Capitulo II. Síntesis de Voz

La Síntesis de voz, como ya lo mencionábamos, es traducir o procesar una voz o señal de audio a partir de un texto dado [Schmandt, 1994], nos sirve como herramienta para apoyar el crecimiento en los avances que hoy en día se están dando. En si es una emulación que realiza el ser humano por medio de las cuerdas vocales.

La síntesis de Voz, realizada a través de los TTS (Text to Speech) o sistemas de procesamiento de texto al habla, son sistemas que transforman texto introducido (ya sea por algún operador o capturado por otro medio, como el OCR, Optical Code Register) en sonidos que podemos reconocer como voz [Meza, 1999].

2.1 Métodos existentes para la síntesis de voz

Sintetizadores Articulatorios

El objetivo de estos sintetizadores es controlar un modelo del aparato fonador, de manera similar como lo hace el cerebro al construir los parámetros circuitales, estos tienen la dificultad para la obtención de los parámetros, es decir, presenta dificultades en el análisis de la posición de los órganos articulatorios de una persona que habla normalmente. Esto hace que no sean desarrollados con frecuencia.

Todo tipo de sintetizador que contenga este tipo de síntesis, están basados en mecanismos naturales del procesamiento de la voz, sus parámetros son, el tamaño de la cavidad oral, la traquea y la posición en la que se encuentre la lengua. [Meza, 1999]. El habla o las producciones de palabras que son generadas por el ser humano, son analizadas de acuerdo con un modelo de producción de la misma, y esta a su vez almacena los valores característicos como secuencias en el tiempo.

Al poder almacenar de esta manera la información, redituara en varias ventajas, ya que se reducirá la cantidad de información así como los parámetros que sean almacenados. Con esto se pude controlar los ritmos, entonaciones y expresiones dentro de las pronunciaciones, pero se corre el riesgo de afectar la naturalidad de la voz, y esto causaría que el sistema no funcionara correctamente.

Dentro de los términos reales, las aplicaciones comerciales no existen, ya que los experimentos han sido realizados para la realización y comprobación de esta técnica, y son muy costosos para poder comercializarse. [Barbosa, 1997].

Sintetizadores por Formantes

Constituidos por filtros que tienen la tarea de modelar la resonancia del tracto vocal, su ventaja es que trabaja de manera directa con los parámetros que mantienen una comunicación directa con el habla, además de que son fácilmente manipulables en el control del sintetizador.

Síntetizadores paramétricos

Es la emulación de la onda sonora que reproduce el ser humano, esto se da cuando se copian los patrones al formarlas, y son líneas y picos de energía que pueden apreciarse en un espectrograma.

Algunas de las resonancias entre las que se encuentran la nasal y la oral, no son mezcladas, pero si existe un cambio en el movimiento dentro de los órganos articulatorios. En particular, donde se encuentran estos órganos, principalmente los articulatorios, existe un formante que produce este cambio en una posición, y se le llama frecuencia fundamental y es denominado como (f1), y así consecutivamente.

El identificador entre persona y persona es conocido como la frecuencia fundamental, ya que este varia dependiendo del modo, énfasis, y expresiones con las que sea pronunciada, pero la magnitud y la relación de las frecuencias de los formantes, es la que facilita que la voz pueda o no ser identificable. [Rowden, 1991].

Sintetizadores por concatenación

Estos sintetizadores intentan reducir al máximo el ruido de la codificación, y se realiza por medio de una concatenación de unidades digitalizadas que son grabadas previamente y es ajustada a la nueva producción de frase por medio de la prosodia original. En especial, dentro de estos sintetizadores se encuentran los que utilizan la selección de unidades por medio de una concatenación dependiendo de sus características prosódicas. Este será nuestro caso y el tipo de sintetizador será en base a lo llamado UNIT SELECTION.

