IMPLEMENTACIÓN DE UN SISTEMA DE CONVERSIÓN DE TEXTO A VOZ, MEDIANTE SÍNTESIS POR REGLA Y COMPOSICIÓN ALOFÓNICA

Luis Fernando D´Haro E.
Oscar Mauricio Agudelo M.

Ingenierías Electrónica y Mecatrónica – Universidad Autónoma de Occidente
Calle 25 Carrera 116 Km 2 Via Cali – Jamundí
E-mail: lfdharo@ieee.org, osmagu@verne.cuao.edu.co
Santiago de Cali, Colombia

Abstract: Este artículo describe el desarrollo de un sintetizador de voz usando el método de síntesis por regla, empleando para ello un archivo de onda con los alófonos del español. Además, se explican las etapas para realizar la conversión texto a voz, el procedimiento de creación e implementación de la base de conocimiento y la técnica de procesamiento digital que varía el tono, y otros rasgos prosódicos, de la voz generada. Finalmente, se hace una descripción del programa elaborado, junto con los resultados obtenidos al emplear el sintetizador en la reproducción de un texto.

Keywords: Síntesis de Voz, Procesamiento de Señales, Herramientas de Software.

1. INTRODUCCIÓN

Los sistemas de Conversión a Voz (TTS: Text To Speech) surgen como un intento por emular en forma algorítmica la manera como los seres humanos hablan; con tal fin se emplean básicamente dos métodos: el primero, que se basa en la teoría acústica, utiliza modelos matemáticos del tracto bocal, junto con otros parámetros tales como el tono fundamental, formantes y espectro de energía de la señal entre otros; ejemplo de este método es el LPC (Linear Predictive Coding). En el segundo, se emplean muestras pre-grabadas de las señales de voz a reproducir, para luego emplearlas ya sea directamente ó modificadas, en el dominio temporal o frecuencial. Las grabaciones realizadas pueden ser: palabras específicas, sílabas, fonemas, alofónos o difonemas. Las dos primeras suelen ser limitadas tanto por la cantidad como por espacio físico de memoria. Las tres últimas son menos extensas y permiten generar cualquier combinación existente de palabras. El software desarrollado empleó el segundo método descrito anteriormente, trabajando con los alófonos de la lengua española.

El proyecto descrito en este documento surge como un intento por desarrollar un sistema de síntesis que permita a personas con discapacidades físicas tanto motrices, como visuales, redactar documentos escritos o comunicarse con los que los rodean. Además, el programa podría llegar a emplearse junto con otras aplicaciones que soporten OLE y que requieran como salida una señal de voz generada por el computador. Por ejemplo, en sistemas de monitoreo

y control (alarmas), opciones de accesibilidad, entre otros.

2. FONEMAS Y ALÓFONOS

Los fonemas son las unidades mínimas del lenguaje utilizables para diferentes enunciados, identificables entre sí mediante el procedimiento de conmutación sucesiva u oposición.

Los fonemas se pueden clasificar según el modo de articulación, es decir por la posición que adoptan los órganos articulatorios en cuanto al grado de abertura o cierre de los mismos, en:

a.) Para las vocales: Cerrados o altos (*I, U*), de mediana apertura (*E, O*) y gran apertura o bajos (*A*).
b.) Para las consonantes: Oclusivos (*P, B, T, D, K, G*), Fricativos (*F, q* (z española), *S, Y, X* (o *J*)), Africados (*Ch*), Nasales (*M, N, Ñ*), líquidos Laterales (*L, Ll*) y Liquidos Vibrantes (*R , RR*).

Junto con la anterior clasificación, los lingüistas emplean los términos sonoro y sordo, para referirse a los sonidos que se producen a partir de la vibración, o no vibración, de las cuerdas vocales respectivamente; dichas vibraciones son las causantes de que los sonidos sonoros presenten un aspecto periódico, en tanto que los sordos no. Finalmente, se puede hacer una clasificación de los fonemas de acuerdo a su nivel de intensidad en: relajados, medios e implosivos.

