Síntesis de Voz por Concatenación de Difonemas para el Español de Colombia

*Claudia V. CORREA, Hoover F. RUEDA, Henry ARGÜELLO**

Escuela de Ingeniería de Sistemas e Informática
Universidad Industrial de Santander
Bucaramanga, Santander, COLOMBIA
e-mail: cato0386@hotmail.com, hfarueda@gmail.com, henarfu@uis.edu.co
*Profesor Asistente Universidad Industrial de Santander

RESUMEN

Las tecnologías del habla, tanto el reconocimiento como la síntesis de voz, han ganado popularidad en los últimos tiempos, actualmente son utilizadas en una amplia variedad de aplicaciones y lugares alrededor del mundo. Dentro de un mismo idioma existen variaciones en la pronunciación, por esta razón se han desarrollado voces sintéticas para diferentes lugares en los que se habla la misma lengua, como es el caso del español en España, y Argentina. Este artículo presenta el desarrollo y evaluación de un sintetizador de voz de dominio ilimitado utilizando la técnica de concatenación de difonemas para el Español de Colombia, el desarrollo incluye desde la concepción del corpus de voz hasta los procesadores implementados, finalizando con los resultados obtenidos.

Palabras Claves: Síntesis de voz, Concatenación de difonemas, Corpus de voz, Español de Colombia.

1. INTRODUCCIÓN

Se entiende por síntesis de voz el proceso de producción de voz artificial partiendo de un texto de entrada. Su objetivo principal es la generación automática por medio de un sistema informático electrónico de una señal acústica que simule la voz humana, normalmente a partir de un texto de entrada [1]. La calidad de las aplicaciones que implementan esta tecnología es medida por diferentes parámetros como: el dominio para el cual fue desarrollada, la naturalidad y la inteligibilidad de la voz resultante y la complejidad del procesamiento [2].

En la concatenación de unidades, la selección de la unidad a utilizar debe contemplar la relación entre el tamaño del conjunto de unidades necesarias para cubrir todo el idioma español (base de datos) y la calidad de voz que se desee obtener. El artículo presenta el proceso realizado por los autores del mismo, para el desarrollo de un sintetizador de voz para el español de Colombia que utiliza la técnica de concatenación de difonemas; éste está distribuido en cuatro grandes secciones: la primera parte hace referencia a la creación del corpus de voz de difonemas; la segunda parte comprende lo concerniente a la síntesis de voz y los procesadores utilizados por el sintetizador desarrollado; posteriormente, en la tercera sección se presentan las pruebas realizadas al sintetizador para evaluar la calidad de la voz y los resultados obtenidos. Por último, las conclusiones del trabajo y las referencias utilizadas.

2. CORPUS DE VOZ

La síntesis de voz por concatenación de unidades requiere un corpus (base de datos) de voz, que contenga todas las unidades de éste tipo pertenecientes a la variante colombiana del idioma Español. Las unidades comúnmente utilizadas son: fonemas, difonemas, semisilabas, silabas, palabras y frases. Este trabajo involucra el uso de difonemas debido a que permiten obtener voz de buena calidad con una base de datos de tamaño razonable. Los difonemas deben su importancia a la inclusión de la coarticulación existente entre el primer y segundo fonema que los componen, reflejando una voz menos robotizada que al utilizar otras unidades de menor tamaño como los fonemas.

Diseño del Corpus

Esta fase comprende la identificación de los difonemas que formarán parte de la base de datos de voz. La Ec. (1) hace referencia a la cantidad de difonemas existentes, el cual viene dado por el número total de fonemas elevado al cuadrado, menos los difonemas que nunca se presentan.

$$difon_{totales} = fonemas^2 - difon_{inexistentes} \quad (1)$$

Teniendo en cuenta esta dependencia, es necesario identificar primero los fonemas correspondientes al Español de Colombia. Los fonemas son abstracciones mentales de los sonidos del habla, por esta razón es posible relacionar cada uno de ellos con una letra del alfabeto. Sin embargo, algunas de ellas no producen sonido como la "h", y en ocasiones, varias letras se asocian a un mismo fonema, así como una letra puede generar varios fonemas dependiendo de las letras que la

rodeen o de su posición dentro de una palabra. La Tabla 1 muestra el listado de los fonemas identificados y las letras que los producen.