Un sintetizador del tipo concatenativo, es decir que “forma la voz pegando unidades de voz digitalizadas”, como lo son los fonemas, di fonemas, silabas, etc. [Meza, 1999]. Dentro de la investigación se encontró que este tipo de sintetizador es usado por sistema o programa de síntesis de voz, llamado Festival TTS, con el cual se trabaja para ofrecer mejores resultados.

Algunos de los segmentos que son utilizados en la síntesis concatenativa, “son almacenadas a partir de grabaciones hechas por algún locutor con el propósito de conservar las propiedades fonológicas de los segmentos” [Barbosa, 1997]. Este sintetizador debe elegir algunos o los mejores candidatos para poder ser concatenados de acuerdo con la trascripción fonética que se hay realizado con anterioridad para después ser concatenada.

La concatenación se puede dar de varias formas:

A partir de fonemas: son las unidades naturales que dan plasticidad a los sistemas de voz, y es costeable por el contenido de unidades. Dentro del español hay 18 consonantes, 23 fonemas y 5 vocales [Uraga, 1999], pero a su vez estas están sometidas a variaciones contextuales, y dado esto se puede tener una mala pronunciación, por la calidad generada.

A partir de di fonemas: Estos, son las unidades que se consideran coarticuladas, ya que dependen del contexto que se encuentre a sus lados, es decir a la derecha o a la izquierda. Específicamente es la unión de dos fonemas, y existen 23^2 posibles di fonemas, que son el resultado de la combinación de las 23 unidades que se manejan en el vocabulario de nuestra lengua.

A partir de trifonemas: Es el tipo de concatenación de mejor calidad ya que las coarticulaciones son generadas a partir de los contextos que contienen una parte derecha y una izquierda, es decir, toma la mitad del primer fonema, el segundo fonema es tomado por completo, y el tercero lo toma solo en su mitad. El inconveniente que se presenta en este tipo de concatenación, es que no todas las frases se pueden representar por este medio, y es cuando recurre a los fonemas y di fonemas para poder realizar la concatenación de una frase.

A partir de la Concatenación de Sílabas: Este tipo de concatenación es usado en el proyecto realizado por Leonardo Flores, lo realiza a partir de las longitudes en la coarticulación, hace uso de los fonemas y trifonemas y hace referencia a que las unidades pueden ser mas grandes y mas completas según [Rownden, 1991].

A partir de la Concatenación de Palabras: Esta concatenación es la del mas alto nivel, ya que se obtiene mayor naturalidad en la voz, y es la que principalmente utilizaremos en el desarrollo de este proyecto, usando un corpus de voz mas grande, realizado por Martín, compañero que realizo la creación del corpus de voz a usar, este corpus se ha etiquetado en su mayoría e incluye dos tipos de voz, es decir de hombre y de mujer,

2.2 Como crear la Voz en una Computadora

El poder convertir un texto, cualquiera que este sea, a una señal de audio, es el propósito fundamental de esta investigación, en la actualidad se cuenta con un sintetizador dentro de laboratorio de automatización de voz (TLATOA) dentro de la UDLA, tomando en cuenta el trabajo de otro compañero, se mejorara el desarrollo del sintetizador ya creado, para obtener una mejor claridad de sonido y pronunciación.

Es necesario conocer los dispositivos que, dentro de los sistemas computacionales nos ayudaran a la creación de dicho sintetizador así como los elementos requeridos para el procesamiento de la voz y señal, así que esto se explicara en este capitulo.

2.2.1 Cómo se Procesa la voz

En las tecnologías actuales existen preocupación por la manipulación de cómo se lleva a cabo el proceso de una voz a través de una computadora, pero para esto esta el reconocimiento y la síntesis de voz, estos dos son esenciales para los medios de comunicación como la comunicación Humano – Computadora.

El hombre produce una señal acústica, que da como resultado lo que llamamos reconocimiento de voz, a diferencia de la síntesis, ya es lo contrario, esta va de la traducción de un texto a la señal de audio.

