Diccionario de difonos del sintetizador de voz SEVEN. Manual de instalación y de uso

Diphone dictionary of speech synthesizer SEVEN. Instructions for installation and use

Nelson Rojas
María Alejandra Blondet
Elsa Mora
Laboratorio de Fonética
Universidad de Los Andes
Mérida, Venezuela
nelsonrojasavendao29@gmail.com
blondetma@gmail.com
elsamora@ula.ve

Resumen

Presentamos en este artículo el diccionario de difonos (en versión escrita y en audio) utilizado por el sintetizador de voz del español venezolano (en adelante, SEVEN), así como la lista y las secuencias de palabras empleadas para construirlo. A través de la lista de palabras el usuario del sintetizador de voz tendrá la oportunidad de conocer las múltiples combinaciones de sonidos del español hablado en Venezuela utilizadas para producir la voz en el computador. Por otra parte en este artículo también ofrecemos al lector la información necesaria para acceder, de forma gratuita, al sintetizador de voz, junto con un didáctico manual de instalación y de uso.

Palabras clave: Sintetizador de voz, diccionario de difonos, manual.

Abstract

In this article, we present the dictionary of diphones (written and audio version) used by the venezuelan Spanish voice synthesizer, as well as the word list and word sequences used to build it. Through this word list, the voice synthesizer's user will have the opportunity of knowing the multiple combinations of the sounds in the Spanish spoken in Venezuela, used to create/produce the voice of the computer. On the other hand, in this article, we also give to

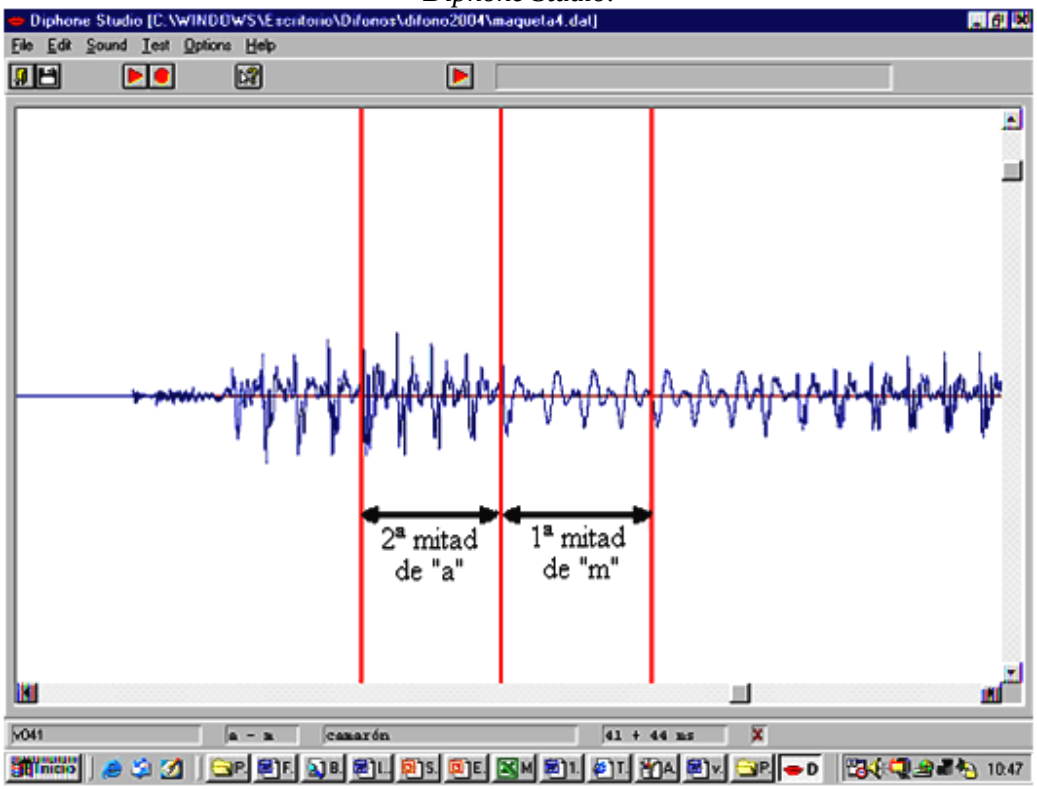
the reader the necessary information to access, in a free way, to the voice synthesizer and also, a didactic installation-and-use manual for it.

Key words: voice synthesizer, dictionary of diphones, manual.

1. INTRODUCCIÓN

La voz de SEVEN es generada utilizando como base un diccionario de difonos venezolanos. Los difonos son los elementos que nacen en la unión de dos sonidos, específicamente, en la transición de dos elementos articulados en el habla oral y que abarcan desde la mitad de un fono hasta la mitad del fono siguiente. La figura 1 ilustra el difono a-m, que inicia en la mitad de "a" (primera línea vertical) y termina en la mitad de "m" (tercera línea vertical) e incluye toda la señal intermedia. La segunda línea vertical indica la frontera entre "a" y "m".

Figura 1. *Segmentación del difono "a-m" extraído de la palabra "camarón" con el programa Diphone Studio.*



Esto quiere decir entonces que la unidad utilizada por el computador para producir la voz no es una palabra, una frase o un fono sino que es una unidad de frontera, el difono. La ventaja en la utilización de esta unidad para la síntesis reside en el hecho de que minimiza los problemas ligados a la concatenación de sonidos, ya que incorpora la mayor parte de transiciones y coarticulaciones entre fonos dentro de la unidad, dentro del difono. El número total de difonos requeridos para sintetizar el habla resulta relativamente pequeño, para el español se necesitan 794 unidades.