Tabla 1. *Fonemas del Español de Colombia*

Fonema	Letras que lo producen
/a/	a
/b/	b , v
/ch/	Combinación c-h
/d/	d
/e/	e
/f/	f
/g/	g*
/i/	i
/j/	j , g*
/k/	c* , k , q
/l/	l
/m/	m
/n/	n
/ñ/	ñ
/o/	o
/p/	p
/r/	r*
/rr/	r* , Combinaciones con r*
/s/	s , z , c*
/t/	t
/u/	u*,ü
/x/	x , Combinación de c*
/y/	y , ll
/A/	á
/E/	é
/I/	í
/O/	ó
/U/	ú

**El fonema se presenta dependiendo de las letras vecinas.*

En Colombia no se hace diferencia al pronunciar una “s”, una “z” o una “c” seguida de “e” o “i”, por esta razón las tres letras están asociadas al fonema /s/, este fenómeno se denomina *seseo*. Situación similar se presenta entre las letras “y” y “ll”, y entre “b” y “v”; en las primeras porque actualmente es muy común el fenómeno del *yeísmo*, que implica la no diferenciación en la pronunciación de estas dos letras, y en las segundas porque según el diccionario panhispánico de dudas de la RAE [3] se asocian al mismo fonema /b/. Con base en la lista de fonemas, se realizan todas las posibles combinaciones entre ellos, obteniendo los difonemas. A partir de este nuevo listado se hace una selección, descartando aquellas combinaciones que nunca se presentan en el idioma.

La existencia de un difonema se confirma si existe al menos una palabra que lo contenga; esta selección se realizó con la ayuda del diccionario de la RAE [4]. De esta manera se obtuvo un total de 590 difonemas existentes en el Español de Colombia, para cada uno de los cuales se tiene una palabra ejemplo que lo contiene. Las palabras ejemplo escogidas para representar cada difonema, en su mayoría, lo contienen en la mitad, pues los difonemas que se encuentran en esta parte de la palabra son pronunciados de manera más natural que los que se encuentran en los extremos. Sin embargo, hay algunas excepciones, como es el caso de las palabras correspondientes a los difonemas de inicio y fin de palabra.

La Tabla 2 muestra una parte de la tabla de difonemas. El primer fonema del difonema está representado por la primera columna de la tabla y el segundo fonema por la primera fila. En cada celda aparece el nombre del difonema formado, y donde aparece únicamente un guión, significa que el difonema no existe. Así por ejemplo, el difonema ch-ch no existe, mientras que el difonema formado por los fonemas d y a se representa d-a. La tabla incluye el fonema “*pau”* y sus combinaciones, que representan las pausas o silencios de inicio y fin de palabra.

Tabla 2. *Fragmento de la Tabla de difonemas*

	pau	**a**	**b**	**ch**	**d**	**e**	**f**
pau	_-_	_-a	_-b	_-ch	_-d	_-e	_-f
a	a-_	a-a	a-b	a-ch	a-d	a-e	a-f
b	b-_	b-a	b-b	b-ch	b-d	b-e	b-f
ch	-	ch-a	ch-b	-	-	ch-e	-
d	d-_	d-a	d-b	-	-	d-e	-
e	e-_	e-a	e-b	e-ch	e-d	e-e	e-f
f	f-_	f-a	-	-	-	f-e	-

Condiciones de grabación

Teniendo los difonemas y las palabras ejemplo, el siguiente paso en el desarrollo del corpus es la definición de las condiciones de grabación. Las características tenidas en cuenta se resumen en la Tabla 3.

