Diseño e Implementación de un Sistema de Síntesis de Voz

K. Palacio[1], J. Auquilla[2], E. Calle[3]
Facultad de Ingenierías – Carrera de Ingeniería Electrónica
Universidad Politécnica Salesiana
Calle Vieja 12-30 y Elia Liut, EC010105, Campus "El Vecino", Cuenca, Ecuador.
kennethpalacio@gmail.com[1], jorgeauquilla@itss.edu.ec[2], ecalleo@cue.ups.edu.ec[3]

Resumen

En este proyecto resume una propuesta en el campo de la Síntesis de Voz: Un sistema de generación de voz artificial que inicialmente se ha desarrollado sobre un computador convencional, pero que ha sido concebido de tal forma (respecto al almacenamiento y procesamiento del sonido digitalizado) que pueda trasladarse sin mayores complicaciones sobre diferentes plataformas de hardware como por ejemplo, basadas en microcontroladores o dispositivos lógicos programables (PLD) con el fin de complementar esta tecnología a sistemas electrónicos ya existentes. El aporte fundamental de este proyecto se centra en primer lugar en el desarrollo de un algoritmo de evaluación y selección de las unidades fonéticas que contiene un texto en formato ASCII y en segunda instancia en la creación de un corpus de voz en español del cual se extraen unidades pregrabadas de audio para generar habla artificial mediante la técnica de concatenación de unidades de voz.

Palabras Claves: *Procesamiento digital de la Voz, Síntesis de voz, concatenación de unidades, habla artificial.*

Abstract

This project resumes an approach in the field of Voice Synthesis; An artificial voice generation system which runs over a personal computer but that has been designed in such way (respect the storage and processing of digital sound) that it could be easily implemented over different hardware architectures, like microcontroller or programmable logic device (PLD) based platforms, without significant changes in order to use this technology in current electronic equipment. The main points of this project are: first, the development of an evaluation and selection algorithm which analyzes the phonetic content of an introduced text in ASCII format, and second, the creation of a voice corpus (in Spanish) from which the system extracts pre-recorded voice units and create artificial voice by using the unit concatenation technique.

Keywords: *Speech processing, Voice synthesis, unit concatenation, artificial voice.*

1. Introducción

El mundo de hoy se fundamenta principalmente en la comunicación; la transmisión de ideas con los semejantes en todos los campos del desenvolvimiento humano es inevitable; el estudio, el trabajo y el hogar, constituyen los escenarios donde este proceso comunicativo tiene lugar, y en general es protagonizado por los seres humanos.

Gracias al desarrollo de la tecnología programable, se ha logrado automatizar casi por completo la mayoría de industrias, el manejo de los electrodomésticos, los juguetes, los automóviles y en sí, cualquier dispositivo que reciba órdenes humanas para realizar algún trabajo específico.

A raíz de la creación de dispositivos que requieren "instrucciones" para funcionar, el hombre se vio obligado a crear interfaces de diferentes tipos, que permitan de cierta forma establecer una "comunicación" con las máquinas, lo que comúnmente se denomina interface hombre-computador, donde se incluyen teclados para introducir los datos, parámetros o variables de cualquier tipo y pantallas que permiten visualizar el estado de los procesos y los valores de las magnitudes que se controlan. Joaquim Llisterri [1], establece que la exploración de las posibilidades de interacción con los computadores mediante la voz, es tal vez uno de los temas privilegiados dentro de la investigación actual en el campo de la comunicación hombre-computador.

Si bien una parte considerable de los trabajos se orienta hacia la comprensión del habla (reconocimiento de voz), la generación automática de un texto oral a partir de una representación escrita "síntesis del habla" ha ganado la atención de muchos investigadores como complemento del proceso comunicativo.

Recibido: Mayo, 2008
Aceptado: Agosto, 2008

2. Síntesis de Voz por Concatenación de Unidades

Existen varias técnicas que se han desarrollado alrededor del mundo para la generación de habla artificial:

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

El presente proyecto se basa en la técnica de Concatenación de Unidades, que utiliza unidades pregrabadas del lenguaje que se concatenan en base a algoritmos computacionales de selección para formar nuevas palabras y oraciones. Las unidades de voz digitalizadas, que son escogidas por el sintetizador, en base a la transcripción fonética del texto que quiere convertirse en habla. Actualmente la mayoría de los conversores utilizan la síntesis por concatenación de unidades, ya que con menos esfuerzo se obtienen mejores resultados [2].

2.1. Unidades utilizadas en los sintetizadores por concatenación

Existen diferentes unidades del lenguaje que pueden ser utilizadas para este tipo de sintetización, y de hecho, la selección adecuada de las unidades fonéticas en el desarrollo de un sintetizador de voz, es el factor determinante en la calidad que se pueda llegar a obtener. Para Gerardo Martínez [2], el tipo de unidad a concatenar es un parámetro crítico: hay que llegar 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.