Los alófonos por su parte, son las variaciones en la producción y emisión, al nivel fonológico, de un fonema con el fin de adaptarlo al contexto de los sonidos que están alrededor de este, sin que por ello se produzca un cambio en el significado del mensaje transmitido; y pueden ser clasificados según el lugar en que se producen en (ver figura 1):

- Bilabiales
- Labiodentales
- Interdentales
- Dentales
- Alveolares
- Palatales
- Velares
- Prevelares y PostVelares

El empleo de los alófonos como unidades grabadas trae consigo el problema de que cada persona puede generarlos distintamente según el país o región en que haya nacido, sumado a que los mismos fonetistas están tan sólo parcialmente de acuerdo en la clasificación de cada uno de los mismos, por lo cual un mismo alófono puede estar clasificado en diferentes grupos al tiempo; para solucionar estos inconvenientes se estudió 4 lingüistas diferentes, tomando sólo aquellos alófonos en que por lo menos 3 de ellos coincidieran en su clasificación; unido a lo anterior, dicha clasificación se particularizó a la forma de hablar de los habitantes del país de origen de los autores (Colombia).

Como ejemplo del proceso anterior se puede citar el caso del fonema G que se presenta en tres alófonos diferentes: como oclusivo en la palabra "Ven**g**o", como fricativo en "Pan**g**o" y como implosivo en "Í**g**neo"; cada uno de estos alófonos se produce en diferentes posiciones a nivel glotal, regidos por el contexto precedente y consecuente del entorno que rodea al fonema. Para deducir las reglas que permitieran identificar cuando emplear un determinado alófono en preferencia de otro, hubo necesidad de realizar un estudio fonético, gramatical y ortográfico de las palabras. De esta manera se obtuvo un total de 70 reglas.

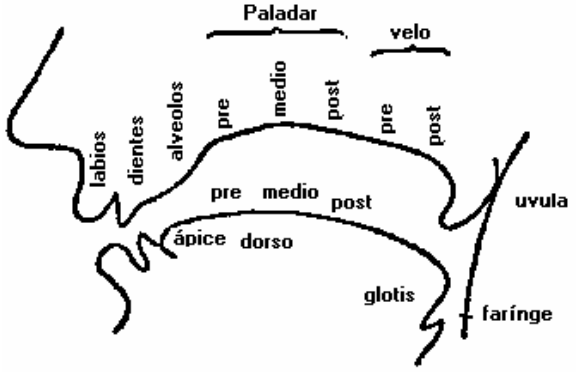

Figura 1.División por regiones de los órganos de las cavidades supraglóticas.

3. DESCRIPCION DEL PROCESO DE CONVERSION TEXTO A VOZ

La gran mayoría de los sistemas de síntesis emplean el esquema utilizado para este proyecto, variando tan solo el número de pasos y los procesos intermedios que se realizan con el fin de otorgar alguna característica especial.

3.1 Normalización del texto

Este módulo se encarga de transformar la entrada de texto en una serie de palabras habladas. Como por ejemplo, realizar las conversiones de números, encargarse del manejo de abreviaturas, símbolos y personalizaciones del usuario.

3.2 Corrección de ambigüedades grafonológicas

Este paso se encarga de corregir las desviaciones patográficas existentes en el idioma, estas desviaciones son básicamente cuatro: Poligrafía de los fonemas (Es el hecho de que a un mismo fonema le pueden corresponder varios fonogramas distintos, ejemplo la *C, X, S* y *Z*), Polifonía de los grafemas (fonemas que poseen una única grafía, pero que se leen en forma distinta según su contexto), Homografía de los morfemas heterófonos (Son palabras que se escriben igual pero que de acuerdo al contexto se pronuncian diferente), Heterografía de morfemas uniformes (son adiciones no relevantes a los morfemas según el contexto en que este se encuentra); de estos cuatro, tan sólo se tuvo en cuenta los dos primeros ya que los dos últimos no son aplicables al español o a los alófonos.

3.3 Prosodia

En este procedimiento se aplican las reglas de conversión de texto a voz para cada una de las letras y su contexto, asignando un determinado alófono a reproducir. Para lograrlo, se hace necesario primero crear una base de conocimientos del tipo ***si.. entonces...,*** en donde las entradas son las letras precedentes y consiguientes, así como otros parámetros definidos por el contexto en que se encuentra la letra a reproducir. El siguiente ejemplo, para la letra d, aclara este punto:

Si LetraAnterior = “R” **o** LetraPosterior = “R” **Entonces** Reproducir D Fricativa dental sonora

De lo contrario, Si LetraPosterior = “N” **o** “L” **Entonces** Reproducir D Oclusiva dental sonora

Si no, Entonces Reproducir D implosiva

En proceso anterior además se asigna un valor en milisegundos a los retardos producidos por las comas, puntos, espacios y otros signos ortográficos, a parte de la utilización de otras reglas para realizar variaciones al tono, la intensidad de las palabras según su contexto, así como algunos otros rasgos prosódicos del habla.