En este artículo se presenta, en primer lugar, el diccionario de difonos del español hablado en Venezuela (EHV) que utiliza SEVEN. Este diccionario o base de datos de los difonos del EHV fue confeccionada por el grupo de investigación al cual pertenecemos. Pretendemos, ofrecer esta lista de palabras y las secuencias de palabras en las frases portadoras del difono con el audio respectivo, a instructores o profesores de lengua española como segunda lengua, con el fin de que muestren a sus estudiantes todos y cada uno de los sonidos de la lengua en contexto.

Este artículo ofrece, en segundo lugar, un manual de instalación y de uso de SEVEN. Este sintetizador ha resultado de valiosa ayuda para las personas con discapacidad visual y con discapacidad auditiva. En personas con discapacidad visual, el sintetizador ha resultado muy útil como herramienta de lectura. Esta función se ve magnificada con la posibilidad de convertir el texto escrito en un archivo de audio con formato *.wav*, que puede soportar casi cualquier *codec* de audio, lo que implica que se puede reproducir en cualquier reproductor multimedia. Las personas con discapacidad auditiva utilizan el sintetizador como una voz mediante la cual pueden establecer comunicación con personas oyentes que no hablan la Lengua de Signos Venezolana (LSV).

2. ¿CÓMO SE CONSTRUYÓ EL DICCIONARIO DE DIFONOS PARA SEVEN. CÓMO INTERPRETAR LA CODIFICACIÓN DE LOS DIFONOS DE SEVEN?

El inventario de difonos del español es el producto de un conjunto de combinaciones de sonidos del español. Para efectuarlo se tomaron en cuenta solo las combinaciones de los sonidos que fuesen efectivamente articulados, producidos en el español de Venezuela**.** Este diccionario incluye todas las consonantes y todas las vocales del español venezolano, acentuadas y no acentuadas, en todas las posiciones silábicas posibles, considerando las características de los fonemas consonánticos en posición inicial o final de sílaba, así como los grupos consonánticos y las glides.

Se localizaron palabras del español venezolano donde estuvieran presentes las combinaciones de los sonidos que conformaban los difonos. De esta forma se construyó una lista de palabras con la fonética típica o característica del español venezolano (Acceda al archivo de Excel desde la Tabla de Contenidos de la revista). Cada una de esas palabras, insertada en una frase portadora, fue pronunciada y grabada por un locutor nativo (Acceda al archivo de audio desde la Tabla de Contenidos de la revista).

Las frases grabadas fueron digitalizadas y el proceso de segmentación y etiquetaje de los difonos se realizó a través de los programas *Praat y el Diphone Studio* (Cf. Rodríguez *et al.* 2006a, 2006b).

Se creó de esta forma la base de difonos o diccionario de difonos, llamada *vz1*, que se encuentra reseñada en la página web www.tcts.fpms.ac.be/synthesis/*Mbrola*.html, puesto que la síntesis de voz fue realizada utilizando la técnica *Mbrola*. Con el fin de dotar a la voz sintetizada de naturalidad fue necesario establecer una serie de criterios fonético-fonológicos. Por ello, en la frontera de los sonidos, se combinaron armoniosamente aspectos físicos como la intensidad, la frecuencia fundamental y el espectro. Simultáneamente, en nivel de la frase global, se aplicaron funciones de duración, intensidad y frecuencia fundamental para simular

los aspectos suprasegmentales del habla natural. Esto se logró por medio del procesamiento digital de los difonos (siguiendo a Dutoit, 1997).

La lista de realizaciones fónicas consideradas en el diccionario de difonos, salvo excepciones, fue codificada utilizando el código SAMPA (Speech Assessment Methods Phonetic Alphabet). Es éste un Alfabeto fonético que posee dos virtudes: la primera, que al funcionar con caracteres ASCII facilita el trabajo en el computador; y la segunda, que cada símbolo representa a un solo sonido del AFI (Alfabeto Fonético Internacional).

A modo de ejemplo, la frase 'voy a la casa' [bojala´kasa] se produce con once difonos:

#b;bo;oj;ja;al;la;ak;ka*;a*s;sa;a#. (transcripción de los difonos utilizando el SAMPA)

Rodríguez *et al.* (2006a) proponen una modificación en el alfabeto SAMPA para adaptarlo a las idiosincrasias del español hablado en Venezuela, de modo que se establece un inventario de 23 sonidos consonánticos, incluyendo glides, y 10 sonidos vocálicos, diferenciando las vocales tónicas de las átonas; también se propone la distinción entre el fonema fricativo alveolar sordo en posición inicial, [s], y su alófono fricativo glotal que se realiza generalmente en posición final de sílaba, [s2].

Inventario de sonidos del español venezolano presentado por Mora *et al.* (2000) y Rodríguez *et al.* (2006a) con su representación SAMPA:

Vocales

Vocales inacentuadas: a, e, i, o, y u.
Vocales acentuadas: a*, e*, i*, o*, y u*.