Las unidades utilizadas para la conversión de texto en habla dependen, en gran medida, de la estrategia elegida para la síntesis. Joaquim Llisterri en su publicación: “La conversión de texto en habla: aspectos lingüísticos" [2], afirma que en este tipo de síntesis existen dos criterios básicos para decidir qué tipo de unidad se elige: el tamaño más adecuado para el almacenamiento de la base de datos y el ruido que se genera al concatenar dichas unidades.

Las unidades que participan en la concatenación, pueden ser de diferente tamaño en un mismo sistema, haciendo todas las consideraciones pertinentes para ello. Las unidades que más frecuentemente se han utilizado en diferentes proyectos alrededor del mundo son cinco:

2.1.1. Fonema

Es la unidad mínima capaz de diferenciar significados en las palabras. Los fonemas agregan mucha flexibilidad a los sintetizadores que los utilizan, y desde el punto de vista del espacio, sus requerimientos de almacenamiento no son exigentes, ya que en el español por ejemplo, se tienen solamente 24 fonemas (un total de 30 alófonos). Una limitante en el uso de fonemas para la sintetización, es su propiedad abstracta que engloba variaciones fonéticas contextuales (los alófonos), lo que desemboca en una mala calidad de la voz sintética [3].

2.1.2. Difonema

Los efectos coarticulatorios tienden a minimizarse en el centro acústico de un fonema, lo cual derivó la propuesta del difonema, el trozo de voz que va desde la mitad de un fonema a la mitad del siguiente fonema, como la unidad más satisfactoria para la concatenación. Llisterri [2], la establece desde la parte estable de un fonema hasta la parte estable del siguiente, incluyendo la transición entre ambas partes que corresponde a la coarticulación del habla natural.

2.1.3. Trifonema

Un trifonema es aquella unidad constituida por un fonema más la mitad del segmento precedente y la mitad del segmento siguiente, con objeto de evitar la concatenación por el segmento que ocupa el centro del trifonema [2]. Su uso dota de más naturalidad a un sintetizador de voz, pero no todas las frases y palabras pueden formarse utilizando esta metodología, por lo que siempre suelen completarse las concatenaciones con fonemas y difonemas [3].

2.1.4. Sílabas y Semisílabas

La semisílaba es aquella unidad formada por la mitad de la sílaba estableciendo el límite en el centro del núcleo silábico (en español, una vocal). El número de unidades depende del número de fonemas, del número de estructuras silábicas y de las combinaciones posibles. La concatenación a nivel de sílabas, implícitamente incluye los conceptos de las unidades fonéticas: fonemas, difonemas y trifonemas. Muchas líneas de investigación han escogido esta unidad como base de sus sistemas. Como el caso del presente proyecto.

5. Palabras

Para Llisterri [2], la palabra es la unidad que proporciona mayor calidad en la síntesis, pero sería necesario disponer de una base de datos con todas las palabras de la lengua donde además, cada palabra se haya grabado en diferentes contextos para solucionar el problema de la coarticulación. Por ese motivo, esta unidad se utiliza en dominios restringidos, como en aplicaciones de diálogo, donde los vocabularios suelen ser limitados.

3. Corpus de Voz

Un Corpus de voz, es un set de grabaciones que contienen frases, palabras y expresiones comunes de una lengua en particular. Constituye una base de datos, que almacena implícitamente todas las propiedades del lenguaje, con el fin de estudiarlas, considerando las variaciones del tipo regional y dialécticas.

La síntesis de voz basada en concatenación de unidades requiere de una base de datos de la cual se extraen dichas unidades para formar la voz sintética, esta base suele denominarse “corpus” y debe incluir el etiquetado de las unidades fonéticas. El objetivo principal de este tipo de sistemas es reproducir el habla con la mayor naturalidad posible, por ello el corpus deber ser grabado por un único locutor, que presente las mejores características acústicas en su voz, como lo afirma Llisterri en su artículo: Corpus Orales para el desarrollo de las tecnologías del habla en español [4].

3.1. Determinación del Corpus de Voz

Las grabaciones de voz, deben englobar las unidades fonéticas necesarias para poder producir la mayoría de las palabras del español hablado en el Ecuador. Se estableció un total de 1000 archivos de sonido .wav con frases y expresiones típicas de la región, y grabadas por un locutor nativo.

El corpus contiene la mayoría de palabras cotidianas, y que se repiten con frecuencia para la mayoría de frases en el lenguaje común. Además, el contexto de cada frase aporta con diferentes palabras que eventualmente se utilizarán en el proceso de sintetización. El corpus contiene implícitamente elementos fonéticos más pequeños, tales como fonemas, difonemas y trifonemas, que servirán para la formación de palabras nuevas. Con el objetivo de reunir todos los elementos fonéticos y sonidos necesarios, se ha tomado un listado de todas las sílabas del español, publicado en internet bajo la referencia [5], para asegurar que todos ellas se incluyan en el contenido de las grabaciones del corpus de voz SVART (nombre que se le dio al corpus del sistema).

