DESIGN AND DEVELOPMENT OF A SPEECH SYNTHESIS SOFTWARE FOR COLOMBIAN SPANISH APPLIED TO COMMUNICATION THROUGH MOBILE DEVICES

DISEÑO Y DESARROLLO DE UN SOFTWARE DE SÍNTESIS DE VOZ PARA EL ESPAÑOL DE COLOMBIA APLICADO A LA COMUNICACIÓN A TRAVÉS DE DISPOSITIVOS MÓVILES

HOOVER F. RUEDA CH.
B.Sc, Master (c), Universidad Industrial de Santander, Bucaramanga, Colombia, hoover.rueda@correo.uis.edu.co

CLAUDIA V. CORREA P.
B.Sc, Master (c), Universidad Industrial de Santander, Bucaramanga, Colombia, claudia.correa@correo.uis.edu.co

HENRY ARGUELLO FUENTES
B.Sc, Ph.D. (c), Universidad Industrial de Santander, Bucaramanga, Colombia, henarfu@uis.edu.co

Received for review: July 16th, 2011, accepted: February 06th, 2012, final version: April 23th, 2012

ABSTRACT: In several scenarios of everyday life, there is a need to communicate orally with other people. However, various technological solutions such as mobile phones cannot be used in places such as meetings, classrooms, or conference rooms without disrupting the activities of people around the speaker. This research develops a tool that enables people to establish a conversation in a public place without disrupting the surrounding environment. To this end, a speech synthesizer is implemented on a personal computer connected to a cell phone, which allows one to establish a mobile call without using the human voice. The speech synthesizer uses the diphone concatenation technique and is developed specifically for the Spanish from Colombia. A mathematical description of the synthesizer shows the decomposition of the synthesizer into various mutually independent processes. Several user-acceptance and quality tests of the obtained signal were performed to evaluate the performance of the tool. The results show a high signal to noise ratio of generated signals and a high intelligibility of the tool.

KEYWORDS: speech synthesis, voice corpus, diphone concatenation, Spanish from Colombia, mobile devices, algorithms

RESUMEN: En diversos escenarios de la vida cotidiana existe la necesidad de comunicarse oralmente con otras personas. Sin embargo, diversas soluciones tecnológicas como la telefonía móvil no pueden ser utilizadas en lugares como reuniones, salones de clase, conferencias, entre otras, sin interrumpir las actividades de las personas alrededor del hablante. Este trabajo de investigación desarrolla una herramienta que permite entablar una conversación de voz en un recinto público sin interrumpir las actividades del medio circundante. Para ello se implementa un sintetizador de voz en una computadora personal comunicada de forma alámbrica con un teléfono móvil, lo cual permite establecer una llamada sin utilizar la voz humana. El sintetizador de voz utiliza la técnica de concatenación de difonemas y es desarrollado específicamente para el idioma español de Colombia. Una descripción matemática del sintetizador muestra su descomposición en diversos procesos independientes entre sí. Se realizaron diversas pruebas de aceptación de usuarios y de calidad de la señal obtenida para evaluar el desempeño de la herramienta. Los resultados muestran una alta relación señal a ruido de las señales generadas y una alta inteligibilidad de la herramienta.

PALABRAS CLAVE: síntesis de voz, corpus de voz, concatenación de difonemas, español de Colombia, telefonía móvil, algoritmos

1. INTRODUCTION

Speech synthesis is the generation of artificial voice from a written text [1,2]. Electronics and software generate acoustic signals to simulate the human voice [3–6]. Each language has its proper phonetic rules to determine the correct pronunciation of the words. Particularly, in Spanish, pronunciation is similar to what is written, but there are some special structures of the language that require special processing [7,8].

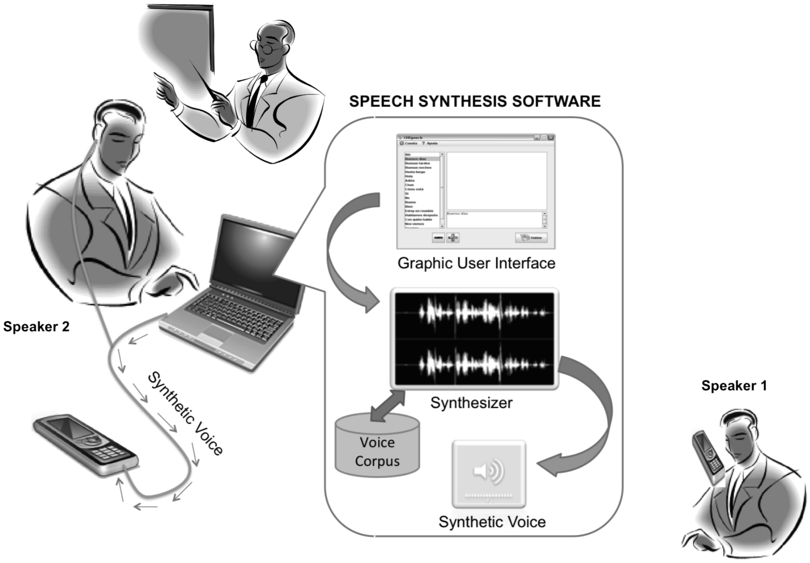


Figure 1. General design and components of the proposed speech synthesis software. Speaker 1 and 2 are in different geographical places. Since Speaker 2 cannot use his voice (he is in a classroom), he can communicate with Speaker 1 by using the speech synthesis software.

