Chapter 9

Esophageal Speech Enhancement Using a Feature Extraction Method Based on Wavelet Transform

Alfredo Victor Mantilla Caeiros and Hector Manuel Pérez Meana

Additional information is available at the end of the chapter

http://dx.doi.org/10.5772/49943

1. Introduction

People that suffer from diseases such as throat cancer require that their larynx and vocal cords be extracted by surgery, 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 suggested, such as: The esophageal speech, the use of tracheoesophageal prosthetics and the Artificial Larynx Transducer (ALT), also known as "electronic larynx" [1, 2].

The ALT, which has the shape of a handheld device, introduces an excitation in the vocal tract by applying a vibration against the external walls of the neck. The excitation is then modulated by the movement of the oral cavity to produce the speech sound. This transducer is attached to the speaker's neck, and in some cases to the speaker's cheeks. The 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 that it is distorted by the ALT produced background noise. Thus, ALT results in a considerably degradation of the quality and intelligibility of speech, problem for which an optimal solution has not yet been found [2].

The esophageal speech, on the other hand, is produced by the compression of the contained air in the vocal tract, from the stomach to the mouth through the esophagus. This air is swallowed and it produces a vibration of the esophageal upper muscle as it passes through the esophageal-larynx segment, producing the speech. The generated sound 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 parts of the speech within a word or phrase resulting 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 [2, 3].

This chapter presents an alaryngeal speech enhancement system, which uses several methods for speech recognition such as voiced and unvoiced segment detection, feature extraction method and pattern recognition algorithms.

The content of this chapter is a follows:

1. Acquisition and preprocessing of esophageal speech: This section explains the acquisition and preprocessing of speech signal including filtering, segmentation and windowing.
2. Voiced/unvoiced segment detection: This section discusses several methods for classifying voiced and unvoiced segments such as:
   - Pitch detection
   - Zero crossing
   - Formant Analysis
3. Feature extraction: The performance of any speech recognition algorithm strongly depends on the accuracy of the feature extraction method. This section exposes some important feature extraction methods such as the Linear Predictive Coding (LPC), the Cepstral Coefficients, as well as a feature extraction method based on an inner ear model, which takes into account the fundamental concepts of critical bands using a wavelet function. The later method emulates the basilar membrane operation, through a multiresolution analysis similar to that performed by a wavelet transform.
4. Clasiffier: The parameter vector obtained in the feature extraction stage is supplied to a classifier. The classification stage consists of neural networks, which identifies the voiced segments present in segment under analysis.
5. Voice synthesis: The voiced segments detected are replaced by voiced segments of a normal speaker and concatenated with unvoiced and silent segments to produce the restored speech.
6. Results: Finally, using objective and subjective evaluation methods, it shows that the proposed system provides a fairly good improvement of the quality and intelligibility of alaryngeal speech signals.

2. Methods

Figure 1 shows a block diagram of the proposed system. It 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 is that the voiced segments have a more significant impact on the speech quality and intelligibility than the unvoiced segments.

The following explains the stages of the system.