Al considerar que todas las palabras del español están formadas por sílabas, si se tiene una base de datos con todas las sílabas, es teóricamente posible formar cualquier palabra. En el corpus SVART, se han incluido todas las sílabas que incluyen este listado con las que se pueden formar las palabras. Es importante recalcar que las sílabas no se han colocado de forma aislada sino que implícitamente están presentes en las palabras del corpus. Adicionalmente se incluyeron sonidos autóctonos que no son parte del español, como por ejemplo aquel que es causado por la combinación de las letras “z” + “h”, que de hecho suena igual que la de las letras “s” + “h”, que son utilizados para palabras extranjeras como *show*, o nativas como “Zhucay", el nombre de una comunidad local.

4. Etiquetación de un Corpus de Voz

Una vez que se ha grabado diferentes archivos de sonido con frases, palabras y expresiones del lenguaje, se ha creado un corpus de voz. Este conjunto de archivos de sonido debe contener el mayor número de unidades del lenguaje posible, con el fin de utilizarlas para la concatenación en la creación de voz sintética; el proceso mediante el cual se identifican dichas unidades en el corpus, se conoce como etiquetación.

La síntesis de voz por concatenación de unidades, puede trabajar con la técnica de ***Unit Selection*** propuesta por Andrew Hunt y Alan Black [7], que permite trabajar con unidades del lenguaje diferentes, por ejemplo palabras y fonemas; lo que 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.

Gracias a este proceso, las unidades del habla no tienen que grabarse de manera aislada, por ejemplo el caso de un archivo de sonido que contenga solamente un fonema, si no que, el etiquetado de un corpus, permite leer las unidades que implícitamente están presentes en las frases grabadas. Este aspecto contribuye considerablemente al resultado que se obtiene tras la concatenación; las unidades se extraen del habla común, por lo que fluyen con naturalidad para diferentes contextos.

Sería absolutamente diferente si se grabaran las unidades individualmente, aunque se disminuiría notablemente el espacio destinado al almacenamiento del audio, esta ventaja no se sobrepone sobre un posible resultado de voz en extremo robotizada o demasiado entrecortada, lo que restaría legibilidad y naturalidad al habla artificial.

4.1. Objetivo general de la etiquetación de un corpus de voz

En el proceso de etiquetación de un corpus de voz, se delimitan las fronteras de las unidades fonéticas presentes en las grabaciones. Las etiquetas son marcas que indican donde comienza y donde termina una unidad fonética, ya sean palabras o fonemas, generalmente son archivos que contienen información asociada a las grabaciones de un corpus. Se alinean en el tiempo la duración de cada unidad del lenguaje, respecto a las ondas de voz presentes en los archivos de sonido. Así, se puede identificar en términos de milisegundos, por ejemplo, el inicio y el fin de un fonema. 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 una unidad determinada para crear una palabra asociándolo con otros fragmentos de sonido.

La etiquetación a nivel de palabras, paralelamente agrega naturalidad a la voz sintética; si se tiene que sintetizar una palabra común o cotidiana, por ejemplo el artículo “las", esta palabra no necesariamente tiene que formarse por concatenación de unidades pequeñas, sino que directamente puede extraerse de un archivo de sonido del corpus, gracias a la etiquetación a nivel de palabras y en base a un análisis previo de contextos.

4.2. Etiquetación del Corpus de Voz grabado para el Sistema

Los procesos de etiquetado se realizan mediante herramientas informáticas desarrolladas en pro de la investigación de las tecnologías de voz, y en efecto, este proyecto se sirve del software CSLU Tool Kit 2.0 (Center for Spoken Lenguaje Understanding) desarrollado por el OGI (Oregon Graduate Institute) Oregon School of Science & Engineering, Oregon Health and Science University, disponible en su sitio web: http://www.cslu.ogi.edu/toolkit/. Se seleccionó esta herramienta por su versatilidad y por tener una licencia de uso libre para fines académicos. El uso de este software es exclusivo para la etiquetación del corpus de voz y es totalmente independiente del sistema desarrollado para la generación del habla artificial.

4.2.1. Niveles de Transcripciones del CSLU Tool Kit 2.0

La guía de etiquetación (The CSLU Labeling guide [6]), establece que la suite de desarrollo de tecnologías de voz, CSLU Tool Kit 2.0, admite dos niveles fundamentales de transcripciones para manejar los datos del habla:

a. Transcripciones ortográficas o a nivel de texto.

Este nivel de transcripción contiene la información de las grabaciones sin una referencia de tiempo, se utiliza ortografía común, en letras mayúsculas y no distingue los signos de puntuación ni los segmentos de señal que no contienen voz (las pausas). Esta información se almacena en archivos de texto (.txt), que reflejan exactamente lo que se dijo en cada grabación.

b. Transcripción ortográfica a nivel de palabras alineadas en el tiempo.