Those structures include e-mail accounts, dates, abbreviations, and phone numbers. This situation is one of the biggest challenges in text-to-speech conversion. This is why different stages need to be taken into account in a speech synthesis system. First, a pre-processing stage analyzes the structures present in the text. Then, the text is divided into many entries for the synthesizer. This process is done by algorithms that apply the rules of the language and identify the separation of words (blank spaces, punctuation marks, written accents, etc.).

There are different parameters for measuring the quality of speech synthesis applications: the naturalness and intelligibility of the voice, the complexity of the process, and the domain for which it was developed [9]. Different techniques for speech synthesis have been developed, each one offering benefits in terms of naturalness or intelligibility compared to others. Some of them, such as synthesis by concatenation, use pre-recorded tokens of voice stored in a database called voice corpus [10–16]; other techniques are based on acoustic mathematical models that generate the artificial voice by the variation of parameters like noise levels, frequency, and the movements of the vocal apparatus [17–21].

This paper presents the design of a software tool that allows for one to make mobile-phone calls by using speech synthesis, specifically diphone concatenation to generate an artificial voice on a computer from an input text and to reproduce it on a mobile device. This software is a solution for people having trouble answering their mobile devices due to situations that limit the direct use of speech. Figure 1 presents a general diagram of the components in the proposed software.

Each section of this paper presents one component of the speech synthesis software tool. The first section is related to the speech synthesizer and the mathematical approach of each of its processors. The second section presents the voice corpus (voice database). The transmission device used to transmit the voice to the mobile device is presented in the third section. Finally, the tests are performed [26], the results obtained, and conclusions are presented.

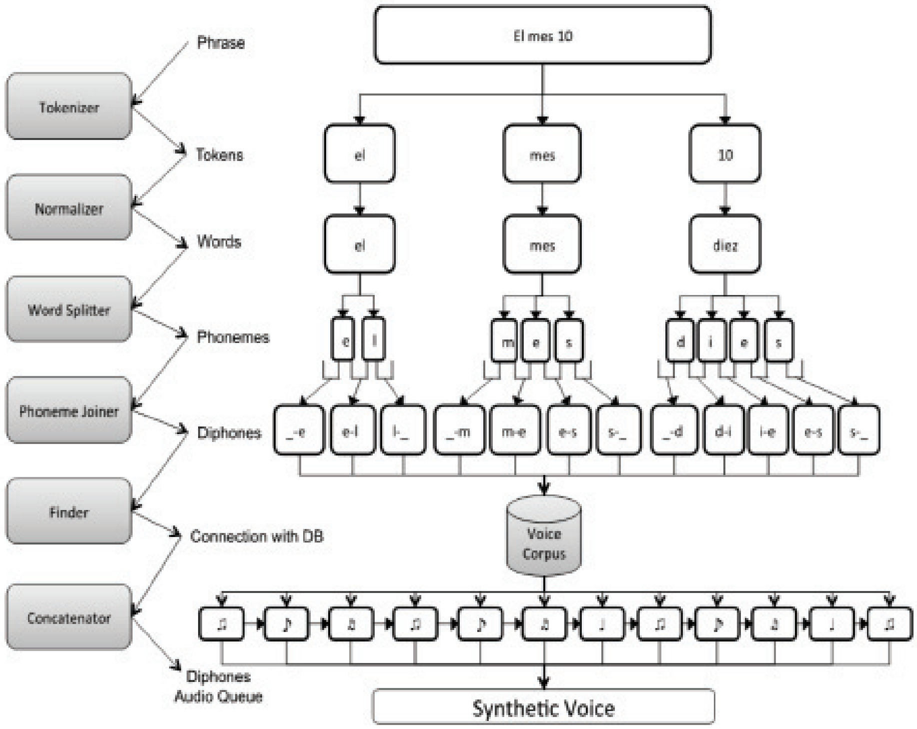


Figure 2. Architecture of the proposed speech synthesizer. A phrase is the input of the system. The tokenizer divides it into tokens using the blank spaces between words; each token is normalized to be represented in words; then, each word is divided into phonemes, which are then grouped by the phoneme joiner to form the diphones; finally, diphone mapping is performed in the voice corpus to extract the audio files, concatenate them, and obtain the synthetic voice.