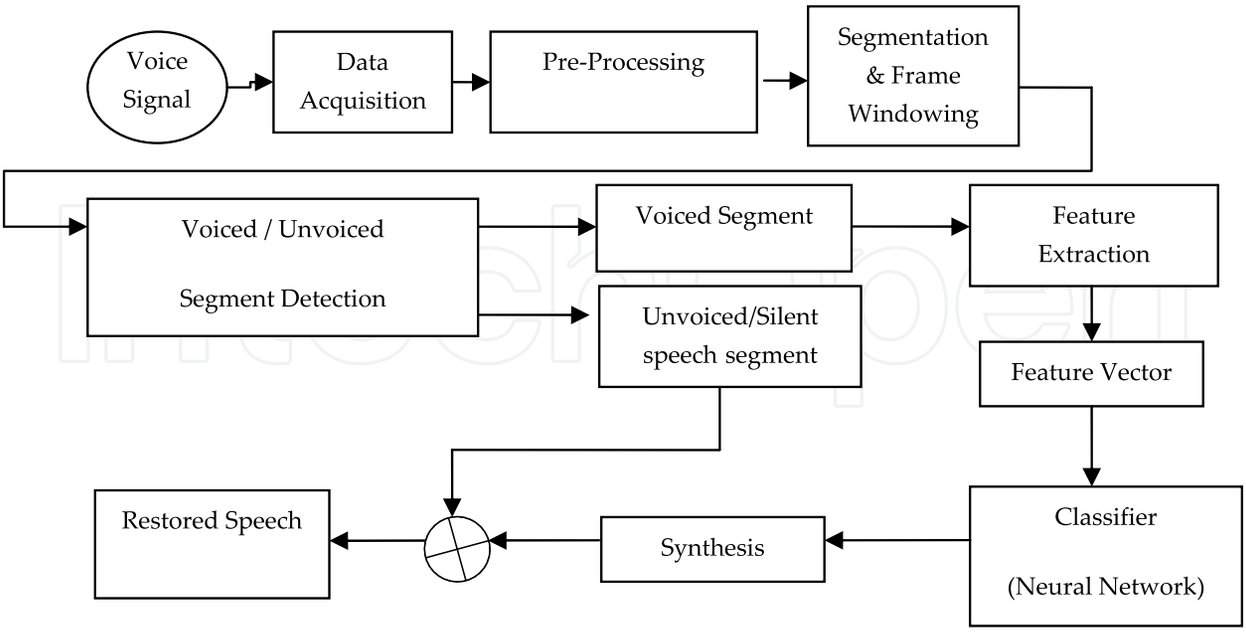

Figure 1. Block Diagram

2.1. Data acquisition

The first stage consists of recording a speech file, from an esophageal speaker. The audio data is kept as a WAV file. All files are monaural and they are digitalized in PCM (Pulse Code Modulation) format using a sampling rate of 8000Hz with a resolution of 8bits.

2.2. Preprocessing

The digital signal is low-pass filtered to reduce the background noise. This stage is implemented by a 200 order digital FIR filter with a cut-off frequency of 900Hz. A common practice for speech recognition is the use of pre-emphasis filter in order to amplify the higher frequency components of the signal with the purpose of emulating the additional sensibility of the human ear to high frequencies. Generally a high pass filter characterized by a slope of 20 dB per decade is used [4].

2.3. Segmentation and frame windowing

The filtered signal is divided into 100ms segments (800 samples) and at the same time each segment is subdivided into 10ms blocks, on which the later processing is going to be realized. The size of these blocks is determined by the quasi-periodic and quasi-stationary character of speech in such interval.

A Hamming window is applied to the segmented signal so that the extreme samples of the segments had less weight that the central samples. The window's length is chosen to be larger than the frame interval, preventing a loss of information which could take place during the transitions from one frame to the next.

2.4. Voiced/unvoiced segment detection

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 such 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 energy average, zero crossing and formant analysis of speech signal for voiced/unvoiced segment classification

2.4.1. Energy average

A first criterion ponders the average power of each frame by comparing it to that of its surroundings. An interval of 100 milliseconds in the neighborhood of the actual frame was selected. As part of the system's initial configuration, two thresholds are fixed. If the quotient between the frame's average power and that of the frame's surrounding is smaller than the lower threshold, the frame is labeled as unvoiced. Otherwise, if the quotient is larger than the higher threshold, the frame would be taken as voiced. For those cases in which the rate of average power lies between both thresholds, the energy criterion is not enough to determine the signal's nature.

2.4.2. Zero crossing

The second criterion is based on the signal periodicity using the number of zero crossings in each frame. Here two thresholds are used to establish that in a noise free speech segment of 10ms a voiced segment has about 12 zero crossings, while in an unvoiced segment has about 50 zero crossings [5, 6]. 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 10ms is approximately 40. Thus an upper threshold of 30 was chosen for voiced/unvoiced classification.

2.4.3. Formant analysis

The third criterion is based on the amplitude of formants which, represents the resonance frequency of the vocal tract. Formants 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 [4].

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) the amplitude of the formant frequency can be obtained.

