A pattern recognition based esophageal speech enhancement system

A.Mantilla-Caeiros[1], M. Nakano-Miyatake[2], H. Perez-Meana*[2],

[1]Instituto Tecnológico y de Estudios Superiores de Monterrey, Campus Ciudad de México
Calle del Puente 222, Ejidos de Huipulco, Tlalpan 14380 Mexico City
[2]ESIME Culhuacán, Instituto Politécnico Nacional
Av. Santa Ana 1000, Col, San Francisco Culhuacán, 04430 Mexico City
*Email hmperezm@ipn.mx

ABSTRACT
A system for improving the intelligibility and quality of alaryngeal speech based on the replacement of voiced segments of alaryngeal speech with the equivalent segments of normal speech is proposed. To this end, the system proposed identifies the voiced segments of the alaryngeal speech signal by using isolate speech recognition methods, and replaces them by their equivalent voiced segments of normal speech, keeping the silence and unvoiced segments without change. Evaluation results using objective and subjective evaluation methods show that the proposed system proposed provides a fairly good improvement of the quality and intelligibility of alaryngeal speech signals.

Keywords: Speech enhancement, esophageal speech, electronic larynx, multilayer perceptron, voiced and unvoiced segments detection, speech synthesis.

RESUMEN
Este artículo propone un sistema para mejorar la calidad e inteligibilidad de la voz de personas laringetomizadas, el cual se basa en el reemplazo de segmentos vocalizados de voz laringetomizada por segmentos equivalentes de voz normal. Con esta finalidad el sistema identifica los segmentos vocalizados de voz laringetomizada usando técnicas de reconocimiento de comandos aislados de voz, y las reemplaza por los segmentos equivalentes de voz normal, conservando sin cambio los segmentos y los no-vocalizados. Resultados obtenidos usando métodos de evaluación tanto subjetivos como objetivos muestran que el sistema propuesto proporciona una mejoría importante tanto en la calidad como en la inteligibilidad de señales de voz laringetomizada.

1. Introduction

Persons that suffer diseases such as throat cancer require their larynx and vocal cords to be extracted by a surgical operation, requiring then rehabilitation in order to be able to reintegrate to their individual, social, familiar and work activities. To accomplish this, different methods have been used, such as the esophageal speech, the use of tracheoesophageal prosthetics and the Artificial Larynx Transducer (ALT), also known as "electronic larynx". Among them, the Artificial Larynx Transducer (ALT), which has been used by alaryngeal patients for over the last 40 years without essential changes, is the most widely used rehabilitation method [1], [2].

ALT, which has the form of a handheld device, introduces an excitation in the vocal track by applying a vibration against the external walls of the neck. This excitation is then modulated by the movement of the oral cavity to produce the speech sound [1]. This transducer is attached to the speaker's neck, and in some cases to the speaker's cheeks. ALT is widely recommended by voice rehabilitation physicians because it is very easy to use even for new patients, although the voice produced by these transducers is unnatural and with low quality, besides, it is distorted by the ALT produced background noise. This results in a considerably degradation of the quality and intelligibility of speech, problem for which an optimal solution has not yet been found [2]-[4].

The esophageal speech, on the other hand, is produced through the compression of the contained air in the vocal track, from the stomach to the mouth through the esophagus. This air is swallowed and, as passing through the esophageal-larynx segment, produces a vibration of the esophageal upper muscle, bringing about the speech. The sound generated is similar to a burp, the tone is commonly very low, and the timbre is generally harsh. As in the ALT produced speech, the voiced segments of esophageal speech are the most affected part of the speech within a word or phrase [2], [5], giving as a result an unnatural speech. Thus, many efforts have been carried out to improve its quality and intelligibility.

Several approaches have been proposed to improve the quality and intelligibility of alaryngeal speech, esophageal as well as ALT produced speech. To reduce the ALT produced background noise, several adaptive filter based speech enhancement algorithms have been proposed [3], [4]. These algorithms reduce the ALT produced background noise, improving the intelligibility, although not the quality, of the ALT produce speech. An analysis-synthesis-based method to improve the alaryngeal speech quality was proposed in [6]. In this method, which uses a vocoder-like approach, the speech is firstly digitalized using a sampling frequency equal to 10 kHz. Next the sampled signal is divided in 2 frequency bands: the lowpass band from 0 to 2.5 kHz and the highpass band from 2.5 kHz to 5 kHz. The analysis highpass filter output is fed into a synthesis highpass filter whose output used to synthesize the highpass band of restored speech. On the other hand, the analysis lowpass filter output is further processed to synthesize the lowpass band of restored speech. To this end, firstly the signal power is estimated and, if it is smaller than a given threshold, the speech segment is considered as unvoiced and fed to a reconstruction lowpass filter; otherwise, the linear predictive coefficients, LPC, of speech segment are estimated and a restored voice segment is synthesized using normal voice pitch information. Finally, the synthesized voiced signal is fed into a reconstruction lowpass filter whose output signal is combined with the highpass filter output to obtain the restored speech signal. This system provides a restored speech with improved quality, similar to that provided by a vocoder speech coder. Other esophageal speech enhancement approach proposed in [7] uses a frequency band extension form 4 kHz to 8 kHz. Here, the highpass band, from 4 to 8 kHz, is estimated from the lowpass band from 0 Hz to 4 kHz. This fact allows having frequency components from 0 to 8 kHz improving in such a way the speech tone. In this system, the generation of the highpass band is based on analysis-synthesis methods using the LPC coefficients of esophageal speech lowpass band. This approach provides an improved signal by adding high frequency components, although the low frequency band remains unchanged. Finally, a very promising approach is based on speech conversion techniques [8-10], which carry out a spectral conversion using vector quantization methods. These approaches perform fairly well although still present some problems because the spectral conversion reduce a continuous spectral space into a discrete code book, which may produce a distortion that still must be reduced.

This paper proposes an alaryngeal speech enhancement system based on pattern recognition methods. In the system proposed, firstly the alayryngeal voice signal is filtered to reduce the background noise. Next, the voiced/unvoiced detection is performed and the voiced segments identified using artificial neural networks (ANN). Next, the voiced segments are replaced by their equivalent normal speech voiced segments, while the unvoiced segments are kept without change. Finally, the estimated voiced, unvoiced and silence segments are used together to produce the restored speech. Evaluation results show that the system proposed provides quite a good

improvement in the quality and intelligibility of esophageal as well as ALT produced alaryngeal speech.