En este tipo de transcripción, se alinean las palabras mencionadas en cada grabación con las ondas sonoras a un nivel temporal, estableciendo el inicio y el fin de cada palabra en términos de milisegundos. Se distinguen además los segmentos que no tienen voz, como pausas. Mediante la aplicación SpeechView del CSLU Tool Kit 2.0 pueden observarse simultáneamente las señales sonoras y las etiquetas a nivel de palabra, como se muestra en la figura 1.

Esta herramienta permite ajustar manualmente las fronteras para las etiquetas a nivel de palabra, en base al contenido del archivo de sonido (que también puede escucharse dentro de SpeechView) y más empíricamente a su espectrograma. Para este nivel de transcripción se utilizan archivos de texto (con extensión .wrd) que se presentan con ortografía normal, en letras mayúsculas y sin distinguir los signos de puntuación.

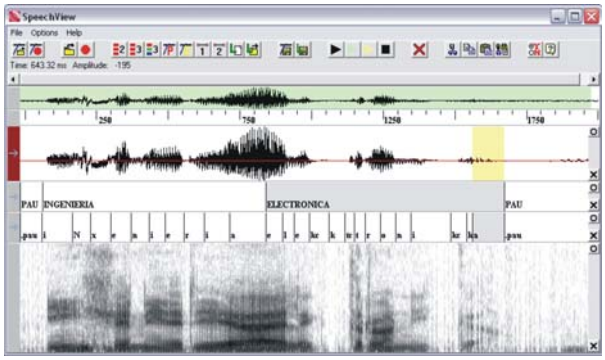

Figura 1. SpeechView utilizado en la etiquetación.

c. Transcripción Fonética alineada en el tiempo.

Corresponde a la representación fonética, detallada, del contenido de una grabación. Una vez que se han delimitado las fronteras de las palabras, éstas se pueden descomponer fonéticamente para obtener las fracciones que las conforman. Estas fracciones corresponden a los sonidos que se enlazan para crear las palabras, y se representan por símbolos, que para el caso del CSLU Tool Kit 2.0 se definen a través del WorldBet (representación ASCII del International Phonetic Alphabet). De manera similar que en el nivel de etiquetación de palabras, la transcripción fonética también se alinea en el tiempo, definiendo el comienzo y el fin de cada unidad fonética. Se manejan, de igual manera, archivos de texto, pero con la extensión .phn (de phoneme), cuyo formato es similar al de los archivos .wrd. SpeechView, permite visualizar y ajustar las fronteras fonéticas también a este nivel. En la figura 1, se puede observar la aplicación SpeechView, con una onda sonora y sus etiquetas a nivel de palabras y fonemas.

5. Diseño del Sistema

Debido al enfoque a futuro de este proyecto, que busca su aplicación principal en el desarrollo de tecnología para personas discapacitadas y su implementación en diversos dispositivos electrónicos, cuyo funcionamiento gire en torno a un micro-controlador, un PLD (programable logic device) o un

procesador digital de señales DSP, se ha encaminado el desarrollo de cada una de las fases del sistema de tal forma que puedan trasladarse sin mayor dificultad a las plataformas antes mencionadas: lo que incluye la adaptación de todos los datos que se manejan para que puedan manipularse a nivel de bytes en una memoria convencional. En su totalidad el sistema está implementado sobre un computador convencional, con sistema operativo Windows XP®, y desarrollado en lenguaje C++, plataforma dev-c++, v4.9.9.2.

5.1. La Técnica de Unit Selection

Propuesta por Hunt y Black [7], que se basa fundamentalmente en la diversidad del tipo de unidades que pueden concatenarse para crear la voz sintética. Esta flexibilidad en el tamaño de las unidades, permite escoger palabras y hasta frases enteras, así como fonemas, difonemas, trifonemas, tetrafonemas, etc.

5.2. Formato de las Grabaciones de Voz

Se parte de una base de datos de voz de la cual se extraen dichas unidades, en base a un análisis fonético del texto ingresado que se desea sintetizar para crear una voz artificial.

El software CSLU Tool Kit 2.0 utilizado en este proyecto para la identificación de las unidades fonéticas que intervienen en el proceso de síntesis, requiere que las grabaciones de voz se encuentren en archivos de sonido del tipo wav. Se crearon un total de 1000 archivos cuyo contenido ha tratado de abarcar la totalidad de las unidades fonéticas del español.

5.3. Tratamiento de los Archivos .wav

Los archivos de sonido del corpus de voz se han grabado en formato .wav, con una frecuencia de muestreo de 16 Khz, una resolución por muestra de 16 bits y con un solo canal de salida monoaural, ya que un archivo stereo ocuparía el doble de memoria de almacenamiento.