To take the decision 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).

$$\begin{aligned} AF_1 &= \left[\frac{AF_{1-1}}{AF_{1Max}} \frac{AF_{1-2}}{AF_{1Max}} ..... \frac{AF_{1-10}}{AF_{1Max}} \right] \\ AF_2 &= \left[\frac{AF_{2-1}}{AF_{2Max}} \frac{AF_{2-2}}{AF_{2Max}} ... \frac{AF_{2-10}}{AF_{2Max}} \right] \\ AF_3 &= \left[\frac{AF_{3-1}}{AF_{3Max}} \frac{AF_{3-2}}{AF_{3Max}} .... \frac{AF_{3-10}}{AF_{3Max}} \right] \end{aligned} \tag{1}$$

The local normalization process is justified for esophageal speakers due to the loss 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 greater than 0.25 are set to one, otherwise are set to zero, as shown in (2).

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

Next an 'and' logic operation is applied 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 criterions 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.

2.5. 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 allow a robust characterization of the speech signal. Some of these methods are: The Linear Prediction Coefficients (LPCs), Formants

Frequencies Analysis, Mel Frequency Cepstral Coefficients (MFCC) among others [5,6]. This section discusses these methods and proposes one based on Wavelet Transform.

2.5.1. Linear Prediction Coefficients (LPCs)

The LPCs methods are based on the fact that the signal can be approximated from a weighted sum of precedent samples [7]. This approximation is given by:

$$s'_n = \sum_{k=1}^{p} a_k s_{n-k} \tag{3}$$

where a_k (1<k<p) is a set of real constants known as predictor coefficients, that must be calculated, and p is the predictor order. The problem of linear prediction resides on finding the predictor coefficients a_k that minimize the error between the real value of the function and the approximated function.

To minimize the total quadratic error is necessary to calculate the autocorrelation coefficients. This is a matrix equation with different recursive solutions, the commonly used is the Levinson recursion.

The developed algorithm takes each segment of 10 milliseconds and calculates its linear prediction coefficients. The number of predictor coefficients is obtained by substituting the sampling frequency value (f_s) in (4).

$$p = 4 + \frac{f_s}{1000} = 4 + \frac{8000}{1000} = 12 \tag{4}$$

The sequence of the minimal error could be interpreted as the output of the H(z) filter when it is excited by the S_n signal. H(z) is usually known as an inverted filter. The approximated transfer function could be obtained if it is assumed that the transfer function S(z) of the signal is modeled as an only pole filter with the form of (5).

$$\hat{S}(z) = \frac{A}{H(z)} = \frac{A}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{5}$$

The LPC coefficients correspond to $\hat{S}(z)$ poles. Therefore, the LPC analysis aims to calculate the filter properties of the vocal tract that produces the sonorous signal.

2.5.2. Formant frequencies analysis

Formants are the envelope peaks of the speech signal spectrum that represent the resonance frequency of the vocal tract.

If the spectrum of a speech signal can be approximated only by its poles, then the formants could be obtained from the $\hat{S}(z)$ poles. The poles of $\hat{S}(z)$ can be calculated making the

denominator of (5) to zero and solving it to find its roots. The S plane conversion is done by substituting z by e^{skT} , where s_k is the pole in the s plane. The resultant roots are generally conjugated complex pairs.

Formants frequencies are obtained from the polynomial roots generated by the linear prediction coefficients. The formant frequency is defined by the angle of the roots closer to the unitary circle. A root, with an angle close to zero, indicates the existence of a formant near the origin. A root whose angle is in close proximity to π indicates that the formant is located near the maximum frequency, in this case 4000Hz. Since the frequency dominion is symmetric with respect to the vertical axis, the roots located in the inferior semi plane of z plane can be ignored.

Let r_P as a linear prediction coefficient root with real part ϕ_p and imaginary part θ_p .

$$r_p = \phi_p + j\theta_p \tag{6}$$

The roots (r) which are located in the superior semi plane near the unitary circle can be obtained using (7).

$$r_p = \begin{cases} 0 & for \quad \theta_p \le 0.01 \\ r_p & for \quad \theta_p > 0.01 \end{cases} \tag{7}$$

By using the arctangent function is possible to obtain the roots angle. Doing this, the roots are mapped into the frequency dominion by using (8) to get the formants.

$$for_p = \vartheta_p \frac{f_s}{2\pi} \tag{8}$$

Once the formants are obtained, they are organized in ascending order, and the first three are chosen as parameters of the speech segment.