2. System proposed

The alaryngeal speech restoration system proposed, shown in Fig. 1, is based on the replacement of voiced segments of alaryngeal speech by their equivalent normal speech voiced segments, while keeping the unvoiced and silence segments without change. The main reason for it is the fact that the voiced segments have more impact on the speech quality and intelligibility than the unvoiced ones.

To achieve this goal, firstly the alaryngeal speech signal is filtered with a low pass filter with cutoff frequency of 900 Hz to reduce the background noise. Then the silence segments are estimated using the time average speech signal power as proposed in [11]. Here, if a silence segment is detected, the switch is enabled and the segment is concatenated with the previous one to produce the output signal. If voice activity is detected, the speech segment is analyzed using the pitch analysis, the zero crossings number and the formant analysis.

Next, if the segment is considered as unvoiced, the switch is enabled and the speech segment concatenated at the output with the previous segments; otherwise, the switch is disabled and the speech segment is identified using pattern recognition techniques. Then the alaryngeal voiced segment is replaced by the equivalent normal speech voiced segment, contained in the codebook, which is finally concatenated with the previous segments to synthesize the restored speech signal.

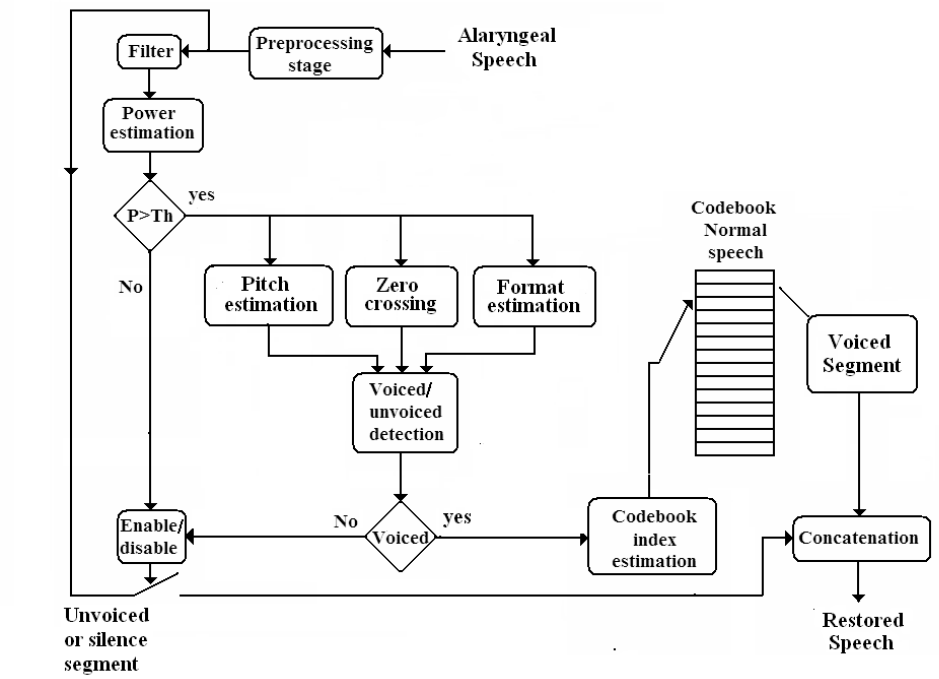


Figure 1. Alaryngeal speech enhancement systems proposed

2.1 Detection of voiced segments of alaryngeal speech.

A voiced (sonorous) segment is characterized by a periodic or quasiperiodic behavior in time, a fine harmonic frequency structure produced by the vibration of the vocal chords, as well as a high energy concentration due to the little obstruction that the air meets in its way through the vocal tract. The vowels and some consonants present this behavior.

Several approaches have been proposed to detect the voiced segments of speech signals. However, the use of a single criterion of decision to determine if a speech segment is voiced or unvoiced is not enough. Thus, most algorithms in the speech processing area use the combination of more than one criterion. The proposed speech restoration method uses the combination of pitch estimation, zero crossing and formant analysis of speech signal for voiced/unvoiced segment classification.

2.1.1 Pitch detection method

The first criterion used for voiced activity detection is the pitch information. To this end, the autocorrelation method [11], [12] is used, in which the speech segment is divided in blocks of 30 ms, with 50% of overlap. Next, a center clipper is applied. Subsequently, the autocorrelation of the resulting signal is estimated. Finally, the pitch is estimated as the time distance between the autocorrelation peak located in the origin $r_{xx}(0)$ and subsequent peak that is larger than $0.7r_{xx}(0)$. Thus, if the second peak exists, the segment is considered as voiced; otherwise, it is unvoiced.

2.1.2 Zero Crossing

The second criterion is based on the signal periodicity using the number of zero crossing in each frame. Here, two thresholds are used which establish that in a noise-free speech segment of 10 ms, a voiced segment crosses by zero about 12 times, while in an unvoiced segment, it crosses approximately 50 times [12, 13]. These values are not fixed and must be adjusted according to the sampling frequency used. In the proposed algorithm, for a sampling frequency of 8 kHz, the maximum value of zero crossings that could be detected in 10 ms is approximately 40. Thus, an upper threshold of 30 was chosen for voiced/unvoiced classification.

2.1.3. Formant Analysis

The third criterion is based on the amplitude of formants which, representing the resonance frequency of the vocal tract, are the envelope peaks of the speech signal power spectrum density. The frequencies in which the first formants are produced are of great importance in speech recognition.

The formants are obtained from the polynomial roots generated by the linear prediction coefficients (LPC) that represent the vocal tract filter. Once the formants, whose frequency is defined by the angle of the roots closer to the unitary circle, are obtained, they are ordered in an ascending form and the first three formants are chosen as parameters of the speech segment. These formants are then stored in the system so that they can be employed to take the voiced/invoiced decision. Using the normalized Fast Fourier Transform (FFT) of speech frame, the amplitude of the formant frequency can be obtained.