2. SPEECH SYNTHESIZER

The principal component of the developed software is an unlimited domain speech synthesizer that uses the diphone concatenation technique. The synthesizer produces a synthetic voice by processing an input text. In other words, it finds the sound representation of a given text. The synthesizer consists of six processors, each one developing a specific task in the synthesis process. Figure 2 shows the architecture of the synthesizer. Note that the processors are executed sequentially. Thus, the output of each processor is the input of the next one.

The mathematical approach, the description of the speech synthesizer, and each processor are presented below. Let Σ be the alphabet that contains the Colombian Spanish letters and numbers, the Greek symbols, punctuation marks, mathematical operators, and other special symbols. Define Σ^* as the set of all the words of finite length formed with elements of Σ. Let S be the set of punctuation marks named *separators,* which are given by $S = \{`\phi' ,\ `,' ,\ `.' ,\ `;' ,\ `:' ,\ `\backslash t' ,\ `\backslash n'\}$, where $S \subset \Sigma$.

2.1. Tokenizer

First, the input phrase is processed by the tokenizer which divides the phrase into tokens. The *blank space* $(\phi | \phi \in S)$ is the element that indicates the end of one token and the beginning of the next one. Define the *phrase* F as a sequence of symbols $\langle a_0, a_1, a_2, a_3, \ldots, a_n \rangle$ where $\{a_i \in \Sigma, \forall i \geq 0\}$ and $F \in \Sigma^*$. A *token* t, such that $\{t | t \subset F\}$, is defined as a sequence of symbols $t = \langle a_k, a_{k+1}, a_{k+2}, \ldots, a_p \rangle$, where a_{k-1} and $a_{p+1} = \phi$. Then F is the set of t detached by ϕ. The function T, called *Tokenizer*, is defined by the equations

$$Y_t = T(F), \qquad (1)$$

$$T(F) = [t_1 \,.\,.\,.\, t_n], \qquad (2)$$

$$n = E(F) + 1, \qquad (3)$$

where F is the input phrase, Y_t is the set of tokens in F, $Y_t \in \Sigma^*$, and $E(F)$ is a function that finds the blank spaces in F.

2.2. Normalizer

The developed synthesizer is classified as unlimited domain [9]. For this reason, it is necessary that it identify different types of special constructions of the language such as numbers, dates, time, phone numbers, e-mail, and web pages. These constructions have a different pronunciation compared to their written representation. For that reason, the normalizer identifies the type of construction that corresponds to the input text and defines the way it is going to be pronounced.

Let $\Sigma_a \subset \Sigma$ be the alphabet with the letters of Colombian Spanish, such that $\Sigma_a = \{a, b, c, \ldots, y, z, á, é, í, ó, ú\}$. Let $\Sigma_a{}^*$ be the set of words of finite length formed with elements of Σ_a. Define $P \subset \Sigma_a{}^*$ as the set of words in Colombian Spanish. The *Normalizer* is represented by a function N given by the equations

$$Y_p = N(Y_t), \qquad (4)$$

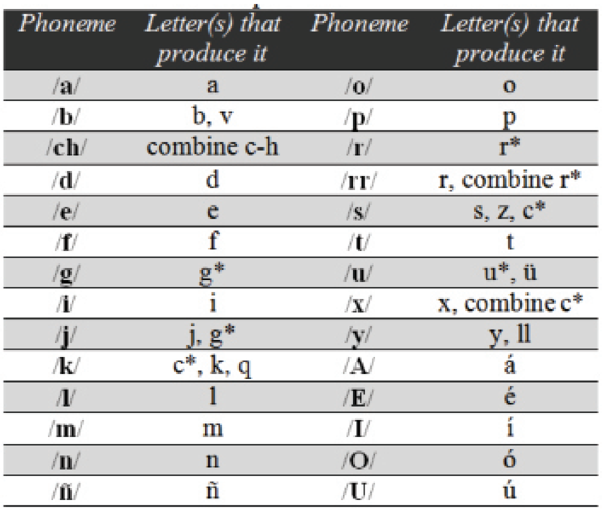
$$N(Y_t) = [w_1 \ . \ . \ . \ w_p] \ , \quad |p| \geq |t|, \qquad (5)$$

where Y_t is the set of tokens in F, $Y_p \in P$ is the set of words that correspond to the tokens in Y_t, and $|.|$ represents the number of elements in a vector. The normalizer uses a set of 16 pre-defined formats to perform the classification of the special constructions of the language. The formats are represented using regular expressions.

2.3. Word Splitter

The normalized words in phrase F are used as the input of the word splitter. This processor divides each word into its corresponding phonemes. Let $\Sigma_f \subset \Sigma_a$ be the set of written representations of the phonemes in Colombian Spanish, which are presented in Table 1.