Tabla 3. *Condiciones de grabación del corpus*

Lugar	Cabinas de radio
Frecuencia de Muestreo	16 KHz
Resolución	16 bits
Número de Canales	1 (mono)
Formato de Grabación	WAV

La utilización de una cabina de radio como lugar de grabación permite que las señales de audio obtenidas contengan menor cantidad de ruido del exterior, y la voz se escuche con mayor claridad. La frecuencia de muestreo hace referencia al número de muestras por unidad de tiempo que se toman de una señal continua para producir una señal discreta, durante el proceso necesario para convertirla de analógica a digital; la resolución es la cantidad de bits utilizados para realizar el muestreo; estos dos valores de establecieron en 16 KHz y 16 bits respectivamente debido a que proporcionan buena calidad de audio, y el tamaño de los archivos relacionados no es muy grande, por lo tanto su procesamiento es rápido, favoreciendo la velocidad de respuesta del sintetizador. El número de canales

especifica la cantidad de sonidos diferentes que pueden reproducirse simultánea e independientemente; este valor se estableció en 1 puesto que solo se requiere un sonido a la vez, que es el de la voz.
Debido a que cada difonema será extraído de una grabación diferente, el conjunto de dichas grabaciones debe cumplir con los siguientes requisitos: Ser realizadas por un mismo locutor, mantener el tono y volumen al hablar, buena pronunciación y vocalización, neutralizar al máximo el acento nativo. Para la realización de las grabaciones se utilizó la siguiente frase portadora:

"Él dijo _____. Yo sé que él dijo _____." [5]

En cada espacio se insertó la palabra ejemplo seleccionada para cada difonema. El uso de la frase portadora permite al locutor mantener la naturalidad de la voz con mayor facilidad que cuando se leen palabras sueltas. También al pronunciar dos veces cada palabra dentro la frase es posible seleccionar la que tenga mejor pronunciación y así obtener una mejor representación del difonema.

Etiquetado y extracción

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 [6]. El proceso realizado en el desarrollo del presente proyecto involucró el etiquetado del corpus para identificar los difonemas. Las etiquetas utilizadas indican el comienzo y el fin del difonema, y la transición entre los fonemas que lo conforman. El etiquetado de las grabaciones se realizó utilizando el software Diphone Studio 1.3. Esta herramienta permite el desarrollo y mantenimiento de un conjunto de difonemas para su uso en síntesis de voz [7]. La Figura 1 presenta el etiquetado de un archivo de audio.

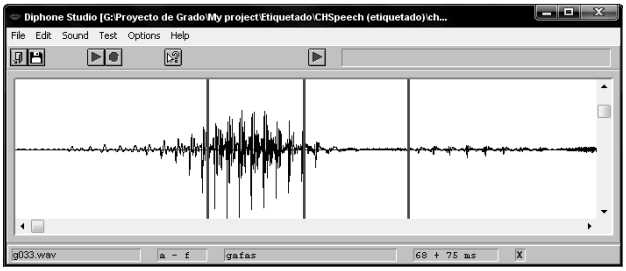

Figura 1. *Etiquetado de un archivo de audio*

La señal de audio de la palabra ejemplo es dividida por 3 líneas verticales que representan las etiquetas insertadas correspondientes al difonema. Con base en las etiquetas de cada difonema se procedió a extraerlo, desechando el resto de la palabra ejemplo. De esta manera, se conformó una base de datos con 590 archivos de audio, uno por cada difonema. Al ser extraídos, se les aplicó un filtro de *"normalización de señal"* con el fin de obtener audio estable en todas las grabaciones y mejor volumen. Este proceso se realizó debido a que algunos difonemas presentaban sonidos muy bajos casi imperceptibles que debían resaltarse para que no perdieran su significado.

3. SÍNTESIS DE VOZ

El proceso de síntesis se realiza a través de un motor de síntesis o sintetizador. La función de éste es convertir un texto de entrada en una salida de audio. Para lograr esto, es necesario que dicho texto pase por una serie de procesadores encargados de realizar diferentes tareas hasta llegar al objetivo.
Lo anterior se refleja en la Figura 2, que presenta un esquema del modelo de caja negra de un sintetizador.

Figura 2. *Modelo caja negra de un sintetizador*

El sintetizador desarrollado está compuesto de 6 procesadores que se ejecutan secuencialmente, es decir, que la salida de un procesador se convierte en la entrada del siguiente. La Figura 3 muestra un esquema del funcionamiento de estos procesadores.