In order to decide whether the segment is voiced or not, the value of the formants amplitude is normalized each 100 millisecond segment. Then the algorithm finds the maximum value of each formant among the 10 values stored for each fragment. Then, each value is divided between the estimated maximum values as shown in (1).

$$AF_1 = \left[\frac{AF_{1-1}}{AF_{1Max}} \frac{AF_{1-2}}{AF_{1Max}} ..... \frac{AF_{1-10}}{AF_{1Max}} \right]$$
$$AF_2 = \left[\frac{AF_{2-1}}{AF_{2Max}} \frac{AF_{2-2}}{AF_{2Max}} ... \frac{AF_{2-10}}{AF_{2Max}} \right] \quad (1)$$
$$AF_3 = \left[\frac{AF_{3-1}}{AF_{3Max}} \frac{AF_{3-2}}{AF_{3Max}} .... \frac{AF_{3-10}}{AF_{3Max}} \right]$$

The local normalization process is justified for esophageal speakers due to the lost of energy as they speak. Once the normalized values are obtained, the decision is made using an experimental threshold value which is equal to 0.25. It can be seen as a logic mask in the algorithm if the normalized values are greater than 0.25, it is set to one, otherwise it is set to zero, as shown in (2).

$$\frac{AFx-N}{AFx_{\max}} = \begin{cases} 0 & \frac{AFx-N}{AFx_{\max}} < 0.25 \\ 1 & \frac{AFx-N}{AFx_{\max}} > 0.25 \end{cases} \quad (2)$$

Next an 'and' logic operation is realized with the three formant array using the values obtained after the threshold operation. Here, only the segments in which the three formants have values over the 0.25 are considered to be voiced segments.

Finally, using the three criteria mentioned above, a window is applied to the original signal which is equal to one if the segment is classified as voiced by the three methods; and it is equal to zero otherwise, such that only the voiced segments of the original signal are obtained.

2.2 Feature vector extraction

The performance of any speech recognition algorithm strongly depends on the accuracy of the feature extraction method. This fact has motivated the development of several efficient algorithms to estimate a set of parameters that allows a robust characterization of the speech signal such as the MEL scale [12-15], the Linear Predictive Coding (LPC) which models the vocal track [12-15], the cepstral coefficients [12,13], etc. Most widely used feature extraction methods, such as those described above, are based on modeling the form in which the speech signal is produced. However, if the speech signals are processed taking in account the form in which they are perceived by the human ear, similar or even better results may be obtained. Thus, the use of an ear model-based feature extraction method may be an attractive alternative because this approach allows characterizing the speech signal in the form that it is perceived [16]. Thus, a feature extraction method, based on an inner ear model taking into account the fundamentals concepts of critical bands, will be developed.

In the inner ear, the basilar membrane carries out a time-frequency decomposition of the audible signal through a multiresolution analysis similar to that performed by a wavelet transform [17]. Thus, to develop a feature extraction method that emulates the basilar membrane operation, it must be able to carry out a similar frequency decomposition, as proposed in the inner ear model developed by Zhang et. al. [17]. In this model, the dynamics of the basilar membrane, which has a characteristic frequency equal to fc, can be modeled by using a gamma-tone filter which consists of a gamma distribution multiplied by a pure tone of frequency fc. Here, the shape of the gamma distribution, α, is related to the filter order

while the scale θ is related to the period of occurrence of the events under analysis when they have a Poisson distribution. Thus, the gamma-tone filter representing the impulse response of the basilar membrane is given by [17]

$$\psi_{\theta}^{\alpha}(t)=\frac{1}{(\alpha-1)!\,\theta^{\alpha}}t^{\alpha-1}e^{\frac{-t}{\theta}}\cos(2\pi t/\theta)\quad t>0 \tag{3}$$

Equation (3) defines a family of gamma-tone filters characterized by θ and α. Thus, to emulate the basilar membrane behavior, it is necessary to look for the more suitable filter bank which, according to the basilar membrane model given by Zhang et. al. [17], can be obtained if we set $\theta=1$ and $\alpha=3$; because with these values (3) provide the best approximation to the inner ear dynamics. Thus, from (3), it follows that [18]

$$\psi(t)=\frac{1}{2}t^{2}e^{-t}\cos(2\pi t)\quad t>0 \tag{4}$$

Taking the Fourier transform of (4), it can be shown that $\psi(t)$ presents the expected attributes of a mother wavelet because it satisfies the admissibility condition given by [19]

$$\int_{-\infty}^{\infty}\frac{|\Psi(\omega)|^{2}}{|\omega|}d\omega<\infty \tag{5}$$

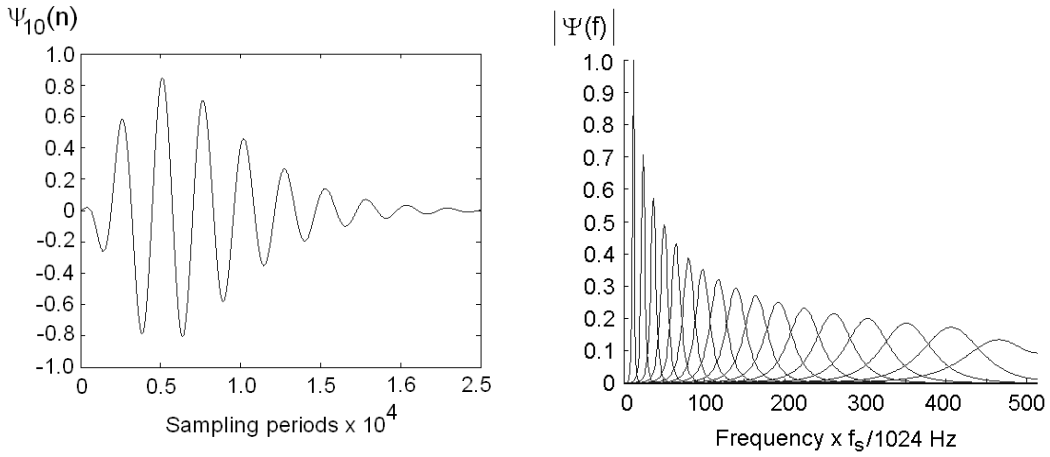
This means that $\psi(t)$ can be used to analyze and then reconstruct a signal without loss of information [19]. That is the functions given by (6) constitute an unconditional basis in L^2**(R)** [19]; and then we can estimate the expansion coefficients of an audio signal *f(t)* by using the scalar product between *f(t)* and the function $\psi(t)$ with translation τ and scaling factor s as follows [19]:

$$C(\tau,s)=\frac{1}{\sqrt{s}}\int_{0}^{\infty}f(t)\psi\left(\frac{t-\tau}{s}\right)dt \tag{6}$$

A sampled version of (6) must be specified because we require characterizing discrete time speech signals. To this end, a sampling of the scale parameter, s, involving the psychoacoustical phenomenon known as critical bandwidths will be used [20].

