis, 1990).

La teoría de la comunicación de Watzlawick, Beavin y Jackson del equipo de Palo alto en Estados Unidos. Considera la comunicación como una relación cualitativamente diferente de las propiedades de los individuos que participan en ella.

La comunicación humana tiene una fuente, una persona con un objetivo o razón para ponerse en comunicación; el propósito de la fuente tiene que ser expresado en forma de mensaje, mediante una traducción de ideas, propósitos e intenciones en un código. También se necesita de un codificador, este es el encargado de tomar las ideas de la fuente y ponerlas en un código, expresando así el mensaje.

En la comunicación de persona a persona la función de codificar es efectuada por medio de la capacidad motora de la fuente, mecanismos vocales, que producen la palabra hablada, los gritos, las notas musicales, entre otros; los sistemas musculares de las demás partes del cuerpo, que originan los gestos del rostro y ademanes de los brazos o las posturas. Ahora se introduce el mensaje en un canal, el cual será el portador del mensaje. La persona o personas situadas en el otro extremo del canal, es llamada receptor, este necesita

de un decodificador para retraducir, decodificar el mensaje y darle forma que sea utilizable por el receptor. El decodificador de códigos es el conjunto de facultades sensoriales del receptor.

En la figura 2.2 se puede observar el proceso de comunicación muy simplificado, el cual se compone de:

- Hablante: compuesto del cerebro y los órganos articulatorios.
- Oyente: compuesto de los órganos auditivos y el cerebro.

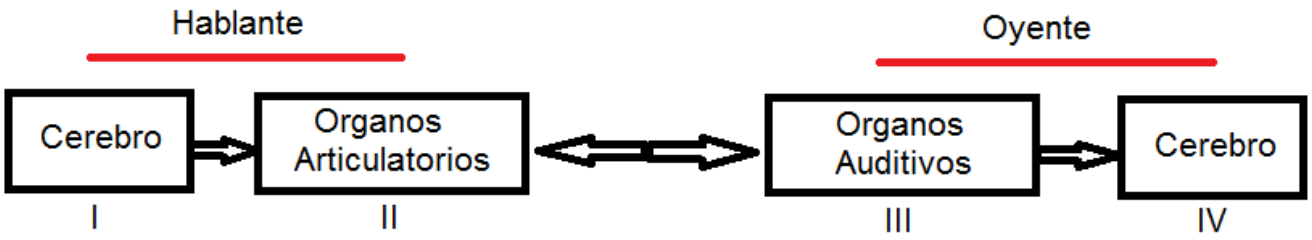


Figura 2.2. Proceso simplificado de comunicación (Quilis, 1990).

El proceso de la comunicación cuenta con un esquema basado en un emisor, en un mensaje que se transmite por un canal y en un receptor al que llega este mensaje. Como se observa en la figura 2.3.

La comunicación oral se divide en cuatro procesos:

1) Generación: es la conversión del significado.
2) Codificación: conversión de la forma en la señal.
3) Decodificación: es la conversión de la señal en forma, por el oyente.
4) Compresión: conversión de la forma en significado, por el oyente.

El emisor transmite una señal de voz al receptor, este recibe la señal de voz, la cual decodifica y crea su propio lenguaje de forma lingüística. La decodificación se realiza tomando una señal para crear el mensaje.

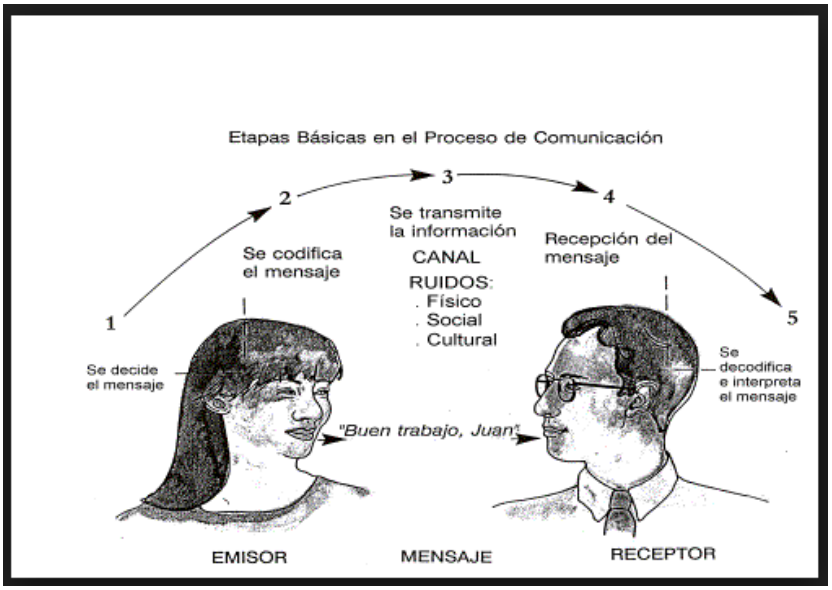


Figura 2.3. Esquema de comunicación (Ancinas, 2004).

2.2.1 Elementos de la comunicación

Existe una serie de factores necesarios para que se lleve a cabo la comunicación, si uno de ellos no aparece, se impide la comunicación o se dificulta. Para que se realice el proceso de la comunicación intervienen los siguientes elementos:

1.- Emisor: Es la persona que posee la información y la transmite.

2.- Mensaje: Información que se desea transmitir

3.- Código: Elementos que se utilizan para transmitir la información, como son el conjunto de símbolos, signos, señales, que se le asignan al mensaje.

4.- Canal: Medio que se usa para transmitir la información

5.- Receptor: Persona que recibe la información o mensaje.

6.- Realimentación: El receptor comunica al emisor que fue recibido el mensaje o no.

7.- Feedback: Designa el conjunto de observaciones, verbales y no verbales que el emisor recibe del receptor sobre los contenidos del mensaje durante el acto de comunicación. Puede adoptar la forma de preguntas, gestos, respuestas, entre otros.

El feedback que recibimos como emisores nos sirve para evaluar nuestra actitud y el impacto de nuestro mensaje en el receptor, gracias a esto se puede corregir sobre la marcha.

8.- Contexto: Es la situación en que se transmite el mensaje y que contribuye a su significado.

La realimentación o mensaje de retorno, no solo puede partir de lo que recibe. El mensaje mismo puede ser fuente de retroalimentación. Algunos problemas que se presentan en el canal son cuantos portadores de mensaje y del mensaje de retorno o feedback, las limitaciones de los canales, los ruidos u obstáculos que evitan que los mensajes lleguen a los destinatarios como era previsto por la fuente.
En la figura 2.4 se muestra un mapa conceptual de los elementos de la comunicación.

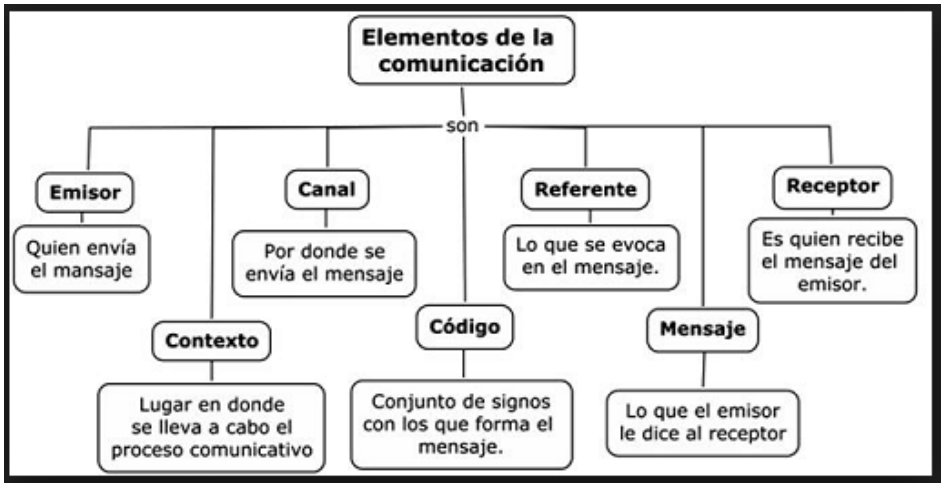


Figura 2.4. Elementos de la comunicación (Ancinas, 2004).

Existen varias clases de comunicación como son:

- Comunicación escrita: Es cuando el emisor produce un texto (obras, cuentos, mensajes, otros.) en papel u otro material y es enviado, el cual llega a varios locutores, los cuales decodificarán para leer el mensaje.
- Comunicación visual: En esta predominan las imágenes en la construcción del mensaje, aunque también se completa con textos, sonidos, locuciones que acotan y precisan su sentido o significado.
- Comunicación oral: Es el dialogo entre dos o más personas, a través de un código lingüístico, lenguaje articulado, sonidos estructurados que dan lugar a las silabas, palabras u oraciones con las que nos comunicamos con las demás tienen como medio de transmisión el aire y como código un idioma.
- Comunicación por gestual: Es un lenguaje corporal o no verbal y sirve para contradecir, complementar o reforzar tanto la comunicación verbal como escrita, proporcionando señales informativas.
- Comunicación de masa: Es un proceso comunicativo intencionado, que emplea a medios tecnológicos que trascienden tiempo y espacio pretendiendo causar impacto global.

2.2.2 Comunicación verbal

La comunicación verbal se refiere a las palabras que utilizamos y a las inflexiones de nuestra voz, es decir, el tono de voz. Esta comunicación se puede realizar de dos formas.

1) Oral: a través de signos orales y palabras habladas.
2) Escrita: por medio de la representación gráfica de signos.

En la comunicación verbal existen propiedades del lenguaje verbal como:

- Arbitrariedad: es cuando se tiene significados diferentes pero los sonidos son similares.
- Dualidad: combinación de fonemas y palabras
- Productividad: número limitado de cosas que se pueden decir en el lenguaje verbal.
- Discreción: los fonemas son distintos uno de otro y forman un conjunto finito y fijo.

La comunicación verbal es un proceso comunicativo que se da mediante el uso de símbolos hablados o escritos que pueden ser comprendidos por las personas. Estas intercambian información y respuestas en un continuo feedback. En esta interacción influye el contexto, por ejemplo la hora, lugar, normas sociales establecidas, el porqué del encuentro; y las características culturales e individuales de ambos.
La forma de comunicación escrita son los ideogramas, jeroglíficos, alfabetos, siglas, grafiti, logotipos, entre otros. Para interpretar correctamente los mensajes escritos es necesario conocer el código, que ha de ser común al emisor y al receptor.

2.3 Lenguaje, Fonética y Fonología

La fonética y la fonología son dos disciplinas de la lingüística encargadas de estudiar los sonidos del lenguaje y configuran el ámbito fónico de una lengua, de modo que ambas disciplinas resultan imprescindibles en el proceso de la comunicación (Hidalgo, 2012).
La fonología parte del fonema, estudia la función y estructura en el sistema de comunicación lingüística y la fonética parte de los alófonos o sonidos, estudia la producción, de su constitución acústica y de su percepción.
Las técnicas de síntesis de voz tradicionales se basan en la fonética y la fonología, estas desempeñan un papel vital en la determinación de cómo aplicar mejor las representaciones y algoritmos.
Para la producción del habla se examina el proceso del ser humano, cómo convierte la lingüística en un mensaje de voz. El proceso inverso, donde los seres humanos determinan el mensaje de voz, se le llama percepción del habla, en conjunto estos constituyen la columna vertebral del campo conocido como fonética.

El ser humano genera un discurso por el uso coordinado de diversos articuladores atómicos, órganos vocales, y la mayoría de los sonidos son creados por el aire en movimiento desde los pulmones y atraviesa los resonadores: la laringe, la cavidad bucal, las fosas nasales.
Las cuerdas vocales son dos repliegues de tejido, los cuales se extienden a través de la laringe. Un hablante puede controlar la tensión en sus cuerdas vocales, de modo que puedan estar completamente cerrados, estrechos o abiertos, todas las vocales y consonantes son sonidos sonoros. El tracto vocal es el término colectivo dado a la laringe, la cavidad oral y la cavidad nasal.

Lenguaje proviene del vocablo latino “lingua” que señala el órgano lingual, contenido en la boca, se relaciona con el fenómeno locutorio. Históricamente la palabra lenguaje se usa por primera vez en el siglo XIII por Gonzalo de Berceo, para significar el conjunto de sonidos articulados con los que el hombre expresa su pensar y sentimientos. En el siglo XVIII el lenguaje ya fue interpretado como idioma hablado por una nación o pueblo.

El lenguaje es un fenómeno esencialmente humano, facultad y actividad del hombre que le permite comunicarse a través de un sistema de signos verbales. Pero, además de ser medio de comunicación, es un fenómeno social, histórico y cultural, instrumento de la transmisión de ideas y vivencias, así como soporte del pensamiento y crisol de una cultura (Quilis, "linguistica española aplicada a la terapia del lenguaje", 1990).

El lenguaje es uno de los principales medios de expresión para el hombre, en el cual este expresa sus pensamientos, sentimientos y la forma de la organización de su mente. El lenguaje es tan solo uno de los códigos que utilizamos para expresar nuestras ideas.

El procedimiento de la creatividad lingüística es un número reducido de unidades mínimas no significativas como son los fonemas, estos se combinan y permutan de determinados modos, sometidos a unas normas, extensión y restricciones concretas y fijas, formando así unidades de orden inmediatamente superior, los morfemas ya comportan un significado de una unidad contigua. Las combinaciones de un lexema, unidad mínima portadora de la base significativa, más un morfema, mínima forma de expresión que delimita la significación del lexema, respondiendo a una categoría, suele ser la estructura de lo que llamamos palabra. Si bien hay palabras que constan de un solo morfema o solo lexema o de varios lexemas y un morfema.

El funcionamiento de una lengua al menos de su código segmental, responde a un múltiple sistema combinatorio y de permutaciones, articulatorio de pocas unidades y de las producidas e admitidas sucesivamente en cada proceso, capaces de permitir a un hablante de dicha lengua producir infinitos mensajes o realizaciones lingüísticas a partir de aquel número reducido de unidades mínimas.

2.4 Funciones del lenguaje

El lenguaje es un hecho social, de intercomunicación humana, que permite plasmar las vivencias del hablante. El significado habitual del término “función”, del latín function-onem, se entiende como un ejercicio o actividad determinada y normal de un órgano o aparato del ser humano con sentido, orden y finalidad concreta.

A continuación se describirán los tipos de funciones del lenguaje:

- En la función referencial se asigna un significante a un significado y de ella nos servimos para relacionar cosas e ideas.
- La función emotiva pone de relieve al emisor del mensaje, manifiesta una intensa afectividad y exige, generalmente, una comunicación directa, actualizada.
- La función apelativa o conativa trata de ganar la atención e interés del interlocutor, de impresionarlo y condicionarlo.
- La función metalingüística es la que se centra en el código, explica el lenguaje con el propio lenguaje. Es una variante de la referencial o representativa. La única diferencia consiste en que utiliza como referentes los conceptos, fenómenos o relaciones lingüísticas.
- La función fática tiene como objetivo establecer, interrumpir, prolongar o dar una transición al mensaje, a la comunicación. Lo cual garantiza la comunicación.
- La función poética, está orientada al mensaje, aparece siempre que la expresión atrae la atención sobre su forma, en cualquier manifestación en la que se utilice el lenguaje con propósito estético (Frías, 2000).

En la figura 2.5 se muestra el esquema de las funciones del lenguaje.

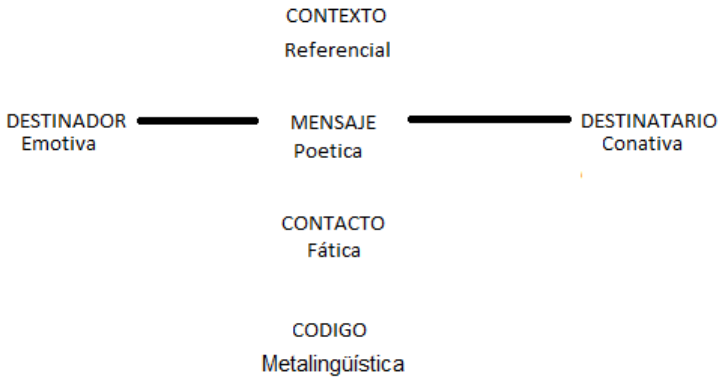


Figura 2.5. Funciones del lenguaje (Quilis, 1990).

La comunicación humana necesita el signo lingüístico, este se conforma de significado y significante. En el signo hay dos planos: la expresión y el contenido, y cada uno de estos consta de una sustancia y una forma.

El conjunto de forma de expresión más la forma de contenido, será el signo. La forma de expresión más la forma de contenido en una función de una estructura, eso es un signo lingüístico. El conjunto de signos con idéntica sustancia forman un código.

En la lengua funcionan varios códigos unidos simbiótica o parásitamente:

- El código segmental o el de las palabras.
- El código suprasegmental (tonos, acentos, tiempo, entre otros).
- El código concomitante (gestos).

La relación entre significado (forma de contenido) y significante (forma de expresión) es convencional. La convencionalidad es la esencia del signo, cuya forma de expresión está parcialmente motivada; los derivados de onomatopeyas.

En el código segmental existen tres conjuntos de unidades:

- Conjunto de unidades con forma de contenido léxico: son los que tienen un significado propiamente dicho.
- Conjunto de pocos elementos que sirven para señalar o indicar los deícticos, que carecen de significado.
- Conjunto de unidades que no significan y solo sirven para establecer algunas relaciones gramaticales.

La lengua se desarrolla en cadenas lineales, es decir, la forma de expresión de la lengua, de carácter articulatorio y acústico, se desarrolla en tiempo unidimensional. La lengua tiene en realidad una manifestación oral; la lengua escrita es como una fotografía que tiene otra sustancia y otra forma de expresión, la lengua hablada solo acomoda a significantes gráficos visuales y convencionales.
Los signos se relacionan entre sí de dos maneras: sintagmáticamente (en cadena) y paradigmáticamente, en su microsistema (paradigma y sintagma).

El lenguaje es imaginado un sistema complejo de relaciones entre los elementos o un sistema de relaciones entre sí. Cada sistema constará de una serie de unidades interrelacionadas que forman un conjunto. La estructura es la red de relaciones de unos huecos funcionales que forman una cadena lingüística gramatical y aceptable en una lengua funcionando dentro de un sistema.

La distribución es el conjunto de los entornos y contextos en que puede aparecer la unidad lingüística. Es evidente que toda unidad lingüística se encuentra en mayor o menor grado condicionada y limitada por los contextos en que puede aparecer. Cuando dos unidades aparecen en la misma clase de contextos decimos que son distribucionalmente equivalentes.

Existen cuatro tipos de correlación distributiva entre las unidades lingüísticas de cualquier nivel:

1) Distribución equivalente: cuando dos elementos tienen el mismo entorno o contexto.
2) Distribución complementaria: cuando las unidades no aparecen en el mismo contexto o entorno.
3) Distribución incluyente: cuando la distribución de un elemento puede incluir la distribución de otro.
4) Distribución por intersección: cuando dos o más unidades pueden aparecer en unos mismos contextos, pero no en todos.

2.4.1 Fonética

Anteriormente se ha mencionado cómo los órganos vocales pueden organizarse para producir una amplia variedad de sonidos del habla. Donde el lenguaje natural es un sistema semiótico simbólico discreto. En el sistema se combinan palabras de varias formas para producir un número casi limitado de frases cada una con un significado distinto.

La fonética es el estudio de los sonidos del habla, incluye la clasificación sistemática de los sonidos de acuerdo a la forma en que se producen y cómo lo escucha el oyente. El fonetista se especializa en la fonética para entender la clasificación sistemática de los sonidos del habla de los diferentes idiomas del mundo (Martinez, 1995).

La fonética clásica estaba basada sobre la observación de los movimientos articulatorios las distintas impresiones auditivas que ellos producían, los fonetistas se apoyan en su habilidad para reconocer, al oírlos, los diversos tipos de sonido, también se apoyan de métodos no tan sofisticados como el método de palatografía. Este sistema consiste en introducir en la boca un paladar artificial con nueve filas de electrodos: tres en la zona alveolar, dos en la palatal y tres más en la velar. Este paladar artificial detecta los contactos de la lengua y envía las señales a un programa de computadora que recoge a la misma vez la onda osciloscópica de lo pronunciado, y la sucesión de contactos a lo largo del tiempo en todo el paladar (Martinez, 1995).

La fonética es el estudio de los sonidos utilizados en las lenguas naturales, y considerando que los sonidos del habla constituyen elemento primordial de nuestro sistema de comunicación, no ha de parecer extraño que al establecer divisiones en el interior de la fonética se tenga en cuenta la naturaleza vocal y auditiva del lenguaje.