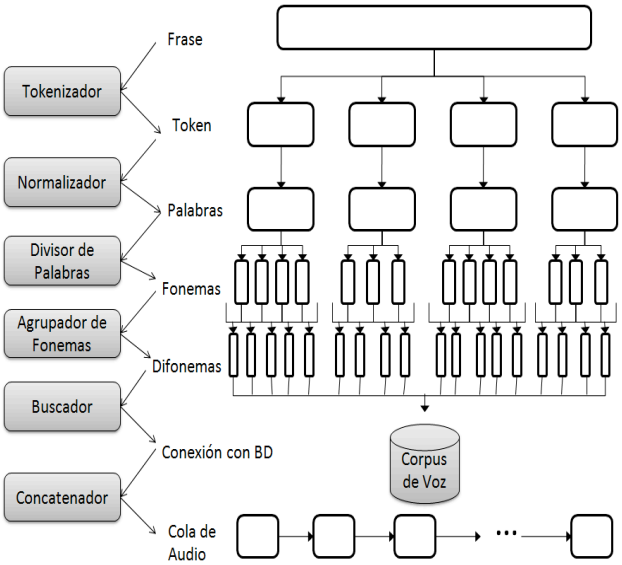

Figura 3. *Procesadores del sintetizador*

A continuación se explica en detalle la función que desempeña cada uno de ellos.

Tokenizador

A este procesador ingresa una cadena de caracteres que se ha denominado frase, y la salida que proporciona es una lista de trozos o "tokens". Su función es realizar particiones en el texto de entrada, tomando como separador de tokens, el espacio en blanco existente entre ellos.

Normalizador

La existencia de este procesador se debe a que el sintetizador desarrollado es de dominio ilimitado, esto implica que debe estar en condiciones de procesar entradas de texto que representen diferentes construcciones del idioma (fechas, horas, teléfonos, correos electrónicos, abreviaturas, etc.). La función que cumple el normalizador es identificar las estructuras ingresadas y a su vez, establecer el modo en que se procesará cada una de ellas. Para esto se definieron una serie formatos que permiten procesar las siguientes estructuras: fechas, horas, números telefónicos (celulares y fijos), números romanos, cantidades de dinero (en pesos), direcciones (residenciales), direcciones web, correos electrónicos, números fraccionarios, números ordinales, números decimales y enteros, o simplemente palabras. Además se incluyeron listados de símbolos, abreviaturas y algunos extranjerismos de uso común. La salida que proporciona el normalizador es una lista de palabras que corresponden a la manera como se leería la estructura ingresada.

Divisor de Palabras

Las palabras obtenidas por el normalizador ingresan al divisor de palabras, que se encarga de descomponer cada una de ellas en sus respectivos fonemas. Al comienzo y fin de cada palabra se inserta una pausa, ésta está representada por un fonema "silencio". La salida que proporciona este procesador es una lista de fonemas. La Figura 4 muestra el proceso que se realiza con la palabra "queso".

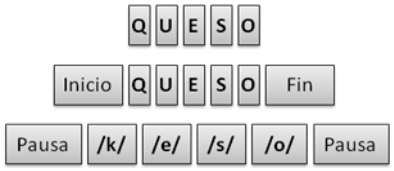

Figura 4. *Funcionamiento del Divisor de palabras*

Agrupador de fonemas

Al agrupador de fonemas ingresa la lista de fonemas generada por el divisor de palabras. La función de este procesador es, como su nombre lo dice, agrupar los fonemas de una palabra de manera que se obtenga esta palabra compuesta por sus difonemas. La Figura 5 presenta el proceso realizado por el agrupador de fonemas, aplicado al ejemplo de la Figura 4.

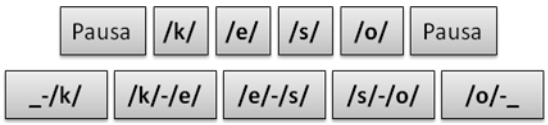

Figura 5. *Funcionamiento del Agrupador de Fonemas*

Buscador

La función que cumple es realizar la conexión con la base de datos, en busca del archivo de audio correspondiente a cada uno de los difonemas de entrada al procesador, obteniendo como resultado una lista con la representación sonora de los difonemas implicados.

Concatenador

Este es el último procesador del sintetizador; es el encargado de la generación de la señal de audio. A este procesador ingresa la lista de archivos correspondiente a los difonemas que se deben concatenar, para generar una única señal compacta que los contenga a todos y pueda reproducirse. Para cumplir su función, el Concatenador utiliza el Java Media Framework (JMF) [8]. Este framework permite que audio, video o cualquier otro tipo de archivo multimedia pueda ser añadido a aplicaciones y applets construidos con la tecnología Java.