The critical bands theory models the basilar membrane operation as a filter bank in which the bandwidth of each filter increases as its central frequency also increases [6, 7]. This requirement can be satisfied using the Bark frequency scale that is a logarithmic scale in which the frequency resolution of any section of the basilar membrane is exactly equal to one Bark, regardless of its characteristic frequency. Because the Bark scale is characterized by a biological parameter, there is not an exact expression for it given as a result several different proposals available in the literature. Among them, the statistical fitting provided by Schroeder et al. [20] appears to be a suitable choice. Thus, using the approach provided in [8], the relation between the linear frequency, *f*, given in Hz and the Bark frequency *Z*, is given by [20]:

$$Z=7\ln\left(\frac{f}{650}+\sqrt{\left(\frac{f}{650}\right)^{2}+1}\right) \tag{7}$$



(a) (b)

Figure 2. (a) Gamma-tone Function derived from an inner ear model. (b) Frequency response of filterbank derived from an inner ear model.

Next, using (7) the *j-th* scaling factor sj given by the inverse of the *j-th* central frequency in *Hz*, f_c, corresponding to the *j-th* band in the Bark frequency scale becomes [19]

$$s_j = \frac{e^{j/7}}{325\left(e^{2/7}-1\right)}, \quad j = 1,2,3,.... \tag{8}$$

The inclusion of the bark frequency in the scaling factor estimation, as well as the relation between (7) and the dynamics of the basilar membrane, allows frequency decomposition similar to that carried out by the human ear. Because the scaling factor given by (8) satisfies the Littlewood Paley theorem since

$$\lim_{j\to+\infty}\frac{s_{j+1}}{s_j} = \lim_{j\to+\infty}\frac{e^{(j+1)/7}\left(e^{2j/7}-1\right)}{e^{j/7}\left(e^{2(j+1)/7}-1\right)} = e^{-1/7} \neq 1 \tag{9}$$

there is not information loss during the sampling process. Finally, the number of subbands is related with the sampling frequency as follows:

$$j_{\max} = \text{int}\left(7\ln\left(\frac{f_s}{1300}+\sqrt{\left(\frac{f_s}{1300}\right)^2+1}\right)\right) \tag{10}$$

Thus, for a sampling frequency equal to 8 KHz, the number of subbands becomes 17. Finally, the translation axis is naturally sampled because the input data is a discrete time signal and then the *j-th* decomposition signal can be estimated as follows [19]:

where T denotes the sampling period. Here, the expansion coefficients *Cj* obtained for each subband are used to estimate the feature vector to be used during the training and recognition tasks. Figure 2 shows *ψ10(n)* and the magnitude of the filter bank frequency response, respectively.

$$c_j(m)=\frac{1}{2}\sum_{n=0}^{\infty}\left[f(n)\left(\frac{(n-m)T}{s_j}\right)^2 e^{-\left(\frac{(n-m)T}{s_j}\right)}\cos\left(\frac{2\pi(n-m)T}{s_j}\right)\right] \qquad (11)$$

Using (11), the feature vector used for voiced segment identification consists of the following parameters [10]: the energy of the *m-th*, speech signal frame, $\overline{x^2}(n)$, where $1\le n\le N$ and N is number of samples in the *m-th* frame, the energy contained in each one of the 17 wavelet decomposition levels of m-th speech frame $\overline{C_j^2}(m)$, where $1\le j\le 17$; the difference between the energy of the previous and actual frames given by

$$d_x(m)=\overline{x^2}(n-mN)-\overline{x^2}(n-(m-1)N) \qquad (12)$$

together with the difference between the energy contained in each one of the 17 wavelet decomposition levels of current and previous frames,

$$\overline{v_j}=\overline{c_j^2}(m)-\overline{c_j^2}(m-1) \qquad (13)$$

where m is the number frame. Then the feature vector derived using the proposed approach becomes

Here, the last eighteen members of the feature vector include the spectral dynamics of speech signal concatenating the variation from the past feature vector to the current one.

2.3 Classification Stage

The classification stage consists of one neural network, which identifies the vowel, in cascade with a parallel array of 5 neural networks, which are used to identify the alaryngeal speech segment to be changed by its equivalent normal speech segment, as shown in Fig. 3. To this end, the estimated feature vector, given by Eq. (14), is fed into the first ANN (Fig. 3) to estimate the vowel present in the segment under analysis. Once the vowel is identified, the same feature vector is fed into the five ANN structures of the second stage, together with the output of the first ANN, to identify the vowel-consonant combination contained in the voiced segment under analysis. Here, the output of the enabled ANN (Fig. 3) corresponds to the codebook index of identified segment. Thus, the first ANN output is used to enable the ANN corresponding to the detected vowel, disabling the other four; while the second ANN is used to identify the vowel-consonant or vowel-vowel combination. The ANN in the first stage has 10 hidden neurons while the ANNs in the second stage has 25.

$$\mathbf{X}(m)=\left[\overline{x^2}(n-mN),\overline{c_1^2}(m),\overline{c_2^2}(m),..,\overline{c_{17}^2}(m),d_x(m),\overline{v}_1(m),\overline{v}_2(m),..,\overline{v}_{17}(m)\right] \qquad (14)$$