Todos estos parámetros se han escogido teniendo en cuenta las características naturales de la voz humana, para su correcta digitalización y posterior reproducción. Cada archivo de sonido .wav tiene 44 Bytes de encabezamiento que almacenan información propia sobre su contenido, tales como su tamaño, el número de canales, la frecuencia de muestreo, etc., y a continuación las muestras correspondientes a la digitalización del sonido en formato Little endian, con 2 Bytes por cada una, ya que se tiene una resolución de digitalización de 16 bits.

Al tener que trabajar con un número considerable de archivos del mismo tipo, toda la información almacenada en el encabezado tiende a ser redundante. El enfoque del proyecto, hace que partir de los archivos de sonido .wav como tales para la extracción de las unidades fonéticas, sea un proceso relativamente complicado y demorado cuando se trabaje con dispositivos de programación electrónicos como un microcontrolador o un DSP.

Se optó por desarrollar un programa que lea los datos útiles de todos los archivos de sonido (las muestras correspondientes a la digitalización) que forman parte del corpus de voz y los adapte a un formato que pueda fácilmente ubicarse sobre una memoria convencional, y ser leído desde cualquier plataforma electrónica. Esa adaptación se manifiesta en la creación del archivo *Corpus.dat*.

5.3.1. Los archivos de Indexación. Los niveles de etiquetado y marcado fonético admitidos por el software de desarrollo de tecnologías del habla CSLU Tool Kit 2.0, son el de palabras y el de fonemas. Tras el proceso de etiquetado del corpus de voz, se obtienen tres tipos de archivos diferentes:

- El archivo de transcripciones ortográficas no alineadas en el tiempo .txt.
- El archivo de transcripciones a nivel de palabras alineadas en el tiempo .wrd.
- El archivo de transcripciones fonéticas alineadas en el tiempo .phn.

Los archivos obtenidos son archivos de texto que contienen las respectivas fronteras iniciales y final, en términos de milisegundos, correspondientes a cada unidad presente en los archivos de sonido, sea esta una palabra o un fonema. Los valores de estas fronteras se producen en función del archivo .wav al que las unidades pertenecen y están referidas al mismo, por lo que existe un archivo de tipo .wrd y de tipo .phn para cada grabación (archivo .wav) del corpus de voz. Los archivos .txt contienen información referente a la transcripción ortográfica del contenido de los archivos de sonido, y de hecho no contienen información sobre la duración de las unidades fonéticas, por lo que solo intervienen la etapa de etiquetación.

Los archivos de etiquetado (.wrd y .phn) juegan un papel muy importante en el desarrollo de este proyecto, ya que es aquí donde se delimitan las unidades dentro de los archivos de sonido para que puedan ser extraídas, y formen parte de nuevas frases y oraciones que se creen a partir del sistema.

5.3.2. Objetivo de los Archivos .wrd y .phn

Tras el análisis del texto a sintetizar, se identifican los elementos fonéticos que componen la frase que se desea reproducir, con el fin de buscarlos en la base de datos de archivos de voz (que es el corpus), extraerlos y concatenarlos. Para ello, debería conocerse cada una de las unidades fonéticas como tales y su ubicación temporal dentro de los archivos de sonido .wav. Los archivos .wrd y .phn contienen esa información y

justamente pueden utilizarse para encontrar las unidades y definir sus fronteras fonéticas. Debido a que estos archivos son de tipo texto, y almacenan simplemente caracteres, surgió la necesidad de adaptar estos datos con el fin de agilitar su obtención y para que puedan manipularse desde dispositivos electrónicos como un microcontrolador.

Así nacieron los archivos de indexación, mediante un programa desarrollado en C++ se lee la información de los archivos de texto .wrd y .phn y se crean los archivos: *defindex.dat* (para la etiquetación a nivel de palabras) y el archivo *defindph.dat* (para la etiquetación a nivel de fonemas). Estos dos archivos identifican las unidades para la concatenación (palabras o fonemas) y sus parámetros, dentro del archivo que contiene las grabaciones *corpus.dat.*

5.3.3. Formato de los Archivos de Indexación. Tienen una estructura común que difiere únicamente en el tipo de unidad que se está tratando. Para almacenar la información correspondiente a cada unidad (indiferentemente de que sean palabras o fonemas), se han destinado 16 Bytes, que forman una trama de información que debe ser interpretada de la siguiente manera, según se observa en la tabla I.

Tabla 1. Archivo de indexación: Trama de datos de Unidad fonética.