Table 1. Phonemes of Colombian Spanish and letters that produce them. 28 phonemes are presented, including vowels with and without an accent, and assuming that the pairs "b, v", and "y, ll" have the same pronunciation



Phoneme	*Letter(s) that produce it*	*Phoneme*	*Letter(s) that produce it*
/a/	a	/o/	o
/b/	b, v	/p/	p
/ch/	combine c-h	/r/	r*
/d/	d	/rr/	r, combine r*
/e/	e	/s/	s, z, c*
/f/	f	/t/	t
/g/	g*	/u/	u*, ü
/i/	i	/x/	x, combine c*
/j/	j, g*	/y/	y, ll
/k/	c*, k, q	/A/	á
/l/	l	/E/	é
/m/	m	/I/	í
/n/	n	/O/	ó
/ñ/	ñ	/U/	ú

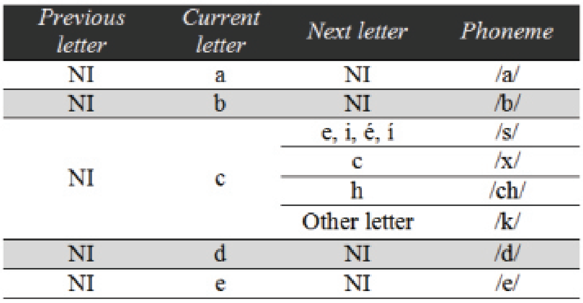
*The phoneme is presented differently depending on the neighboring letters

Let DP be a function called *Word Splitter* defined by the equation,

$$Y_f = DP(Y_p), \quad |f| > |p|, \qquad (6)$$

where $Y_f \in \Sigma_f$ is the set of phonemes that correspond to each word. The function DP uses a word processing algorithm based on the location of each letter in the word and the neighboring letters. It assigns the phonemes that correspond to each letter based on those criteria. Table 2 presents a portion of the conditions for assigning a phoneme to a letter.

Table 2. Portion of the table of conditions for assigning phonemes to letters. Assigning a phoneme to a letter depends on its location in the word and its neighboring letters.



Previous letter	*Current letter*	*Next letter*	*Phoneme*
NI	a	NI	/a/
NI	b	NI	/b/
NI	c	e, i, é, í	/s/
		c	/x/
		h	/ch/
		Other letter	/k/
NI	d	NI	/d/
NI	e	NI	/e/

NI = Not important

The output of the word splitter is represented by the equation

$$Y_f = \lfloor H \;\; DP(w_1) \;\; H\, DP(w_2)\, H \ldots DP(w_p)\; H \rfloor,$$

$$|f| > |p|, \qquad (7)$$

where $H \in \Sigma_f$ represents the *'pause'* phoneme (blank space between words) and $DP(w_x)$ are the phonemes of the respective word w_x, $1 \leq x \leq p$.

2.4. Phoneme Joiner

This processor takes the list of phonemes in a phrase and obtains its representation in diphones. Let $\Sigma_f{}^*$ be the set of all possible combinations of elements in Σ_f (diphones, triphones, etc.). Define $\Sigma_d \subset \Sigma_f{}^*$ as the set of diphones in Colombian Spanish (a portion is presented in Table 3). Also, define the function AF as the *Phoneme Joiner* which is given by

$$Y_d = AF(Y_f), \quad |d| = |f| + 1, \qquad (8)$$

where $\mathrm{Y_d}$ represents the set of diphones that correspond to each word of F. The function AF uses an algorithm for concatenating two consecutive phonemes, which can be written as

$$Y_d = \left[\left(Y_f^1\, Y_f^2\right) \;\; \left(Y_f^2 \;\; Y_f^3\right) \ldots \left(Y_f^{i+(j-1)}\, Y_f^{i+j}\right) \right]$$

$$i, j \geq 0\,, \qquad j = |f|, \qquad (9)$$

where Y_f^i corresponds to the phoneme in location i of the list of phonemes Y_f.

2.5. Finder

Based on the list of diphones that represent phrase $\mathbf{f}$, the finder connects the synthesizer with the voice corpus (database) and links each diphone with its sound. Define C_d as the voice corpus of diphones containing the set of sound representations of the elements in Σ_d . Also, define function B as

$$Y_b = B(Y_d), \qquad (10)$$

where Y_b is the set of sound representations that correspond to the search in C_d of the set of diphones Y_d. In this way, the representation of the input phrase in terms of the audio files matched to the diphones is obtained.