Las ramas en las que tradicionalmente se divide la fonética son:

- Fonética articulatoria: también llamada fonética fisiológica, se centra en la clasificación estática de los sonidos en función de las diversas partes del aparato fonador que intervienen en su producción.
- Fonética acústica: se basa en las propiedades físicas de los sonidos del habla, considerándolos como ondas sonoras. Estas propiedades se derivan del modo en que se producen los movimientos que tienen lugar en el tracto vocal, donde permite tanto su transmisión como su percepción. El interés de esta fonética surgió más con las técnicas de síntesis y reconocimiento de voz.

- Fonética perceptiva: se encarga de la audición del habla, examinando como las ondas sonoras llegan hasta el oído y se transforman en impulsos nerviosos que se transmiten al cerebro. También estudia la interpretación que se le da a estos impulsos, percepción.

La fonética estudia los sonidos del habla, excluyendo todos aquellos fenómenos sonoros que no forman parte del repertorio presente en las lenguas naturales. Se diferencia la fonética general y la fonética de las lenguas particulares.

- Fonética general: su objetivo es la caracterización de los mecanismos que intervienen en la producción y la percepción de los sonidos de las lenguas naturales.
- Fonética de las lenguas particulares o fonética descriptiva: describe como se utilizan estos mecanismos en las distintas lenguas que conocemos.

La prosodia tiene un dominio muy amplio que comprende el estudio de diversos fenómenos asociados al acento, al ritmo y la entonación, así como a sus manifestaciones físicas producto de las variaciones de la duración, de la frecuencia fundamental y de la intensidad. La prosodia se puede definir como un estudio de los hechos fónicos no segméntales que contribuyen a organizar tanto el léxico como la sintaxis. Estos hechos fónicos tienen una función específica en la interpretación semántica de los enunciados y del discurso.

2.4.2 Fonología

La fonología estudia los sonidos del lenguaje desde el punto de vista de su función en el sistema comunicativo. Los sonidos adquieren valores distintos según la función que ocupen en un contexto. Los sonidos que componen una palabra son las unidades mínimas que la hacen diferente de otra palabra, estos son los fonemas.

2.4.3. Fonema

La unidad más pequeña en que se puede dividir un conjunto fónico recibe el nombre de “fonema” (una sola letra). El fonema no tiene en sí mismo significado, pero hace cambiar el significado de una palabra cuando es conmutado por otro. Un fonema puede tener diferentes realizaciones fonéticas, de acuerdo con el contorno que este situado.

El alófono o sonido es la realización, la materialización del fonema en un hablante, en un momento determinado. Ni tiene significado, ni cambian el significado de las palabras al ser conmutados entre ellos.

La grafía o letra es la representación, más o menos afortunada, de un fonema en la escritura. Se debe advertir que los fonemas siempre se representan por barras oblicuas (//), y los alófonos o sonidos entre corchetes ([]).

Para proceder a la identificación de los fonemas de una lengua es necesario emplear el procedimiento de la conmutación sucesiva, es decir sustituir cada uno de los fonemas de una palabra por otros, con el fin de encontrar diferencias en su significado.

La relación que existe entre dos fonemas conmutables recibe el nombre de oposición. Teóricamente sería necesario conmutar todos los fonemas de una lengua para realizar su inventario fonológico, pero en la práctica es suficiente con conmutar los fonemas que ofrecen características similares.

2.5 El español de México

A finales del siglo XV, con la unión de las monarquías de Castilla y Aragón, las cuales extendieron su dominio por gran parte de la península, el castellano se impuso sobre los demás idiomas y dilectos. La colonización española del XVI llevo la legua a alas Américas, a los estados federales de Micronesia, Guam, Marianas, Palaos y Filipinas. El dialecto de castilla o español de castilla, fue poco a poco convirtiéndose en la lengua estándar, por el dominio político de castilla en el siglo XIII. La mayor parte de las palabras del español derivan del latín, pero también tenemos unas que se derivan de otras lenguas pre latinas, como: euskera o celta.

La real academia de la lengua española se fundó en 1713. Establecía los criterios para castigar los neologismos y para la incorporación de palabras de ámbito internacional. La gramática española se normalizó en este periodo y la literatura española fue muy prolífica.

El siglo XX ha sido testigo de cómo ha ido cambiando el uso del español, en la lengua han entrado multitud de neologismos, alimentados por los avances tecnológicos y científicos.

En los siglos XV al XVII, España formó un gran imperio abarcando diversas partes del mundo, lo cual hizo que el idioma español adquiriera gran importancia teniendo un mayor número de hablantes. Hay una pequeña diferencia entre el español que se habla en España, que el que es hablado en otros países. Este idioma ha ido adquiriendo diferentes características como las fonológicas, que se refiere a la pronunciación específica de cada región, por ejemplos las diferentes entonaciones de un argentino a un norteño, y las léxicas que se refieren al vocabulario.

El español de México tienen vocablos específicos, como los indigenismos o los nahualismos que son palabras de origen náhuatl, también hay otros vocablos por su entonación y pronunciación.

El español hablado en México no es homogéneo, tiene modalidades distintas según la región en donde se habla. La pronunciación de las lenguas no está reflejado de manera coherente en la grafía, ya que no corresponde al mismo sonido en el español, también ocurre que dos grafías correspondan al mismo sonido, como se describe en las siguientes características propias del español de México.

Características propias del español de México son las siguientes:

- Se utiliza “siempre” con el significado de “a fin de cuentas” o “definitivamente”
- Fonéticamente la “c” y la “z” se pronuncian igual que las “s”. la “y” y la “ll” se pronuncian igual.
- Se conservan algunos arcaísmos como: mucho muy rico, mucho muy grande, mucho más bonita, entre otros.
- Se utiliza el vocablo disque como significado de duda o negociación.
- La “b” es el mismo sonio de “v”
- La “j” es el mismo sonido de “g”

Capítulo 3

En este capítulo se explicará la síntesis de voz, su arquitectura y las diferentes técnicas para los sintetizadores de voz.

3.1 Síntesis de voz

La comunidad científica ha buscado la manera de encontrar una máquina capaz de hablar, buscando un sistema que emule de la mejor manera posible el aparato fonador del humano. Un sistema de síntesis de voz es un sistema que simula el proceso humano de leer en voz alta, es decir, convierte una entrada de texto a una salida hablada.

Las tecnologías del habla son un conjunto de conocimientos y técnicas de procesamiento de la señal de voz, su objetivo es aumentar la capacidad de comunicación, permitiendo el control y acceso a la información de computadoras, comunicación entre personas que tienen un diferente idioma, mejorar la comunicación en los ambientes ruidosos, etc.

La señal de voz desde el punto de vista acústico consiste en una onda sonora que se propaga en la misma dirección de la vibración. El origen de esta onda se genera en una corriente de aire, procedente de los pulmones y modulada por los órganos de la laringe y del tracto vocal. La señal de excitación (pasa a través del tracto vocal) y va a determinar principalmente las características personales de la voz, como son:

- Entonación (tono), transmite información sobre el estado de ánimo, intenciones del locutor y sobre la estructura del mensaje.
- La estructura de formantes (envolvente espectral), va a transmitir fundamentalmente información sobre la naturaleza de los alófonos.

Los sintetizadores de voz en general, se componen de dos módulos, los cuales interactúan para realizar la síntesis de voz como se muestra en la figura 3.1. La función del módulo 1 es recibir el texto y separarlo en segmentos, la función del módulo 2 es convertir los segmentos a sonidos, generando una voz artificial, el cual va interpretar el texto de entrada.

Figura 3.1. Sistema de síntesis de voz (Gaspar, 2006).

Un sistema ideal de síntesis de voz debe ofrecer:

- Calidad elevada de la voz sintética (tanto la naturalidad como la inteligibilidad de la voz).
- Ser capaz de reproducir cualquier mensaje.
- Conseguir una voz agradable para el usuario
- Permitir transmitir emociones con el discurso
- Procesamiento de voz simple.

El proceso de conversión texto a voz otorga a las máquinas la capacidad de producir mensajes orales no grabados previamente. Tomando como entrada un texto, los sistemas de conversión texto a voz realizan el proceso de lectura de forma clara e inteligible, con una voz más natural y humana posible. La síntesis de voz conforma la interfaz oral de la comunicación entre una máquina y el usuario de la misma. La síntesis de voz es un instrumento diseñado para poder aislar y combinar rasgos acústicos sustituyendo unos por otros y comprobando en qué medida estos cambios afectan a la respuesta del oyente, de esta forma se introducen variaciones controladas, lo que en el lenguaje natural no es posible.

Existen para los sintetizadores de voz de forma general dos preocupaciones técnicas principales que son:

- La conservación del contenido del mensaje en la señal de voz.
- La representación de la señal de voz en una forma que sea conveniente para la transmisión y almacenamiento, o de forma que sea flexible para poder hacer modificaciones a la señal de voz, sin perder el contenido del mensaje.

3.2 Arquitectura de un sintetizador de voz

Como se había mencionado anteriormente la conversión de texto-voz es la generación a través de medios automáticos de una secuencia de sonidos, que produce una persona al leer un texto en voz alta. El objetivo de la conversión de texto a voz es que se pueda lograr una voz inteligible y natural. Debe quedar claro que la conversión de texto-voz no es síntesis de voz a partir de conceptos, es decir, la conversión texto-voz siempre trabaja a partir de un texto previamente escrito, no incluye la capacidad de generar el texto respondiendo a condiciones variables (Montiel, 2006)

El proceso de síntesis de voz se compone generalmente de dos bloques (Furui, 1989):

1) Procesamiento del lenguaje natural: en este bloque a partir del texto de entrada se genera una descripción fonética, fonemas que intervienen en cada una de las palabras del texto de entrada cuando se pronuncian y prosódica que establece el ritmo adecuado que se le dará a la voz artificial de salida, entonación, emoción, entre otros. Este bloque consta de tres módulos:
   - Analizador de texto: la función de este módulo es tomar el texto de entrada y darle el formato adecuado para ser entendido por el convertidor de texto a fonemas.
   - Convertidor de texto a fonemas: la función de este módulo es asignar la pronunciación adecuada a cada palabra para poder generar la señal de voz
   - Generador prosódico: la función de este módulo es asignar la duración y entonación adecuada.

2) Procesos de síntesis: en este bloque se transforma la información del bloque anterior en una voz de salida. Produciendo la voz sintetizada o artificial y también es generada la pronunciación del texto. En este proceso existen dos enfoques:
   - En el primero se permite modelar el mecanismo de producción de la voz. En este enfoque están los sintetizadores articulatorios y los sintetizadores por formantes.
   - En el segundo se intenta modelar la señal de voz. En este enfoque están los sintetizadores por concatenación de unidades.

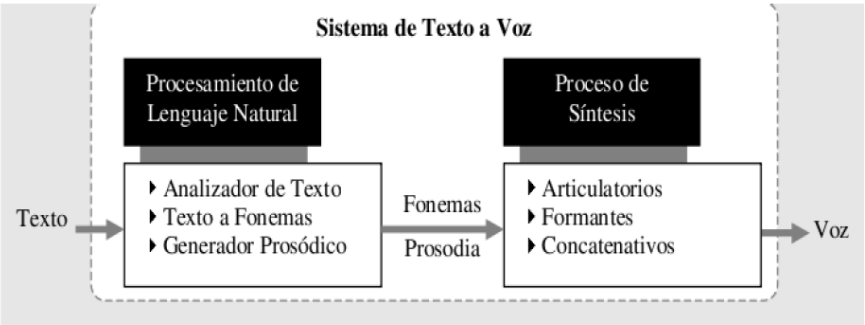


Figura 3.2. Síntesis de voz (Furui, 1989).

Hay una gran variedad de síntesis de voz con una complejidad variada, pero la mayoría de estos sistemas comparten la misma arquitectura, como se muestra en la figura 3.2.

Existen diversas técnicas que se han desarrollado para la generación de habla artificial:

- Sintetizadores articulatorios
- Sintetizadores por formantes
- Sintetizadores derivados de las técnicas de predicción lineal (LPC)
- Sintetizadores de concatenación de unidades.

3.3 Sintetizadores Articulatorios

El propósito de este sintetizador es controlar un modelo del aparato fonador de una forma similar al que nuestro cerebro lo hace, es decir, intentan modelar de manera amplia los movimientos mecánicos de los articuladores y las distribuciones resultantes del volumen de velocidad y la presión del sonido en los pulmones, laringe y tractos tanto nasal como vocal. En éstos se realiza una analogía entre parámetros relativos a los órganos articulatorios y sus movimientos con parámetros circulares, los cuales pueden proporcionar una calidad alta pero es difícil obtener y controlar parámetros para un sintetizador de este tipo.

Este tipo de sintetizadores utilizan los siguientes parámetros:

- tamaño de la cavidad oral
- tráquea
- posición de la lengua
- otras variables.

Estos factores se relacionan entre sí para producir una voz muy similar a la voz humana, la cual aplica señales armónicas a la señal sonora y establece una analogía entre parámetros relacionados con los órganos articulatorios, sus movimientos y parámetros. Los sintetizadores articulatorios proporcionan voz sintética de alta calidad, pero tiene un inconveniente, el cual es que sus parámetros son muy difíciles de obtener y controlar automáticamente.

El problema principal de los modelos articulatorios es la enorme cantidad de parámetros internos de control que precisan y dificultan la coordinación y derivación de los parámetros internos de control disponibles a la entrada del sintetizador. Por otro lado, también se presenta la gran cantidad de información que se necesita obtener analizando, en un espacio tridimensional, la posición y el movimiento de los órganos articulatorios de una persona que habla normalmente.

3.4 Sintetizadores por formantes

En este tipo de sintetizador se realiza una simplificación del aparato fonador, el cual consiste en realizar un filtrado de una señal de excitación introduciendo resonancias, ya que los alófonos se distinguen por el tipo de excitación, por los valores de la frecuencia y ancho de banda de sus resonancias. La síntesis se puede hacer mediante varios filtros de segundo orden, conectados en serie o paralelo. Una de las ventajas de este modelo es que son de bajo costo y que no ocupan mucha memoria.

Estos sintetizadores se basan en la teoría acústica de producción de voz ya que es posible ver la voz como resultado de la excitación de un filtro lineal por una o más fuentes sonoras (Vega, 2007). Son una serie de filtros los que modelan el tracto vocal, excitados por fuentes que simulan las cuerdas vocales, estos gozan de gran difusión y se ilustra en la figura 3.3.

Figura 3.3. Sintetizador por formantes (Vega, 2007).

Las fuentes primarias del sonido son tonos, los cuales se producen por la vibración de las cuerdas vocales y ruido turbulento causado por la diferencia de presión a través de una construcción.

Existe una gran cantidad de configuraciones de sintetizadores propuestos, de las cuales solo dos son las más sobresalientes.

1) Sintetizador de formantes en paralelo, los resonadores por formantes que estimulan la función de transferencia del tracto vocal se encuentran conectados en paralelo. Cada resonador de formantes es predicho por un control de amplitud, la cual determina la amplitud relativa pico espectral (formante), en el espectro de salida de los sonidos de voz vocalizados como no vocalizados.
2) Sintetizador de formantes en cascada, los sonidos son sintetizados utilizando un conjunto de resonadores de formantes en cascada, ambas configuraciones se ilustran en la figura 3.4.

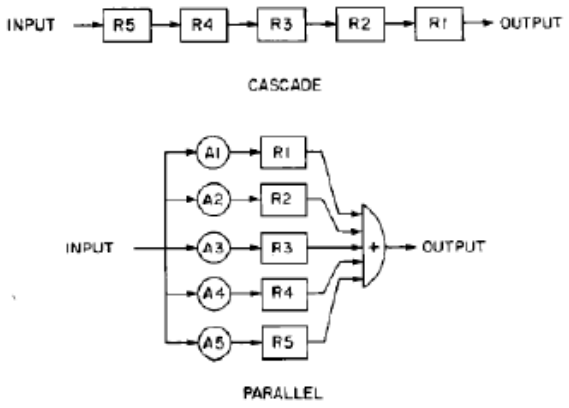


Figura 3.4. Sintetizadores en paralelo y cascada (Klatt, 1980).

La función de transferencia del tracto se puede simular por un conjunto de resonadores digitales R conectados en cascada (la salida de uno conecta al siguiente), o por un conjunto de resonadores conectados en paralelo. La ventaja de la conexión en cascada es que la amplitud relativa de los picos de los formantes para las vocales está perfectamente ajustada, sin la necesidad para los controles en amplitud individuales para cada formante, otra ventaja de este modelo es que es un modelo más adecuado de la función de transferencia del tracto vocal durante la producción de sonidos no nasales. En ciertas ocasiones, las funciones de transferencias de ciertas vocales son difíciles de realizar al utilizar un sintetizador de formantes en paralelo, sin embargo es útil para la generación de estímulos que violan las relaciones de amplitud normal entre los formantes, ya que la mayor parte de la energía de la señal de voz se encuentra contenida en frecuencias comprendidas entre los 80 y 800 Hz.

Universidad Autónoma de la Ciudad de México

Nada humano me es ajeno

3.5 Sintetizadores LPC (Linear Predictive Coding o codificación predictiva lineal)

Son sintetizadores de análisis-síntesis, en los que los parámetros para controlan la función de transferencia del filtro que simula el tracto vocal. LPC es una técnica de mayor alcance de análisis del habla y uno de los métodos más útiles para codificar el habla. La codificación por predicción lineal es un procedimiento matemático que permite predecir los sucesos futuros de un sistema lineal.

El objetivo principal de este tipo de sintetizador es modelar el tracto vocal, con una serie de cilindros huecos de diámetro variable. En el sistema de síntesis por LPC se siguen los siguientes pasos:

1) Elección de las unidades (fonemas, difonemas, palabras) que se deseen utilizar
2) Codificación de las unidades
3) Almacenamiento en una ROM de los parámetros obtenidos
4) Decodificación por medio de un sintetizador LPC cuando debe producirse un mensaje.

La predicción lineal es un procedimiento que dada una señal de habla, permite definir la función de transferencia del filtro que la ha generado. Basado en la redundancia de las señales de habla, periodicidad y variación relativamente lenta, que permite la predicción de la señal muestreada a partir de muestras anteriores. El modelo de análisis por LPC es una predicción de una señal periódica a partir de valores anteriores en un sistema sin separaciones, los parámetros del análisis LPC son: orden del modelo que son los números de picos o polos a partir de los cuales se modela la función de transferencia de la señal analizada, también se relaciona el ancho de banda la frecuencia de muestreo de la señal y el ancho de banda de la ventana de análisis, la cual contienen el número de muestras de la trama, superposición de ventanas para evitar la pérdida de información en las transiciones. Las posibilidades del LPC son análisis que es la detección de formantes a partir de los picos o polos del espectro y determinación del envolvente espectral y la síntesis que es la codificación de unidades de síntesis para la conversión de texto a habla, codificación de frases para sistemas de respuesta vocal.

3.6 Sintetizadores por concatenación de unidades

Este sintetizador se basa en la obtención un conjunto de pequeños segmentos de voz tomados de un locutor, los cuales se van concatenando en base a algoritmos computacionales de selección para formar nuevas palabras y oraciones.

El tipo de unidad a concatenar es un parámetro crítico para conseguir una buena calidad de voz sintetizada, se llega a un compromiso entre la calidad intersegmental posible (a mayor longitud de los segmentos, menos puntos de concatenación y por lo tanto mayor calidad) y la cantidad de memoria necesaria para almacenar las unidades pregrabadas.

El audio grabado por el locutor no pueden ser solo palabras por dos motivos fundamentales. El primero, la pronunciación de una frase es muy diferente a la de una secuencia de palabras recitadas aisladamente, ya que en una frase las palabras tienen una duración más corta que cuando están aisladas y el ritmo, entonación y sintácticos son totalmente antinaturales cuando se concatenan palabras grabadas aisladamente. El segundo, Innumerables palabras existentes en un idioma.

La síntesis de voz basada en concatenación de unidades tiene como objetivo reproducir el habla con la mayor naturalidad posible e inteligibilidad, reuniendo todos los elementos fonéticos y sonidos necesarios. Para esto se requiere de una base (“corpus”) de datos, grabada por un solo locutor que presente las características acústicas: tono, intensidad, duración, timbre de voz, del cual se van a extraer unidades para formar la voz sintética, la base de datos contiene implícitamente elementos fonéticos más pequeños, como fonemas, difonemas y trifonemas, los cuales se van concatenando con algoritmos computacionales para formar nuevas palabras y oraciones.