CÓDIGO UNIDAD	SEPARADOR	FRONTERA INICIAL	FRONTERA FINAL	NUMERO DE UNIDAD	NUMERO DE .WAV
4 BYTES	1 BYTE	4 BYTES	4 BYTES	1 BYTE	2 BYTES
00 0E D2 BA	00	00 02 97 5A	00 02 B6 5A	03	00 02

Cada campo de la trama de datos correspondiente a cada unidad fonética, tiene una importancia vital para el proceso de concatenación de unidades.

a. Código Unidad. Corresponde a un número de 32 bits equivalente a una palabra o a un fonema determinado. Este número es generado en base a un algoritmo propuesto para agilitar la búsqueda de las unidades fonéticas. Durante el proceso de concatenación de unidades, cuando un fonema o palabra tenga que encontrarse dentro de toda la base de datos de unidades del corpus, una búsqueda que implique la comparación de caracteres tardaría mucho más tiempo que la comparación de un valor numérico concreto.

Se desarrolló un algoritmo (una función) que recibe como argumento una palabra en formato *string* y devuelve un valor numérico de 32 bits, equivalente única y exclusivamente a dicha cadena de caracteres. Matemáticamente este algoritmo se expresa como:

$$ValorPalabra = \sum_{i=0}^{i=Palabra.length()} \left[(Palabra[i])^3 \cdot (i+1)\right]$$

Donde **i** corresponde un índice que indica cada carácter de la variable string. El Código generado para cada unidad, palabra o fonema, se basa en que cada caracter tiene un valor numérico correspondiente al código ASCII. Gracias a que todas las unidades presentes en los archivos defindex.dat y defindph.dat se han definido como valores numéricos, se las ha podido ordenar ascendentemente, lo que permitió la implementación del algoritmo de búsqueda binaria para la búsqueda de candidatos para unidades fonéticas.

b. Separador. Es un espacio de 1 Byte que preliminarmente no desempeña función alguna en el proceso de síntesis. Puede utilizarse para agregar un parámetro adicional en aplicaciones futuras.

c. Frontera Inicial. Corresponde a la dirección (32 bits) del archivo *corpus.dat* en donde empiezan los datos correspondientes a la primera muestra de la digitalización de la unidad fonética analizada.

d. Frontera Final. Corresponde a la dirección (32 bits) del archivo *corpus.dat* en donde terminan las muestras de la digitalización de la unidad fonética correspondiente.

e. Número de Unidad. Este campo de 1 Byte se refiere a la ubicación que la unidad fonética tiene dentro de cada archivo wav. Permite evaluar la contigüidad de las unidades

f. Número de Wav. Número de 16 bits indica a que grabación (número de archivo wav del corpus de voz) corresponde la unidad analizada.

Los campos **e** y **f** pueden interpretarse de manera más objetiva con un ejemplo:

- El archivo NU-001.test.wav tiene grabado el texto: “pau universidad politécnica salesiana pau” (“pau” hace referencia a espacios sin sonido). La unidad: “politécnica" correspondería al archivo wav: 00 01 (campo f) y ocuparía la posición 03 (campo e).

Cada campo interviene en el proceso de selección de las mejores unidades para generar el habla artificial. La posición que cada unidad ocupa en los diferentes archivos de sonido, y el número de archivo wav, son parámetros que agilitarán la selección de unidades contiguas con el objeto de mejorar la coarticulación y agregar naturalidad a la señal de salida.

6. Principios Generales del Proceso

La técnica de Unit Selection permite escoger diferentes tipos (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 un ejemplo, se prefiere tomar difonemas, trifonemas o tetrafonemas en comparación fonemas aislados e independientes. Mientras más grande sea el tamaño de la unidad, mayor será la calidad de la voz sintética, lo que agrega flexibilidad a tal punto que pueden tomarse hasta palabras e inclusive frases como unidades de concatenación. En los archivos de indexación, se dispone de unidades fonéticas etiqueta-das a nivel de palabras y fonemas.

La búsqueda inicial de elementos a partir del texto ingresado a sintetizar se hace a nivel de palabras. Por obvias razones, no se ha abarcado la totalidad de las palabras del español, por lo que el sistema debe tener la capacidad de resolver la presencia de una palabra inexistente con un algoritmo que la cree a partir de sus componentes fonéticos que pueden estar presentes en otras de las palabras del corpus de voz. La figura 2 ilustra la idea del proceso.

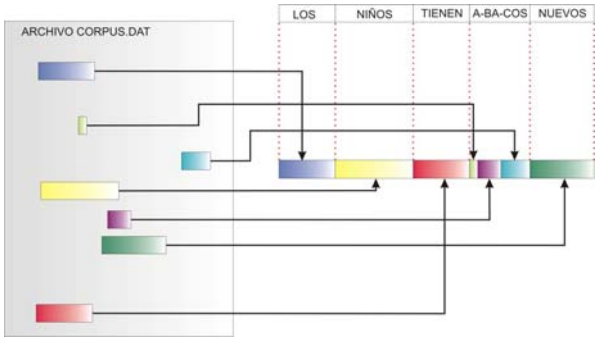

Figura 2. Módulo de construcción de salida de audio.