2.6. Concatenator

The concatenator is the last processor of the synthesizer. It generates the output audio signal. This task is performed using the list of audio files obtained in the *Finder*. Define the function called *Concatenator* which is represented by

$$Y = \left[Y_b^1 \quad Y_b^2 \quad Y_b^3 \;\; \ldots \;\; Y_b^d \; \right],$$

$$|d| = |f| + 1, \qquad (11)$$

where Y represents the audio signal obtained at the end of the synthesis process and Y_b^i represents the diphone (sound unit) in location i of the set of sound representations Y_b. In this way, it is possible to reproduce a signal that contains all the input text represented in sound units.

3. VOICE CORPUS

Unit concatenation speech synthesis requires a database from which the audio units are extracted to form the synthetic voice. The database is called a "corpus" and includes labeled phonetic units [22–25]. The data stored in the corpus corresponds to audio files recorded previously by a natural speaker and depend on the selected units for the synthesizer. In this case, the corpus has 590 audio files with the diphones of Colombian Spanish.

A matrix was developed for the identification of the diphones. The number of rows and columns correspond to the identified phonemes, including the phoneme *"pau"* (which represents the blank space). The matrix has 29 rows and 29 columns (a total of 841 diphones). Since not all the combinations of phonemes correspond to a real diphone in the Spanish language (for instance: ñ-ñ), the final number of diphones identified in this work is 590. Table 3 presents a portion of the matrix developed for the identification of the diphones (a dash indicates that the diphone does not exist).

Table 3. Portion of the diphones matrix. The rows and columns correspond to the identified phonemes, including the phoneme “pau”, which represents the blank space and is denoted by “_”. When a diphone does not exist, a dash is presented. The total number of diphones is 590.

	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	-

All possible combinations of phonemes in Colombian Spanish were tested to determine if a diphone is valid or not. Then the diphones were recorded to obtain the voice corpus. The block-diagram of the process in the development of the voice corpus, after identifying the diphones, is presented in Fig. 3.

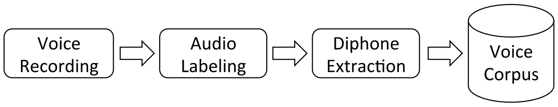


Figure 3. Stages in the development of the voice corpus. First, phrases containing each of the diphones were recorded. Then, the beginning and the end of the diphones in the audio files were labeled. Finally, the diphones were extracted in individual files and stored in the voice corpus.

The previous stages were performed sequentially. First, phrases containing the diphones at least once were recorded. Then, the phrases were labeled to extract and store each diphone in the voice corpus.

4. TRANSMISSION DEVICE

As was shown in Fig. 1, the developed application runs on a computer (laptop or desktop). The audio signal is generated there by the synthesizer and needs to be transmitted to the mobile phone during a call. For that reason, a transmission device is required to send the synthetic voice from the computer to the mobile phone. In other words, the synthesizer will speak for the user during the call. That means that the transmission device has to be connected to the microphone of the mobile phone and, at the same time, must allow the person to use the earpiece.

Since there is no such commercial device, a hardware piece with the above features had to be designed. The principal element of the device designed is the headset that comes with almost every mobile phone model. This cable can access both the microphone and the earpiece of the phone. However, this characteristic does not allow transmission between the mobile phone and the computer. For that reason, this is the principal element to modify.

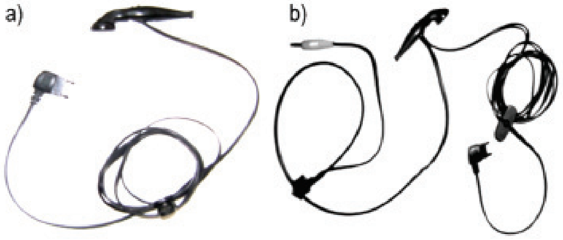


Figure 4. a) Original headset b) Modified headset

Fig. 4.a) presents a common headset cable. It has an earphone connected to the earpiece of the mobile phone, a microphone connected to the microphone of the phone, and two cables through which the signals are transmitted. The proposed transmission device consists of connecting the cable from the microphone of the headset to the audio output of the computer. Hence the synthetic voice is sent to the microphone of the mobile phone. The cable obtained is shown in Fig. 4b).

5. TESTS AND RESULTS

In our previous work [27], the performance of the software tool was evaluated mainly in response time and its execution in other operative systems. In this paper, the quality of the voice is evaluated by using performance measures such as peak signal-to-noise ratio (PSNR), percentage of correct pronounced words, and intelligibility.

5.1. Evaluation of the output of the synthesizer

A comparison was made between the output of the synthesizer and the same phrases recorded by a person. The objective was to obtain a quantitative measure of the quality of the synthetic voice. The chosen measure is the PSNR, because it provides a sense of the behavior of the synthetic voice compared to the natural voice. The calculation was performed using the equation

$$PSNR = 20 * log_{10}\left(\frac{n * max(O)}{\sqrt{\sum_{i=1}^{n}\left(O(i) - S(i)\right)^{2}}}\right), (12)$$

where *O* represents the original signal, *S* the output signal of the synthesizer (synthetic), *max* is a function that calculates the maximum value of the signal, and *n* is the number of values.

