Capítulo 2 Sintetizadores y Corpus de voz

2.1 ¿Qué es un Sintetizador de Voz?

Con el paso del tiempo la tecnología del habla ha englobado conocimientos que pertenecen a distintas áreas, entre ellas la lingüística, la acústica, la psicología, el procesado de señal y la inteligencia artificial. Todos estos campos están relacionados, por lo que utilizaré conceptos de ellos para la realización de un sintetizador de voz, en el cual centro mi trabajo.

La finalidad del sintetizador es realizar la conversión automática de un texto a la voz sintetizada. El progreso en este campo ha sido posible gracias a importantes avances en la teoría lingüística, en el modelado acústico-fonético de los sonidos, en la generación de voz artificial y en el diseño de las computadoras, tanto a nivel software como hardware.

Se puede definir la transformación de texto a voz como un proceso compuesto de dos pasos [Montero, 2002]:

Primero: Analizar el texto de entrada para determinar la estructura de la frase y la composición fonética de cada palabra (procesado lingüístico-prosódico)

Segundo: Transformar esta representación lingüística abstracta en voz (procesado acústico).

Los sistemas de síntesis de voz se suelen clasificar en función del método seguido para la reconstrucción de la voz.

2.2 Tipos de Sintetizadores

Los sintetizadores están divididos por su forma de trabajar y como fueron realizados para la reproducción de un mejor sonido y no sea tan robotizado. Entre ellos tenemos los articulatorios, los sintetizadores por formantes, derivados de las técnicas de predicción lineal (LPC) y los sintetizadores por concatenación de forma de onda.

.

2.2.1 Sintetizadores articulatorios

Modelos físicos basados en los mecanismos fisiológicos del aparato fonador, véase en la siguiente figura.

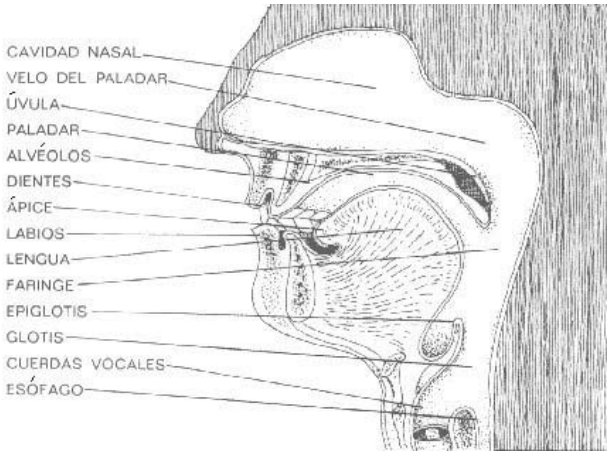

Figura 2.1 Partes del Aparato Fonador.

Los sintetizadores articulatorios utilizan parámetros como el tamaño de la cavidad oral, la tráquea y la posición de la lengua, entre otras variables. Estos factores se relacionan entre sí para producir una voz que se asemeje en mayor medida de lo posible a la voz humana [Cuétara, 2002]. Ésta aplica señales armónicas a la señal sonora y establecen una analogía entre parámetros relacionados con los órganos articulatorios, sus movimientos y parámetros. Los sintetizadores articulatorios proporcionan voz

sintética de alta calidad, pero su inconveniente es que los parámetros son muy difíciles de obtener y controlarlos automáticamente.

El problema principal de los modelos articulatorios es, por un lado, la enorme cantidad de parámetros internos de control que precisan y dificultan la coordinación y derivación de los parámetros de control disponibles a la entrada del sintetizador; y por otro lado, 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, cosa muy difícil de medir en estas condiciones [Rodríguez, 2002].

2.2.2 Sintetizadores por formantes

Modelan el tracto vocal a través de un conjunto de filtros, excitados por fuentes que simulan las cuerdas vocales. Este tipo de sintetizadores tiene una amplia difusión pero la calidad de la voz sintetizada es menor [Montero, 2002].

Estos sintetizadores están basados en la teoría acústica de producción de voz, según la cual, una o más fuentes sonoras excitan un filtro lineal, el tracto vocal, dando como resultado la señal de voz. La excitación del tracto vocal es debida a la vibración de las cuerdas vocales, las cuales producen un obstáculo al paso del aire originando los sonidos.