4. PROGRAMA DESARROLLADO: AUDIOTEXTO

El software realizado se hizo con el lenguaje de programación Visual Basic Versión 6.0, empleando además tecnología ActiveX a fin de permitir la interacción del mismo con otros programas de edición de texto. Para el manejo de los recursos del sistema, así como de la tarjeta multimedia y otras funciones adicionales se emplearon librerías dinámicas (DLL) y junto con ellas las funciones API´s asociadas. Además se crearon algunas macros para Microsoft Word en el lenguaje de programación Visual Basic para Aplicaciones a fin de permitir que este editor de textos pudiera emplear el programa desarrollado, para reproducir cualquiera de sus documentos.

Los requerimientos del programa son los siguientes: Ordenador personal Pentium a 200 MHz, sistema operativo Windows 95 o Windows 98, tarjeta de sonido y parlantes, Memoria RAM de 32 MB o superior, espacio en disco duro de 3 MB y para obtener mayores ventajas en el uso del programa, se recomienda tener instaladas las aplicaciones Microsoft Word y WordPad para Windows.

4.1 Archivo base para la generación de la voz

Para realizar el archivo de audio que contiene los alófonos, se realizó una lista de los mismos, así como de diversas palabras en las que estos se presentaran, a fin de disponer de varias muestras del alófono en cuestión y poder así seleccionar el mejor de ellos. Posteriormente se seleccionó el formato de grabación del archivo, la velocidad de grabación, el número de bits de resolución y el número de canales, escogiéndose para ello el formato WAV (PCM) por ser estándar a todas las tarjetas de sonido existentes en el mercado; la velocidad de 44100 Hz (estándar de los discos compactos), 16 bits de resolución y sonido estéreo. Lo anterior a fin de garantizar una alta

fidelidad y una relación señal a ruido (SNR) de por lo menos 96.3 dB (valor obtenido a partir de la formula 1).

$$\frac{S}{N} = 6.02 \times n + a \qquad (1)$$

donde n representa el número de bits de resolución del sistema y α es una constante que se adiciona para encontrar el valor pico o el valor promedio según sea el caso.

Para realizar las grabaciones de las palabras se seleccionó a un especialista en fonética, con el objetivo de garantizar que los diferentes alófonos escogidos fueran correctamente generados al hablar. Y a fin de obtener la mayor nitidez y ausencia de ruido posible, se hicieron las grabaciones en un estudio profesional.

Una vez obtenidas las grabaciones se realizó el proceso de extracción manual de cada uno de los alófonos de las palabras en que estos se encontraban, tomando únicamente el estado estacionario de la forma de onda y reemplazándolos en las otras palabras no seleccionadas a fin de escuchar el resultado final, escogiendo el que mejor se adaptara y sonara más natural; después se retocaba el alófono seleccionado y se adicionaba al archivo base etiquetándolo cuidadosamente. El mismo procedimiento se realizó para cada uno de los alófonos, añadiendo otros sonidos no planificados, con el objetivo de obtener mejores resultados en algunas palabras.

4.2 Entorno del programa

En esta sección se mostraran algunos de los formularios que se despliegan al ejecutar el programa desarrollado (En la figura 2 se muestra la ventana principal del programa).

Figura 2. Aspecto del formulario de inicio del programa.

El formulario que se presenta en la figura 3 es el que permite al usuario variar algunos de los rasgos prosódicos de la voz generada, en este caso la frecuencia de reproducción, permitiendo realizar modificaciones en un rango entre 8000 Hz y 44100 Hz, en donde el valor por defecto es este último. Se aprecia además, el control que permite variar el volumen o intensidad de la voz generada.

Figura 3. Formulario de configuración de algunos parámetros prosódicos de la voz del sistema.

El otro parámetro es el de la velocidad, que permite realizar variaciones al tono de la voz; el método empleado es en el dominio temporal, mediante la adición o supresión de muestras del archivo base de alófonos, en forma similar al procedimiento seguido por el algoritmo TD-PSOLA (Time Domain Pitch Synchrounous Over-Lap Add), pudiéndose variar el

tono entre un rango de 0.1 a 2 veces la frecuencia fundamental de la voz, con lo que se consiguen sonidos más agudos o más graves según la configuración deseada por el usuario.