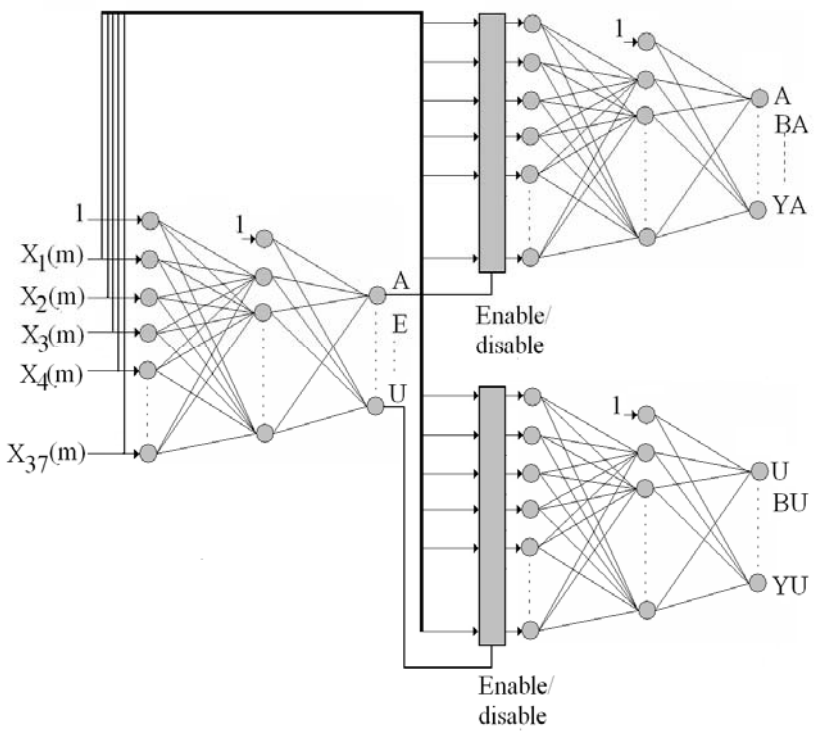


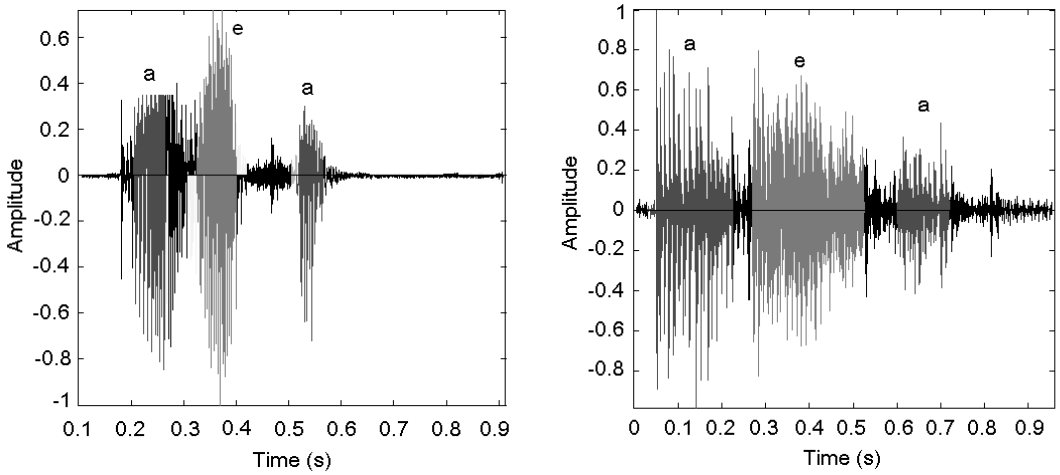
Figure 3. Pattern recognition stage. The first ANN indentifies the vowel present in the segment and the other 5 ANNs identify the consonant-vowel combination.

The ANN training process is carried out in two steps: First, the ANN used to identify the vowel contained in the speech segment is trained in a supervised manner using the backpropagation algorithm. After the convergence is achieved, the enabled ANN in the second stage, used to identify the vowel-consonant or vowel-vowel combination is also trained in a supervised manner using the backpropagation algorithm; while the coefficients vectors of the other 4 ANNs are kept constant. In all cases, 650 different alaryngeal voiced segments with a convergence factor equal to 0.009 are used, achieving a global mean square error of 0.1 after 400,000 iterations. [16].

3. Evaluation results

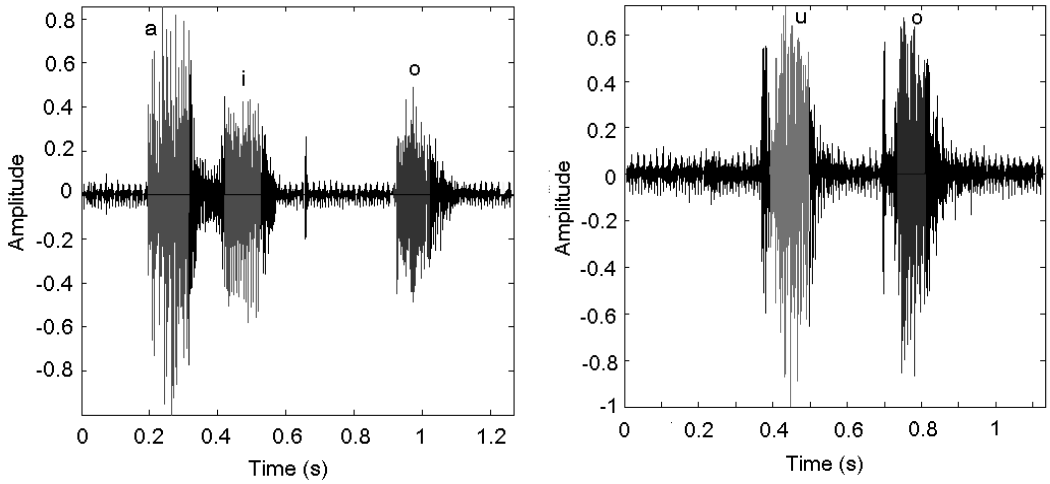
Figure 4 shows the plot of mono-aural recordings of the Spanish word “abeja”, pronounced by normal and esophageal speakers with a sample frequency of 8 kHz, respectively, including the detected voiced segments. Figure 5 shows the plot of the Spanish word “adicto” pronounced by an esophageal speaker together with the plot of the Spanish word “cupo”, in both cases the detected voiced segments are shown. These figures show that a correct detection is achieved using the combination of several features, in this case zero crossing, formats and pitch period.

Figure 6 shows the ALT produced speech signal corresponding to the Spanish word “cachucha” (cap) together with the restored signal obtained using the system proposed, while the corresponding spectrograms of both signals are shown in Fig. 6(b). Figure 7 shows the ALT produced and restored signals, respectively, corresponding to the Spanish word “hola” (hello), together with the corresponding spectrograms.