La fuente de voz usada en los sintetizadores ha ido variando a lo largo del tiempo. Los modelos más antiguos utilizaban trenes de impulsos o dientes de sierra. Más tarde se pasó a modelos matemáticos cada vez más complejos que permiten

controlar los parámetros principales de la señal glotal: frecuencia fundamental, amplitud, tiempo de apertura de la glotis en un período, etc. La glotis es un espacio limitado entre las cuerdas vocales verdaderas, y por lo tanto no esta irrigado, ni inervado.

2.2.3 Sintetizadores derivados de las técnicas de predicción lineal

Son sintetizadores de análisis-síntesis que trabajan con parámetros Lineal Predictive Coding (LPC) para controlar la función de transferencia del filtro que simula el tracto vocal. LPC es una de las técnicas de mayor alcance del análisis del habla y uno de los métodos más útiles para codificar el habla con calidad en un índice binario. Proporciona estimaciones extremadamente exactas de los parámetros del habla, el cual es relativamente eficiente.

El LPC comienza asumiendo que la señal del habla es producida por un zumbador en el extremo de un tubo. La glotis (el espacio entre las cuerdas vocales) produce el zumbido, que es caracterizado por su intensidad y la frecuencia (pitch). La zona vocal (la garganta y la boca) forma el tubo, que es caracterizado por sus resonancias, que se llaman "*los formantes*".

El LPC analiza la señal del habla estimando los formantes, quitando sus efectos de la señal del habla, y estimando la intensidad y la frecuencia del zumbido restante. Al proceso de quitar los formantes se le llama *filtración inversa*, y la señal restante se llama *el residuo.* Los números que describen los formantes y el residuo se pueden almacenar o transmitir en alguna parte.

El LPC sintetiza la señal del habla invirtiendo el proceso: utiliza el residuo para crear la señal fuente, utiliza los formantes para crear un filtro (que represente el tubo), y funcione la fuente a través del filtro, dando como resultado el habla.

Ya que las señales varían con el tiempo, este proceso se hace en los segmentos cortos de la señal del habla, que se llaman *frames*. Generalmente con 30 a 50 frames por segundo y una buena compresión se puede hacer una buena pronunciación.

El problema básico del sistema del LPC es determinar los formantes de la señal del habla. La solución básicamente es una ecuación diferencial, que expresa cada muestra de la señal como combinación lineal de muestras anteriores. Tal ecuación se llama *lineal predictor*, por esa razón se le llama Lineal Predictive Coding. [Howitt, 1995]

2.2.4 Sintetizadores por concatenación de forma de onda

Estos sintetizadores concatenan unidades pregrabadas para generar frases. Pretenden mejorar la calidad de la voz sintetizada minimizando el ruido de codificación. Tienen una complejidad variable pero la voz generada es de alta calidad [Montero, 2002].

Principalmente se basan en tener un conjunto de pequeños segmentos de voz tomados de un hablante, que se van concatenando para formar el discurso deseado. La unidad elegida para la concatenación es el parámetro clave de estos sintetizadores. Para decidir el tamaño y número de estas unidades hay un compromiso entre la calidad de la voz que se quiere sintetizar y limitaciones de memoria de datos.

Los fragmentos grabados no pueden ser palabras por dos razones fundamentales. La primera es que la manera de pronunciar una frase es muy distinta a la de una secuencia de palabras aisladas, ya que en una frase la duración de las palabras es más corta. Además, la concatenación de palabras grabadas aisladamente produce una voz sintética de baja calidad, poco natural. La segunda razón es el elevado número de palabras existentes en un idioma concreto, lo que implicaría una gran cantidad de memoria para almacenar las unidades pregrabadas.

La sílaba es una unidad lingüísticamente interesante pero tiene el inconveniente del número de sílabas diferentes que hay. La información sobre la división en sílabas de la palabra es necesaria para poder determinar la acentuación fonética. Además, influye en la decisión de los alófonos, controla las reglas de entonación y pausado, e indica cómo deben tratarse las palabras ilegibles (siglas o palabras no válidas). Por esa razón se ha considerado la importancia de comenzar a realizar este proyecto con el lenguaje para niños, ya que, con él comenzamos a hablar todos los seres humanos.

Se ha fijado el objetivo de utilizar Unit Selection, el cual es un algoritmo que concatena archivos de tipo texto (.txt), palabra (.wrd) y fonema (.phn). Este algoritmo lo desarrolló Leonardo Flores en el 2001 como tesis de licenciatura el cual se mencionará más adelante. También se ha considerado una cantidad de sílabas que podría abastecer el protocolo de grabación, así como el corpus de voz para la realización más eficiente, debido a que se considera que agrupan casi todos los posibles sonidos del lenguaje (Véase Apéndice B). El fonema es otra unidad a considerar, en español hay cinco vocales con sus variantes fonéticas, cuarenta y dos consonantes y once signos diacríticos (Véase Apéndice A).