El sistema separa las palabras que tienen que sintetizarse, busca los posibles candidatos y escoge los mejores haciendo una evaluación profunda en base a su posición dentro del corpus de voz, y a su estado como unidades contiguas. Posteriormente, tanto los mejores candidatos escogidos para las palabras que se encontraron en el corpus de voz, como aquellas que no se encontraron pasan al módulo de construcción de la salida de audio, en el cual se inicia el Módulo de construcción de palabras nuevas.

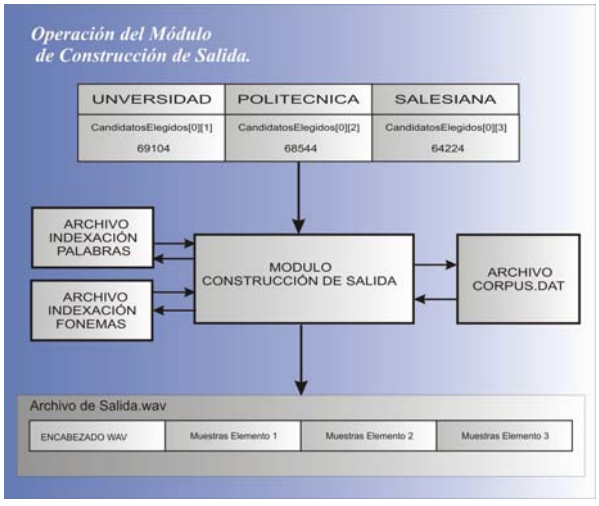

Figura 3. Diagrama de bloques Módulo de construcción de salida de audio.

Finalmente todos los candidatos escogidos copian los segmentos de voz del archivo corpus.dat ordenadamente a un archivo de sonido de salida que se reproduce. La figura 3 ilustra un diagrama de bloques de la operación de módulo de salida.

6.1. Proceso de Construcción de palabras nuevas

Se requiere de un proceso de transcripción fonética para determinar los sonidos que componen una palabra cuyos símbolos equivalentes están establecidos por el WorldBet [8], en un conjunto de caracteres en formato ASCII. Conocidos estos elementos, y siguiendo los principios que se aplicaron para las palabras, podría directamente invocarse un módulo de búsqueda y subsecuentemente uno de evaluación de características similares a los utilizados para palabras, sin embargo, los resultados que se obtienen no son los mejores y su calidad decrece proporcionalmente al tamaño de las palabras, ya que cada vez, se hace más complicado encontrar fonemas contiguos a medida que su número aumenta en una palabra lo que obliga a la utilización de unidades aisladas.

Por esta razón se segmenta las palabras en unidades de tamaño intermedio con el fin de reducir la complejidad y el tiempo de procesamiento requerido para encontrar fonemas contiguos. Un diagrama de bloques general de la operación de creación de palabras nuevas puede apreciarse en la figura 4.

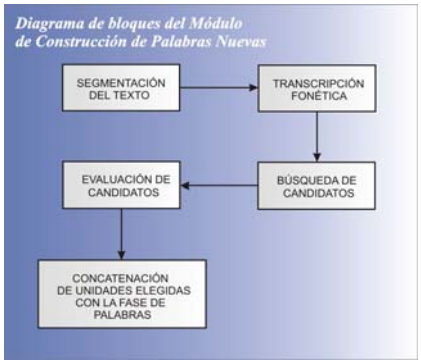

Figura 4. Diagrama de bloques del módulo de construcción de palabras nuevas.

Recordando que estas unidades intermedias nacen de la unión de fonemas y de fracciones de ellos, se ha propuesto extender el concepto de su delimitación a utilizar los componentes fonéticos cuya pronunciación sea estable. Conociendo que cada letra de una palabra indica cómo debe pronunciarse y que implícitamente incluye la información fonética necesaria para comprender su constitución, se optó por un sistema de segmentación silábico.

6.2. Evaluación de candidatos

Encontrados los posibles candidatos para cada fonema, el algoritmo realiza una búsqueda orientada al mayor esfuerzo, es decir encontrando unidades contiguas dentro de las grabaciones y complementando la agrupación más grande encontrada con unidades aisladas, la figura 5 muestra un listado de candidatos para una sílaba de 3 letras. Por ejemplo si se necesita crear la silaba "sil"; si se encontrara solamente unidades contiguas con "si" el sistema complementa la búsqueda con un fonema aislado "l". Finalmente los mejores candidatos se copian al archivo de salida conjuntamente con aquellos escogidos a nivel de palabras y sílabas subsecuentes de la misma palabra nueva.

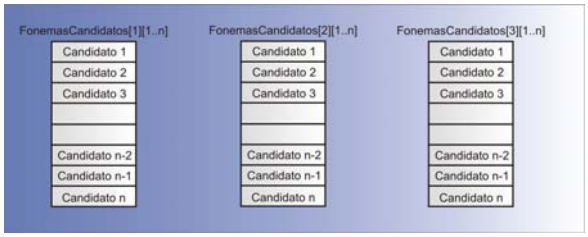

Figura 5. Listado de fonemas candidatos.