A total of 70 phrases were used for the test, including all of the special language constructions mentioned in Section 1.2. Figure 5 presents the average PSNR for each phrase. In general, the average PSNR for the phrases was 56.68 dB. The results show the high correlation between the synthetic signal and phrases pronounced by the human voice.

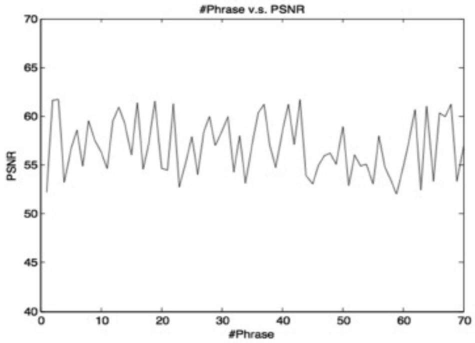


Figure 5. Results of the comparison between the original voice with the output of the synthesizer. Values of PSNR for the 70 phrases are between 50 and 62 dB, which shows high correlation between the signals.

Also, a frequency analysis was performed. A comparison between the spectrum of the original and the synthetic voices is shown in Fig. 6. It can be seen that the spectrum of both signals is similar. Since the number of samples and the amplitude of the synthetic signal are higher than those of the original signal, there are some differences between their spectra. Processing the output signal of the synthesizer does not include filtering or the modulator to normalize the amplitudes.

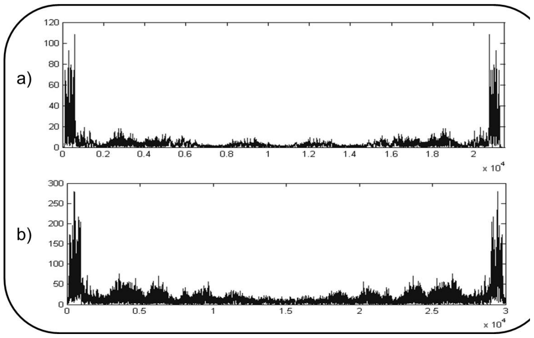


Figure 6. Frequency comparison between a) Natural voice, b) Synthetic voice

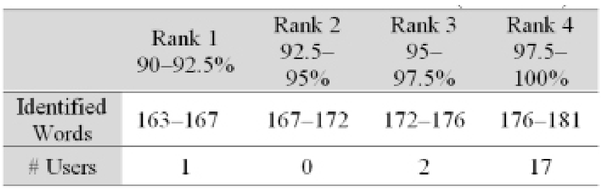
5.2. User Testing

A test was designed for users to evaluate the performance of the software. The features included in the evaluation were the intelligibility of the voice, the correctness of the outputs according to the inputs for the synthesizer, and the transmission device.

5.2.1. Intelligibility

Users listened to ten phrases with a total of 181 words that were pronounced by the synthesizer. The phrases used in the test included some of the pre-defined formats (dates, abbreviations, etc.). The results of the test are divided into four categories according to the quantity of words correctly identified by the users. This is shown in Table 4.

Table 4. Results of the intelligibility test with users. From 181 words pronounced, only one person identified between 163 and 167 words (>90%), two people identified between 172 and 176 words (>95.1%) and 17 identified between 176 and 181 (>97.51%).



	Rank 1 90–92.5%	Rank 2 92.5– 95%	Rank 3 95– 97.5%	Rank 4 97.5– 100%
Identified Words	163–167	167–172	172–176	176–181
# Users	1	0	2	17

Only one person is classified in the first category. The minimum percentage of correctly identified words was 90.6% of the test words. Most of the users (95%) identified more than 95% of the testing words. Table 4 presents the results for all the categories. Figure 7 shows the statistics of the results.

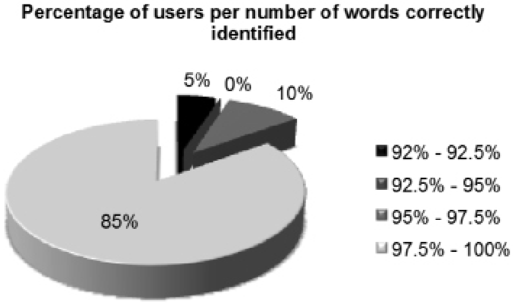


Figure 7. Results of intelligibility test with users (percentages). 95% of the users identified more than 95% of the words pronounced by the synthesizer.