Un sintetizador de concatenación de unidades va formando la voz sintética pegando unidades de voz digitalizadas como son (Placio, 2003):

- Fonemas: es una unidad mínima capaz de diferenciar significados en las palabras
- Difonemas: es un trozo de voz que va de la mitad de un fonema a la otra mitad del siguiente fonema.
- Trifonemas: es la unidad constituida por un fonema más la mitad del segmento procedente y la mitad del segmento siguiente. Es mejor para un sintetizador ya que contiene mayor naturalidad, aunque no se pueden formar todas las frases con trifonemas.
- Silabas: incluyen los conceptos de las unidades fonéticas.
- Palabras: es la unidad que proporciona mejor calidad en la síntesis.

La técnica de síntesis por concatenación de unidades en los últimos años ha tenido una mejora de la calidad acústica de la señal sintética, se concatenan segmentos para producir la señal de salida deseada. La prosodia de la señal de salida debe ser calculada de acuerdo a los modelos prosódicos adecuados. Para generar una transición natural entre los segmentos de voz y conseguir la prosodia deseada, para manipular la señal de voz con alta calidad se han utilizado distintos algoritmos como PSOLA, MBROLA, HNM y modelos senoidales.

En la actualidad los sintetizadores por concatenación son los sistemas de mayor calidad, debido a que es posible contar con una gran cantidad de memoria a precio razonable, pero no tienen mucha naturalidad.

3.7 Procesamiento Digital de la Señal

En los últimos años se desarrollaron algoritmos para la generación de la prosodia y la producción de voz natural, esto ha llevado a una mejora de la calidad acústica, la mayor parte de los algoritmos están basados en técnicas de "Overlap-Add" síncronas con la frecuencia fundamental, como el algoritmo PSOLA, existen diversas variantes según se trabaje en el dominio del tiempo (TD-PSOLA), o en el de la frecuencia (FD-PSOLA) (Moulines, 1989) esta tiene una forma más flexible de modificar las características espectrales de la señal de voz, evitando gran parte de distorsión introducida por esta.

El procesado segmental trata de asegurar la continuidad de los parámetros de síntesis, de manera que la entonación, la intensidad y el timbre de los sonidos no se rompa al pasar de una unidad a otra, también la generación de coarticulación cuando no esté incluida en la propias unidades de síntesis, como sucede con los difonemas, ya que generalmente la continuidad de la frecuencia fundamental viene determinada por el patrón entonativo.

3.7.1 Algoritmo PSOLA (Pitch Synchronous Ovelap And Add o solapamiento y suma sincronizada con la frecuencia fundamental).

Es una técnica para el suavizado en la concatenación de segmentos de audio usado para la síntesis del habla, modifica la frecuencia y duración de la señal de habla. El funcionamiento de PSOLA es dividiendo la onda en pequeños segmentos superpuestos, para cambiar la frecuencia de la señal, los segmentos serán acercados o alejados, para cambiar la duración de la señal. En la figura 3.5 se muestra el funcionamiento de PSOLA con una ventana de Hanning.

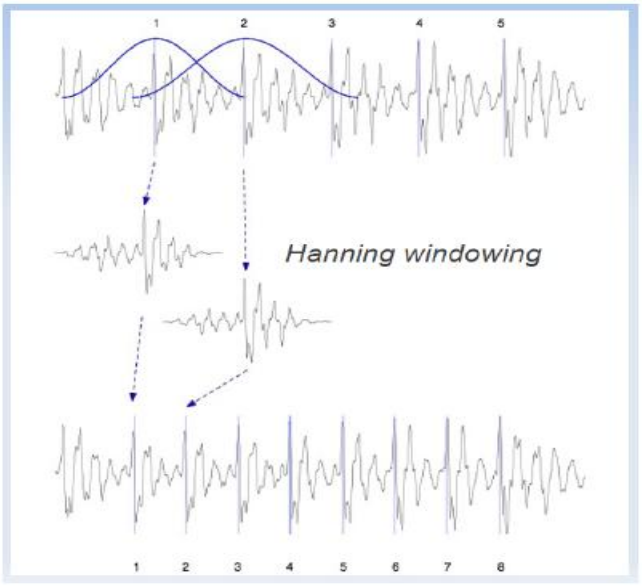


Figura 3.5. Funcionamiento de PSOLA (Bonafonte, 2012).

Este proceso termina suavizando las uniones entre audios, de esta manera también disminuimos los ruidos causados por tres tipos de discontinuidad: de fase, de tono y de espectro (Bonafonte, 2012).

Este método es un Sistema de codificación en el dominio del tiempo, ya que permite modificar la frecuencia fundamental de la señal, el cual fue desarrollado en France Telecom CNET. Actualmente no es un método de síntesis sino que regrabando muestras de voz concatenadas se controla el tono y la duración de la señal.

El funcionamiento del algoritmo PSOLA consta de tres etapas: (Bullón, 1994)

1) Análisis de la onda original, para obtener la representación no paramétrica de la misma, es decir la señal original se descompone en una serie de unidades de corta duración superpuestas denominadas señales ST (Short Term) de análisis, la cual se obtiene multiplicando la señal por una secuencia de ventanas h_m(n) como se muestra en la figura 3.6 y se describe en la ecuación 1.

$$S_m\ n\ = h_m(t_m - n)s(n) \qquad (1)$$

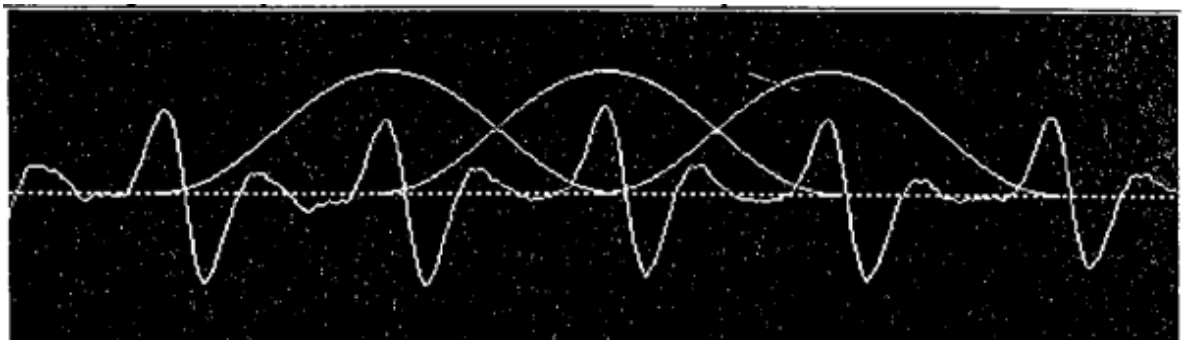


Figura 3.6. Ventana de análisis sobre la onda sonora (Bullón, 1994).

2) Modificación prosódica, a partir de esta representación, las señales de la etapa anterior son procesadas para producir otro conjunto de señales ST de síntesis. Mediante la superposición y suma de estas últimas se genera la onda sintetizada, esta señal de voz sintética puede obtenerse mediante un proceso de solapamiento y suma de señales ST de síntesis. Como se expresa en la ecuación 2.

$$S'(n)= \frac{{}_{q}\alpha q\, S'q\ \ n\ \ h'q(t'q-n)}{{}_{q}h'q^{2}(t'q-n)} \qquad (2)$$

3) Producción de la señal sintética, construida a partir de la representación intermedia modificada. Se asocia cada unidad de ST de síntesis con la de análisis que debe ser copiada en su lugar y los valores de t_q determinan los retardos que deben ser introducidos entre unidades sucesivas y se representa con la ecuación 3.

$$S'_q(n) = S_m(n- t_m+ t'_q) \qquad (3)$$

Si la duración y frecuencia de la señal deben ser modificadas por un factor β, la relación entre las señales ST de análisis y las de síntesis serán de uno a uno, es decir en este caso el algoritmo sólo copiará las unidades de análisis en el tiempo de síntesis, ajustando el retardo entre ellas según el factor.
En las figuras 3.8 y 3.9 se observa el efecto que se produce en la onda sintetizada la variación del grado de solapamiento de las señales ST de análisis, en la figura 3.7 se muestra la onda original.

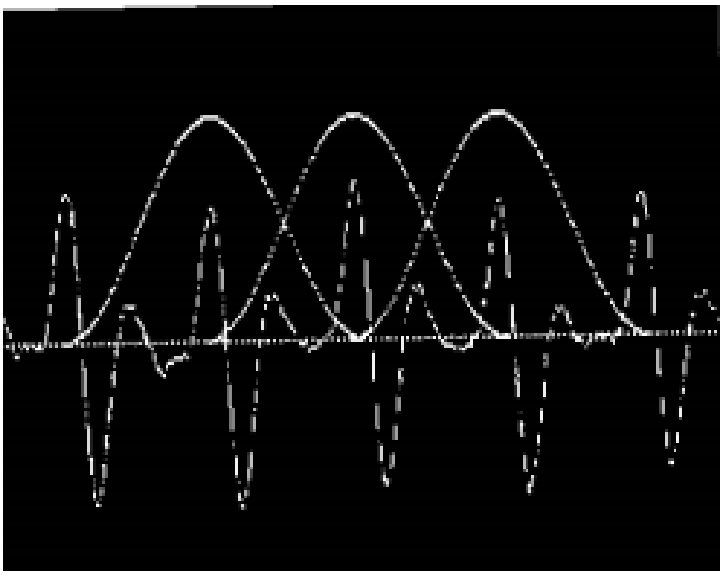


Figura 3.7.señal original (Bullón, 1994).

En la figura 3.8 se observa que se ha producido una separación de las ventanas, originando una disminución de la frecuencia en la señal sintetizada.

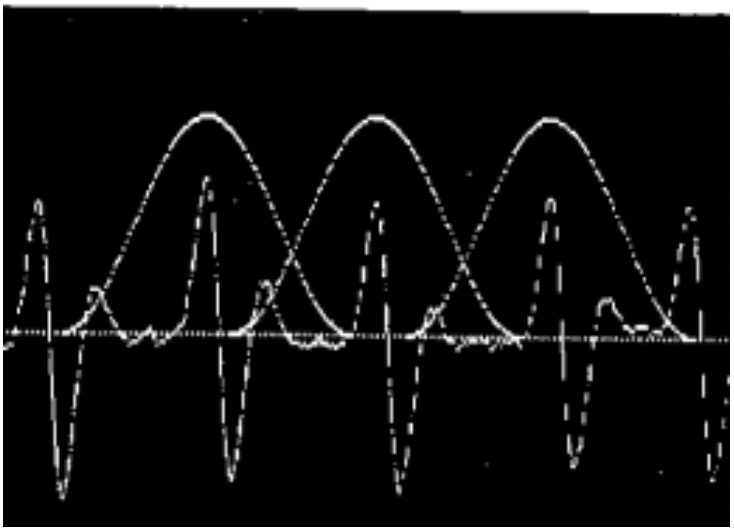


Figura 3.8. Frecuencia baja (Bullón, 1994)

En la figura 3.9 se muestra el proceso inverso, es decir se aproximan las ventanas con lo que se aumenta la frecuencia pero se disminuye la duración.

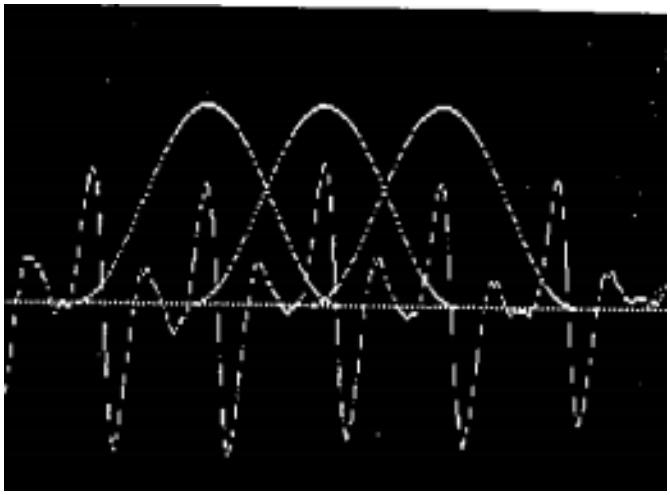


Figura 3.9. Frecuencia alta (Bullón, 1994).

Este algoritmo es específico para la síntesis por concatenación de unidades de la señal, el cual permite modificar arbitrariamente la frecuencia fundamental y la duración de los segmentos sin necesidad de parametrizar la señal, esto ha permitido dar flexibilidad a los sistemas de concatenación de formas de onda, donde el principal inconveniente era la modificación de la prosodia. Es decir consiste en la separación de las distintas unidades y superponiendo las unidades para que no tengan saltos cuantitativos a la hora de la unión. Se utiliza para aumentar la naturalidad de la voz sintética.

3.7.2 Algoritmo MBROLA (Multi-Band Resynthesis Overlap and add, solapamiento y adición de resíntesis de multibanda)

Este algoritmo es una herramienta de conversión de texto en habla, basado en la concatenación de difonemas y tiene una alta calidad, este toma como entrada una lista de fonemas junto con la información prosódica (duración de los fonemas y una descripción del tono de los mismos) y produce una salida de audio a 16 bits, por lo tanto no es un conversor texto-voz dado que no acepta texto como entrada, este algoritmo fue desarrollado en los laboratorios TCTS de la facultad politécnica de Mons (Bélgica) cuyo objetivo fue obtener un conjunto de sintetizadores que funcionen con el mayor número de idiomas posibles y que pueda usarse sin ningún tipo de restricción.

El funcionamiento de MBROLA se divide en cuatro pasos (Bonafonte, 2012):

1) El archivo .pho
2) Selección de la base de datos
3) Procesado PSOLA
4) Concatenación con alisado de bordes

En la figura 3.10 se muestra la descripción de estas etapas, formando la frase con MBROLA.

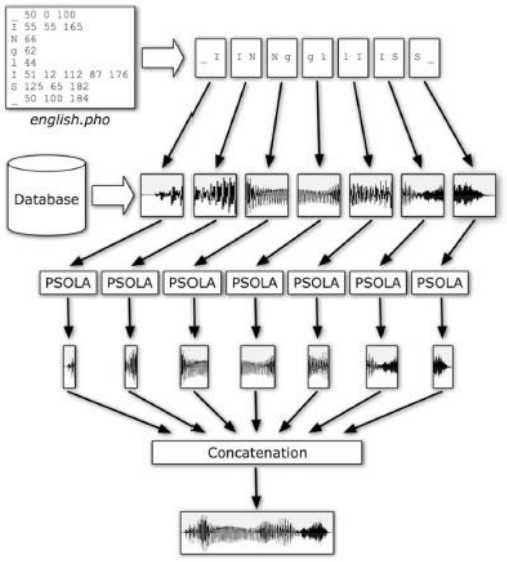


Figura 3.10. Descripción de la formación de la frase con MBROLA (Bonafonte, 2012).

Este programa da la información de los valores de duración y de la frecuencia fundamental (F_0) de cada alófono considerado, el cual acepta hasta 20 valores máximo de la F_0 para cada alófono, sin embargo aunque los valores de la F_0 se puedan utilizar para dibujar una curva melódica, esto implica la existencia de un modelo prosódico. El algoritmo MBROLA ha sido usado con fines de evaluación de los resultados obtenidos sobre generación automática de prosodia para la conversión de texto en habla.

Capítulo 4

En este capítulo se explicará cómo se segmentaran las palabras y fonemas para construir el corpus del sintetizador de voz.

4.1 Corpus de voz

Un corpus lingüístico es un conjunto de audios almacenados en formatos electrónicos y agrupados con el fin de estudiar una lengua o una determinada variedad lingüística. La realización de un corpus es una actividad fundamental para el desarrollo de sistemas del lenguaje hablado, ya que el corpus contiene frases, palabras y expresiones comunes en la lengua particular (Llisterri, 2005). El objetivo es construir un corpus de voz con elementos de referencias para el estudio de una fase concreta en un cierto aspecto de la lengua.

Existen dos tipos de corpus que son:

- Textuales: se dividen en el corpus de la lengua general y corpus de un sub-lenguaje. El corpus textual, recoge íntegramente todos los textos de los documentos que los constituyen (textos, series o párrafos coherentes).
- Orales: se dividen en corpus para el estudio del lenguaje oral, para fines específicos, para el desarrollo de las aplicaciones tecnológicas del habla y para el desarrollo de aplicaciones específicas.

La idea del corpus de voz nace del origen de la comunicación en los seres humanos. Gracias a los avances tecnológicos del siglo XX, ha sido más fácil hacer grabaciones de mejor calidad. Iniciaron los elementos magnetofónicos estos protagonizaron la industria, sin embargo para las ultimas décadas se desarrollaron técnicas de digitalización, que incrementaron aún más la calidad de las grabaciones y se implementó el uso del procesamiento digital de señales (Llisterri, 2005).

La síntesis de voz basada en la concatenación de unidades requiere de una base de datos de la cual se extraen dichas unidades para formar la voz sintética. Un corpus de este tipo está enfocado a un sistema de síntesis de voz, el cual debe contener la información necesaria para poder hacer la concatenación de los segmentos como: fonemas y difonemas, esto incluye un correcto etiquetado de las unidades fonéticas.

El propósito fundamental de estos sistemas es reproducir el habla con la mayor inteligibilidad y naturalidad posible, por ello el corpus debe ser grabado por un único locutor, y este debe presentar las mejores características acústicas en su voz.

Un corpus de voz que puede ser utilizado para el desarrollo de un sintetizador de voz concatenativo debe tener (aparte de las unidades del habla en sí mismas) los siguientes elementos:

- Elementos fonéticos que acojan todas las variaciones alofónicas presentes en la lengua.
- Características de la entonación y duración de las palabras.
- Fonemas.
- Difonemas.
- Trifonemas
- Silabas y Semisílabas
- Palabras

En la concatenación de unidades, la selección de la unidad a utilizar debe contemplar la relación entre el tamaño del conjunto de unidades necesarias para cubrir el idioma español (base de datos) y la calidad de la voz que se desee obtener. Entre más grandes son los segmentos más calidad tendrá el sintetizador, la desventaja de esto es que entre más longitud tenga la unidad mayor número de posibles combinaciones tenemos por lo que el tamaño de la base de datos aumenta considerablemente.

Las unidades grabadas en el presente proyecto fueron obtenidas de acuerdo al análisis realizado en la tesis doctoral: Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el Corpus del Español Mexicano Contemporáneo de México (Urrea, 2003). De este trabajo se obtienen la lista de las palabras más comunes del español contemporáneo de México (ver apéndice A), así como algunas terminaciones.

4.2 Grabación de base de datos de audio

En la grabación de la base de datos son importantes los siguientes factores:

- Inteligibilidad en la voz
- Naturalidad de la voz
- Calidad de la voz

Los puntos antes mencionados, influyen positiva o negativamente en el resultado final del sistema de sintetización del habla por concatenación de unidades. Si las unidades fonéticas que están grabadas en el corpus están bien pronunciadas sin que exista algún otro sonido ajeno al contenido fonético de cada texto, se puede conseguir una voz sintética de calidad e inteligibilidad considerable.

Para realizar las grabaciones de un locutor se tomaron en cuenta las siguientes consideraciones:

1) Leer lentamente cada frase que se deba grabar, una o dos veces antes de hacerlo en voz alta delante del micrófono, para asimilar todo el contenido fonético que la frase incluye.
2) Mantener una amplitud constante en la voz, desde la primera, hasta la última frase de cada sesión de grabación, así como a lo largo de todas las sesiones de grabación.
3) Leer los textos pausadamente, es decir a una velocidad que permita una vocalización adecuada de las palabras pero sin perder naturalidad, sin exagerar el énfasis que puede dársele a ciertas letras o silabas durante la pronunciación de las palabras.
4) Mantener una posición similar y una distancia uniforme respecto del micrófono durante todas las sesiones de grabación, para evitar posibles variaciones en la amplitud de los sonidos grabados.
5) Evitar todos los sonidos que puedan provenir de la respiración del locutor, así como los de manipulación de papeles u otros objetos que incluyen el pedestal del micrófono, en espacios de voz que no pueden evitarse.
6) Cuando surge alguna equivocación de pronunciación se repite la frase.
7) Evitar hablar directamente al micrófono, ya que existen algunas letras, como es el caso de la p. que requiere una pequeña explosión de aire para ocurrir, lo que conlleva a un sonido de soplido capturado en la grabación.

Las grabaciones fueron realizadas con una grabadora digital marca SONY a una frecuencia de 8 KHz de 16 bits y sonido estéreo.

En este trabajo se usaran archivos de audio grabados por un locutor que cumple con todas las características acústicas, para realizar la construcción de un corpus oral que se ocupará en el sintetizador de voz por concatenación de unidades.