2.5.3. Mel Frequency Cepstral Coefficients (MFCC)

The cepstral coefficients estimation is another widely used feature extraction method in speech recognition problems. These coefficients form a very good features vector for the development of speech recognition algorithms, sometimes better than the LPC ones.

Cepstrum is defined as the inverse Fourier transform of spectrum module logarithm (9)

$$c(t) = F^{-1}\left[\log\left(\mathrm{S}(\omega)\right)\right] \tag{9}$$

Developing de above equation it obtains:

$$c(t) = F^{-1}\left\{\left[\log \mathrm{E}(\omega)\right] + \left[\log \mathrm{H}(\omega)\right]\right\} \tag{10}$$

The above equation indicates that Cepstrum of a signal is the sum of Cepstrum excitation source and the vocal tract filter. The vocal tract information is of slow variation, and it appears in the first cepstrum coefficients. For speech recognition application the vocal tract information is more important than excitation source. The cepstral coefficients can be estimated from the LPC coefficients applying the following recursion:

$$\begin{aligned} c_o &= \ln \sigma^2 \\ c_m &= a_m + \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k} \qquad 1 \le m \le p \\ & c_m \sum_{k=1}^{m-1} \left(\frac{k}{m} \right) c_k a_{m-k} \qquad m > p \end{aligned} \tag{11}$$

where *Cm* is the n-th LPC-Cepstral coefficients, *a* is the i-th LPC coefficients and m is the Cepstral index.

Usually the number of cepstral coefficients is equal to the number of LPC ones to avoid noise. A representation derived from the coefficients cepstrum are the Mel Frequency Cepstral Coefficients (MFCC) whose fundamental difference with Cepstrum coefficients is that the frequency bands are positioned according to a logarithmic scale known as MEL scale, which approximates the frequency response of the human auditory system more efficiently than Fast Fourier Transform(FFT).

2.5.4. Feature extraction method based on Wavelet Transform

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 into account the form in which they are perceived by the human ear, similar or even better results may be obtained. Thus using an ear model-based feature extraction method might represent an attractive alternative, since this approach allows characterizing the speech signal in the form that it is perceived [8]. This section proposes a feature extraction method based on an inner ear model, which takes into account the fundamentals concepts of critical bands.

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. Thus to develop a feature extraction method that emulates the basilar membrane operation, it should be able to carry out a similar frequency decomposition, as proposed in the inner ear model developed by Zhang et. al [9]. In this model the dynamics of basilar membrane, which has a characteristic frequency equal to f_c, can be modeled by using a gamma-tone filter which consists of a gamma distribution multiplied by a pure tone of frequency f_c. The shape of the gamma distribution α, is related to the filter order while the scale θ, is related to 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 (12)

$$\psi_{\theta}^{\alpha}(t)=\frac{1}{(\alpha-1)\,!\,\theta^{\alpha}}t^{\alpha-1}e^{\frac{-t}{\theta}}\cos(2\pi t/\theta)\quad t>0 \tag{12}$$

Equation (12) 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 [9], can be obtained if we set $\theta=1$ and $\alpha=3$, *since* such values (12) result in the best approximation to the inner ear dynamics. From (12) we have:

$$\psi(t)=\frac{1}{2}t^{2}e^{-t}\cos(2\pi t)\quad t>0 \tag{13}$$

Taking the Fourier transform of (13)

$$\Psi(\omega)=-\frac{(\omega-2\pi)^{2}}{\left[1+j(\omega-2\pi)\right]^{3}}+\frac{(\omega+2\pi)^{2}}{\left[1+j(\omega+2\pi)\right]^{3}} \tag{14}$$

It can be shown that $\psi(t)$ presents the expected attributes of a mother wavelet since it satisfies the admissibility condition given by (15)

$$\int_{-\infty}^{\infty}\frac{|\Psi(\omega)|^{2}}{|\omega|}d\omega<\infty \tag{15}$$

This means that $\psi(t)$ can be used to analyze and then reconstruct a signal without loss of information [10]. That is the functions given by (13) constitute an unconditional basis in $L^2(\mathbf{R})$ 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:

$$C(\tau,s)=\frac{1}{\sqrt{s}}\int_{0}^{\infty}f(t)\psi\left(\frac{t-\tau}{s}\right)dt \tag{16}$$

A sampled version of (16) 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 [11].