Dado que todo este proceso necesita ser llevado a cabo en lo mejor posible, necesitamos poder interpretar claramente lo que es la señal de voz. Para ello contamos con la digitalización del audio con las transformaciones de Fourier. Estas transformaciones están dadas por unos sinusoides que integran la señal, se da en una muestra de estos mismos para que puedan ser representadas.

El teorema de Nyquist [Witten, 1986], nos dice que el muestreo de la frecuencia necesaria para poder convertir una señal análoga a digital se necesita el doble de la señal de voz para que esta pueda ser procesada, es decir necesitamos entre (8 Khz.) si es que la señal contiene de (0 a 4 khz.) [Vargas, 2001], y para que pueda reproducirse en un sistema se alta calidad son necesarios 16 khz.

Ya que la voz que hemos procesado ha sido totalmente digitalizada, es necesario que se lleve a cabo un proceso de codificación, esto es, debemos encapsular toda la información que se ha ido almacenando de todas las muestras tomadas, sin este proceso la información puede llegar a un punto donde se pierda la señal, y desarrollando el proceso, podemos regenerar la señal, si es que se ha perdido.

Esto se realiza con una técnica llamada PCM, que es la modulación por codificación de los pulsos, en ella se realiza la sinterización y al predicción lineal, [Rownden, 1991], estas son tecnologías que realizan acciones basadas en los parámetros, como lo es el espectro de la voz, que es lo que influye para que la voz pueda ser producida. Una vez que se haya visto el tema de síntesis de voz podemos retomar y poner en claro los procesos que tenemos que tomar en cuenta para la producción de voz.

En algunas ocasiones la síntesis de voz no tiene implicaciones, esto se da para idiomas en los cuales su estructura gramatical o fónica esta bien establecida, ya que o tiene variaciones en sus fonemas como las hay en el español.

Tenemos que tener claro que lo que estamos desarrollando es un sintetizador, lo cual implica los siguientes pasos que son importantes para la producción del mismo:

- El ambiente en que se trabajara.
- Las unidades que se van a emplear.
    - Fonemas
    - Di fonemas
    - Trifonemas
    - Silabas
- La evaluación del costo de beneficio – perdida.
- Mecanismo utilizado para generar la voz.
- Arquitectura del Sintetizador.

Dentro del punto en el que hay que evaluar el costo de lo que se conoce como trade off, es necesario aclarar que esto se da solo para saber que nivel de calidad tiene nuestro sistema.

Lo siguiente a evaluar es la metodología que se utilizara para generar la voz. Después de este elemento le seguirá la representación de la arquitectura que el sintetizador de texto al audio debe te contener para la realización de este proceso.

Dentro de este trabajo, la búsqueda de las unidades se dará principalmente por palabra (.wrd) y por fonema, (.phn) de esta manera, forzaremos que dentro de la búsqueda de las palabras a sintetizar se den los candidatos a concatenar, en caso de que no sean encontradas las palabras, entonces se creara una lista de candidatos de fonemas y de ahí tomaremos la mejor unidad para la construcción de la palabra a sintetizar.

2.3 Arquitectura del sistema de un TTS

En al actualidad los sistemas de procesamiento de texto al habla, conllevan una parte muy similar en cuanto a su arquitectura. Y están constituidos en parte por el NLP (Natural Language Processing), que es el encargado de tomar un texto y darle un significado, originar su trascripción fonética, así como darle la entonación necesaria. Otra parte de la constitución de la arquitectura es el Proceso de Síntesis, que es el que modifica la información dentro del NLP para darle una salida al habla. [Barbosa, 1997].