Para la construcción del corpus se ocupara un software científico llamado Praat, creado por Paul Boersma y David Weenink, del instituto de ciencias fonéticas de la universidad de Ámsterdam, el cual está diseñado para el estudio fonético del habla y es usado en la lingüística, este software permite diferentes tipos de análisis y manejo de audio acústico como: segmentación, etiquetado, manipulación del habla natural y creación de estímulos sintetizados. Con la ayuda de este software se realiza la selección y extracción del audio a usar, es decir se selecciona el archivo de audio grabado por el locutor, se habré la ventana mostrando el espectrograma, esto se observara en la sección 4.4, se selecciona el fragmento de audio que se necesita

grabar y se guarda con extensión .wav. El mismo proceso se realizará para la selección de todas las posibles combinaciones existentes de difonemas y se descartaron aquellas combinaciones que no se presentan en el español (Apéndice B), palabras (Apéndice A) y terminaciones (ción) más comunes en el español de México, para la construcción del corpus o base de datos de este trabajo.

4.3 Organización de la base de datos del corpus

La grabación del corpus oral para el sintetizador de voz por concatenación de unidades se organizara en carpetas llamadas difonemas, palabras y terminaciones (ver figura 4.1).

En cada una de las carpetas están los archivos .wav, que se van a ocupar para la arquitectura del sintetizador de voz por concatenación de unidades para formar la voz artificial. Se nombraron así para que sea más fácil invocar desde el programa que se realizará en el software Dev-C++ en lenguaje de programación C, al ingresar una frase, el programa lo buscara por carpeta, es decir, palabra, difonema o terminación según sea el caso para que el sintetizador pueda reproducirlo. Este proceso se explicará a detalle en el capítulo 5.

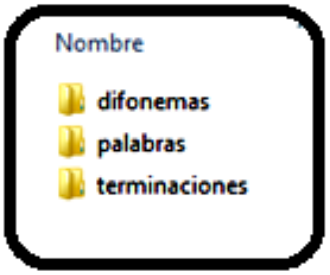


Figura 4.1. Base de datos.

4.4 Espectrograma de la voz

Los espectrogramas son gráficas que permiten visualizar la evolución temporal y frecuencial de un espectro, es decir, se puede observar las variaciones en el tiempo y frecuencia de la voz humana, y de sonidos que se hayan adherido a las grabaciones.

La voz se puede modelar como la respuesta de un sistema lineal e invariante con el tiempo, tracto vocal: es un sistema de transmisión acústica que se caracteriza por sus frecuencias naturales, denominadas formantes, que corresponden a la resonancia de su frecuencia. La voz es claramente una señal no estacionaria, sin

embargo se puede suponer que las características de la señal permanecen esencialmente constantes en intervalos de tiempo del orden 30 o 40 ms. El contenido de la señal de voz puede abarcar hasta los 15 KHz o más, pero es latamente, inteligible e incluso con bandas de frecuencia limitada a unos 3 KHz. La mayoría de la información fonética está por debajo de 8 KHz.

El espectrograma se construye mediante la transformada de Fourier dependiente del tiempo. Su proceso es calcular el espectro de frecuencia de pequeños segmentos de la señal de sonido que se desplazan a medida que pasa el tiempo, evaluando el contenido espectral, hasta que la señal termine. El espectrograma es la representación de dos dimensiones de la transformada de Fourier X [n, λ], con ω en el eje de las ordenadas, n en las abscisas y el módulo de [X [n, λ]] según una escala de grises (o de colores) establecida para indicar un tercer eje de amplitudes del contenido espectral para un instante y frecuencia determinados.

En la figura 4.2, se muestra el proceso de un espectrograma.

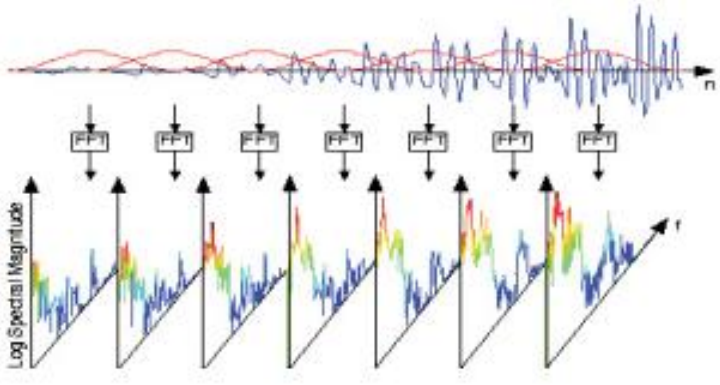


Figura 4.2. Proceso de un espectrograma (D, 2005).

En una señal de voz se pueden distinguir dos contribuciones:

- La del tracto vocal, responsable de la estructura de formantes que tiene una variación lenta a lo largo del tiempo.
- La de la excitación que proporciona la estructura fina, armónica en el caso sonoro, del espectro cuyas variaciones son más rápidas.

La identificación de unidades fonéticas basada en la lectura de un espectrograma es un proceso complejo que requiere de mucha práctica, ya que precisa del conocimiento del comportamiento temporal del contenido espectral, para los diferentes tipos de sonidos que conforman el lenguaje.

Como se mencionó en la sección 4.2 el proceso que se realiza para segmentar las unidades (palabras, difonemas, terminaciones) se ilustran en la figura 4.3 se muestra el espectrograma de la vocal “a”, en la figura 4.4 se muestra el espectrograma de la palabra “cada”, en la figura 4.5 se muestra el espectrograma del difonema “ba”, estos son ejemplos de cómo se van a ir segmentando las unidades que se ocuparan para el diseño del corpus en este trabajo.

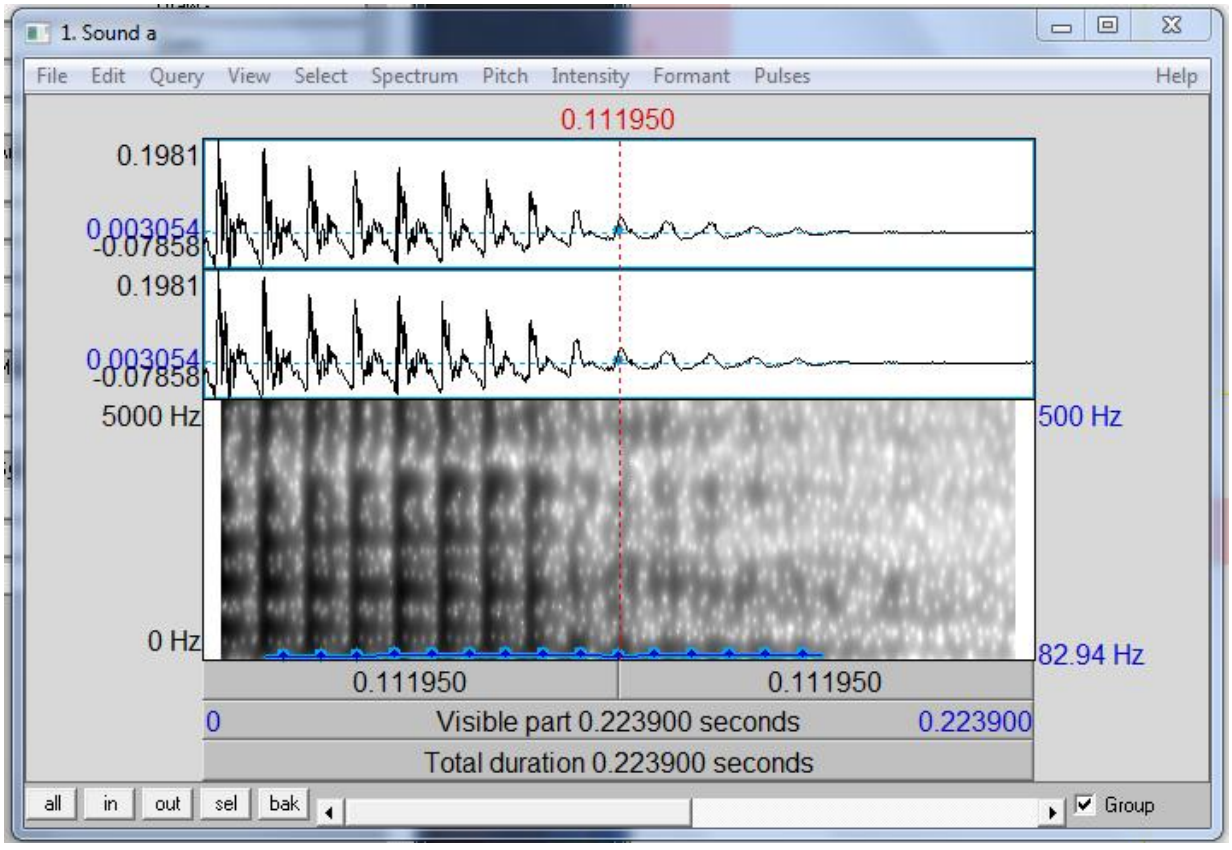


Figura 4.3. Espectrograma del difonema _a (realizada en Praat).

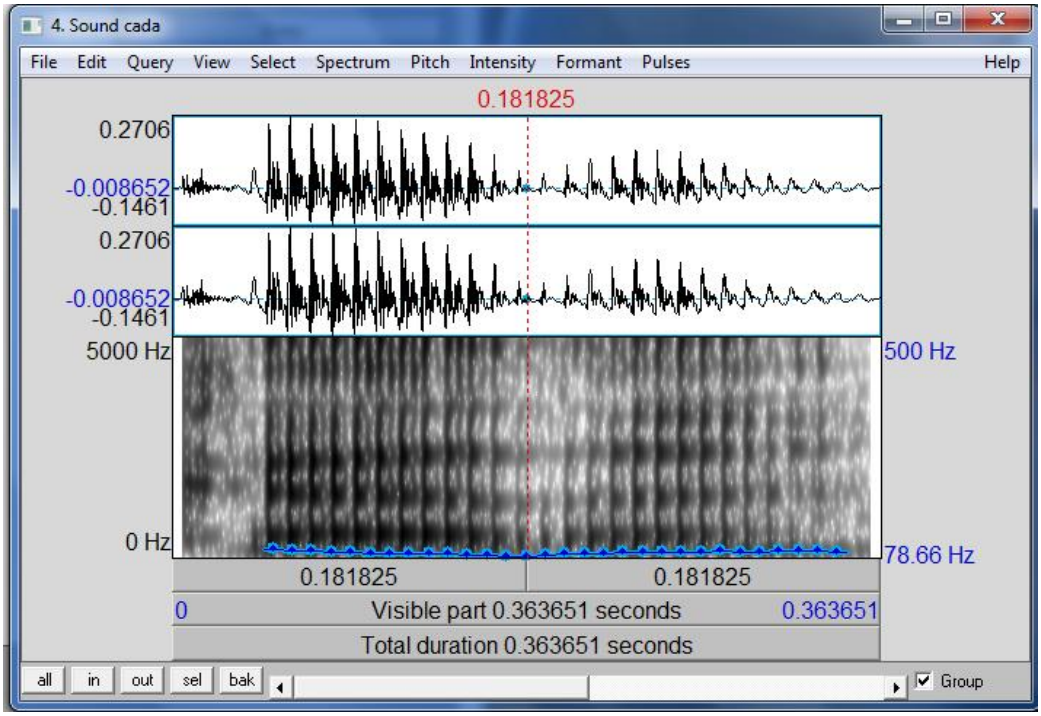


Figura 4.4. Espectrograma de la palabra cada (realizada en Praat).

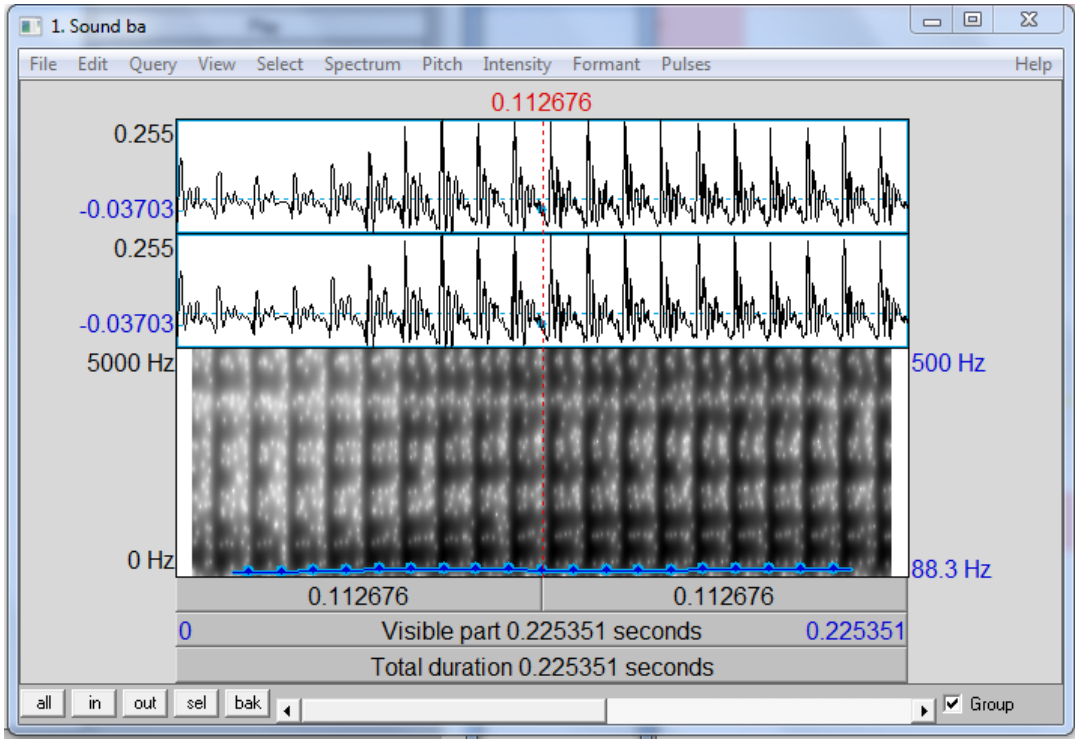


Figura 4.5. Espectrograma del difonema “ba” (realizada en Praat).

En la tabla 4.1 se muestran algunos de los difonemas, en la tabla 4.2 se muestran algunas palabras que conformaran el corpus de voz de este trabajo para la construcción del sintetizador de voz por concatenación de unidades.

ai	en	ni	te
al	es	no	tu
an	ir	oi	un
as	la	pa	ya
da	le	ps	yo
de	lo	se	ca
he	me	si	ce
el	mi	su	ci

Tabla 4.1. Difonemas.

años	esa	mas
antes	ese	mayor
ahora	eso	Me
has	esos	medio
hace	esta	mejor
hacer	estaba	menos
así	estado	mil
hacia	están	mis

Tabla 4.2. Palabras

4.5 Etiquetado de los archivos de audio

Como se había mencionado antes, es necesario que una vez finalizada las grabaciones de los diferentes archivos de sonidos de frase, palabras, fonemas y difonemas, se haya creado un corpus de voz. Este conjunto de archivos de sonido debe contener el mayor número posible de unidades del lenguaje de español México posibles, con el fin de utilizarlas para la concatenación de unidades en la creación de una voz sintética. El proceso mediante el cual se identifican dichas unidades en el corpus se conoce como etiquetado.

El etiquetado conlleva un enriquecimiento del corpus mediante información adicional introducida por el usuario en función de sus objetivos, y lo más importante la interpretación lingüística de los materiales recogidos.
En el proceso de etiquetación de un corpus de voz se delimitan las fronteras de las unidades fonéticas presentes en las grabaciones a cada uno de sus límites se le llama etiqueta. Las etiquetas son marcas que indican dónde comienza y dónde termina una unidad fonética ya sean palabras o fonemas, generalmente son archivos que contienen información asociada a las grabaciones de un corpus. Como se muestra en la figura 4.7. En el proceso de etiquetación, se alinean en el tiempo la duración de cada unidad del lenguaje respecto a las ondas de voz presentes en los archivos de sonido, de esta forma se puede identificar en términos de milisegundos (al inicio y fin del fonema).

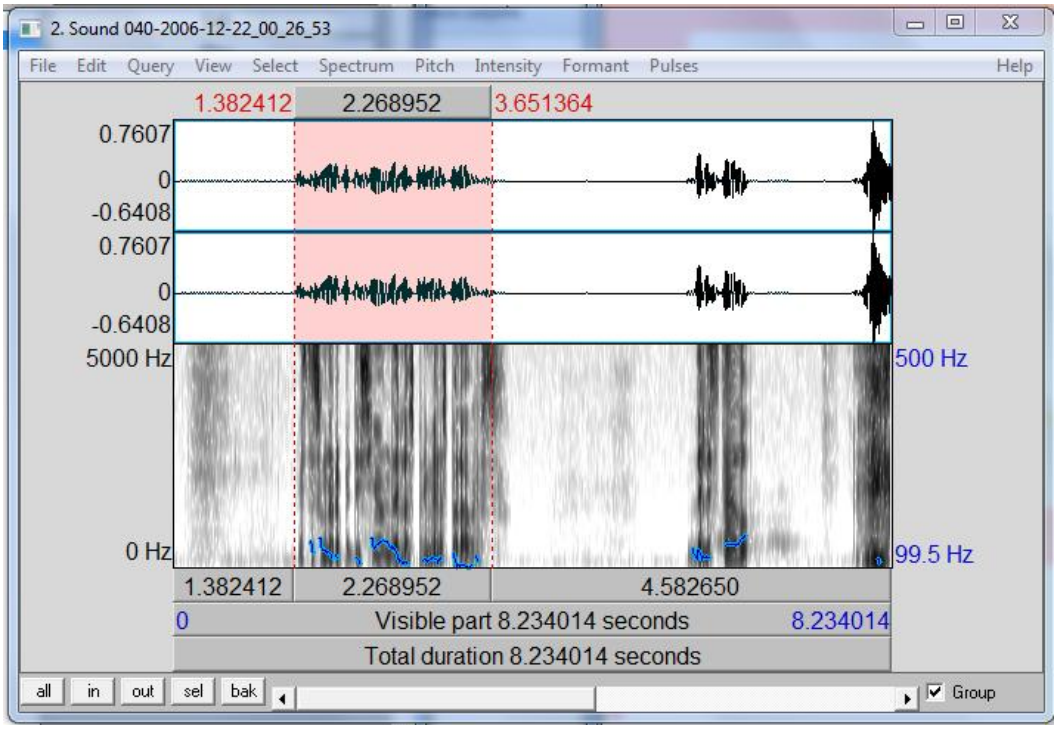


Figura 4.7. Etiquetado (realizada en Praat).

Gracias a las etiquetas a nivel de fonemas durante el proceso de concatenación de unidades se puede extraer literalmente un fragmento de un archivo de sonido del corpus y utilizarlo como unidad determinada para crear una palabra asociándolo con otros fragmentos de sonido. En este trabajo no se realiza un etiquetado de las unidades para mantener la simplicidad, quedando como trabajo futuro para mejorar la calidad del sintetizador.

Capítulo 5

En este capítulo se utilizara el corpus construido en el capítulo 4 para el desarrollo de las etapas del programa para el sintetizador de voz por concatenación.

5.1 Diseño del sintetizador de voz por concatenación de unidades

Debido al enfoque de este proyecto que busca una aplicación para el desarrollo de dispositivos o robots que ayude a personas discapacitadas con HMI. En este trabajo se desarrolló un código en el software Dev-C++ en lenguaje de programación C, en la plataforma de Windows, este programa inicia una búsqueda a partir del texto ingresado a sintetizar a nivel de palabra, en caso de no encontrar la palabra, realiza el concatenado de los difonemas para formarla y que reproduzca el audio de la voz sintética de la frase ingresada, tomando en cuenta que cada formato de audio .wav contiene una cabecera, mismo que se explica a detalle en la sección 5.2, en la sección 5.3 se explicara las cuatro etapas para la implementación del sintetizador de voz por concatenación de unidades.

5.2 Selección de Unidades (Unit selection)

En la técnica unit selection se deben escoger diferentes tamaños de unidades de concatenación con el objetivo de lograr la máxima naturalidad de coarticulación posible. Se da preferencia a aquellos candidatos con el mayor tamaño de unidad que puedan utilizarse como fonemas, difonemas y palabras. Mientras más grande sea el tamaño de la unidad y realizando una selección adecuada de las unidades fonéticas, será mayor la calidad de la voz sintética (Placio, 2003).