Consonantes

Oclusivas sordas: p, t, y k (en posición inicial de sílaba).
Oclusivas sonoras: b, g, y d (en posición inicial de sílaba).
Fricativas sonoras: B, D, (en posición inicial de sílaba) y G (corresponde a la realización de "g" antes de consonantes líquidas, como "l" o "r", antes de vocales abiertas como "a", "e", "o", o antes vocales cerradas precedidas por "u", como "ui"o "ue").
Africada: tS (realización que corresponde a la pronunciación de "ch").
Fricativas sordas: f, s, s2 (un alófono en distensión de la "s"), y h (realización para la cual no hay código SAMPA. Corresponde a la pronunciación de "j" y "g" antes de "e" e "i", así como a la pronunciación del fonema/s/en posición final de sílaba).
Nasales: m, n, J (realización que corresponde a la pronunciación de "ñ").
Laterales: l, L (realización que corresponde al fonema fricativo palatal sonoro para el cual no hay código SAMPA. Corresponde a la pronunciación de "ll" de "llave" y a la "y" de "ayer"), r, y rr.
Glides: j, w (correspondientes respectivamente a "i" y "u" en diptongos, como en "quieto" y "jueves").

Pausa: _.

De esta forma se creó el diccionario de difonos del habla venezolana, lo que equivale a decir, un inventario de todos los sonidos articulados en esta variedad lingüística (Anexo 2).

3. EL SINTETIZADOR TEXTO A VOZ

Con el diccionario de difonos del español venezolano creado fue posible "alimentar" el sintetizador texto a voz. Esta segunda sección del artículo está destinada a guiar al lector durante el proceso de instalación de éste. En este apartado se enseña al usuario cómo al teclear en su computador una frase esta puede ser enunciada "sintéticamente" en español venezolano.

El programa *tafv.pl* (Texto a Fonema Venezolano) es un programa en el lenguaje de programación *Perl* para la conversión automática de texto ortográfico a un archivo *.pho* para la síntesis con el método *Mbrola* del español venezolano que utiliza la base de datos *vz1*, de difonos venezolanos.

El programa consta de un gran número de subrutinas, cada una de las cuales ejecuta una pequeña parte del procesado lingüístico del texto de entrada. Entre otras cosas realiza la transcripción ortográfico-fonética del texto de entrada y coloca el acento correspondiente al acento prosódico.

A continuación, el programa da información sobre palabras no acentuadas, palabras monosilábicas que portan acento prosódico y palabras excepcionales que terminan en *-mente* con un solo acento. Seguidamente se aplica una regla de silabificación, de tal manera que sólo las sílabas posibles sean reconocidas por el convertidor.

El programa establece la función de entonación en dos partes[1]: primero hace una asignación de entonación simbólica, utilizando las etiquetas ToBI (Tones and Break Indices), y después, en función de los símbolos ToBI, asigna una entonación numérica a fonemas puntuales. Finalmente, entre los valores puntuales, el programa *Mbrola* realiza una interpolación lineal. En los actuales momentos el sintetizador, además de hacer la transcripción del texto, maneja: dígitos cardinales, números romanos, siglas, abreviaturas, símbolos especiales (%, $, €, etc.).

4. ADQUISICIÓN, INSTALACIÓN Y MANUAL DE USO DEL SINTETIZADOR

El sintetizador de español venezolano puede ser descargado gratuitamente de la siguiente dirección electrónica: http://webdelprofesor.ula.ve/ingenieria/hourcade/. En esta página se encuentra un vínculo que dice "bajar sistema de conversión de texto a voz venezolana", al presionar sobre él estaremos descargando una carpeta denominada *Síntesis Completa* en cuyo interior encontraremos los dispositivos necesarios y los requerimientos de sistema para hacer funcionar el sintetizador de español venezolano. La carpeta de instalación *Síntesis Completa* trae los siguientes dispositivos:

1) La plataforma de *Active-Perl, ActivePerl-5.8.6.811-MSWin32-x86-122208*
2) El juego de herramientas *Mbrola Tools35*
3) El juego de herramientas *Praat4301_win*
4) La base de datos de difonos venezolanos, *vz1*

[1] La arquitectura del programa prevé la asignación de la entonación, sin embargo vale decir que aún no se encuentra disponible.

5) Un *readme* especificando la licencia de uso de este sistema
6) Un juego de programas y archivos acompañantes, *Tyna-win, venezuelan_rules1, venezuelan_rules2, símbolos, deletrear*
7) Un pequeño archivo *Word* con frases cortas para probar el sistema, *Frases para probar*.

El sistema para operar estas herramientas debe ser Windows, incluyendo Windows98, WindowsNT, WindowsXP, Windows Vista y Windows 7. Se hace imprescindible dispositivos de salida de audio.

Las diferentes funciones que cumple este juego de programas son las siguientes:

1) Los archivos acompañantes (*Tyna-win, venezuelan_rules1, venezuelan_rules2, símbolos, deletrear*) son los que hacen la transcripción ortográfico-fonética. Estos están escritos en lenguaje *Perl*, por esta razón hay que instalar la plataforma *Perl*.

2) Por otra parte, se crea un archivo *.pho* en el formato *Mbrola*, por lo cual se hace necesario instalar el juego de herramientas del programa *Mbrola*, cuyo fin es la concatenación de los difonos que en este caso se toman del banco de difonos venezolanos, *vz1*.