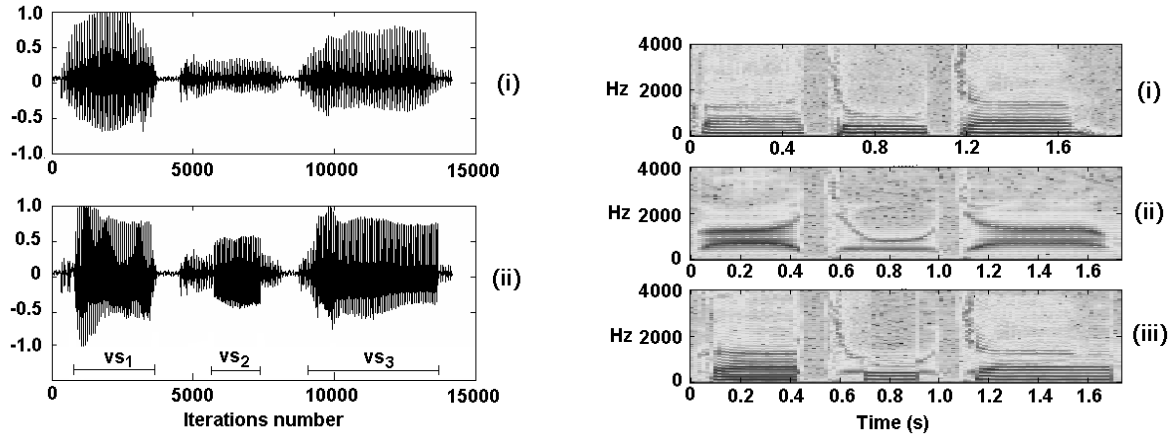
(a) (b)

Figure 4. Detected voiced and unvoiced segments of (a) the normal speech signal of the Spanish word “abeja” together with the detected vowels a, e, a, and (b) the esophageal Spanish word “abeja”, together with the detected vowels a, e, a.



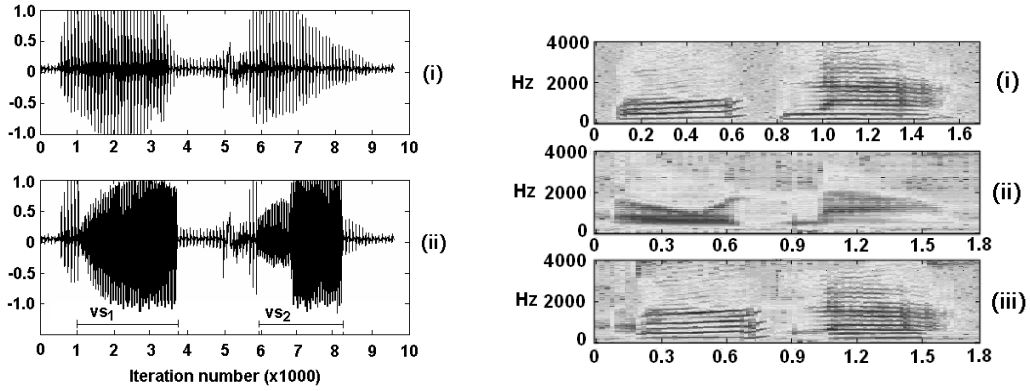
(a) (b)

Figure 5. Detected voiced and unvoiced segments of the esophageal speech signal of (a) the Spanish word “adicto” together with the detected vowels a, i, o, and (b) the Spanish word “cupo” together with the detected vowels u, o.



(a) (b)

Figure 6. (a) Waveforms trace corresponding to the Spanish word “Cachucha”, (Cap). i) ALT produced speech, ii) restored speech. (b) Spectrograms trace corresponding to the Spanish word “Cachucha” (Cap). i) Normal speech, ii) ALT produced speech, ii) restored speech.



(a) (b)

Figure 7. (a) Waveforms trace corresponding to the Spanish word “hola” (hello), i) ALT produced speech, ii) restored speech. (b) Spectrograms trace corresponding to the Spanish word “hola” (hello). i) Normal Speech, ii) ALT produced speech, iii) speech restored.

To evaluate the actual performance of the system proposed, two different criteria were used: the bark spectral distortion (MBSD) and the mean opinion scoring (MOS). The bark spectrum *L(f)* reflects the ear's nonlinear transformation of frequency and amplitude, together with the important aspects of its frequency and spectral integration properties in response to complex sounds. Using the Bark spectrum, an objective measure of the distortion can be defined using the overall distortion as the mean Euclidian distance between the spectral vectors of the normal speech, $L_n(k,i)$, and the processed ones, $L_p(k,i)$, taken over successive frames in an utterance as follows [20], [21]:

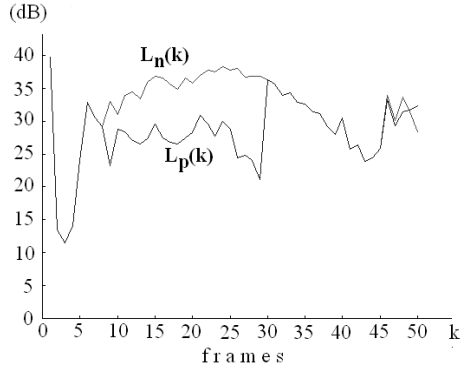


Figure 8. Bark spectral trace of normal, Ln(n), and enhanced, Lp(n), speech signals of the Spanish word "hola".