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. 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 [11], appears to be a suitable choice. Thus using the approach provided, the relation between the linear frequency, f in Hz and the Bark frequency Z, is given by

$$Z = 7\ln\left(\frac{f}{650} + \sqrt{\left(\frac{f}{650}\right)^2 + 1}\right) \tag{17}$$

Using (17) the *j-th* scaling factor s_j 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

$$s_j = \frac{e^{j/7}}{325\left(e^{2/7} - 1\right)}, \quad j = 1,2,3,.... \tag{18}$$

The inclusion of bark frequency in the scaling factor estimation, as well as the relation between (17) and the dynamics of basilar membrane, allows frequency decomposition similar to that carried out by the human ear. Since the scaling factor given by (18) satisfies the Littlewood-Paley theorem (19)

$$\lim_{j\to+\infty}\frac{s_{j+1}}{s_j} = \lim_{j\to+\infty}\frac{e^{(j+1)/7}\left(e^{2j/7} - 1\right)}{e^{j/7}\left(e^{2(j+1)/7} - 1\right)} = e^{-1/7} \neq 1 \tag{19}$$

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

$$j_{\max} = \text{int}\left(7\ln\left(\frac{f_s}{1300} + \sqrt{\left(\frac{f_s}{1300}\right)^2 + 1}\right)\right) \tag{20}$$

Therefore, 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

$$C_j(m) = \sum_{-\infty}^{\infty} f(n)\psi_j(n-m) \tag{21}$$

where

$$\psi_j(n) = \frac{1}{2}\left(\frac{nT}{s_j}\right)^2 e^{-\left(\frac{nT}{s_j}\right)}\cos\left(\frac{2\pi nT}{s_j}\right) \quad n > 0 \tag{22}$$

In (22) T denotes the sampling period. The expansion coefficients C_j obtained for each subband are used to estimate the feature vector to be used during the training and recognition tasks.

Using (21), the feature vector used for voiced segment identification consists of the following parameters:

a. 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.
b. The energy contained in each one of the 17 wavelet decomposition levels of *m-th* speech frame $\overline{C_j^2}(m)$, where $1 \le j \le 17$
c. The difference between the energy of the previous and actual frames given by (23)

$$d_x(m) = \overline{x^2}(n - mN) - \overline{x^2}(n - (m-1)N) \tag{23}$$

d. The difference between the energy contained in each one of the 17 wavelet decomposition levels of current and previous frames given by (24)

$$\overline{v_j} = \overline{c_j^2}(m) - \overline{c_j^2}(m-1) \tag{24}$$

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

$$\mathbf{X}(m) = \left[\overline{x^2}(n - mN), \overline{c_1^2}(m), \overline{c_2^2}(m), .., \overline{c_{17}^2}(m), d_x(m), \overline{v}_1(m), \overline{v}_2(m), .., \overline{v}_{17}(m) \right] \tag{25}$$

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.

2.6. 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 Figure 2. At this point, the estimated feature vector, given by (25), is feed into the first ANN (Figure 2) to estimate vowel present in the segment under analysis. Once the vowel is identified, the same feature vector is feed into the five ANN structures of the second stage, along with the output of first ANN, to identify the vowel-consonant combination contained in the voiced segment under analysis. The output of enabled ANN 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 have 25.

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 convergence is achieved, the enabled ANN in the second stage is used to identify the vowel-consonant or vowel-vowel combination and is also trained in a supervised manner using the backpropagation algorithm , while the coefficients vectors of the other 4 ANN 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.