También existen 3 cajas de texto que permite al usuario ingresar el tiempo, en milisegundos, que desea que el programa le asigne a la reproducción de los signos ortográficos: coma, punto y espacio entre palabras. Finalmente, existe una caja de comprobación que permite seleccionar si se desea que el programa lea textualmente los signos ortográficos, es decir que si el texto contiene el siguiente símbolo "*(* ", se reproduzca como *"abre paréntesis…*" y así con los demás signos ortográficos.

La figura 4, muestra el menú y los comandos disponibles del AUDIOTEXTO que aparecen en la barra de tareas del programa Microsoft Word una vez que el programa ha sido instalado. Este menú permite emplear el AUDIOTEXTO como lector de documentos de Word, así como realizar configuraciones al sistema, iniciar (play), pausar (pause) y detener (stop) la reproducción de la voz que el programa genera. Para conseguir la comunicación entre las dos aplicaciones se empleó la tecnología ActiveX (clases y objetos) que ofrece Visual Basic, así como la programación de macros en el lenguaje de programación Visual Basic para Aplicaciones que emplea Microsoft Office y el uso de las automacros de Word.

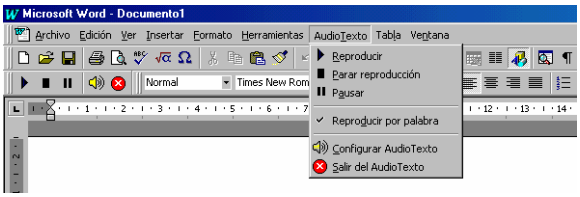

Figura 4. Menú desplegable en Microsoft Word para el enlace con el AUDIOTEXTO.

5. PRUEBAS Y RESULTADOS FINALES

Con el fin de evaluar la capacidad del programa de reproducir cualquier palabra o texto, se le sometió a la lectura de diversos documentos escogidos al azar, así como la lectura de palabras no congruentes, mal escritas, o aún de otros idiomas, observándose que gracias a su fuerte base de conocimientos y de alófonos grabados el sistema podía leerlas o en caso contrario deletrearlas. Además se probó la capacidad de variar el tono de voz empleando una misma palabra pero con acentuación diferente, apreciándose el efecto en casi todas las pruebas. Se le sometió a la lectura de frases enteras, pero el sistema mostró falencias al no poder ir variando el tono de la voz a medida que se lee la oración. Por último, se pidió a varias personas que escucharan la reproducción de diversas palabras generadas y que luego dijesen cual había sido leída, el resultado final mostró que las personas comprendían entre un 70 a 90 % las palabras, pero que esta les resultaba un tanto artificial.

Las figuras 5 y 6, permiten comparar visualmente, por medio de un ejemplo, los resultados obtenidos en la reproducción de una palabra dicha por el locutor que hizo las grabaciones y la misma palabra generada por el programa. Se observa que en términos generales presentan una forma similar, así como una duración casi igual. Sin embargo, se observan diferencias en el inicio de las palabras, pues en la frase original se presenta una ascensión exponencial, mientras que en la palabra sintetizada el inicio es abrupto. Esto como consecuencia del empleo de un único alófono "L", el cual el programa usa indistintamente al inicio, intermedio o final de palabra.

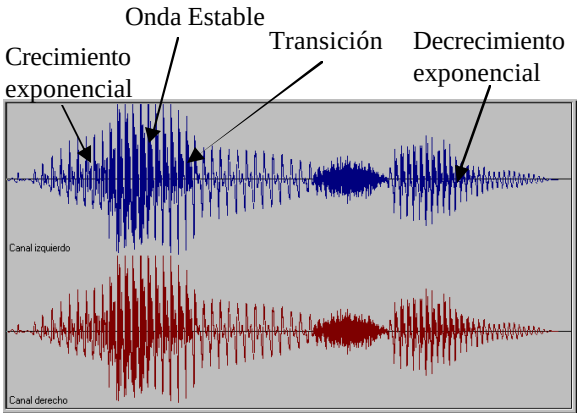

Figura 5 Aspecto de la onda original para la palabra LANZA

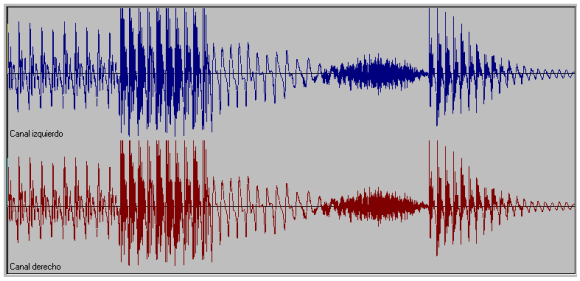

Figura 6. Aspecto de la onda generada por el sistema para la palabra LANZA.