3) Finalmente, el programa *Mbrola* genera un archivo de sonido, y para escucharlo se requiere el software de reproducción, que cualquier sistema Windows normalmente trae incorporado, y dispositivos de salida de audio. Todo este software es gratuito y solo se pide hacer el reconocimiento debido a los autores.

Para la instalación del Sintetizador de español venezolano se recomienda primero hacer una copia de la carpeta *Síntesis Completa* en su computador, la cual puede ser proporcionada en un *CD* o cualquier otro dispositivo de almacenamiento o directamente de la dirección electrónica antes mencionada. Luego de que la carpeta esté guardada en el computador se debe abrir y seguir los siguientes pasos:

1) Hacer doble clic sobre icono *Active-Perl* 

y seguir las instrucciones para instalar la plataforma *Perl* en el computador.

2) Hacer doble clic sobre *Mbrola Tools35* 

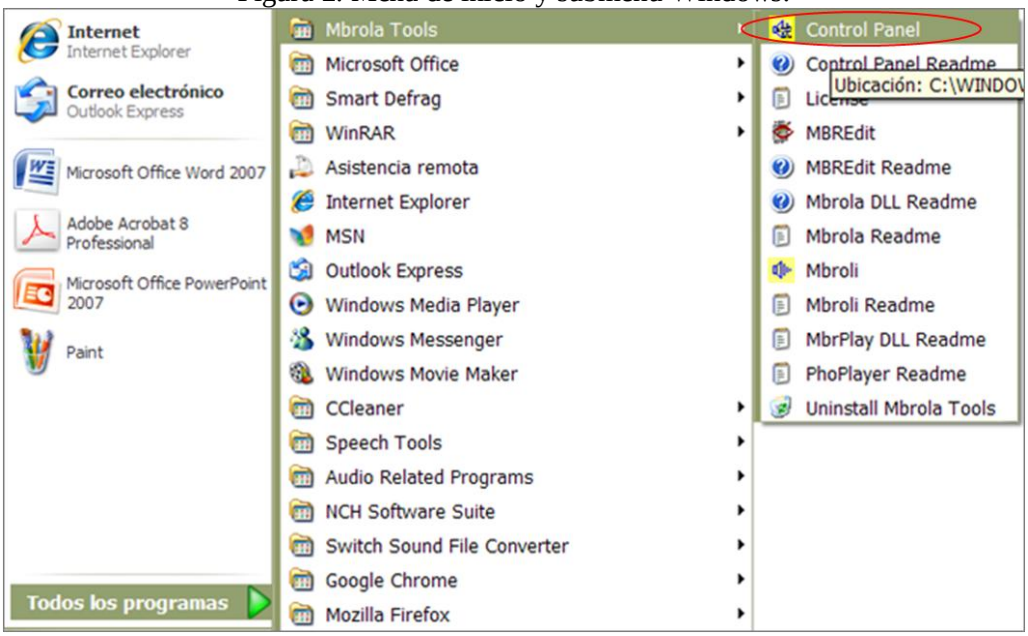
y seguir las instrucciones para instalar el juego de herramientas *Mbrola* en el computador.

3) Hacer doble clic sobre *Praat4301_win* 

y seguir las instrucciones para instalar el juego de herramientas *Praat* en el computador. Esta herramienta entre otras cosas sirve para reproducir archivos de sonido.

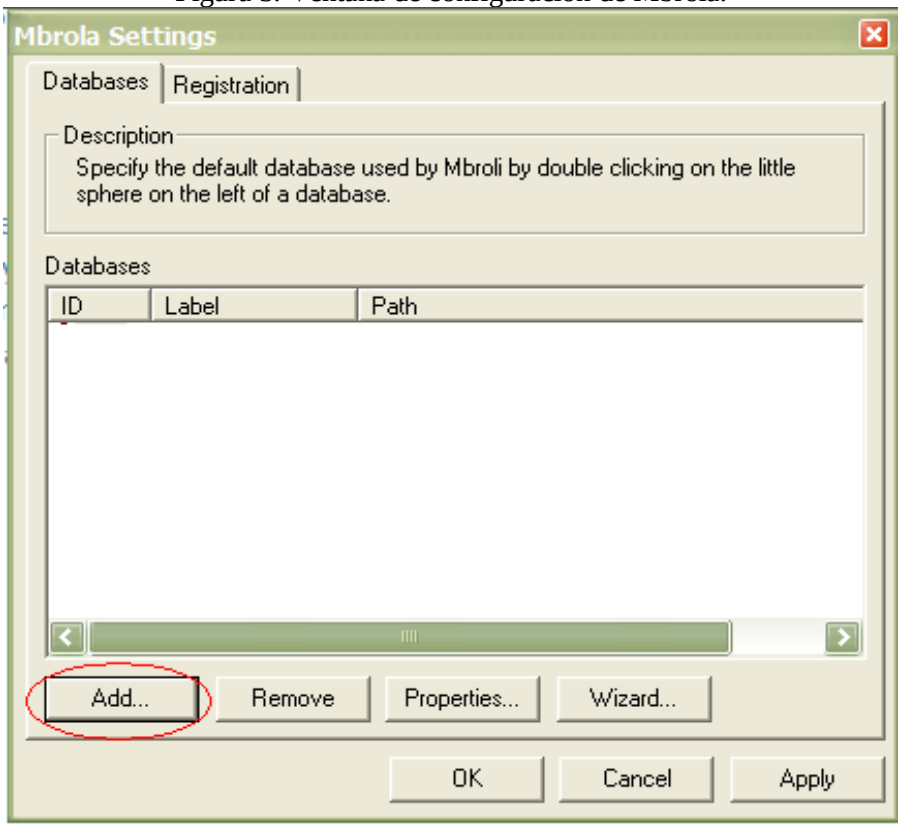
4) Ya instalados los programas se debe ir a *Inicio*, seleccionar *Todos los programas* y buscar la carpeta *Mbrola Tools*; se desplegará entonces un submenú en el cual se muestra el icono *Control Panel* de *Mbrola* como se ve en la figura 2.