Based on the previous results, a percentage of intelligibility for the synthesizer was calculated. Taking into account that for a total of 20 users, 3620 words were pronounced by the speech synthesizer (181 words per person), the percentage of intelligibility is calculated by the equation

$$I = \left(\sum_{i=1}^{20} \frac{P_a(i)}{P_t} \right) * 100\%, \qquad (13)$$

where I is the intelligibility percentage (between 1 to 100%), P_a is the number of identified words and P_t is the total number of pronounced words. By replacing the number of total words and the correctly identified words in Eq. (13), the resulting intelligibility percentage is 98% which denotes that the users can identify a high percentage of the words pronounced by the speech synthesizer.

5.2.2. Synthesizer Output

The synthesizer was tested by the users with a set of phrases randomly proposed by them. A total of 4153 words were used for the test. The users concluded that 36 of them were incorrectly pronounced by the synthesizer. The percentage of correctly pronounced words is calculated as

$$P_c = 100\% - P_i \,, \qquad (14)$$

where P_c represents the percentage of words pronounced correctly and P_i the percentage of words pronounced incorrectly. The results show that 99% of the words were pronounced correctly.

These tests are useful for identifying failures in the processors of the synthesizer, for future improvement. The high percentage of correct words shows that the synthesizer can pronounce most of the words in Colombian Spanish when the pre-defined formats are used.

5.2.3. Transmission Device

Finally, users tested the software and the transmission device to answer a call in a mobile phone. All of the users (100%) said that they heard and understood the voice through the designed transmission device. That means that there was no perceptible loss in the signal during transmission and that the voice was intelligible.

6. CONCLUSIONS

This work integrates two technologies: speech synthesis and mobile phones. The implementation of this type of software allows users to answer phone calls on their mobile devices despite some limitations in the use of their voice.

Studying speech synthesis techniques and the phonetics of the language allowed the development of a first prototype of unlimited-domain speech synthesizer based on diphone concatenation technique for Colombian Spanish including a voice corpus. Nonetheless, the tests performed show that future work should focus on the improvement of the quality of the synthetic voice, specifically in terms of naturality. More comprehensive studies on prosody and linguistics could lead to a more natural voice by adding new diphones.

The tests with users for the transmission device showed positive results. However, future work should investigate different wireless technologies, such as Bluetooth, for the transmission of the synthetic voice to the mobile phone.

ACKNOWLEDGMENTS

This work was sponsored by *Vicerrectoría de Investigación y Extensión* of the *Universidad Industrial de Santander,* under the research project #5537.

REFERENCES

[1] Flores, L.; Vargas, A.; Olivier, A.; Kirschning, I; and Cervantes, O. “Síntesis en Español Mexicano con el Método de Selección de Unidades de Longitud Variable”*, ENC*, 601-610, Sept. 2001.

[2] O’Shaughnessy, D. “Modern Methods of Speech Synthesis”, *IEEE Circuits and Systems Magazine*, Vol. 7, No. 3, pp. 6–23, 3rd Quarter 2007.

[3] Laboratorio de fonética ULA. “Tutorial de fonética: Síntesis de habla”, Universidad de los Andes, Mérida, Venezuela, 2005.

[4] Black, A.; Lenzo, K. “Flite: a small fast run-time synthesis engine”, *Proceeding of the 4th ISCA Workshop on Speech Synthesis*, 2001.

[5] Black, A.; Taylor, P. “The Festival Speech Synthesis System: system documentation”, *Technical Report HCRC/TR-83*, Human Communications Research Centre, University of Edinburgh, Scotland, UK, January 1997.

[6] Pitrelli, J.; Bakis R.; Eide E.; Fernandez R.; Hamza W.; Picheny M. “The IBM expressive text-to-speech synthesis system for American English”, *TSAP*, Vol. 14, No. 4, pp. 1301 – 1312, Jul. 2006.

[7] “Otros estudios sobre el español de Colombia”, *Publicaciones del Instituto Caro y Cuervo*, Santafé de Bogotá, 2000, pp. 31.

[8] Mora, S.; Lozano, M.; Ramírez, R.; Espejo, M. B.; Duarte, G. E. “Caracterización léxica de los dialectos del español de Colombia según el “ALEC””, *Publicaciones del Instituto Caro y Cuervo*, Bogotá, 2004, 325 págs.

[9] LListerri, J. “La síntesis de habla”, *I Jornadas de Tecnología del Habla*, Departamento de Lengua Inglesa, Universidad de Sevilla – Departamento de Electrónica y Tecnología de Computadores, Universidad de Granada, Sevilla, Noviembre 7, 2000.

[10] Rodríguez, M.; Mora, E. “Síntesis de voz en el dialecto venezolano por medio de la concatenación de difonos”, *Revista Ciencia e Ingeniería*, Vol. 27 No. 1, 2006, pp. 17-24.