Uno de los problemas más importantes es la coarticulación, ya que éste fenómeno tiende a minimizarse en el centro acústico de un fonema. Debido a esto, Peterson ayuda a realizar la concatenación por difonemas, es decir el fragmento de voz que va desde la mitad del fonema a la mitad siguiente, dicho de otra manera, la transición entre sonidos. Ejemplos de difonemas podían ser: [a-b], [r-a], [o-l].

Los sintetizadores por formantes permiten manipular las características de la fuente de voz. Por el contrario, en los sintetizadores por concatenación la fuente de voz es única y corresponde a la grabación de los difonemas, lo que debe realizarse por un locutor capaz de controlar y mantener constante la calidad de la voz para evitar cambios repentinos.

En cuanto a la calidad de la voz sintética, con el método de concatenación se consiguen mejores resultados. Además, la síntesis por concatenación permite alcanzar un alto grado de naturalidad.

2.3 Sintetizadores de Hoy

OGI-Festival

El CSLU (Center for Spoken Languages Understanding) del Oregon Graduate Institute (OGI) es uno de los grupos de trabajo sobre procesamiento de lenguaje natural más importantes del mundo.

Sintetizador Festival http://cslu.cse.ogi.edu/tts/

Bell Labs

Síntesis de voz de los Laboratorios Bell http://www1.bell-labs.com/project/tts/

ATLAS

Sintetizador ATLAS (Applied Technologies on Language and Speech).

http://www.atlas-cti.com/es/demotts.htm

Loquendo

Sintetizador de Loquendo http://www.loquendo.com/es/demos/demo_tts.htm

2.4 ¿Que es un corpus de voz?

Un corpus lingüístico es un conjunto de textos almacenados en formatos electrónicos y agrupados con el fin de estudiar una lengua o una determinada variedad lingüística [Montero, 2002].

El objetivo es construir un corupus de voz con elementos de referencias para el estudio de una fase concreta en un cierto aspecto de la lengua. En nuestro caso, este corpus se realizará con grabaciones buscando tener todos los posibles sonidos del lenguaje español hablado en México.

Básicamente, existen dos tipos de corpus llamados textuales y orales, cada uno de ellos se dividen en grupos según su finalidad. Los textuales se dividen en corpus de la lengua general y corpus de un sub – lenguaje. Los orales se dividen en corpus para el

estudio del lenguaje oral, para fines específicos, para el desarrollo de aplicaciones tecnológicas del habla y para el desarrollo de aplicaciones especificas, de los cuales daremos mas adelante un ejemplo con una explicación de su finalidad.

2.4.1 Corpus textuales

Es un Corpus que recoge íntegramente todos los textos de los documentos que lo constituyen. Se entiende como textos las series de frases y/o párrafos coherentes, homogéneos estilísticamente y completos en sí mismos [Torruella, 1999].

2.4.1.1 Corpus de la lengua general con fines generales.

El objetivo de este tipo de corpus es construir una fuente de información textual del español para fines diversos, como por ejemplo el CREA, "Corpus de referencia del español actual" desarrollado por el instituto de lexicografía de la Real Academia de la Lengua Española, ésta contiene textos literarios, periodísticos, científicos, técnicos, transcripciones de grabaciones de la lengua oral y de medios de comunicación desde el año de 1975 a la actualidad, este corpus cuenta con un total hasta el año de 1997, de cien millones de frases intermedias y su tamaño final fue de doscientos millones en el año 2000, en la actualidad se encuentra en desarrollo.[RAE, 2003]

2.4.1.2 Corpus de la lengua general con fines específicos.

Este corpus pretende dar respuesta a problemas específicos, como el estudio de determinados aspectos de la gramática o léxico de la lengua, la extracción de datos estadísticos, el desarrollo y evaluación de sistemas de procesamiento del lenguaje, etc.

Por ejemplo CorVerifSDGEE, “Corpus de Verificación del Sistema de Diccionarios y Gramáticas Electrónicos del Español”. Éste está directamente relacionado con el sistema de diccionarios y gramáticas electrónicas del español, se ha desarrollado en la Universidad Autónoma de Barcelona y se encuentra en constante aplicación. En estos momentos cuenta con un tamaño de tres millones de palabras y se encuentra en desarrollo [GIL, 2002].