Esta técnica busca las unidades que posean características similares o iguales a las de la frase que se desea sintetizar del corpus de voz. Fundamentalmente se basa en la diversidad del tipo de unidades que se pueden concatenar para crear la voz sintética más natural. La síntesis de voz por concatenación de unidades puede trabajar con la técnica de Unit Selection, ya que se pueden concatenar las formas sonoras de diferentes estructuras gramaticales como: los fonemas, difonemas, palabras. De igual forma se consideran las características de prosodia, identidad fonética e incluye unidades de diferentes tamaños. Este hecho abre la posibilidad de que un corpus pueda ser etiquetado a diferentes niveles, estimando las posibles unidades que van a intervenir en el proceso de concatenación (Hunt, 1996).

Como se mencionó en el capítulo 4, en este proyecto se realizaron bases de datos de archivos de voz grabada como punto de partida para la implementación de unit selection en la creación de voz sintética basada en concatenación de unidades. El corpus está compuesto de segmentos de voz que posteriormente serán concatenados para producir frases de cualquier longitud, partiendo de una base de datos de voz de la cual se extraerán dichas unidades, con base en un análisis fonético del texto ingresado que se desea sintetizar, para crear la voz artificial.

Para este trabajo el corpus se creó con archivos de sonido en formato wav, formato diseñado por Microsoft e IBM para almacenar archivos de audio. Cada archivo de sonido .wav tiene 44 bytes de encabezado que almacena información sobre su contenido, tales como su tamaño, el número de canales, la frecuencia de muestreo, entre otros como se muestra en la figura 5.1

El formato wav está organizado de acuerdo a la estructura RIFF (Resource Interchange File Format formato de archivo de intercambio de recurso), (Jimenez, 1996)

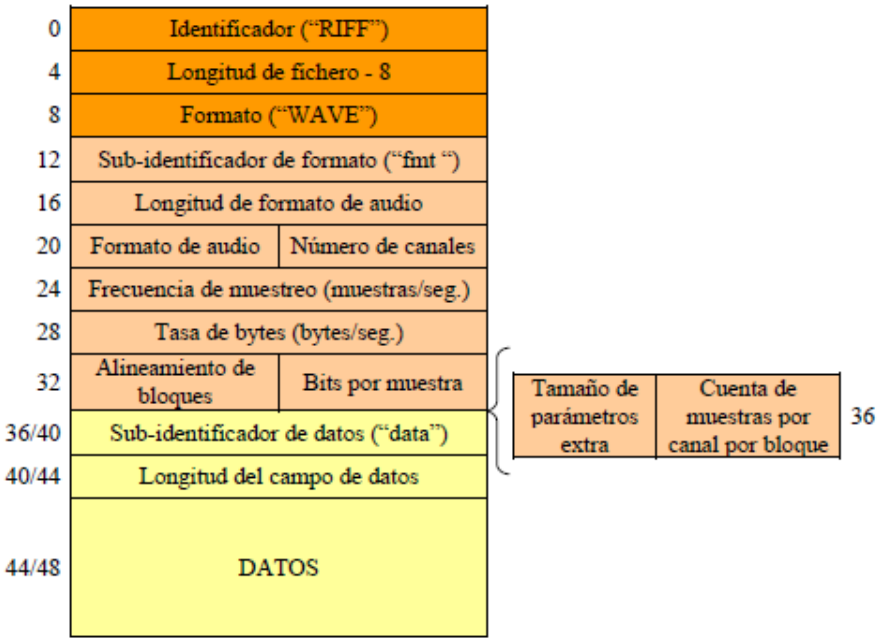


Figura 5.1 formato de archivo .wav (Jimenez, 1996).

Descripción del formato de audio:

- Identificador en código ASCII “fmt”
- El campo longitud del formato de audio (a partir de esa posición). En concreto será de 16 para codificación PCM y de 20 para IMA ADPCM.
- Formato de audio: Indica cómo se encuentran representadas las muestras de audio en los datos. Puede ser 0x01 para PCM o 0x11 para IMA ADPCM. Este campo es de sólo dos bytes.
- Número de canales. Uno para mono y dos para estéreo.
- Frecuencia de muestreo. En el caso de IMA ADPCM, tan sólo son válidas las frecuencias de 8000 Hz, 11025 Hz, 22050 Hz y 44100 Hz
- Tasa de bytes por segundo. Indica el número de bytes de datos que se deben de leer por segundo para reproducir el audio.
- Alineamiento de bloques.
- Bits por muestra.
- Tamaño de parámetros extra.
- Cuenta de número de muestras por canal y por bloque.
- Por último están almacenados los datos, con un identificador previo del campo de datos (la cadena “data”) y otro campo indicando la longitud en bytes de los propios datos.

Para la implementación del algoritmo en lenguaje de programación C del sintetizador por concatenación de unidades, se diseñó una función para ir concatenando los archivo de audio, ya que al ir uniendo los audios para formar la frase o palabras ingresadas en el sintetizador de voz por concatenación de unidades, estas se sobrescribían y a la salida no se obtenía el audio de la voz sintética requerido, esto se debía a que se encimaban las cabeceras de los audios se sobrescribían. Por lo tanto, esta función tiene el trabajo de ir reescribiendo la cabecera de los audios para que realice el concatenado y se escuche adecuadamente el audio final. Para esto se usó un tipo de referencia de estructura para manipular los datos del encabezado del archivo .wav, como se muestra en la figura 5.2.

```
typedef struct swav
{
  char Riff[4];
  long longRiff; //valores+36
  char Wave[4];          //
  char Fmt[4];
  long lonFmt;           // LONGITUD DE LOS DATOS
  //QUE VIENEN A CONTINUACION
  short int catego;          //   2
  short int canales;             //  2
  unsigned int frecuencia;  //2
  int alinea;
  short int e1;
  short int e2;
  char bloque[4];
  long cuerpo; //valores
} TipoWav;
```

Figura 5.2. Estructura para el encabezado del archivo .wav.

5.3 Etapas del sintetizador

Para el diseño del código del sintetizador de voz por concatenación de unidades se identificaron las siguientes etapas necesarias para la implementación del este: procesado de texto, identificación del tipo de unidad, generación de la lista de archivos y concatenación de archivos.

1) Procesado de texto: Dado que en español algunas letras suenan igual, se utilizó el mismo archivo de audio para los difonemas con esas letras, por ejemplo: “va” y “ba”, por lo que se diseñó una función que procesara el texto original para dejarlo en función de los difonemas existentes en el corpus, esta función realiza algunos cambios de literales como son:

- Cambiar la ‘z’ por ‘s’, ya que es una característica propia del español en México.
- Cambiar ‘q’ por ‘k’, ya que su sonido es similar, se quita la “q” y la “u” y se remplaza por una “k”.
- Cambiar ‘v’ por ‘b’, ya que es una característica del español porque tienen el mismo sonido.
- Quitar los acentos de las vocales.
- Quitar las h en caso de que existan ya que este es un sonido mudo.
- Cambiar ‘ll’ por ‘y’, en caso de que la ‘y’ se encuentre al final de la palabra se cambiara por ‘i’
- En caso de encontrar una ‘c’ y después una ‘h’ se cambiaría por ‘C’, si se encuentra una ‘c’ y después una ‘i’ o una ‘e’ se cambia por ‘c’, en cualquier otro caso se cambia por ‘k’.

2) Identificación del tipo de unidad: Dado que el corpus contiene las carpetas mencionadas en el capítulo 4, que son difonemas, palabras y terminaciones, es necesario distinguir el tipo de unidades que se van a utilizar, para esto se ingresa una palabra y esta etapa comienza la búsqueda primero en la carpeta de palabras, si no se encuentra realiza la búsqueda en la carpeta de difonemas para formar la palabra, en caso de que la palabra ingresada tenga terminación que este guardada en la carpeta la despliega, en este caso solo se considera la terminación “cion” por ser la más común como otra unidad posible.

3) Generación de la lista de archivos: una vez seleccionado el tipo de unidad a utilizar, se almacena en un arreglo el nombre del archivo u archivos correspondientes a las palabras o difonemas necesarios en el orden a ser concatenados.

4) Concatenación de archivos: se ingresa a la lista de los difonemas que se deben concatenar y se va generando un solo archivo agregando los audios indicados en la lista, para así obtener un solo audio por cada palabra, el cual a su vez se concatena en el archivo de salida de toda la frase.

En este trabajo no se implementó ningún algoritmo de procesado ya que implicaba un tiempo considerable para su desarrollo. Hasta la fecha ningún algoritmo de modificación prosódica en el dominio de la frecuencia ha sido propuesto como una solución eficiente para realizar un sistema de síntesis en tiempo real.

Capítulo 6

En este capítulo se retomaran las etapas del sintetizador realizadas en el capítulo 5 para diseñar el programa con las funciones requeridas para formar la concatenación de las unidades para la reproducción de la voz sintética en una computadora embebida Raspberry Pi B.

6.1 Implementación del Sistema en una Microcomputadora Embebida

La tarjeta Raspberry Pi B es una placa (SBC) de bajo costo y un mínimo consumo eléctrico, desarrollada en Reino Unido por la fundación Raspberrry Pi, es excelente para desarrollo de software y su sistema operativo es Linux. Su objetivo es estimular la enseñanza de la tecnología. Para su funcionamiento necesitamos un medio de almacenamiento (utiliza tarjetas de memoria SD o microSD) y conectar a la corriente. La placa tiene una base de 85 por 54 mm, ésta contiene un chip Broadcom BCM2835 con procesador ARM hasta 1 GHz de velocidad, GPU VideoCore IV y 512 Mbytes de memoria RAM. Esta tarjeta cuenta con una salida de vídeo y audio a través de un conector HDMI, con lo que conseguiremos conectar la tarjeta tanto a televisores como a monitores que tengan esta conexión. Para el vídeo cuenta con una salida compuesta y una salida de audio a través de un minijack. En cuanto a la conexión de red, disponemos de un puerto Ethernet 10/100 o podemos recurrir a utilizar cualquier adaptador inalámbrico WiFi compatible.
En este proyecto se utiliza la tarjeta Raspberry Pi B, la cual se muestra en la figura 6.1.

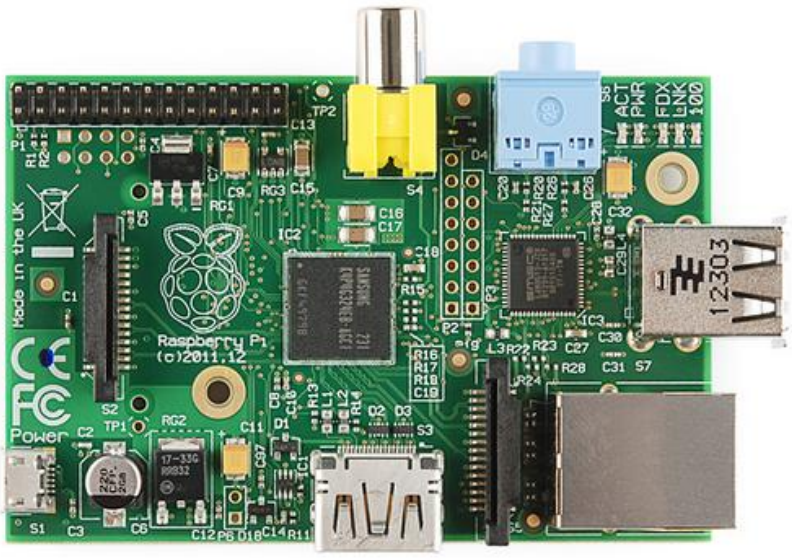


Figura 6.1. Tarjeta Raspberry Pi (Dev 11546).

En un principio para este trabajo se pensaba ocupar la tecnología Bluetooth mediante el módulo de radio RN41XV, este módulo dotaría de comunicación inalámbrica a la tarjeta Raspberry Pi B mediante el puerto serial. Sin embargo, dado que la tarjeta Raspberry Pi B cuenta con un puerto Ethernet (IEEE 802.3), se decidió usar un adaptador de red TP-LINK WN225N de tecnología Wi-Fi (IEEE 802.11) ya que cuenta con mejores características para el desarrollo de este trabajo que Bluetooth.

Los programas en Linux se pueden escribir en cualquier editor de textos de GNU, se guarda el archivo con extensión .c, los cuales son compilados en GNU/Linux utilizando el compilador GNU de C, llamado gcc y además se apega al estándar ANSI, permitiendo la portabilidad de estos códigos, el compilador se invoca con el comando gcc.

GCC es un compilador que está integrado en GNU para C y es capaz de recibir un programa fuente y generar un programa ejecutable. Las siglas GCC significan “GNU Compiler Collection”, antes era “GNU C Compiler”. Este compilador cuenta con varias opciones, las cuales van acompañadas de un guion como se muestra en la tabla 6.1.

-c	Realiza el procesamiento y compilación obteniendo el archivo código objeto, no realiza el enlazado
-e	Realiza el procesamiento enviando el resultado a la salida estandar
-o	Archivo, indica el nombre del archivo de salida cualesquiera sean las etapas cumplidas
-l	Especifica la ruta hacia el directorio donde se encuentran los archivos de biblioteca con el código objeto de las funciones referenciadas en el programa fuente
-v	Muestra los comandos ejecutados en cada etapa de compilación y la versión del compilador

Tabla 6.1. Opciones para GNU.

6.2 Implementación del sintetizador

Como ya se mencionó en el capítulo 5, se utilizó lenguaje de programación C para la creación del sintetizador de voz, la figura 6.2 corresponde con el diseño de la función principal del programa para el sintetizador.

Para la identificación del tipo de unidad se implementaron las siguientes funciones:

- Función busca: recibe pbus[], hace un recorrido sobre un arreglo, en el cual están las palabras más frecuentes del español en México, estas se encuentran en el corpus. Recibe la palabra que se va a buscar, en caso de ser encontrada devuelve 1, en caso contrario devuelve 0. (figura 6.3)

- Función validaTerm: esta función recibe la palabra a validar p[] y la L longitud de esta, verifica si la palabra que se ingresó contiene la terminación “cion” en el caso de que tenga la terminación devuelve 1, si no regresa 0 (figura 6.4).

El procesamiento del texto se realiza mediante la función de restricciones, esta recibe pbus[] y no regresa nada, esta función se encarga de modificar el texto para que al separar en difonemas, estos coincidan con los nombres de los archivos en el corpus, es decir, hace cambio de literales en algunos casos respecto a las características propias del español en México (figura 6.5).

Para la implementación de la obtención de difonemas se construyó la función separDifonemas, recibe f que es la frase y df son los difonemas y regresa j, esta función actúa en caso de no ser encontrada la palabra ingresada. Los difonemas se van separando tomando cada dos caracteres de la palabra a sintetizar con desplazamientos de uno en uno; para los difonemas inicial y final se agrega la letra más “_” (antes si es inicial y después si es final), todos los difonemas se van almacenando en un arreglo (figura 6.6).

En la figura 6.2 se observa el diagrama de flujo de la función principal del programa para el sintetizador de voz por concatenación de unidades, en la figura 6.3 se muestra el diagrama de flujo de la función busca, en la figura 6.4 se muestra el diagrama de flujo de la función valdaTerm, en la figura 6.5 se muestra el diagrama de flujo de la función restricciones y en la figura 6.6. Se muestra el diagrama de flujo de la función separDifonemas.

```
main(int a,char *args[]){ /*función principal del programa*/
int k;
 char frase[500]={0};
 char *ptr, *mensaje;
 int nd,r,L,i,j; printf("<h3>Reproduciendo...</h3><br>") for(j=1;j<a;j++){
mensaje=args[i]; strcat(frase,mensaje);/*concatenacion de cadena*/}
printf("frase:%s\n",frase);
ptr=strtok(frase," "); /*separa la frase en palabras*/  creaArchivo();
 do{
 char df[100][7]={0}
 char p[M]={0};
 nd=0;/*número de difonemas */    printf("tok:%s\n",ptr);
strcpy(p,ptr);
L=strlen(p);/*longitud de la cadena*/ restricciones(p);
 if(busca(p)){
printf("%s.wav",p);
 strcpy (df[0], p);/* (destino, origen)*/ }
else {
 if(validaTerm (p,L)==1) { /*eliminar la terminación*/puts("tiene
terminacion");
 for (j=L-1;j>L-4;j--){/*(inicio;condición;incremento)*/
 p[j]=0; }
nd=separDifonemas(p,df);  /*número de difonemas*/
strcpy(df[nd],"cion.wav"); /*copia el audio de la terminación*/   }
 else
  nd=separDifonemas (p,df);
 for(i=0;i<=nd;i++)
 puts(df[i]);/*imprime la cadena*/   }     }
 if(!nd) {strcpy(subd,"palabras/"); }
 else{ strcpy(subd,"dif/"); }
copiaArchivo("salida",df[0]);
if (nd){
k=1;
while( k<nd)
strcpy(subd,"dif/"); /* la función strcpy sirve para copiar una cadena*/
agregaArchivo ("salida", df[k]);
 k++;  }  }
 printf("archivo creado");
 agregaArchivo("final","salida");  }
 while( (ptr = strtok(NULL," ")) != NULL );/*strtok permite partir la
cadena en subcadenas, en este caso es el separador de la palabras*/
system ("cmxplayer CORPUS/final.wav");
/*audio*/
puts(frase);/*imprime la cadena*/
          }
```

Figura 6.2. Código de la función principal.

```
int busca(char pbus[]){     /*la función busca palabras
completas recibe pbus[] y regresa true en caso de ser
encontrada, en caso contrario regresa false*/
 int j,a=0
 char res;
  res=tolower(pbus[0])-97; //devuelve carácter en minúsculas
  for (j=0;j<N;j++){          if(!strcmp(palabras[res][j],pbus)  {
                                a=1;}  }
 return a;
```

Figura 6.4. Código de la función buscar.

```
int validaTerm(char p[],int L) /*esta función recibe la palabra
a validar p[] y L la longitud de la palabra y regresa true si la
palabra tiene terminacion "cion", en caso contrario regresa
false*/ {
 if (p[L-1]=='n'
 && p[L-2]=='o'
 && p[L-3]=='i'
 && p[L-4]=='c'
  return 1;
 else
  return 0;  }
```

Figura 6.5. Código de la función validaTerm.

```
void restricciones (char pbus[]){/*Esta función recibe pbus[] y no regresa nada*/
int x,i,y,z,Lo;
unsigned char n;
Lo=strlen(pbus); /*Lo=longitud de (pbus)*/
for (x=0;x<Lo;x++){n=pbus[x];              switch (n){
case 'z':
pbus[x]='s';
break;
case 'v':
pbus[x]='b';
break;
case 'h':  {
for(z=x;z<Lo;z++) pbus[z]=pbus[z+1]; }    }
break;
case 'q':    {
pbus[x]='k';
for(i=x+1;i<Lo;i++) {
pbus[i]=pbus[i+1];  }  }
break;
case l:          pbus[x+1]='l';          pbus[x]='y';     for(i=x+1;i<Lo;i++) {
pbus[i]=pbus[i+1];} }  break;}}
case 'y':
x==Lo-1)
pbus[x]='i';
break;
case 'c':
pbus[x+1]=='h'
pbus[x]='C';
pbus[x+1]=='e'|| pbus[x+1]=='i')
(pbus[x]='s');
else pbus[x]='k';
break;
/*se eliminan los acentos de las vocales*/
case 160:
pbus[x]='a';
break;
case 130:
pbus[x]='e';
break;
case 161:
pbus[x]='i';
break;
case 162:
pbus[x]='o';
break;
case 163:
pbus[x]='u';
break; }
}

}}
```

Figura 6.6. Código de la función restricciones.

```
int separDifonemas(char f[50],char df[100][7]){ /*la función
separDifonemas recibe f que es la frase y df que son los
difonemas y regresa j que es el difonema*/
 int i=0,j=1;
df[0][0]='_';
df[0][1]=f[0];
while (f[i+1]!=0){
df[j][0]=f[i];
df[j][1]=f[i+1];
 if (f[i+2]==' ')  i+=2
 i++;j++;   }
df[j][0]=f[i];
 df[j][1]='_';  j++;
                                return j;   }
```

Figura 6.7. Código de la función separDifonemas

Para iniciar la tarjeta Raspberry Pi B se descargó la imagen ISO del programa Raspbian en la memoria SD y se instaló el software libre Win32DiskImager, este hace que la SD sirva como memoria de arranque para la tarjeta Raspberry Pi B. Después se ingresó a la memoria SD el programa que se realizó en lenguaje de programación C, el cual se compiló con el comando gcc – o, pero como se mencionó al inicio del capítulo 5 el programa se realizó en la plataforma de Windows y la tarjeta Raspberry Pi trabaja en Linux, por lo tanto se hicieron algunos cambios de sintaxis para poder compilar el programa en la tarjeta y este funcionara.

1) Se cambiaron la \\ por /
2) Se eliminaron las bibliotecas de windows.h y conio.h
3) Para producir el audio en Linux se cambió la utilería PlaySound que para Windows por Omplayer.