Figura 2. Menú de inicio y submenú Windows.



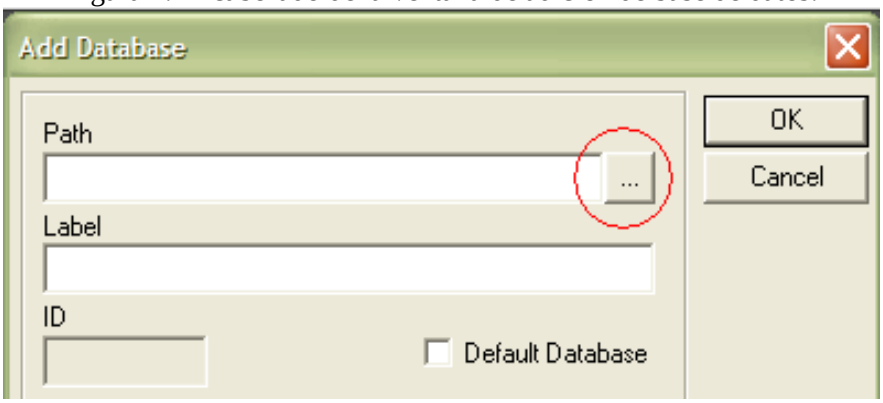
Al tocar este icono aparece una ventana como la que se muestra en la figura 3. En ella debemos introducir la base de datos de difonos *vz1*.

Figura 3. Ventana de configuración de Mbrola.



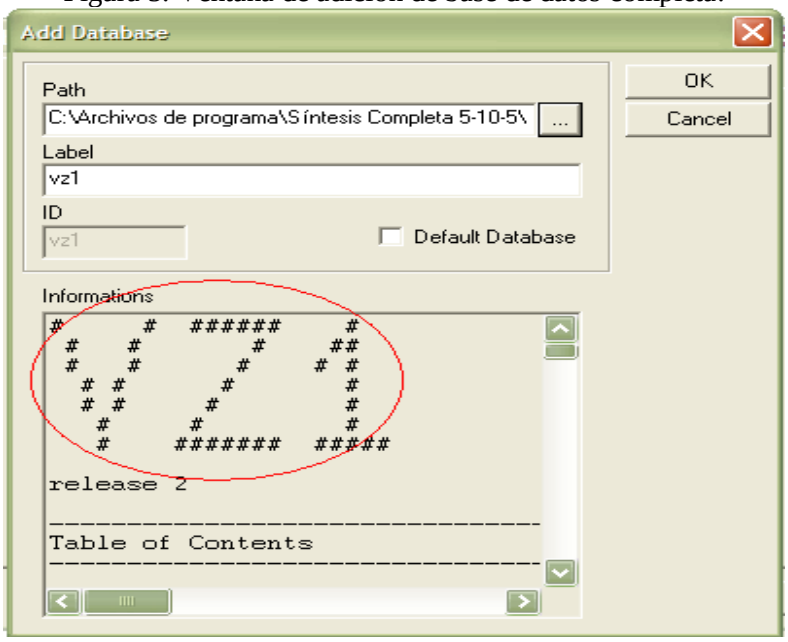
Esto se hace pulsando sobre *Add*, luego de lo cual aparece otra ventana en la que pulsaremos el botón situado a la derecha de la fila *Path*, tal como se muestra en la figura 4.

Figura 4. Encabezado de la ventana de adición de base de datos.



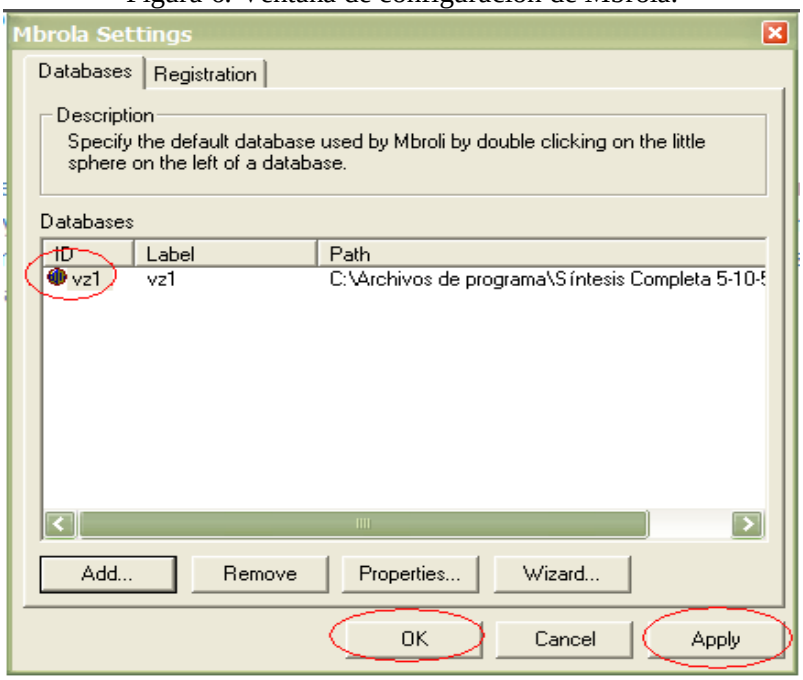
Esto nos lleva a *Mis documentos*, donde buscaremos nuevamente la carpeta *Síntesis Completa* y una vez abierta buscaremos la base de datos de español venezolano *vz1* y le ordenaremos abrir. La ventana nos indicará que la base de datos está lista para cargarse cuando aparezca en el recuadro *Información* los caracteres *VZ1*, como se muestra en la figura 5.