[11] Rodríguez, M.; Mora E. “Conversor texto a voz en el dialecto venezolano por medio de la concatenación de difonos”, *Revista Ciencia e Ingeniería,* Vol. 27, No. 2, 2006, pp. 79-87.

[12] Iriondo, I.; Martí, J.; Oliver, J.; Guaus, R.; Moure H. “Hacia una síntesis concatenativa de alta calidad para aplicaciones de conversión texto-habla”, *Procesamiento del Lenguaje Natural,* Sociedad Española para el procesamiento del Lenguaje Natural, No. 25, Sept. 1999, pp. 109-113.

[13] Rodríguez Banga, E.; Campillo Díaz, F. “Sistema de conversión texto-voz en lengua gallega basado en la selección combinada de unidades acústicas y prosódicas”, *Procesamiento del Lenguaje Natural,* Sociedad Española para el procesamiento del Lenguaje Natural, No. 29, 2002, pp. 153-158.

[14] Hunt, A. and Black, A. “Unit selection in a concatenative speech synthesis system using a large speech database”, *Proceedings of ICASSP 96*, Vol. 1, pp. 373-376, Atlanta, Georgia, 1996.

[15] Lewis, E. AND Tatham, M. “Word and Syllable Concatenation in Text-To-Speech Synthesis”, *Sixth European Conference on Speech Communications and Technology*, pp. 615-618, ESCA, September 1999.

[16] Zapata, C. And Carmona N. “El experimento Mago de Oz y sus aplicaciones: una mirada retrospectiva”, *Revista Dyna*, No. 151, pp. 125-135,2007.

[17] Guzmán Arreola, M. A. “Sintetizador de voz para la enseñanza de la lectura a niños mexicanos”, *Tesis Licenciatura Ingeniería en Sistemas Computacionales*, Departamento de Ingeniería en Sistemas Computacionales, Escuela de Ingeniería, Universidad de las Américas Puebla, 2004.

[18] Barra-Chicote, R.; Yamagishi, J.; Montero, J. M.; King S.; Lufti, S.; Macias-Guarasa, J. “Generación de una voz sintética en castellano basada en HSMM para la evaluación Albayzín 2008: Conversión Texto a voz”, *V Jornadas en Tecnología del Habla,* Noviembre 2008, Bilbao, España, pp. 115-118.

[19] Tokuda, K.; Masuko, T.; Miyazaki, N. and Kobayashi, T. “Hidden Markov Models Based on Multi-Space Probability Distribution for Pitch Pattern Modeling”, *IEEE International Conference on Acoustics, Speech and Signal Processing*, Vol. 1, Phoenix, Arizona, USA, March pp. 15-19, 1999.

[20] Yoshimura, T. “Simultaneous modeling of phonetic and prosodic parameters, and characteristic conversion for HMM-based Text-To-Speech systems”, *PhD dissertation*, Nagoya Institute of Technology, 2002.

[21] Tokuda, K.; Zen, H.; Black A. W. “An HMM-based speech synthesis system applied to English”, *Proceedings IEEE 2002 Workshop on Speech Synthesis*, Santa Monica, USA, Sept. 2002.

[22] Palacio Baus, K. S.; Auquilla Peralta J. V. “Diseño e Implementación de un Sistema de Síntesis de Voz”, *Tesis Ingeniería Electrónica, Facultad de Ingenierías*, Carrera de Ingeniería Electrónica, Universidad Politécnica Salesiana, Cuenca, Ecuador, 2007.

[23] Campbell, N. “Developments in Corpus-Based Speech Synthesis: Approaching Natural Conversational Speech”, *IEICE Trans. Inf. & Syst.*, Vol. E88–D, No. 3, March 2005.

[24] LListerri, J.; Machuca, M. J.; De la mota, C.; Riera, M.; Ríos A. “Corpus orales para el desarrollo de las tecnologías del habla en español”, *Oralia, Análisis del discurso oral,* Vol. 8, pp. 289-325, 2005.

[25] Mora, E. “Discapacidad y comunicación: Una experiencia de fonética aplicada”, *EFE, ISSN 1575-5533,* XVII, 2008, pp. 317-329.

[26] Aguilar, L.; Fernández J. M.; Garrido J. M.; LListerri J.; Macarrón A.; Monzón L.; Rodríguez, M. A. “Diseño de pruebas para la evaluación de habla sintetizada en español y su aplicación a un sistema de conversión de texto a habla”, en *Actas del X Congreso de la Sociedad Española para el Procesamiento del Lenguaje Natural*, Córdova, 20-22 de Julio de 1994.

[27] Correa, P.; Rueda, H.; Arguello, H. “Síntesis de voz por concatenación de difonemas para el español de Colombia”, *Revista Iberoamericana en Sistemas, Cibernética e Informática,* Vol. 7, No. 1, pp. 19-24, 2010.