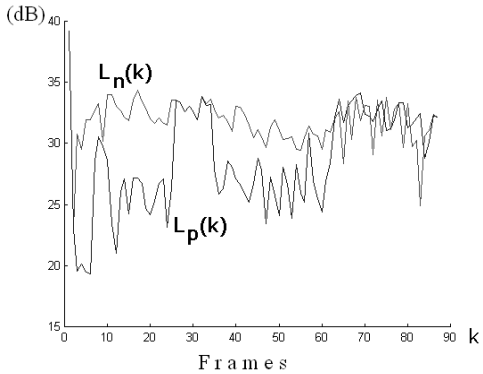


Figure 9. Bark spectral trace of normal, Ln(n), and enhanced, Lp(n), speech signals of the Spanish word "mochila".

$$MBSD = \frac{\sum_{k=1}^{N}\sum_{i=1}^{M}\left[L_n(k,i) - L_p(k,i)\right]^2}{\sum_{k=1}^{N}\sum_{i=1}^{M} L_n^2(k,i)} \qquad (15)$$

where $L_n(k,i)$ is the Bark spectrum of the kth segment of the original signal, $L_p(k,i)$ is the Bark spectrum of the processed signal and M is the number of critical bands. Figures 8 and 9 show the Bark spectral trace of both, the ALT produced and enhanced signals, respectively, corresponding to the Spanish words "hola" (hello) and "mochila" (bag). Here, the MBSD during voiced segments was equal 0.2954 and 0.4213 for "hola" and "mochila", respectively, while during unvoiced segments the MBSD was 0.6815 and 0.7829 for "hola" and "mochila", respectively. Here, the distortion decreases during the voiced periods as suggested by Eq. (15). Finally, an average MBSD equal to 0.7575 and 0.4213 were obtained for esophageal speech during unvoiced and voiced segments, respectively, using the Spanish word "Coca" (Coke), whose Bark spectral trace $L_n(k,i)$ and $L_p(k,i)$ are shown in Fig. 10. Evaluation results using the Bark spectral distortion measures show that a good enhancement can be achieved using the method proposed.

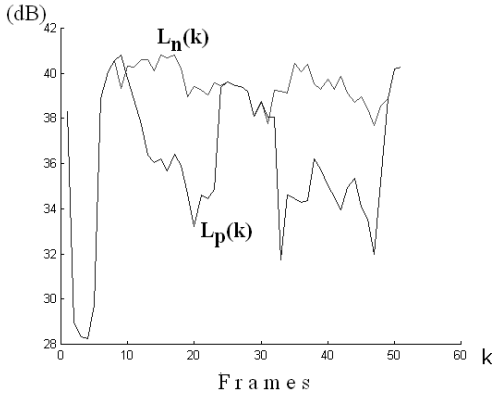


Figure 10. Bark spectral trace of normal, Ln(n), and enhanced, Lp(n), speech signals of the Spanish word "Coca".

A subjective evaluation was also performed using the Mean Opinion Scoring (MOS) in which the system proposed was evaluated by 200 normal speaking persons and 200 alaryngeal ones (Table 1 and Table 2), from the point of view of intelligibility and speech quality. Here, 5 is the highest score and 1 is the lowest one. In both cases the speech intelligibility and quality evaluation without enhancement are shown for comparison. These evaluation results show that the system proposed improves the performance of [2] which reports a MOS of 2.91 when the enhancement system is used and 2.3 without enhancement. These results also show that, although the improvement is perceived by the alaryngeal and normal speakers, the improvement is larger in the opinion of alaryngeal speakers. Thus, the system proposed is expected to have quite a good acceptance among the alaryngeal speakers because the system proposed allows synthesizing several kinds of male and female speech signals.

Finally, about 95% of alaryngeal persons participating in the subjective evaluation reported preferring to use the system during conversation. The last result is quite similar to that reported in [8] and [7]. Subjective evaluation shows that quite a good performance enhancement can be obtained using the system proposed.

	Normal listener		Alaryngeal listener	
	Quality	Intelligibility	Quality	Intelligibility
MOS	2.30	2.61	2.46	2.80
Var	0.086	0.12	0.085	0.11

Table 1. Subjective evaluation of esophageal speech without enhancement.

	Normal listener		Alaryngeal listener	
	Quality	Intelligibility	Quality	Intelligibility
MOS	2.91	2.74	3.42	3.01
Var	0.17	0.102	0.16	0.103

Table 2. Subjective evaluation of proposed alaryngeal speech enhancement system

The performance of the voiced segments classification stage was evaluated using 450 different alaryngeal voiced segments from which the system failed to classify correctly 22 segments, which represents a misclassification rate of about 5% using a network as identification method, while a misclassification of about 7% was obtained using the HMM. The comparison results are given in Table 3. Finally, to evaluate the behavior of the method proposed, it was compared with the performance of several other wavelet functions whose evaluation results are shown in Table 4. Evaluation results show that the methods proposed perform better than other wavelet based feature extraction methods.

Identification Method	**Normal Speech**	**Alaryngeal Speech**
ANN	98%	95%
HMM	97%	93%

Table 3. Recognition performance using two different identification methods using the feature extraction method proposed.

	Method proposed	Daub 4 wavelet	Haar wavelet	Mexican hat wavelet	Morlet wavelet
Recognition rate	95%	75%	40%	79%	89%

Table 4. Performance of different wavelet based feature enhanced methods when an ANN is used as identification method.

4. Conclusions

This paper proposed an alaryngeal speech restoration system suitable for esophageal and ALT produced speech based on a pattern recognition approach in which the voiced segments are identified and replaced by equivalent segments of normal speech contained in a codebook. The voiced segments are estimated using an ANN whose input vector is obtained using wavelet functions derived using the inner ear model developed by Zhang [17]. Evaluation results provided in Figs. 4 to 10 show the fairly good voiced segments detection and restoration performance of the algorithm proposed. It follows from the fact that the spectrograms of enhanced and normal speech signals are quite similar. Objective and subjective evaluation results, using the MBSD and the MOS criteria, are given which shows that the system proposed provides a good improvement in the intelligibility and quality of alaryngeal speech signals. Evaluation results also show that the feature extraction method proposed provides better detection performance than other widely used methods when used with ANN as well as HMM. Evaluation results show that the system proposed, which presents a flexible structure that allows it to enhance esophageal as well as artificial laryinx produced speech signals [2] without further modifications, is an attractive alternative to enhance alaryngeal speech signals. The system proposed could be used to enhance alaryngeal speech in several practical situations such as in telephone and teleconference systems, improving in such way the voice and life quality of alaryngeal persons.

References

[1] Barney H., Hawork H. & Dunn F., An experimental transitorized artifcial larynx,.Bell System Technical Journal, Vol. 38, 1959, pp. 1337-1356.

[2] Aguilar G., Nakano-Miyatake M. & Perez-Meana H., Alaryngeal Speech Enhancement Using Pattern Recognition Techniques, IEICE Trans. Inf. & Syst. Vol. E88-D, No. 7, 2005, pp. 1618-1622.

[3] Espy-Wilson, C., Chari V. & Huang C., Enhancement of alaryngeal speech by adaptive filtering, Technical report, Boston University, Boston, MA, 2000.