Figura 5. Ventana de adición de base de datos completa.



Luego se pulsa *OK* y se abre otra ventana como la mostrada en la figura 6. Ella nos indica que la base de datos está cargada y lista para funcionar; solo resta pulsar sobre *Apply* y luego *OK*.

Figura 6. Ventana de configuración de Mbrola.

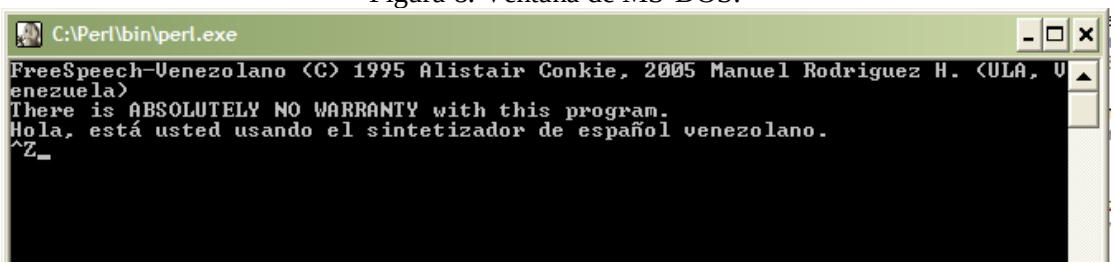


Luego de todos estos pasos debería estar instalado el sintetizador de español venezolano, lo cual podemos comprobar si abrimos nuevamente la carpeta *Síntesis Completa* y presionamos el icono *Tyna-win* cuyo logo, como se aprecia en la figura 7, es una lagartija.

Figura 7. Icono Tyna-win.

Debe aparecer una ventana negra de *MS-DOS*, o su simulación, con un mensaje que comienza con *Free Speech venezolano...* En esta ventana, tal como lo enseña la figura 8, se puede comenzar a escribir el mensaje que queremos expresar mediante el sintetizador. Al término de introducir el texto, se debe teclear *Enter+crtl+z* (de manera simultánea)+*Enter*.

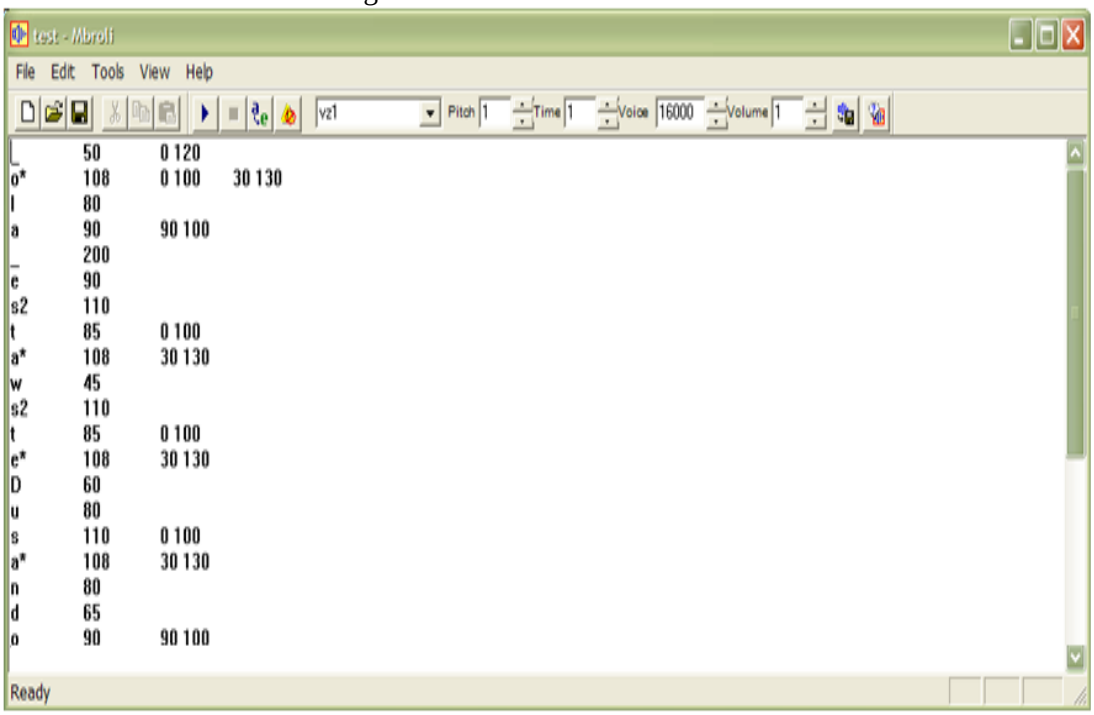
Figura 8. Ventana de MS-DOS.