Otra de las divergencias son las transiciones entre alófonos que se observan en la voz natural y que se producen por la inercia de los órganos articulatorios al pasar de un sonido al otro; en este proyecto, dichas coarticulaciones no se tuvieron en cuenta, por lo que el sonido final no presentó tanta naturalidad como se hubiese deseado. Esta carencia de transición se puede observar más claramente entre el paso de los sonidos de *L* y *A* y de *A* a *N* según se muestra en las figuras 5 y 6.

6. CONCLUSIONES

Dado que los sistemas de conversión de texto a voz presentan grandes posibilidades en diversos campos como la informática, robótica, control, y en general en la vida diaria, su estudio y desarrollo es apremiante. Este proyecto se constituye en un intento por dar una solución fáctica que propicie el avance de nuevas y mejores técnicas. Partiendo del trabajo realizado hasta el momento, se espera mejorar el programa a fin de otorgarle más naturalidad a la voz generada, teniendo en cuenta las transiciones entre los alófonos, bien sea generándolos, mediante técnicas inteligentes (Fuzzy) o ampliando la base de reglas y grabaciones del programa. Otra de las mejoras, que ya se están estudiando, es poder variar de manera más eficaz el tono de la voz al leer frases de diferentes tipos (exclamativas, afirmativas, interrogativas, etc.) y de diversos tamaños. Finalmente quedan por realizar mejoras generales al programa, que permitan perfeccionar el proceso de normalización del texto (manejo de abreviaturas, siglas, símbolos) a fin de garantizar una lectura más acorde al contexto.

REFERENCIAS

Alarcos Llorach, Emilio. (1983). *Fonología española*. Gredos, Madrid, España.

Alconchel, José Dominguez. (1997). *Superutilidades para Visual Basic*. Osborne-McGraw Hill, España.

Alvarez, Luis Eduardo. (1977). *Fonética y Fonología del Español*. La Cátedra, Bogotá, Colombia.

Ceballos, Francisco Javier. *Microsoft Visual Basic aplicaciones para Windows*. Addison-Wesley Iberoamericana.

Deller, John R. Jr, Proakis, John G., Hansen, y John H.L. (1993). *Discrete-time processing of speech signals*. MacMillan, USA.

González Quintero, Alberto José y Plata Ramos, María Fernanda. (1990). *Diseño e implementación de una aplicación que reproduzca vocalmente palabras en español*. Tesis: Pontificia universidad Javeriana. Facultad de ingeniería de sistemas. Santiago de Cali, Colombia.

Llorens camp, María Jose. (1995). *Ortografía práctica*. M.E. Editores, España.

Lopeda Vargas, Gerardo y Osorio Saenz, Ernesto. (1997). *Diseño de software para enseñanza de lenguaje hablado a niños sordos entre 5 y 6 años de edad*. Tesis: Universidad del Valle, departamento de electricidad. Santiago de Cali, Colombia.

Malmberg, Bertil. (1964). *La fonética*. Eudeba, Buenos Aires, Argentina.

Massone, Maria I, Borzone de Manrique, A.M. (1985). *Principios de transcripción fonética*. Macchi, Córdoba, Argentina.

Proakis, John G, Manolakis, Dimitris G. (1998). *Digital signal processing: principles, algorithms and aplications*. Prentice Hall, USA.

Rosch, Winn L. (1996). *Todo sobre el multimedia*. Prentice Hall, México.

Sadaoki, Furui. (1992). *Digital speech processing, synthesis and reconigtion*. Marcel Dekker INC, New York.

Sommerstein, Alan H. (1977). *Fonología moderna*. Cátedra S.A, Madrid, España.

Violaro, Fabio y Böefford, Oliver (1998). A hybrid model for text to speech synthesis. In: *IEEE transactions on speech and audio processing*, Vol. 6 No. 5 septiembre 1998, 426 – 434.