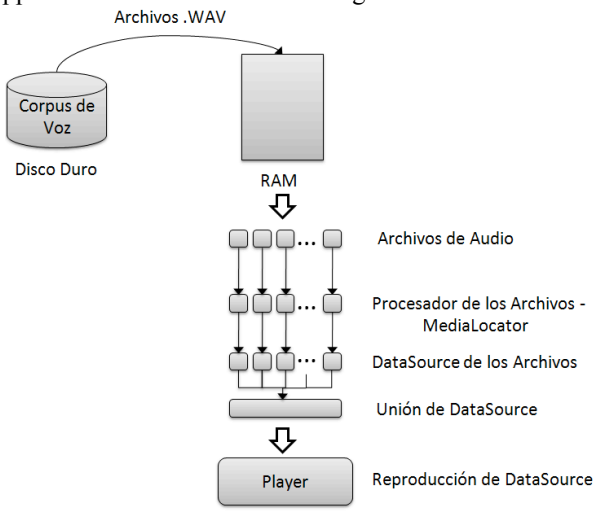

Figura 6. *Actividades del Concatenador*

El Concatenador realiza las siguientes actividades principales (Figura 6):

- Carga los archivos de audio físicos a memoria
- Crea un procesador (JMF Processor) para cada uno de los archivos de audio cargados
- Extrae el DataSource de cada uno de los archivos de audio a través de sus Processor (el DataSource es la fuente principal del audio)
- Realiza la concatenación de todos los DataSource, convirtiendo la cola de audio en un único DataSource que será entregado al Player, encargado de reproducirlo

4. PRUEBAS Y RESULTADOS

Comparación de señales

Se realizó una comparación entre la representación gráfica de dos señales: la primera es la voz del locutor pronunciando la frase "*Yo sé que él dijo",* durante la realización de las grabaciones para la creación del corpus de voz y la segunda es la señal obtenida como salida del sintetizador de voz desarrollado para la misma frase.

La Figura 7 presenta la señal original (a) y la señal de la voz sintética (b), respectivamente. En las dos gráficas se introdujeron líneas verticales que representan el inicio y fin de cada palabra de la frase. La Figura 7. (b), contiene además, tres elipses que identifican las secciones en las cuales se hace notoria la caída de la señal, lo que implica un problema en la naturalidad de la voz sintética.