Para la reproducción del audio se utiliza la utilería “omxpalyer”, que es un reproductor de audio y video para la tarjeta Raspberry Pi B, omxplayer viene instalado por defecto en Raspbian.

6.3 Procesamiento de la voz sintetizada

El análisis de la voz consiste en la extracción de características relevantes para el proceso de comunicación. Se deben tomar en cuenta los siguientes puntos para un buen análisis:

- En la decodificación y el reconocimiento la eliminación de la redundancia es importante para una representación eficiente de la señal de voz, así como la simplificación del procesado.
- En la síntesis hay que tomar en cuenta la continuidad de la voz, al tener varias representaciones (dominio temporal, frecuencial y representaciones tiempo-frecuencia).

Para obtener una voz más natural en los sintetizadores de voz, se pueden implementar los algoritmos mencionados en el capítulo 3. Ya que estos algoritmos trabajan la prosodia de la voz, lo cual dota una mejor calidad de la voz sintética generada con el sintetizador. Para implementar los algoritmos mencionados con anterioridad se necesita calcular el valor de pitch.

Para realizar el análisis en el dominio de tiempo se procesa directamente la forma de onda con cálculos, es decir se transforma la señal en una o más señales que varían más lento que la original, por lo cual se consigue una reducción de la velocidad/ancho de banda. Estas señales se obtienen mediante la extracción de los parámetros de la señal de voz en cada trama y creando una secuencia temporal con ellos (Cerviño, 2012).

El pitch es la frecuencia fundamental a la que las cuerdas vocales vibran (F0). Se considera que las características de la frecuencia fundamental son una de las principales portadoras de la información sobre las emociones. El valor medio del pitch expresa el nivel de excitación del locutor. Una media elevada de F0 indica un mayor grado de excitación. El rango del pitch es la distancia entre el valor máximo y mínimo de la frecuencia fundamental. Refleja también el grado de exaltación del locutor. Un rango más extenso que el normal refleja una excitación emocional o psicológica.
Las fluctuaciones en el pitch descritas como la velocidad de la fluctuaciones entre valores altos y bajos y si son abruptas o suaves son producidas psicológicamente. En general, la curva de tono es discontinua para las emociones consideradas como negativas, miedo, enfado, y es suave para las emociones positivas, por ejemplo la alegría. El pitch es un parámetro importante en las aplicaciones de procesamiento de voz.
La detección del pitch se puede calcular con la función de auto correlación, la cual permite medir matemáticamente el parecido existente entre una señal y una versión retrasada en el tiempo de la misma señal.
La función de auto correlación se puede graficar en el software de Matlab, el cual se encuentra en la página de Matlab central, descargas el archivo que realiza esta función, lo instalas en tu máquina y corres el programa, primero con el botón derecho superior seleccionas la palabra y puedes seleccionar las muestras que quieras y las tramas, para finalizar seleccionas el botón run detector pitch.

En este trabajo se descargó el programa de la función de auto correlación, este mostrara el graficó del pitch de la palabra bueno que está contenida en el corpus (ver figura 6.8).

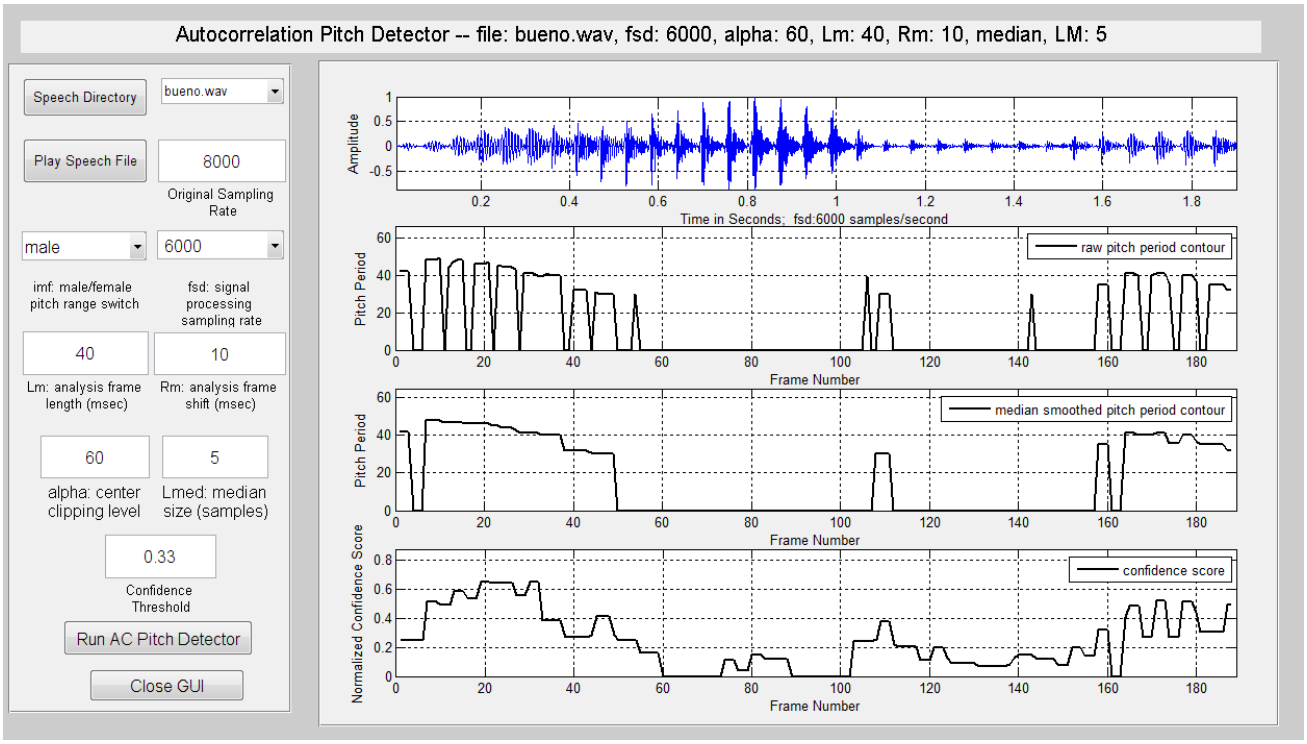


Figura 6.8. Periodo del pitch.

Para aplicaciones de procesamiento digital es necesario trabajar con tramas de la señal, para esto necesitamos ventanas. Después de obtener el pitch que es la frecuencia fundamental, se escoge un tipo de ventana que se describen a continuación y se muestran en la figura 6.9: (Pablo, 2007)

- Ventana rectangular: esta tiene un valor de uno en todo intervalo del periodo y de cero para cualquier otro caso.
- Ventana Hanning: esta ventana es la suma de una ventana rectangular y otra con igual amplitud y forma de coseno, también puede ser descrita como un periodo de una función seno cuadrado.
- Ventana Hamming: es una modificación de la ventana hanning, su forma es similar a la onda coseno.
- Ventana Blackman: es útil para la medición de componentes de bajo nivel en presencia de una señal de entrada larga.

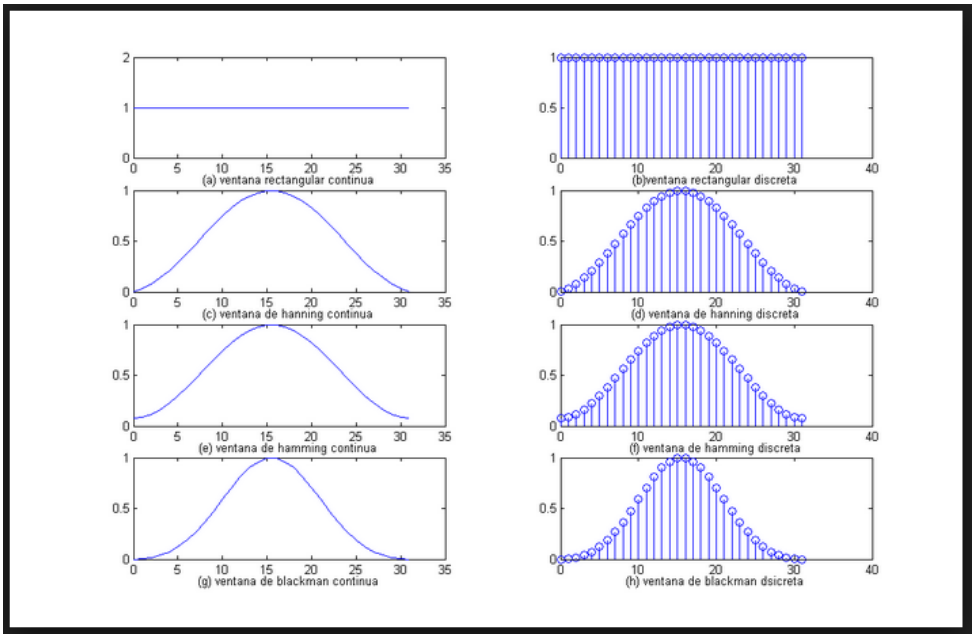


Figura 6.9. Ventanas rectangular, Hanning,Hamming y Blackman. (Pablo, 2007)

En la figura 6.10 se muestra el efecto de cada ventana y su comparación de estas en frecuencia.

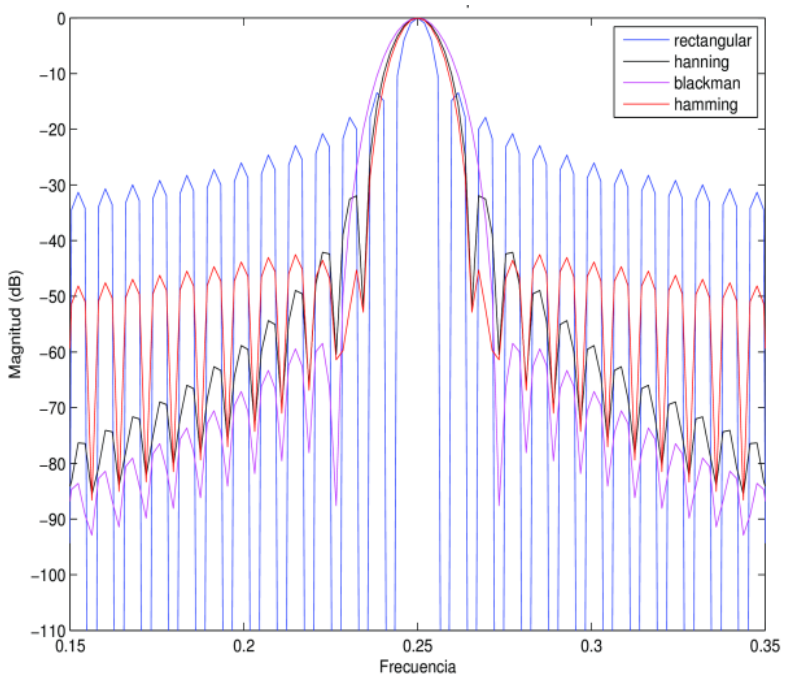


Figura 6.10. Comparación de ventanas en frecuencia (http://www.emus.edu.uy, 2011)

Universidad Autónoma de la Ciudad de México

Nada humano me es ajeno

La ventana de W(n) ha de tener un valor próximo a cero en sus extremos y normalmente es simétrica respecto de la misma. La multiplicación de la ventana tiene dos efectos:

1) Atenuar de forma gradual la señal a ambos lados de la trama seleccionada
2) Produce una convolución de la transformada de Fourier de la función de la ventana y del espectro de la señal.

Debido al segundo efecto, la ventana debe satisfacer dos características para reducir la distorsión espectral introducida por la ventana.

- Lóbulo principal estrecho y agudo, con buena resolución en alta frecuencia
- Gran atenuación de los lóbulos secundarios.

Se elige la ventana y se implementa la ecuación del algoritmo (ver capítulo 3) que se escogió para el suavizado de la señal de voz.

Para el trabajo futuro de este proyecto se recomienda usar la ventana de Hanning para la implementación del algoritmo PSOLA, ya que esta ventana tiene un efecto en el dominio del tiempo y de la frecuencia. En el dominio de tiempo la ventana disminuye la amplitud de la señal cerca de los bordes de la ventana lo cual ayuda a eliminar las discontinuidades. (Genoveva, 2008)

6.4 Sistema de comunicación HMI

En la actualidad, dado que las máquinas y procesos en general están implementados con controladores y otros dispositivos electrónicos que dejan disponibles puertos de comunicación, es posible contar con sistemas HMI bastante poderosos y eficaces, además que permiten una conexión más sencilla y económica con la máquina.

HMI es una interfaz que permite la interacción entre humano y máquina, la cual permite al usuario del sistema interactuar con los procesos.
Hay dos componentes necesarios para la interfaz hombre maquina:

1) Entrada, un usuario humano necesita de algún medio para decirle a la máquina qué hacer, hacerle peticiones o ajustarla.
2) Salida, permite a la maquina mantener al usuario actualizado acerca del avance del proceso o la ejecución de comandos en un espacio físico.

Una adecuada interfaz HMI busca obtener el estado del proceso de un vistazo, es decir, necesita:

- Captar la situación en forma rápida
- Crear condiciones para la toma de decisiones correctas
- Que los equipos se utilicen en forma óptima y segura
- Garantiza la confiabilidad al máximo
- Cambiar con facilidad los niveles de actividades del operador

Interfaz de manipulación directa es el nombre de una clase general de interfaces de usuario, que permiten a los usuarios manipular los objetos que se les presenten, con las acciones que correspondan al menos vagamente con el mundo físico.

Los siguientes tipos de interfaz de usuario son los más comunes (Cobo, 2013):

- Las interfaces graficas de usuario (GUI) aceptan la entrada a través de un dispositivo como el teclado de la computadora y el ratón, estas proporcionan una salida grafica en la pantalla de la pc. Hay por lo menos dos principios diferentes utilizados en el diseño de interfaz gráfica de usuario: orientada a objetos e interfaces a aplicaciones.
- Las interfaces basadas en web de usuario o interfaces de usuario web (IUF), son una subclase de interfaces graficas de usuario que aceptan una entrada y proporcionar una salida mediante las páginas web que se transmiten a través de internet y vistos por el usuario mediante un navegador web.
- Las pantallas táctiles son dispositivos que aceptan una entrada a través del tacto de los dedos o un lápiz. Se utiliza en una amplia cantidad de dispositivos móviles y muchos tipos de punto de venta, procesos industriales y máquinas de autoservicio, etc.,
- Las interfaces de línea de comandos, donde el usuario proporciona la entrada al escribir una cadena de comando con el teclado de la computadora y el sistema proporciona una salida de impresión de texto en la pantalla de la computadora. Utilizado por programadores y administradores de sistemas, en los ambientes científicos y de ingeniería y por los usuarios de computadoras personales de tecnología avanzada.
- La interfaces de voz del usuario, que acepta la entrada y proporcionar una salida mediante la generación de mensajes de voz. La entrada del usuario se realiza pulsando teclas o botones, o responder verbalmente a la interfaz
- La multi-pantalla de interfaces, el empleo de múltiples pantalla para proporcionar una interacción más flexible. Esto se emplea a menudo en la interacción de juegos de computadora, etc.

Las HMI son ampliamente utilizadas en la sociedad actual, su uso puede ir desde el control de un video juego hasta el control de una planta de producción. Para controlar una misma aplicación se pueden utilizar diferentes tipos de HMI solo que unos proveen un mejor control e interfaz que otros.

En este trabajo se implementó una interfaz Web para el sintetizador de voz por concatenación de unidades que pueda interactuar con las personas a través de dispositivos electrónicos a través de una HMI ya que es más rápido sencillo y económico.

Capítulo 7

En este capítulo se describe la tecnología inalámbrica Wi-Fi y la una conexión remota realizada para interactuar con el sintetizador de voz por concatenación de unidades.

7.1 Comunicación Wi-Fi

Un estándar es un conjunto de normas o reglas establecidas con el fin de proporcionar un marco común de trabajo. El estándar IEEE 802.11 establece las características de la capa física y la capa de enlace del modelo OSI (Open System Interconnection), este estándar es llamado de varias formas: Wi-Fi (marca comercial), Wireless-Fidelity, Wireless LAN y WLAN, es usado para cualquier red de área local inalámbrica que utilice ondas de radio como portadora, y IEEE 802.11x, se refiere al grupo de estándares dentro del IEEE 802.11, b, a, g, u otros. En 27 de junio de 1997 la IEEE añadió el estándar IEEE 802.11, el cual definía las redes de área local inalámbricas. Esta primera versión utilizaba una transmisión de infrarrojos, la cual no tuvo buena aceptación, después aparecieron otras dos versiones que usaban radiofrecuencias en la banda 2.4 GHz, la única diferencia fue el método de transmisión una usaba FHSS (Frequency Hopping Spread Spectrum o difusión por salto de frecuencia) y la otra DSSS (Direct Sequence Spread Apectrum o difusión por secuencia directa).
En 1999 aparecieron tres versiones más y en este mismo año se creó la asociación WECA (Wireless Ethernet Compatibility Alliance o Alianza de compatibilidad Ethernet Inalámbrica) (Carballar, 2008).

Existen varias extensiones del estándar 802.11, las cuales se describen a continuación:

- IEEE 802.11a, surgió en el año 1999, utiliza OFDM (Ortogonal Frequency División Multiplexing), la cual divide una señal de datos a través de 48 subcarriers separado con un canal de 20 MHz para transmitir en rangos de 6,9,12,18,24,36,48 o 54 Mbps, opera en la banda de 5 GHz, alcance limitado a 50 m, con 12 canales.
- IEEE 802.11b, surgió en el año 1999, utiliza DSSS con modulación CCK (Complementary Code Keying) y opera en la banda de 2.4 GHz, los rangos de datos que soporta son 1, 2, 5.5 y 11 Mbps, tienen 3 canales de 22 MHz.
- IEEE 802.11g, surgió en el año 2001, permite transmitir datos de 20-54 Mbps, este estándar es compatible con la extensión 802.11b, utiliza OFDM Y DSSS.

- IEEE 802.11n, este estándar incorpora varias antenas para usar varios canales simultáneamente, esto se conoce como MIMO (Multiple Input – Multiple Output “Múltiple entrada – Múltiple salida”), su transmisión es de 450-600 Mbps, canales de 40 MHz.
- IEEE 802.11e, se aprobó a finales del 2005, este estándar tienen las características de la calidad del servicio de las redes inalámbricas y mejoras las técnicas PCF Y DCF con una nueva función HCF (Hybrid Coordination Function, función hibrida de coordinación).

El estándar IEEE 802.11 se ha ido modificando para optimizar el ancho de banda o para especificar componentes de mejor calidad para mayor seguridad y compatibilidad.
Wi-Fi hace referencia al estándar IEEE 802.11b, hoy en día las redes inalámbricas que se instalan cumplen con el estándar 802.11g o 802.11a, y se siguen llamando Wi-Fi porque son compatibles con el estándar 802.11b (Escudero, 2007).

En la tabla 7.1 se muestran algunas versiones del estándar IEEE 802.11.

NORMA	802.11	802.11b	802.11a	802.11g
FECHA	1997	1999	2000	2003
BANDA	2.4 GHz	2.4 GHz	5.8 GHZ	2.4 GHz
DATE RATE	1.22 Mbps	1,2,5.5 Mbps	Hasta 54 Mbps	Hasta 54 Mbps
TECNOLOGIA	DSSS,FHSS	DSSS	OFDM	OFDM
COMPATIBILIDAD		Compatible 802.11	No compatible 802.11,802.11b	Compatible 802.11b
ESTADO	Obsoleta	La más difundida	Emergente	Futuro

Figura 7.1. Versiones del estándar 802.11

7.1.1 Topología de red IEEE 802.11

El estándar IEEE define el concepto de conjunto básico de servicios (BSS, Basic Service Set), el cual consiste en dos o más nodos inalámbricos o estaciones que se reconocen una a la otra y pueden transmitir información entre ellos (Leon, 2004). Una BSS puede intercambiar información de dos modos diferentes (Stalings, 2004):

1) Ad Hoc o IBSS (Independent Basic Service Set), cada nodo se comunica con el otro en forma directa y sin ninguna coordinación, este modo solo permite la transmisión entre los nodos inalámbricos y no resuelve el problema de extender una LAN cableada. (ver figura 7.1)

Figura 7.1. Modo Ad Hoc.