7. Evaluación del Sistema

7.1. Evaluación Subjetiva

Nace de la naturaleza propia de la voz humana y su percepción sensorial por el aparato auditivo. La subjetividad que arrastra consigo la apreciación que pueda tenerse de una voz artificial depende de cada individuo. Se realizó una encuesta entre un grupo de 16 personas, con diversidad en edad, género, y ocupación.

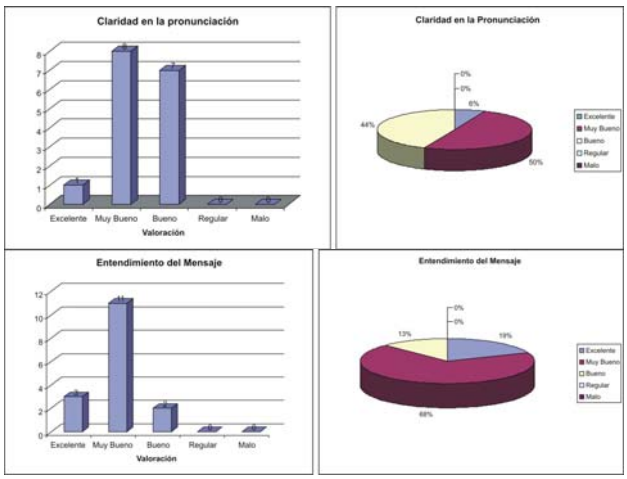

Figura 6. Resultados de la Evaluación Subjetiva.

7.2. Evaluación Objetiva.

Esta evaluación normalmente se centra en determinar la exactitud en la determinación de los mejores candidatos escogidos para la síntesis, en los cuales se obtuvo una eficacia del 100%.

8. Futuras líneas de desarrollo.

La etapa siguiente al proyecto se centra en su traslado a una plataforma de hardware diferente a un computador, dado que los datos digitalizados y el algoritmo se crearon pensando en ello.

Una línea de trabajo interesante contribuye la modificación de la tonalidad de la voz, para expresar emociones por ejemplo. Esto se llama prosodia y podría lograrse con técnica avanzadas de filtrado y modulación mediante procesamiento digital de señales

Estudiar la adaptación de la información de audio a un formato comprimido, como mp3.

Evaluación por internet, para extender los resultados subjetivos encuestando a un mayor número de personas y de diferentes orígenes.

9. Conclusiones

El aporte concreto del proyecto en desarrollar un sistema de Síntesis de Voz en español utilizando un locutor nativo. Lo que profundiza un análisis en las diferencias fonéticas respecto a sistemas que hablan inglés, donde existe una mayor cantidad de alófonos.

Se utilizó el CSLU Tool kit 2.0 como una herramienta de soporte para el proceso de etiquetado de las unidades fonéticas.

El algoritmo desarrollado presenta una estructura compleja y un diseño modular que estudiarse por etapas. La principal virtud de este sistema es su perspectiva genérica, que puede fácilmente cambiarse de plataforma.

10. Referencias

[1] LLISTERRI Joaquim, WEST M., “La conversión de texto en habla: aspectos lingüísticos", Actas del II Congreso de Lenguajes Naturales y Lenguajes Formales, Blanes, Girona, Universitat Autónoma de Barcelona

[2] MARTINEZ Salas Gerardo., “Síntesis de voz con emociones", Proyecto de fin de carrera, Departamento de Ingeniería Electrónica, Universidad Politécnica de Madrid.

[3] FLORES Toscano Leonardo, “Síntesis de Voz mediante la implementación de Unit Selection", Tesis de Licenciatura en Ingeniería en Sistemas Computacionales. Universidad de las Américas de Puebla, Cholula-Puebla, México, Mayo de 2001.

[4] LLISTERRI Joaquim y otros., “Corpus Orales para el desarrollo de las tecnologías del habla en español.", Oralia, Análisis del discurso oral 8 (en prensa). Departament de Filología Espanyola-Universitat Autónoma de Barcelona, 2005.

[5] ARMARIO Toro Jerónimo, “Un listado de las sílabas del español", Cuadernos Cervantes, 2006.

[6] LANDER T., \The CSLU Labeling Guide", Center for Spoken Language Understanding Oregon Graduate Institute, 1997.

[7] HUNT Andrew J. BLACK Alan W., \Unit selection in a concatenative speech synthesis system using a large speech database", in Proceedings of ICASSP 96, vol 1, pp 373-376, Atlanta Georgia. ATR Interpreting Telecommunications Research Labs, Japan, 1996.

[8] HIERONYMUS James L., \ASCII Phonetic Symbols for the World's Languages: Worldbet, AT&T Bell Laboratories, Murray Hill, NJ 07974, USA", http://www.ling.ohiostate.edu/_edwards/worldbet.pdf