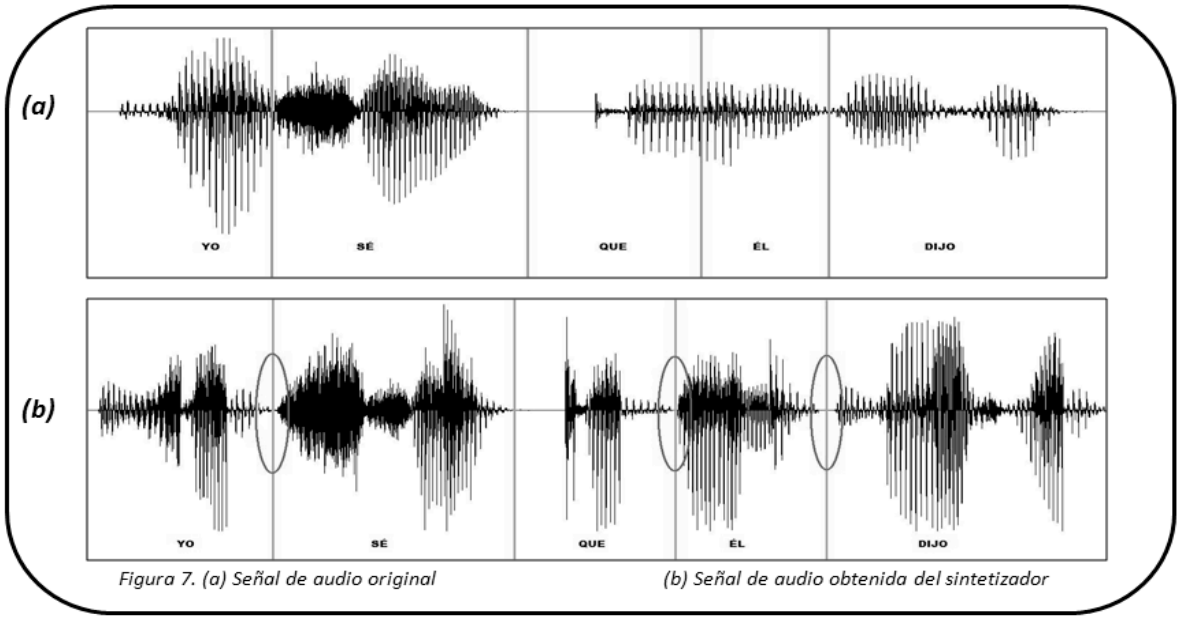

Figura 7. (a) Señal de audio original (b) Señal de audio obtenida del sintetizador

Estas caídas de señal corresponden a pausas (silencio) que existen entre las palabras; una duración prolongada de estos segmentos de silencio hace más notoria la transición entre una palabra y otra, de esta manera la voz se escucha menos fluida y más robótica.

La comparación de las formas de las dos gráficas permite establecer que existe similitud en la forma de las señales. Sin embargo, la señal de la voz sintética presenta picos mayores que la señal original debido al proceso de normalización aplicado a las grabaciones de los difonemas.

Prueba de funcionamiento del sintetizador

Se probó el funcionamiento del sintetizador desarrollado, específicamente del procesador "normalizador", utilizando cinco frases ejemplo para cada uno de los formatos definidos previamente (16 formatos), obteniendo un total de 80 frases de prueba. Se realizó una evaluación subjetiva al audio de salida del sintetizador, comprobando si las salidas del normalizador coincidían con el texto de entrada, es decir, si se obtenían las palabras adecuadas de acuerdo a la pronunciación del texto. El resultado obtenido fue favorable, puesto que en la medida que se utilizaron los formatos preestablecidos, el normalizador estuvo en capacidad de proporcionar las palabras correctas de acuerdo a la construcción ingresada.

Prueba de ejecución en diferentes plataformas

El sintetizador se ejecutó en tres equipos con similares características de hardware usando diferentes sistemas operativos (Windows Vista, Fedora 8 y Mac OS X). En cada uno de ellos se puso a prueba el sintetizador con conjuntos de diez frases de 5, 10, 15, 20, 25, 30, 35, 40, 45 y 50 palabras cada uno, y se realizó una medición del tiempo que tardó el sintetizador en proporcionar una salida de audio. La Tabla 4 muestra los resultados correspondientes al tiempo de respuesta promedio para cada uno de los conjuntos en cada plataforma.

Tabla 4. *Comparación tiempos de respuesta en diferentes sistemas operativos*

No. palabras	Tiempo promedio (ms)		
	Windows Vista	Mac OS X	Fedora 8
5	291	346,7	335,7
10	469,4	637,2	616
15	482,6	632	645,8
20	483,3	633,2	756,5
25	530,3	622,4	808,5
30	574,6	625,4	811
35	644,5	639,7	817,3
40	664,7	673	738,4
45	695,1	643,2	805,4
50	706,7	646,6	764,7

La Figura 8 muestra una representación gráfica de la comparación de los tiempos de respuesta para los tres equipos.

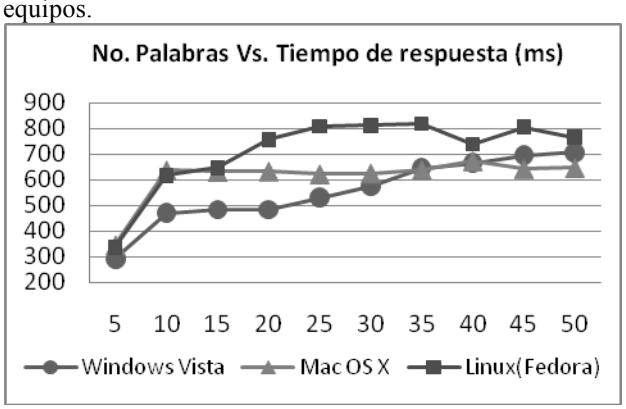

Figura 8. *Comparación de tiempos de respuesta*

Los resultados de estas pruebas muestran que aún para textos extensos (30-50 palabras), el sintetizador logra emitir una respuesta de forma rápida, sin sobrepasar, en el peor de los casos los 900 milisegundos. Además es importante resaltar el hecho de que la diferencia entre el tiempo de respuesta de un texto largo y uno corto no es significativa, aproximadamente existió una diferencia entre 100 y 200 milisegundos.