En ese momento, debe desaparecer la ventana negra y debemos pulsar el icono *Test*, cuyo logo es una corneta sobre un fondo amarillo[2]. Este se encuentra en la misma carpeta *Síntesis Completa*. A continuación, aparecerá una ventana con el contenido del archivo y automáticamente debe reproducir la frase correspondiente al texto introducido, como se ve en la figura 9.

[2] Para un acceso más rápido a estos dos iconos (*Tyna-win* y *Test*) es recomendable hacer un acceso directo a escritorio.

Figura 9. Ventana de Test- Mbrola.



En la ventana *Test* se puede modificar algunos valores predeterminados en el sintetizador como frecuencia, velocidad, volumen. Esto, como se aprecia en la figura 10, se hace mediante la barra de herramientas ubicada en el cintillo superior de la ventana.

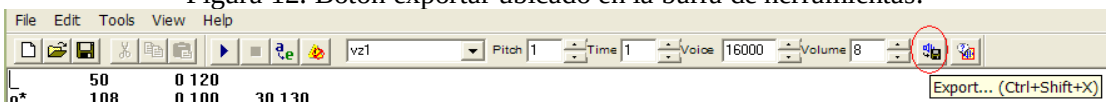
Figura 10. Valores de frecuencia, velocidad y volumen.

Cabe destacar que el sintetizador tiene la capacidad de albergar una gran cantidad de texto. Por ejemplo, se puede copiar de un archivo de texto algo que se desee reproducir y pegarlo en la pantalla, para lo cual debemos pulsar sobre *Tyna-win* y abrir la ventana de *MS-DOS* que se muestra en la figura 11. Una vez abierta, pulsamos sobre el ícono de la esquina superior izquierda, bajo el cual se desplegará un menú, en este menú se selecciona *Editar* y luego *Pegar*.

Figura 11. Ventana de MS-DOS.

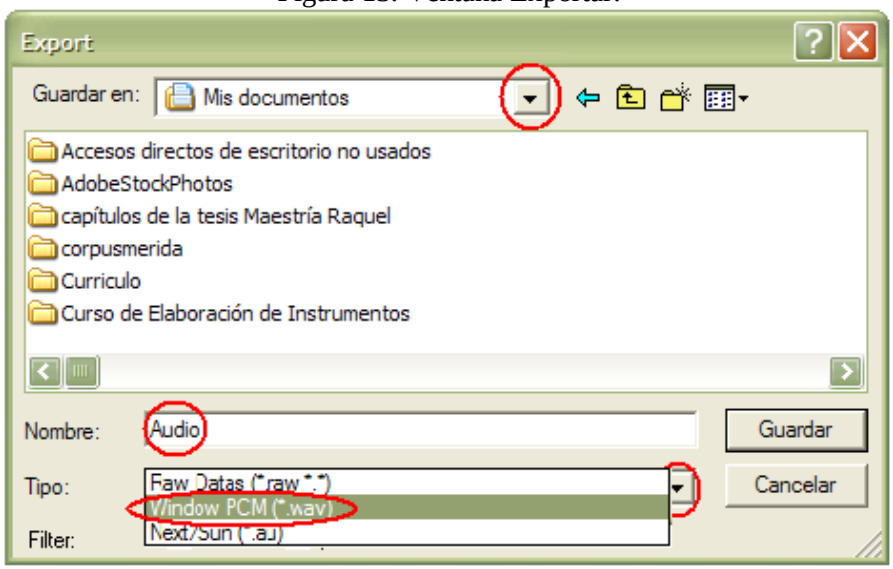
Este texto reproducido por el computador puede guardarse luego en formato de sonido *.wav*, cuya reproducción es posible en la mayoría de los reproductores de sonido de cualquier computador e incluso teléfonos celulares si es transformado a formato *MP3*. Esta función resulta muy útil para aquellas personas con discapacidad visual o auditiva y además es muy sencilla de realizar. Para hacerlo debe estar abierta la ventana *Test* y en la barra de herramientas se debe presionar el botón *exportar*, mostrado en la figura 12, o las teclas *Ctrl+Shif+X*.

Figura 12. Botón exportar ubicado en la barra de herramientas.



Luego, como se muestra en la figura 13, se selecciona la carpeta donde se guardará el archivo, se le pone nombre y se selecciona el formato, en este caso, el formato *.wav*.

Figura 13. Ventana Exportar.



Ciertamente, existen en el mercado otros sintetizadores con características similares, sin embargo el aquí presentado tiene la ventaja de haber sido realizado con los sonidos del español venezolano, además de poderse obtener de forma gratuita.