[4] Becerril H., Nakano-Miyatake M. & Perez-Meana H., Development of an adaptive system for voice enhancement in persons with artificial larynx using DSP, Cientifica, Vol. 8, No. 2, April 2004, pp. 12-20.

[5] Cole D., Sridharan S. & Geva M., Application of noise reduction techniques for alaryngeal speech enhancement, IEEE TECON Speech and Image Processing for Computing and Telecommunications, 1997, pp. 491-494.

[6] K. Matsui and N. Hara, Enhancement of esophageal speech using format synthesis, IEEE International Conference on Acoustic, Speech and Signal Pprocessing , Vo1. 1, 1999, pp. 81-84.

[7] Gorrits M. & Valiere J. , Low-band extension of telephone-band speech, IEEE International Conference on Acoustic, Speech and Signal Processing, 2000, pp. 1851-1854.

[8] Bi N. & Qi Y., Speech conversion and its application to alaryngeal speech enhancement, Proc. of The International Conference on Signal Processing, 1997, pp. 1586-1589.

[9] Bi N. & Qi Y., Application of speech conversion to alaryngeal speech enhancement, IEEE Trans. Speech and Audio Processing, Vol. 5, No. 2, March 1997, pp. 97-105.

[10] Aguilar G., Perez-Meana H., Nakano-Miyatake M. & Becerril H., Speech enhancement of voice produced by an electronic larynx, IEEE Midwest Symposium on Circuit and Systems, Vol. III, August 2004, pp. 37-40.

[11] Rabiner L. & Gold B., Digital processing of speech signals, Prentice Hall, Englewood Cliffs NJ, 1975.

[12] Rabiner L. & Juang B., Fundamentals of Speech Recognition, Prentice Hall, Piscataway, USA, 1993.

[13] Rabiner L., Juang B. & Lee C., An Overview of Automatic Speech Recognition, in Automatic Speech and Speaker Recognition: Advanced Topics, C. H. Lee, F. K. Soong and K. K. Paliwal editors, Kluwer Academic Publisher, 1996, pp. 1-30.

[14] Junqua J. & Halton J., Robustness in Automatic Speech Recognition, Kluwer Academic Publishers, Norwell MA, 1996.

[15] Suarez-Guerra S. & Oropeza-Rodriguez J., Introduction to Speech Recognition, in Advances in Audio and Speech Signal Processing; Technologies and Applications, H Perez-Meana editor, Idea Group Publishing, 2007, pp. 325-347.

[16] Mantilla-Caeiros A., Nakano-Miyatake M. & Perez-Meana H., A New Wavelet Function for Audio and Speech Processing, IEEE Midwest Symposium on Circuit and Systems, August 2007, pp. 101-104.

[17] Zhang X., Heinz M., Bruce I. & Carney L., A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression, Acoustical Society of America, Vol. 109, No.2, 2001, pp. 648-670.

[18] Mantilla-Caeiros A., Nakano.MIyatake M. & Perez-Meana H., Isolate speech recognition based on time-frequency analysis methods, Lecture Notes in Computer Science, vol. LNCS 5856, pp. 297-304.

[19] Rao R. & Bopardikar A., Wavelets Transforms, Introduction to Theory and Applications, Addison Wesley, New York, 1998.

[20] Schroeder M., "Objective measure of certain speech signal degradations based on masking properties of the human auditory perception", Frontiers of Speech Communication Research, Academic Press, New York, 1979.

[21] Wang S., Sekey A. & Gersho A., "An objective measure for predicting subjective quality of speech coders," IEEE Journal on Selected Areas in Comm., Vol. 10, No. 3, June 1992, pp. 819-829.

Acknowledgments

We thank the Consejo Nacional de Ciencia y Tecnología (CONACyT) for the support provided during the realization of this research. Also, we would like to thank Dr. Xochiquetzal Hernandez from the Instituto de la Comunicación Humana of the Centro Nacional de la Rehabilitación of Mexico for her assistance during the subjective system evaluation.

Authors´ Biography

Hector PEREZ-MEANA

He received his M.S. degree in electrical engineering from the Electro-Communications University of Tokyo, Japan in 1986 and his Ph.D. degree in electrical engineering from the Tokyo Institute of Technology, Tokyo, Japan, in 1989. From March 1989 to September 1991, he was a visiting researcher at Fujitsu Laboratories Ltd, Japan. From 2006 to 2009, he was the Dean of Graduate Section of the ESIME Culhuacan of the Instituto Politécnico Nacional of Mexico. In 1991, 1999 and 2000, he received the IEICE excellent Paper Award, the IPN Research Award and the IPN Research Diploma, respectively. In 1998 and 2009, he was the general chair of the ISITA and the MWSCAS. Prof. Perez-Meana is a senior member of the IEEE, member of The IEICE, member of the Mexican Researcher System, level 2, and member of The Mexican Academy of Science. His principal research interests are adaptive systems, image processing, pattern recognition watermarking and related fields.

Mariko NAKANO-MIYATAKE

She received the M.E. degree in electrical engineering from the University of Electro-Communications, Tokyo, Japan in 1985, and her Ph.D. degree in electrical engineering from The Universidad Autónoma Metropolitana (UAM), Mexico City, in 1998. From July 1992 to February 1997, she was in the Department of Electrical Engineering of the UAM, Mexico. In February 1997, she joined the Graduate Department of The Mechanical and Electrical Engineering School of the Instituto Politécnico Nacional of Mexico, where she is now a professor. Her research interests are in information security, image processing, pattern recognition and related field. Dr. Nakano is a member of the IEEE, RISP and the National Researchers System of Mexico.

Alfredo Victor MANTILLA-CAEIROS

He was born in May, 1966 in Havana, Cuba. He received the Ph.D. degree in communications and electronics from the Instituto Politécnico Nacional in 2007, He obtained the master of science degree in computer engineering from the Instituto Politécnico Nacional in 2000 and the B.Sc. degree in electronic engineering from the Instituto Superior Politécnico de la Habana, Cuba, in 1989. From January 2001 to present, he has been a full-time professor in the Mechatronics Department of the Instituto Superior de Estudios Superiores de Monterrey in Mexico City. His interest areas are digital signal processing, speech recognition and digital systems.