2) Infraestructura, existe un AP (Access Point) que coordina la transmisión entre los nodos inalámbricos, permite vincular la red inalámbrica con la red cableada ya que el AP actúa como bridge (puente de red) entre las dos redes. La existencia de varios AP conectados a un sistema DS (Distribution System), que puede ser una LAN cableada es lo que se denomina EBSS (Extended Basic Service Set). (ver figura 7.2)

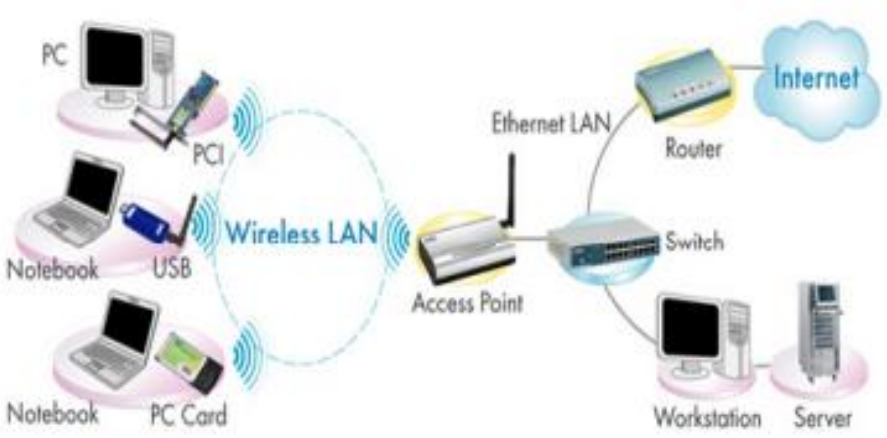


Figura 7.2. Modo de Infraestructura.

En este trabajo se implementó la topología de infraestructura para la conexión inalámbrica Wi-Fi del sintetizador de voz por concatenación de unidades.

7.1.2 Dirección IP

Una dirección IP es un número que identifica de manera lógica y jerárquicamente a una interfaz de un dispositivo dentro de una red que utilice el protocolo IP. Las direcciones IP se pueden expresar como números de notación decimal, los cuales se dividen los 32 bits de la dirección en cuatro octetos, el valor decimal de cada octeto puede ser entre 0 y 255 (Agramunt, 2014).

Existen dos sistemas de asignación IP (protocolo de internet) que son:

1) Fija o estática: no cambia, el equipo o dispositivo al que se le asigna tienen siempre la misma, ya sea internet (red pública) o en una red doméstica. Estas IP´s son asignadas por el usuario después de haber recibido la información del proveedor o bien dadas por el proveedor en el momento de la primera conexión. Esto permite al usuario montar páginas web, correo, ftp, etc. Algunas de las ventajas son: facilidad de identificar al usuario que está utilizando esa IP, permite tener servicios dirigidos directamente a IP y nunca cambia. Y algunas desventajas es que tiene más vulnerabilidad al ataque y es más costosa.

2) Dinámica: es asignada mediante un servidor DHCP (Dynamic Host Configuration Protocol, protocolo de configuración dinámica) al usuario. La IP que se obtiene tiene una duración máxima determinada, actualmente ofrece la mayoría de operadores sin gasto adicional. Estas suelen cambiar cada vez que el usuario reconecta por cualquier causa.
   Algunas de sus ventajas son: Es más difícil identificar al usuario que está utilizando esa IP, Reduce los costos de operación a los proveedores de servicios internet (conocidos como ISPs) y Para los ISP los equipos son más simples.
   Sus desventajas serian que obliga a depender de servicios que dirigen un host a una IP y es ilocalizable en unas horas pueden haber varios cambios de IP.

En este trabajo se dotó con la tecnología Wi-Fi a la tarjeta Raspberry Pi B, su configuración se muestra en la sección 7.2. Una vez que se conoce la dirección IP asignada a la tarjeta Raspberry Pi, se puede tener acceso a ella a través de todas las computadoras o dispositivos móviles que se encuentren en la red a la cual está conectada (ver figura 7.3).

Figura 7.3. Red Wi-Fi.

7.2 Configuración

Para este trabajo se configuró la comunicación WiFi en la tarjeta Raspberry Pi B con un adaptador de red y una conexión USB TP-LINK WN225N y se actualizaron los paquetes necesarios para activar el Wi-Fi, ver tabla 7.2.

apt-get update	Actualiza el listado de paquetes disponibles
rpi-update	Actualiza las direcciones que tiene el paquete
sudo apt-get install	Instala los programas deseados
Sudo apt-get upgrade	Actualiza solo los paquetes ya instalados que no necesitan, como dependencia, la instalación o desinstalación de otros paquetes.
upgrade	Actualiza el sistema operativo y también actualiza las aplicaciones de las direcciones de los paquetes
reboot	Reinicia la tarjeta

Tabla 7.2. Paquetes para Wi-Fi.

En la figura 7.4 se muestra el funcionamiento de la tarjeta Raspberry Pi B con el adaptador de red.

Figura 7.4. Raspberry Pi con comunicación Wi-Fi.

7.3 Acceso remoto a la tarjeta Raspberry Pi B

En la actualidad los seres humanos interactúan y se relacionan con las máquinas, con el fin de aprovechar las tecnologías informáticas, esto se obtienen a través de interfaces que permiten al usuario ingresar y recibir información de una manera fácil, rápida y lo más cercano a una comunicación humana (HMI).

El World Wide Web (WWW), es la forma más popular de manejar información en internet gracias a su facilidad de manejo. La WWW es un sistema que contiene documentos llamados páginas web, estas poseen una dirección electrónica única y la posibilidad de enlazarse unas con otras sin limitaciones, con códigos HTML (Hyper Text Mark up Language): es un lenguaje de programación que se utiliza para el desarrollo de páginas de internet, fue desarrollado por la Organización Europea de Investigación Nuclear (CERN) en el año 1945, y su finalidad fue crear un sistema de almacenamiento donde las cosas no se perdieran y que pudieran ser conectadas a través de hipervínculos, para esto se diseñó el lenguaje HTML que es simple. Las páginas web pueden contener: texto, imágenes, sonido, animaciones, videos u otros y su importancia reside en que es una de las formas más difundidas de comunicación en la actualidad. PHP (Hypertext Preprocessor) es el lenguaje de programación de código abierto popular especialmente adecuado para el desarrollo web y que puede ser incrustado en HTML, lo que distingue a PHP del lado del cliente es que el código es ejecutado en el servidor, generando HTML y enviándolo al cliente. Un servidor es una computadora que formando parte de una red provee servicios a otras computadoras denominadas clientes, es decir, está encargado a múltiples clientes

que hacen peticiones de algún recurso administrativo, el servidor cuenta con un software y hardware especial. Un cliente es el proceso que permite al usuario formular los requerimientos y pasarlos al servidor.
En este trabajo se configuró el servidor apache en la tarjeta Raspberry Pi B, ya que es el más utilizado a nivel mundial. Su funcionamiento básico es ejecutando un proceso padre y tantos procesos hijos como peticiones reciba para atender a cada cliente. Su objetivo es servir o suministrar páginas web (en general, hipertextos) a los clientes web o navegadores que lo solicitan. La arquitectura utilizada es cliente/servidor, es decir el equipo cliente hace una solicitud o petición al servidor y éste la atiende.

En el cliente se ejecuta una aplicación llamada “navegador o cliente web” que:

- Sirve de interfaz con el usuario: atiendes sus peticiones, muestras los resultados de las consultas y proporciona al usuario un conjunto de herramientas que facilitan su comunicación con el servidor.
- Se comunica con el servidor web: transmite las peticiones de los usuarios

El protocolo utilizado para la transferencia de hipertexto es HTTP (Hiper Text Transfer Protocol), este se encarga del envío de mensajes y establece un conjunto de normas, a través de estas se envían peticiones de acceso a una web y la respuesta de esa web.

En este trabajo, Raspberry PI B sirve como servidor y el cliente será un dispositivo móvil, laptop, pc u otros dispositivos. El cliente podrá ingresar desde cualquier dispositivo a la página web diseñada solo poniendo en el URL (Uniform Resource Locator o localizador uniforme de recursos) la dirección IP que se le haya asignado a la tarjeta Raspberry Pi B, ya que tiene acceso a la página web, el cliente ingresa una frase en la barra de texto (ver figura 7.5), después de unos segundos la voz se produce localmente en la tarjeta Raspberry Pi B y si es mandada de una laptop tarda más tiempo en reproducir la voz sintética. La comunicación que se utiliza para la página web es HTML y PHP, código de página web se muestra en el apéndice C. En la figura 7.6 se muestra un diagrama a bloque de la petición del cliente al servidor. Es decir, sólo el acceso al sintetizador de voz es remoto.

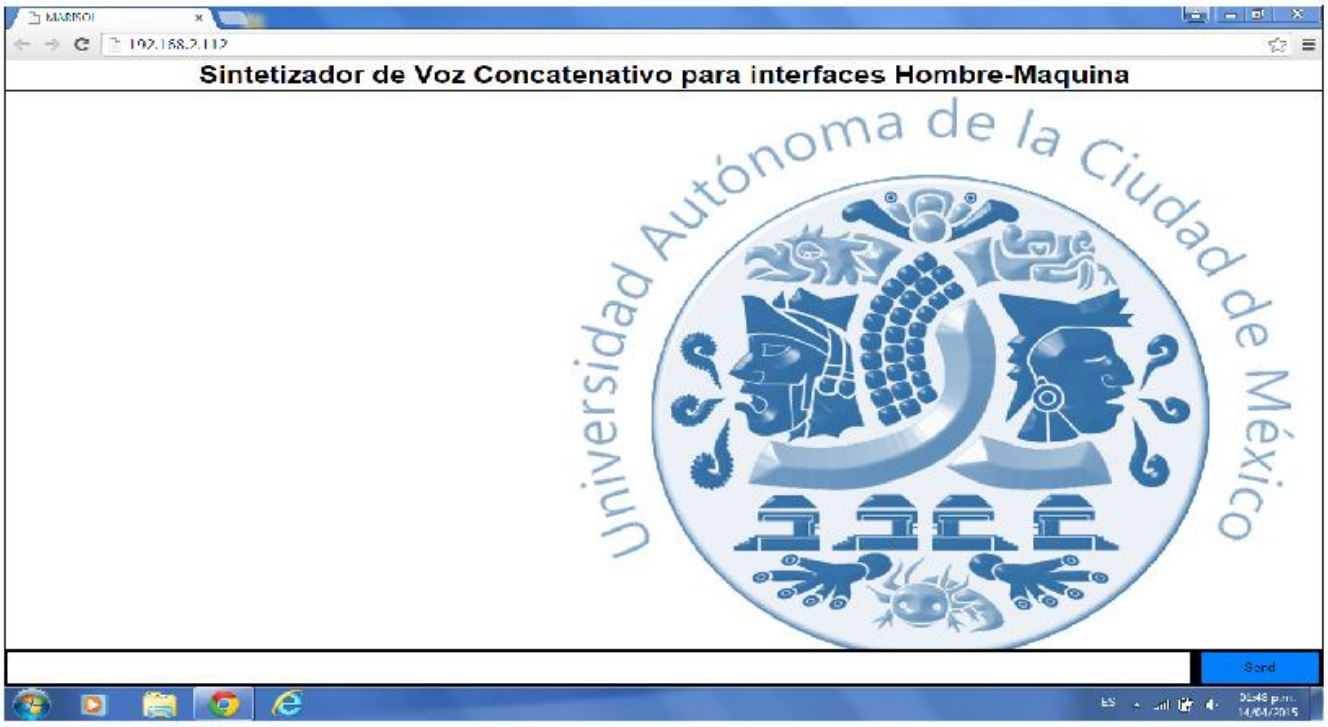


Figura 7.5. Diseño de página web

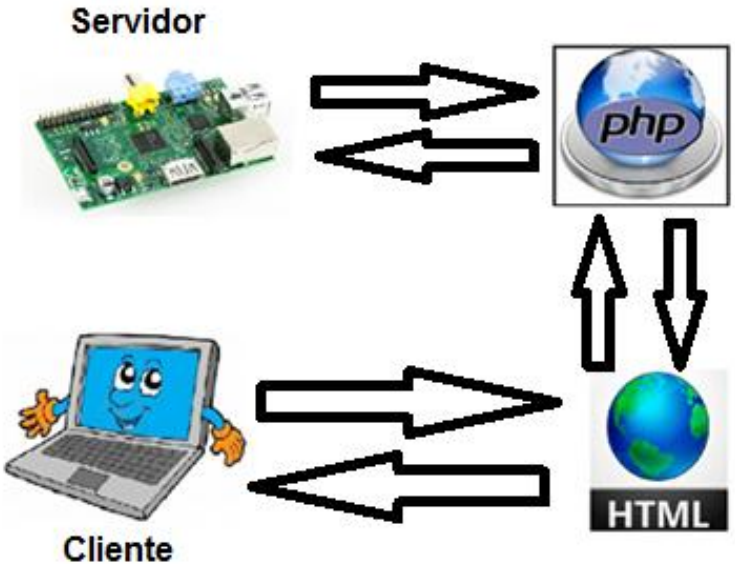


Figura 7.6. Comunicación cliente servidor.

Capítulo 8

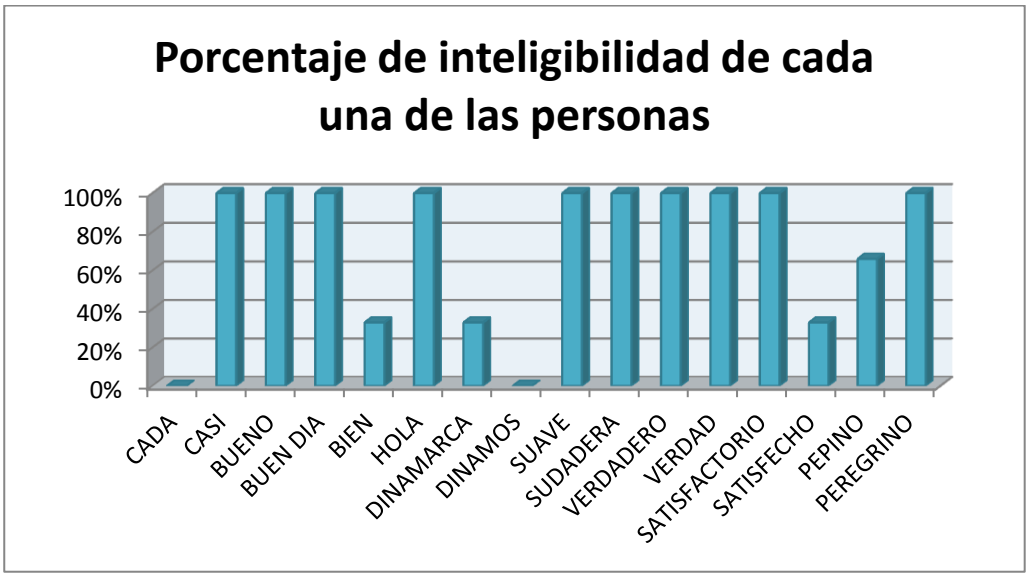
Resultados y Conclusiones

En este trabajo se presentó el diseño e implementación de un sintetizador de voz por concatenación de unidades en una computadora embebida Raspberry Pi B con sistema operativo linux, con tecnología Wi-Fi para acceso remoto, el cual fue concebido como una propuesta de interfaz hombre máquina para aplicaciones en robótica. A continuación se presentan los resultados obtenidos a las pruebas de inteligibilidad realizadas así como las conclusiones obtenidas.

Resultados.

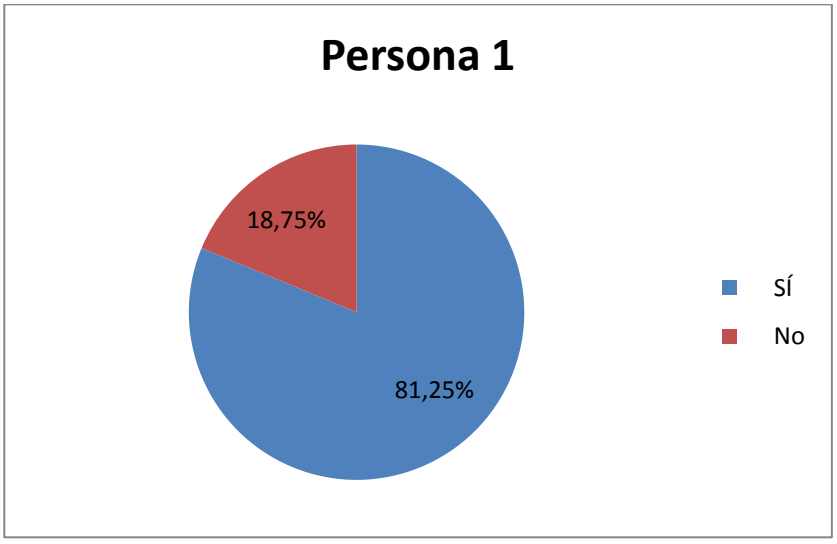
Para evaluar la inteligibilidad del sistema se realizaron pruebas de percepción de la voz sintética que a continuación de describen.

Prueba 1. Se genera la señal de voz sintética de una lista de 16 palabras, las cuales son escuchadas por tres personas diferentes, después de reproducir cada una de las palabras se les pregunta si el audio escuchado es inteligible, la figura 8.1 muestra los porcentajes totales de inteligibilidad para cada una de las palabras.



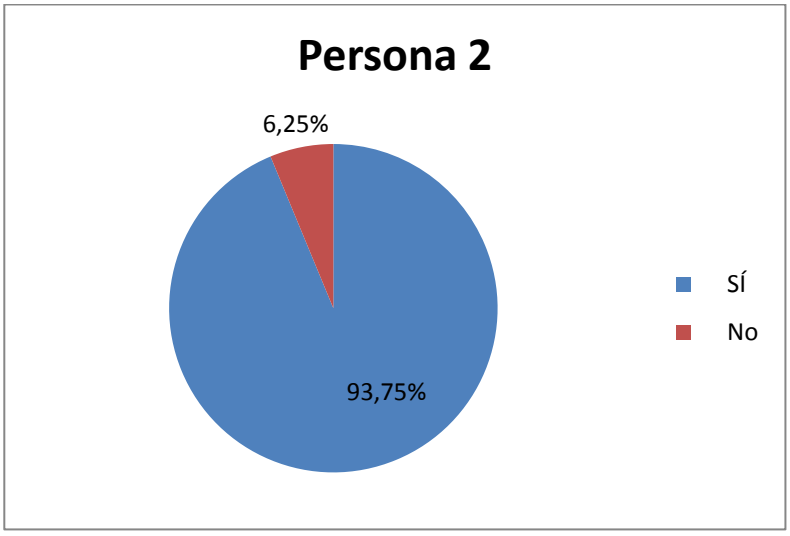
Gráfica 8.1. Inteligibilidad de las palabras.

En la gráfica 8.2 se ilustra el porcentaje de inteligibilidad percibido por la persona 1 al escuchar las 16 palabras.



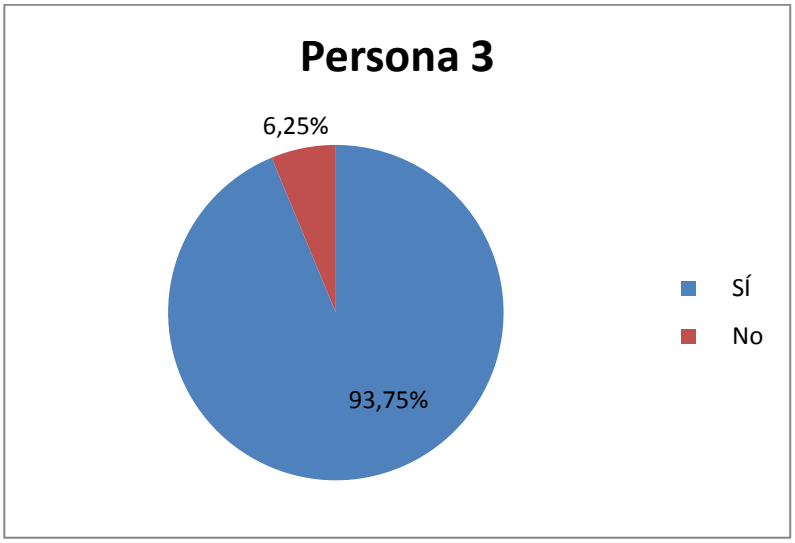
Gráfica 8.2. Porcentaje de inteligibilidad percibido por la persona 1.

En la gráfica 8.3 se ilustra el porcentaje de inteligibilidad percibido por la persona 2 al escuchar las 16 palabras.