5. LA EVALUACIÓN DE LA CALIDAD DE LA VOZ SINTÉTICA DE SEVEN

La *calidad* en el ámbito de las tecnologías de habla, específicamente en el marco de la síntesis de habla, está en relación con la *lengua natural* (*lenguaje natural*), y en la evaluación de la calidad de la síntesis de habla dos dimensiones fundamentales deben ser consideradas:

1) La *inteligibilidad*
2) La *naturalidad*[3]

La primera se refiere sobre todo a la adecuada forma de encadenar los diferentes sonidos lingüísticos. Si observamos, en el habla espontánea no yuxtaponemos un sonido [p] a un sonido [a], ambos conforman una sílaba [pa], es decir una unidad que se interrelaciona mediante transiciones entre ambos elementos por coarticulación; a su vez esta sílaba se puede unir a otra sílaba, [pá] por ejemplo, y así construir el signo [papá]. Es por ello que, como afirma Llisterri (2007): "las unidades a partir de las cuales se construye un sistema de síntesis no suelen ser sonidos aislados sino combinaciones de sonidos". El objetivo de la inteligibilidad es, por lo tanto, evaluar la adecuada unión entre las combinaciones de sonidos. Estas combinaciones suelen llamarse semisílabas o difonos. Las semisílabas son fragmentos compuestos bien por el inicio y el centro de la vocal de una sílaba y su final, bien por el centro de la vocal de una sílaba y su final. Los difonos, por su parte, son segmentos acústicos formados por: 1) la parte estacionaria de un fono, 2) la transición de éste a un segundo fono y, 3) la parte estacionaria del segundo.

La naturalidad, por su parte, está asociada al grado de similitud que tiene la síntesis con respecto a la *prosodia* de una *lengua natural*.

La evaluación de la inteligibilidad de esta voz sintética (Mora, 2008:325) se realizó a partir de un test de pares mínimos cuyo principio de constitución se basó en los ocho rasgos que aparecen en la tabla 1 (Cf. Mora *et al.*, 2005):

Tabla 1. Rasgos y oposiciones del test de pares mínimos en español venezolano transcrito en SAMPA.

Sonoro/sordo	b/p, d/t, g/k
Grave/agudo	f/s, p/t, b/d, m/n, m/ñ
Compacto/difuso	k/p, g/b
Interrupto/no interrupto	rr/l, c/s, k/h
Tenso/relajado	rr/r
Vocálico/no vocálico	l/d
Nasal/no nasal	m/b, n/d
Estridente/no estridente	c/y

[3] Cf. Llisterri *et al.* (s/f).

Los resultados arrojaron algunos inconvenientes en la percepción de los sonidos [+nasales], [+graves] y [+compactos]. Los fonemas afectados a nivel perceptivo fueron: /m,ñ,t,b/. Es importante resaltar que en ningún caso el porcentaje de percepción correcta fue menor al 67%.

En cuanto a la evaluación de la naturalidad, reflejada en el aspecto prosódico, es importante señalar que en el nivel global la comprensión es aceptable, pero a un nivel más restringido de patrones declarativos o interrogativos aislados, el nivel de comprensión no es favorable. Sin embargo, dado que el uso de este sintetizador actualmente está en niveles comunicativos generales, esta restricción resulta menos significativa.

6. AGRADECIMIENTOS

Damos las gracias especialmente al Grupo de Investigación en Ciencias Fonéticas (GICIFO) de la Facultad de Humanidades y Educación de la Universidad de Los Andes. Mérida, Venezuela.

7. REFERENCIAS BIBLIOGRÁFICAS

Dutoit, Thierry .1997. *An introduction to Text to Speech Synthesis*. Dordrecht: Klewer.

Llisterri, Joaquim, Carme Carbó, María Jesús Machuca, Carme de la Mota, Montserrat Riera & Antonio Ríos. La conversión de texto a habla: aspectos lingüísticos. http://liceu.uab.es/~joaquim/publicacions/Llisterri_Carbo_Machuca_Mota_Riera_Rios_04_Conversion_Texto_Habla.pdf (27 de febrero de 2012).

Llisterri, Joaquim. 2007. La conversión de texto en habla. http://www.raco.cat/index.php/quark/article/viewFile/54860/66190 (27 de febrero de 2012).

Mora, Elsa, Daniel Hirst & Christian Cavé. 2000. Développement et évaluation d'un systéme de synthése pour l'espagnol vénézuélien: projet et état d'avancement. *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence 19*, 91-98.

Mora, Elsa, Lourdes Pietrosemoli, Christian Cavé, Enrique Obediente & Erwin La Cruz. 2005. Un corpus de pares mínimos para el español de Venezuela. *Revista Lengua y Habla 9*, 117-122. Mérida: Centro de Investigación y Atención Lingüística. Universidad de Los Andes.

Mora, Elsa. 2008. Discapacidad y comunicación: una experiencia de fonética aplicada. *Estudios de Fonética Experimental 17*, 317-329. Barcelona: Universitat de Barcelona. Laboratori de Fonètica.

Rodríguez, Manuel, Elsa Mora & Christian Cavé. 2006a. Síntesis de voz en el dialecto venezolano por medio de la concatenación de difonos. *Revista Ciencia e Ingeniería 27 (1)*, 17-24. Mérida: Facultad de Ingeniería. Universidad de Los Andes.

Rodríguez, Manuel & Elsa Mora. 2006b. Conversor texto a voz en el dialecto venezolano por medio de la concatenación de difonos. *Revista Ciencia e Ingeniería 27 (2)*. 79-87. Mérida: Facultad de Ingeniería. Universidad de Los Andes.