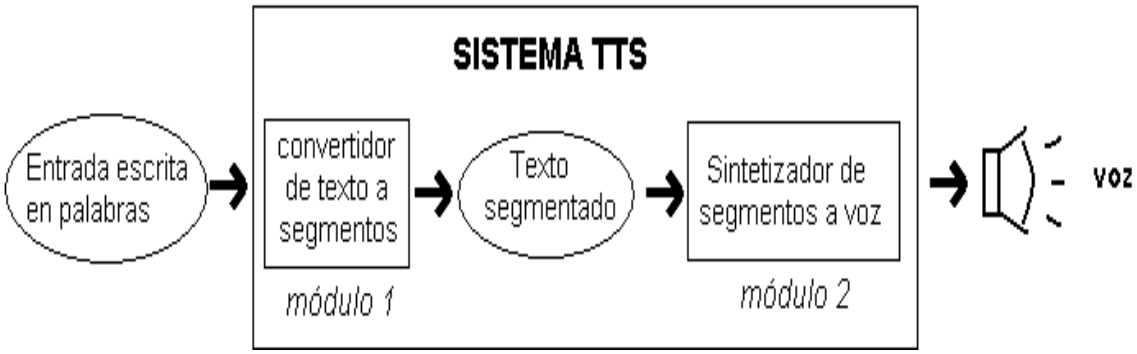

Fig. 2.1 Arquitectura General de Un TtS

La arquitectura está constituida por un analizador de texto, un convertidor de texto a fonemas y un generador prosódico. El analizador, toma cualquier texto y le da el formato necesario para que pueda ser procesado por la siguiente etapa, que es la que realiza la trascripción fonética del texto, y es la reciprocidad que hay entre una palabra, y los fonemas que componen a la palabra.

Tabla de Transcripciones fonéticas de algunas palabras	
novecientos	n o v e s i e n tc tos
Gobierno	gc go v i e r n o
Recursos	rr e kc k u r s o s

Fig. 2.2 Ejemplo de trascripción fonética.

Cuando los fonemas llegan al generador prosódico, se les asigna la duración y entonación, esta información es proporcionada al proceso de síntesis, que es el que nos regresara el audio, o la transformación del texto al habla.

2.4 Comparación De Los Sistemas De Síntesis Más Utilizados

Los sintetizadotes por medio de onda, presentan una ventaja que hace que los diferencie de los demás, ya que permiten operar de manera adecuada las características presentadas por la señal de voz. A su vez los que utilizan la sintetizacion por concatenación, su fuente es primordial, es decir, la voz, y que es pregrabada con di fonemas, frases u oraciones para formar el corpus de voz, para ello cabe mencionar que la persona que sea participe de esta grabación deber de tener la misma entonación para poder mantener una alta calidad de voz, esto es para que no se susciten cambios de manera usual dentro de las silabas.

Con el método de concatenación, se obtienen los mejores resultados ya que este permite la naturalidad dentro del proceso de sintetizacion, es decir es mas claro y de mejor calidad.

2.5 Aplicaciones

Sabemos que en la actualidad, nuestro país ha ido creciendo a pasos agigantados, y también sabemos que existen personas con deficiencias físicas las cuales no les permiten desarrollar las mismas habilidades, es por eso que una de las aplicaciones de la síntesis de voz sea destinada para ayudar a estas personas, es decir podemos aplicarla en los correos, lectura automática de algún texto, entre otras. Dentro de la industria podemos localizar una aplicación que es de gran utilidad para el ser humano, como lo son los reportes de algunas fallas en el desarrollo de aplicaciones, esto se da por medio de mensajes de errores producidos.

2.6 Resumen

Dentro del desarrollo del capitulo, nos dimos cuenta de cómo se lleva a cabo el proceso de sintetización, codificación y los tipos de sintetización que podemos encontrar. Vimos como es que se pueden formar las palabras y como es que se pueden sintetizar, como lo es el caso de los fonemas trifonemas, palabras, etc. Encontramos lo que se conoce como la flexibilidad de la calidad de la voz, recordemos que nuestro propósito es encontrar la mejor unidad dentro de las palabras y los fonemas. Además de la estructura que debemos tomar en cuenta para lograr tener un buen sintetizador.