2.4.1.3 Corpus de un sub-lenguaje con fines específicos.

El Corpus de IBM España, contiene a una gran variedad de tipos de textos y cuya finalidad ha sido la extracción de datos estadísticos para el modelo de lenguaje utilizado en el proyecto TANGORA. Este proyecto se trata de un sistema dependiente del locutor para grandes vocabularios. Su principal interés es un proceso de adaptación a un nuevo locutor que requiere 20 minutos para leer 100 frases de 1,200 palabras, 700 de las cuales son distintas.

2.4.1.4 Corpus de un sub-lenguaje con fines generales.

El “Corpus textual del español periodístico” actualmente está en desarrollo en la Universidad de Barcelona.

2.4.2 Corpus orales

Este tipo de corpus están dedicados al estudio fonológico.

2.4.2.1 Corpus para el estudio del lenguaje oral

El corpus tiene como objetivo principal caracterizar desde un punto de vista lingüístico la lengua del habla y pueden ser generales ó para fines específicos. Para los estudios con

fines generales se hizo un proyecto en la Universidad Autónoma de Madrid llamado corpus de referencia del español contemporáneo. En el caso de los estudios con fines específicos los corpus orales sobre la difusión del español son por radio, televisión y prensa: unidad y diversidad de la lengua.

2.4.2.2 Corpus para el desarrollo de aplicaciones tecnológicas del habla.

El objetivo de este corpus, es desarrollar aplicaciones para el entrenamiento y evaluación de sistemas de reconocimiento. En este tipo de corpus encontramos el corpus Fraga desarrollado en la Universidad de las Américas en Puebla, en el centro de investigación Tlatoa.

2.4.2.3 Corpus para el desarrollo de aplicaciones específicas

Reconocimiento del habla para aplicaciones telefónicas, como el corpus "Tlatoa Common Questions Corpus" es una base de datos de voz grabada por teléfono de personas que hablan español mexicano, con un número de 400 locutores, usando la lengua nativa del español, dependiendo de la region del país en que se encuentre. Fue diseñado para cubrir adecuadamente vocabularios comunes, como son los dígitos, números naturales, si/no, horas, días, meses, fechas, nombres y apellidos, etc. [Olivier, 2000]

2.5 Corpus GAMA

Este corpus es llamado Gama, contiene 750 frases, 1750 palabras y una gramática numérica usando frases, chistes, refranes, ciencia, tecnología, política, deportes, etc.,

con el fin de no aburrir al locutor y tener una voz continua y clara, logrando grabar un total de 1400 archivos de audio con voz clara y uniforme. Con éste podremos comparar el corpus Fraga realizado con documentos de periódicos y revistas, que no puede reproducir las combinaciones necesarias para cubrir el lenguaje infantil.

Este corpus estará grabado con la voz del Lic. Adolfo Madrid el cual se identifica con el nombre de Timo y la Lic. Lizy Arlette Barclay con el nombre de Lizy. La grabación de voz femenina y masculina es necesaria para este trabajo, debido a que el colegio de México [Perissinotto, 1975] realizó un estudio fonológico por género. El cual arroja resultados como que existen diferencias de pronunciación significativas entre el hombre y la mujer. Este punto es importante para la realización de este corpus, ya que psicológicamente algunos niños podrían preferir el tono de voz masculina o femenina y así lograr su mayor atención y concentración para su aprendizaje.

Con esta información podemos darnos cuenta que el tipo de corpus que se realizará es para generar un protocolo de grabación con fines generales, asimismo, tener una mejor cantidad de sonidos y lograr una herramienta específica utilizando un sintetizador por concatenación llamado Unit Selection, basado en la tesis y sistema, realizados por el Ing. Leonardo Flores egresado de la Universidad de las Américas – Puebla [Flores, 2001].

2.6 Corpus Fraga

El corpus de voz con el nombre de Fraga, fue grabado por un locutor profesional en la cabina de sonido de la UDLA, bajo condiciones excepcionales que lo libran de ruido. Está constituido por 800 frases obtenidas de revistas, periódicos y documentos políticos,

con calidad de 48,000 muestras por segundo (48Khz). Estas frases fueron leídas de manera monótona y además fueron grabadas junto con una señal adicional proveniente de un laringógrafo. Las palabras y frases del corpus fueron cuidadosamente seleccionadas por el Dr. Andrew Cronk del OGI (Oregon Graduate Institute) con el propósito de hacer estudios de prosodia, duración e investigación en la técnica de Unit Selection [Flores, 2001].