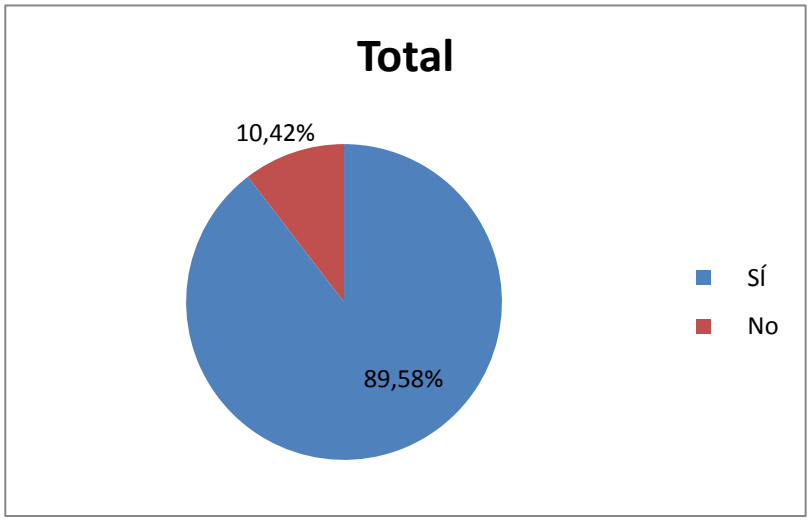
Gráfica 8.3. Porcentaje de inteligibilidad percibido por la persona 2.

En la gráfica 8.4 se ilustra el porcentaje de inteligibilidad percibido por la persona 3 al escuchar las 16 palabras.



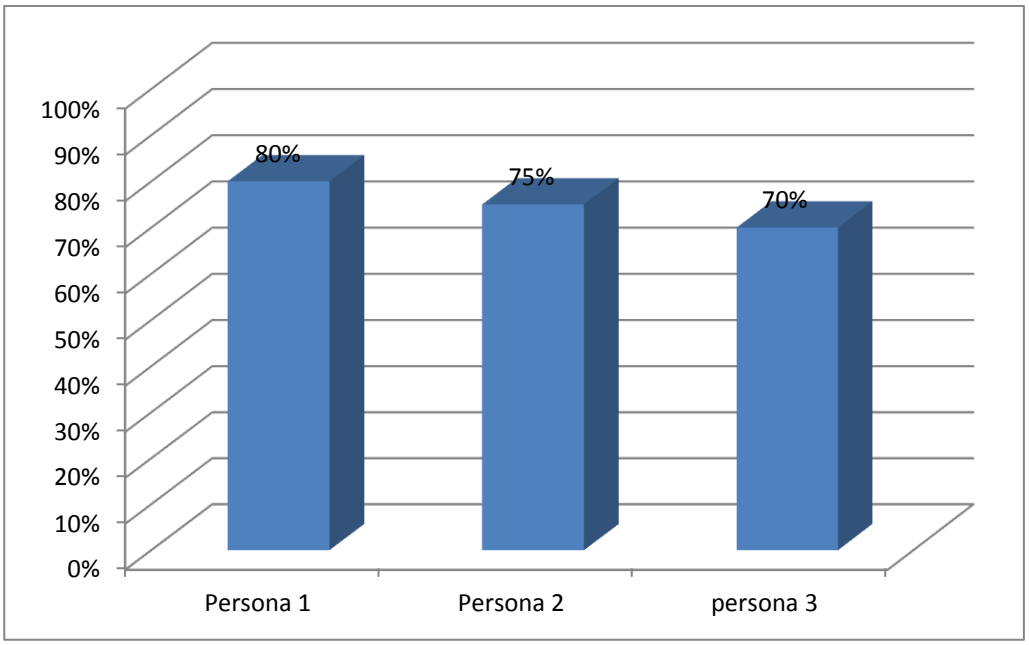
Gráfica 8.4. Porcentaje de inteligibilidad percibido por la persona 3.

En la figura 8.5 se muestra el porcentaje total de la inteligibilidad de las tres personas.



Gráfica 8.5.Porcentaje total de inteligibilidad.

Prueba 2. Se realizó a otro grupo de tres personas que escucho una lista de 20 palabras parecidas entre ellas en donde el oyente tenía que repetir la palabra escuchada, en este caso se anotan los errores y los aciertos, los resultados para cada una de las personas se muestran en la gráfica 8.6.



Gráfica 8.6. Porcentaje de aciertos en las palabras escuchadas.

Al observar las gráficas anteriores se concluye que el sintetizador de voz por concatenación de unidades construido en este trabajo es inteligible en un alto porcentaje.

Entre los resultados obtenidos podemos destacar. El sintetizador de voz por concatenación de unidades implementado es capaz de construir cualquier mensaje para el español de México y reproducir una voz sintética inteligible en tiempo real.

Al usar un corpus que contiene difonemas, mismos que tienen un comienzo y un final estables acústicamente, que por lo tanto genera una distorsión en la concatenación más baja que otras unidades, se obtuvo un sintetizador que genera una voz más natural.

La HMI generada permite una comunicación amigable entre el usuario y la máquina ya que es un sistema bastante poderoso y eficaz que también permite una conexión más sencilla y económica.
La síntesis remota implementada en este trabajo responde prácticamente en tiempo real.

Conclusiones.

Unos de los mayores retos que se presentaron en la conversión texto a voz fue dotar a la voz sintética de la suficiente naturalidad, la manera más sencilla de lograrlo y la menos costosa en cuanto a procesamiento, considerando que el sistema se implementó en una computadora con recursos muy limitados, es obteniendo unidades de concatenación grandes, idealmente palabras o frases, desafortunadamente la cantidad de combinaciones es prácticamente infinita lo que necesitaría una memoria para almacenar las unidades igualmente infinita, como alternativa a este problema se buscaron unidades de concatenación grandes y que fueran comunes en el español de México para el cual fue diseñado el sintetizador. La selección de las unidades grabadas de acuerdo al análisis realizado en la tesis doctoral: Investigación cuantitativa de afijos y clíticos del español de México: Glutinometría en el Corpus del Español Mexicano Contemporáneo de México (Urrea, 2003), generó una mejor inteligibilidad en la voz obtenida ya que no solo se obtuvieron unidades de concatenación más grandes lo que por sí solo genera más calidad sino también la probabilidad de usar estas unidades continuamente es grande pues son palabras y terminaciones muy comunes del español hablado en México de acuerdo a lo reportado en la tesis consultada.

Así mismo la tarea de buscar un locutor que tuviera buenas características acústicas como timbre, tono e intensidad para obtener una voz sintética de buena calidad resultó relevante, ya que una vez realizadas las grabaciones de donde se obtuvieron las palabras, terminaciones, fonemas y difonemas la segmentación de éstas resulte más sencilla y se puedan identificar claramente en los espectrogramas los segmentos buscados, este es uno de los pasos críticos en el diseño del sintetizador pues comúnmente resulta complicado identificar las fronteras entre las unidades buscadas provocando errores al momento de segmentarlas lo que disminuye la inteligibilidad y calidad de la voz sintética aun teniendo unidades grandes.

Otro aspecto que permite mejorar la calidad de la voz es la manipulación de la prosodia mediante técnicas especializadas de procesamiento digital de señales, que hace más natural e inteligible la voz sintética, en este trabajo no se realiza quedando como trabajo futuro.

El dotar al sistema con comunicación inalámbrica Wi-Fi, permite el uso del sintetizador en tiempo real de forma remota y desde cualquier lugar utilizando dispositivos electrónicos comunes (pc, teléfono, tablet) lo cual aumenta la posibilidad de ser utilizado en aplicaciones diferentes al HMI como pueden ser sistemas de ayuda a personas con discapacidad, medios informativos, entre otros.

El uso de una microcomputadora embebida como la tarjeta Raspberry Pi B, resultó una ventaja para el proyecto ya que cuenta con todas las características de una PC con sistema operativo Linux en un reducido tamaño. La potencia de su microprocesador y memoria RAM permiten ejecutar y compilar el programa para el sintetizador de voz, además de poder instalar un servidor web para la comunicación remota. Por otro lado, su bajo costo permitió la implementación del sistema, sin incluir los periféricos, con alrededor de $1,000.00 M.N. lo cual dio como resultado un producto portátil y de bajo costo.

Entre las implicaciones sociales relevantes que este proyecto tiene podemos mencionar:

- Permite lograr el desarrollo creativo de las personas que apliquen este tipo de metodologías, de igual manera se incrementara el ser comunicador de toda persona disponiendo de este sintetizador, ya sea persona – persona o persona - máquina.
- Establecer nuevos entornos de comunicación conocidos que dan nuevas formas de interacción de los usuarios con las maquinas donde uno y otro desempeñan roles diferentes.
- Hacer uso de este tipo de sistemas para mejorar la calidad de vida de los seres humanos creando HMI más accesibles y confortables.

Aún cuando los resultados de este trabajo fueron satisfactorios, y en general se cumplieron los objetivos propuestos, se propone el siguiente trabajo a futuro:

- La implementación de algunos de los algoritmos mencionados en el capítulo 3 para la manipulación de la prosodia del corpus, ya que estos algoritmos permiten modificar la duración, altura tonal y el volumen de segmentos de voz sin modificar el mensaje lo que mejoraría notablemente el desempeño del sintetizador.

Apéndice A Palabras más comunes del español en México

A	durante	cuatro	Pesos
Había	he	la	Poco
Agua	hecho	las	Por
Ahí	ejemplo	le	Porque
Aquí	el	les	Pues
Al	en	lo	Primer
Algo	entonces	los	Primera
Algunos	entre	luego	Puede
Han	era	lugar	Pueden
Año	es	manera	Se
Años	esa	mas	Sea
Antes	ese	mayor	Señor
Ahora	eso	Me	Ser
Has	esos	medio	Si
Hace	esta	mejor	Sido
Hacer	estaba	menos	Siempre
así	estado	mil	Sin
hacia	están	mis	Sino
hasta	estas	misma	Sobre
aunque	este	mismo	Solo
halla	esto	mucho	Son
allí	estos	muchos	Su
van	ella	muy	Sus
ver	ellos	mujer	Tal
verdad	forma	mundo	También
ves	fue	nada	Tanto
veces	fueron	nacional	Te
vida	gran	ni	Tengo
bien	grandes	no	Tenia

voy	y	nos	Tiempo
bueno	ir	nosotros	Tiene
da	general	nuestro	Tienen
de	gente	nunca	Toda
debe	cada	o	Todas
del	caza	hoy	Todo
dentro	casi	hombre	Todos
desde	caso	hora	Trabajo
decir	que	otra	Tres
después	quien	otras	Tu
dia	como	otro	Un
días	con	otros	Una
digo	cosa	papa	Uno
dijo	cosas	país	Unos
dice	creo	para	Usted
donde	cual	parte	Ya
dos	cuando	pero	Yo

Apéndice B Combinaciones de difonemas existentes en México

aa	ua	qb	oc	od	qe	uf	hh	ip	kj	xk	my	ñr	ps	tr	vw
ab	Va	rb	pc	pd	re	vf	hi	iq	lj	yk	mz	ñs	pt	ur	vx
ac	Wa	sb	qc	qd	se	wf	hj	ir	mj	zk	nm	ñt	pu	vr	vy
ad	Xa	tb	rc	rd	te	xf	hk	is	nj	ll	ñm	ñu	pv	wr	vz
ae	Ya	ub	sc	sd	ue	yf	hl	it	ñj	lm	om	ñv	pw	xr	wv
af	Za	vb	tc	td	ve	zf	hm	iu	oj	ln	pm	ñw	px	yr	xv
ag	Bb	wb	uc	ud	we	gg	hn	iv	pj	lñ	qm	ñx	py	zr	yv
ah	Bc	xb	vc	vd	xe	gh	hñ	iw	qj	lo	rm	ñy	pz	ss	zv
ai	Bd	yb	wc	wd	ye	gi	ho	ix	rj	lp	sm	ñz	qp	st	ww
aj	Be	zb	xc	xd	ze	gj	hp	iy	sj	lq	tm	oñ	rp	su	wx
ak	Bf	cc	yc	yd	ff	gk	hq	iz	tj	lr	um	pñ	sp	sv	wy
al	Bg	cd	zc	zd	fg	gl	hr	ji	uj	ls	vm	qñ	Tp	sw	wz
am	Bh	ce	dd	ee	fh	gm	hs	ki	vj	lt	wm	rñ	up	sx	xw
an	Bi	cf	de	ef	fi	gn	ht	li	wj	lu	xm	sñ	vp	sy	yw
añ	Bj	cg	df	eg	fj	gñ	hu	mi	xj	lv	ym	tñ	wp	sz	zw
ao	Bk	ch	dg	eh	fk	go	hv	ni	yj	lw	zm	uñ	xp	ts	xx
ap	Bl	ci	dh	ei	fl	gp	hw	ñi	zj	lx	nn	vñ	yp	us	xy
aq	Bm	cj	di	ej	fm	gq	hx	oi	kk	ly	nñ	wñ	zp	vs	xz
ar	Bn	ck	dj	ek	fn	gr	hy	pi	kl	lz	no	xñ	qq	ws	yx
as	Bñ	cl	dk	el	fñ	gs	hz	qi	km	ml	np	yñ	rq	xs	zx
at	Bo	cm	dl	em	fo	gt	ih	ri	kn	nl	nq	zñ	sq	ys	yy
au	Bp	cn	dm	en	fp	gu	jh	si	kñ	ñl	nr	oo	Tq	zs	yz
av	Bq	cñ	dn	eñ	fq	gv	kh	ti	ko	ol	ns	op	uq	tt	zy
aw	Br	co	dñ	eo	fr	gw	lh	ui	kp	pl	nt	oq	vq	tu	zz
ax	Bs	cp	do	ep	fs	gx	mh	vi	kq	ql	nu	or	wq	tv	
ay	Bt	cq	dp	eq	ft	gy	nh	wi	kr	rl	nv	os	xq	tw	
az	Bu	cr	dq	er	fu	gz	ñh	xi	ks	sl	nw	ot	yq	tx	
ba	Bv	cs	dr	es	fv	hg	oh	yi	kt	tl	nx	ou	zq	ty	
ca	Bw	ct	ds	et	fw	ig	ph	zi	ku	ul	ny	ov	qr	tz	
da	Bx	cu	dt	eu	fx	jg	qh	jj	kv	vl	nz	ow	qs	ut	
ea	By	cv	du	ev	fy	kg	rh	jk	kw	wl	ñn	ox	qt	vt	
fa	Bz	cw	dv	ew	fz	lg	sh	jl	kx	xl	on	oy	qu	wt	
ga	Cb	cx	dw	ex	gf	mg	th	jm	ky	yl	pn	oz	qv	xt	
ha	Db	cy	dx	ey	hf	ng	uh	jn	kz	zl	qn	po	qw	yt	
ia	Eb	cz	dy	ez	if	ñg	vh	jñ	lk	mm	rn	qo	qx	zt	
ja	Fb	dc	dz	fe	jf	og	wh	jo	mk	mn	sn	ro	qy	uu	
ka	Gb	ec	ed	ge	kf	pg	xh	jp	nk	mñ	tn	so	qz	uv	
la	Hb	fc	fd	he	lf	qg	yh	jq	ñk	mo	un	to	Rr	uw	
ma	Ib	gc	gd	ie	mf	rg	zh	jr	ok	mp	vn	uo	rs	ux	

na	Jb	hc	hd	je	nf	sg	ii	js	pk	mq	wn	vo	Rt	uy
ña	Kb	ic	id	ke	ñf	tg	ij	jt	qk	mr	xn	wo	ru	uz
oa	Lb	jc	jd	le	of	ug	ik	ju	rk	ms	yn	xo	rv	vu
pa	Mb	kc	kd	me	pf	vg	il	jv	sk	mt	zn	yo	rw	wu
qa	Nb	lc	ld	ne	qf	wg	im	jw	tk	mu	ññ	zo	rx	xu
ra	Ñb	mc	md	ñe	rf	xg	in	jx	uk	mv	ño	pp	ry	yu
sa	Ob	nc	nd	oe	sf	yg	iñ	jy	vk	mw	ñp	pq	rz	zu
ta	Pb	ñc	ñd	pe	tf	zg	io	jz	wk	mx	ñq	pr	sr	vv

Apéndice C Código de la página web

```
<html>
<head>
  <meta name='viewport' content='width=device-width; initial-scale=1.0;user-scalable=no'>
<title>  MARISOL</title>
  <style>
      * { margin: 0; padding: 0; box-sizing: border-box; background-image:url("uacm1.png"); background-repeat:
no-repeat; background-position: 550px 40px;}
 body { font: 13px Helvetica, Arial; }
    form { background: #000; padding: 3px; position: fixed; bottom: 0; width: 100%;}
    form input { border: 0; padding: 10px; width: 90%; margin-right: .5%; }
    form button { width: 9%; background: #0080FF; border: 0; padding: 10px; }
    #messages { list-style-type: none; margin: 0; padding: 0; }
    #messages li { padding: 5px 10px; }
    #messages li:nth-child(odd) { background: #fff; }
</style>
</head>
<body><center>
<h1>Sintetizador de Voz Concatenativo para interfaces Hombre-Máquina</h1>
<hr color="#07190B"/>
</center>
    <ul id="messages"></ul>
    <form name="formulario"method="get"action="">
    <input id="m" autocomplete="off" name="nombre" />
    <button>Send</button>
    </form> <?php
   $mensaje = $_GET['nombre'];
   $result = exec("./listo $mensaje");
   echo $mensaje  ?>
  <embed src ="CORPUS/salida.wav" hidden=true autostart=autoplay>
  </body>
  </html>
```

Bibliografía

Agramunt, V. (2014). *Direccionamiento IP, calculo de redes TCP/IP.*

Ancinas, M. (2004). *"Habilidades de la comunicacion y estrategias asistenciales en el ambito sanitario".* Andalucia: ALCALA.

Bonafonte, A. (2012). *sintesis de voz aplicada a la traduccion voz a voz.* Barcelona: UPC.

Bullón, L. (1994). *Conversión de texto a voz en castellano aplicando algoritmoi PSOLA.* Valencia: DSIC.

Carballar, A. (2008). *Wi-FI,instalación, seguridad y aplicaciones.* México: Alfaomega.

Cerviño, I. (2012). *estudio e implementación de un codificador de voz .* México.

Cobo, R. (Abril de 2013). *http://www.aie.c*. Recuperado el 16 de febrero de 2015, de http://www.aie.c: http://www.aie.cl/files/file/comites/ca/abc/hmi.pdf

Escudero, A. (2007). *Estandares en tecnologias inalambricas.* Sudafrica: TRICALCAR.

Frías, X. (2000). *"Introducción lingüística".* España: Ianua.

Furui. (1989). *"Panorama de los sistemas de texto a voz".*

Genoveva, V. R. (2008). *Sistema de reconocimiento de voz.* Guatemala.

Hidalgo, A. (2012). *La voz del lenguaje fonética y fonologia del español.* Valencia: Tirant Humanidades.

http://www.emus.edu.uy. (2011). Recuperado el 10 de septiembre de 2014, de http://www.eumus.edu.uy/eme/ensenanza/electivas/dsp/presentaciones/clase06.pdf

Hunt, A. (1996). *"Unit selection in aconcatenative speech synthesis system using a large speech database".* Japon: ATR.

Huse, B. (3 de Diciembre de 2011). *http://www.robotics.org/.* Obtenido de Robotics on line: http://www.robotics.org/content-detail.cfm/Industrial-Robotics-Industry-Insights/How-Robots-Will-Affect-Future-Generations/content_id/834

Jimenez, J. (1996). *Formato Wav.* México: Hunab.

Klatt, D. (1980). *Software for a cascade,parallel formant synthesizer.* Acoust.

Leon, A. (2004). *"Communication Networks".* Catalan: Mc GrawHill.

Llisterri, J. (2005). *"Corpus Orales para el desarrollo de las tecnologías del habla en español".* Barcelona: Espanyola.

Martinez, E. (1995). *"Fonetica experimental: teoria y practica".* España: SINTESIS.

Montiel, S. (2006). *"Sintetizador basico de voz".* México.

Morales, A. (2011). *tecnicas de reconocimiento robusto de la voz basados en el pitch.* Granada.

Moulines, F. (1989). *Pitch synchronous waveformprocessing techniques for text-to-speech syntheis using difhones.* Europa.

Placio, K. (2003). *"Diseño e Implementación de un sistema de síntesis de voz".* Ecuador: Universidad Politecnica Salesiana.

Quilis, A. (1990). *"Lingüística Española aplicada a la terapia del lenguaje".* España: GREDOS.

Stalings, W. (2004). *"Comunicaciones y redes de computadoras".* España: Pearson prentice hall.

Urrea, A. (2003). *"Investigación cuantitativa de afijos y cliticos del español de México: Glutinometría en el corpus del español México contemporaneo".* México: Colegio de México Centro de estudios lingüísticos y literarios.

Vega, L. (2007). *diseño de un sisntetizador de voz del español hablado en México.* México: facultad de ingenieria.