Por otra parte, la gráfica permite notar la similitud entre los tiempos de respuesta de los tres equipos. En el caso de las frases compuestas por 35 palabras o menos, el equipo con Windows Vista tardó menos tiempo que los otros dos (50-100 milisegundos comparado con Mac OSX y 100-200 contra Fedora). Pero en el caso de las frases de más de 35 palabras, fue el equipo con Mac OS X el que menos tiempo tardó en proporcionar una respuesta. La diferencia de tiempos de respuesta del orden de milésimas de segundo no es relevante ya que son imperceptibles al oído humano. La configuración de los equipos con que se realizaron las pruebas se muestran en la tabla 5, todos contaban con 2 gigabytes de memoria RAM.

Tabla 5. *Características de los equipos utilizados*

Equipo	S.O.	Procesador
1	Windows Vista	AMD Athlon 64 X2 Dual Core 5600+ 2.9 GHz
2	Fedora 8	
3	Mac OS X	Intel Core 2 Dúo 2.16 GHz

Pruebas con Usuarios

Se realizaron pruebas a la herramienta software producto de este artículo con usuarios, a quienes se les puso a escuchar diez frases desconocidas para ellos, con el fin de evaluar la inteligibilidad de la voz generada. Se evaluó entonces el número de palabras dichas correctamente por el usuario sobre el total de palabras evaluadas, arrojando como porcentaje de inteligibilidad un 98%.

5. CONCLUSIONES

Se desarrolló un corpus de voz para el español de Colombia teniendo en cuenta los sonidos propios de éste para lograr una voz sintética de acento neutro que puede utilizarse en varias regiones del país. Se obtuvo como producto de este trabajo un primer prototipo de sintetizador de dominio ilimitado basado en concatenación de difonemas para el español de Colombia, que puede utilizarse en diversas aplicaciones, y a su vez, sirve de punto de partida para el avance en el estudio de este tema en el país.

A través de las pruebas realizadas pudo notarse que las señales de audio obtenidas con el sintetizador son similares (en términos cualitativos mas no cuantitativos) a las señales de audio producidas por un hablante natural; sin embargo aún existen secciones en las que pueden notarse las fallas en la coarticulación entre palabras, lo que disminuye la calidad de la voz respecto a su naturalidad. Los tiempos de respuesta de la herramienta obtenidos resultan buenos, ya que en el peor de los casos analizados, éste no fue mayor a los 0.9 segundos, ni mayor a 0.7 segundos con el equipo de mejores características de hardware. Es importante que futuros trabajos tengan en cuenta medidas cuantitativas respecto a la calidad del sintetizador. Análisis del dominio de la frecuencia podrían dar luz en términos de la similitud de las señales y de esta manera conocer puntualmente los aspectos a mejorar.

6. REFERENCIAS

[1] Tutorial de fonética – Síntesis de habla. Universidad de los Andes. Mérida, Venezuela.

[2] Llisterri, J. La síntesis de habla – Generalidades.

[3] Diccionario panhispánico de dudas de la Real Academia Española de la Lengua (RAE).

[4] Diccionario de la Real Academia de la Lengua Española (RAE).

[5] *M. Rodríguez, **E. Mora. Síntesis de voz en el dialecto venezolano por medio de la concatenación de difonos. *Departamento de Electrónica y Comunicaciones, Facultad de Ingeniería, **Departamento de Lingüística, Facultad de Humanidades, Universidad de los Andes, Mérida, Venezuela.

[6] K. Palacio, J. Auquilla, E. Calle. Diseño e Implementación de un Sistema de Síntesis de Voz. Facultad de Ingenierías – Carrera de Ingeniería Electrónica, Universidad Politécnica Salesiana.

[7] Documentación del software para el etiquetado de difonemas, Diphone Studio.

[8] Java Media Framework, Programmers Guide, Overview.