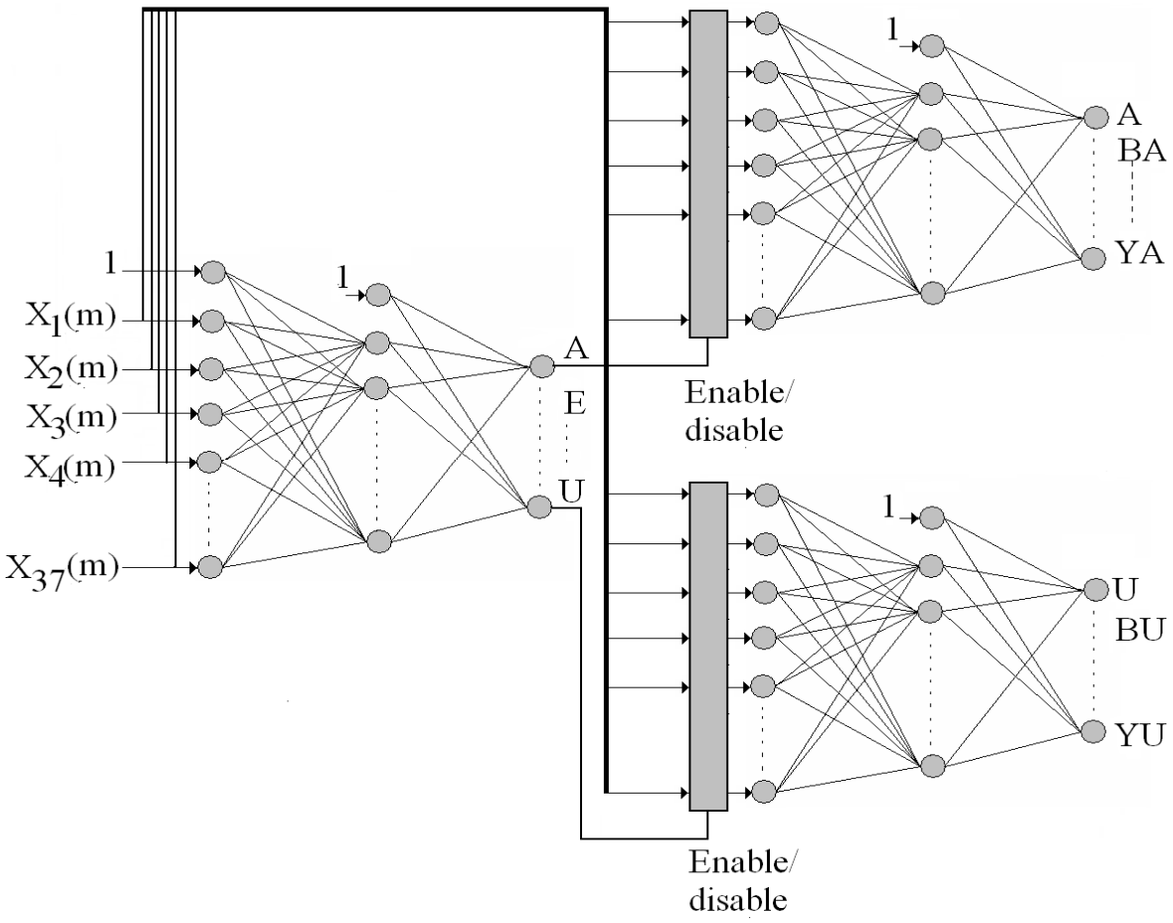

Figure 2. Pattern recognition stage. The first ANN indentifies the vowel present in the segment and the other 5 ANN identify the consonant-vowel combination.

2.7. Synthesis stage

This stage provides the restored speech signal. According to Figure 1, if a silence or unvoiced 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 energy analysis, the zero crossings number and the formant analysis explained in section 2.4. If a voiced segment is detected, it is identified using pattern recognition techniques (ANN). 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.

3. Results

Figure 3 shows the plot of mono-aural recordings of Spanish words “abeja” (a), “adicto” (b) and “cupo” (c), pronounced by an esophageal speaker with a sample frequency of 8 kHz, respectively, including the detected voiced segments. Figure 3 shows that a correct detection is achieved using the combination of several features, in this case zero crossing, formants analysis and energy average.

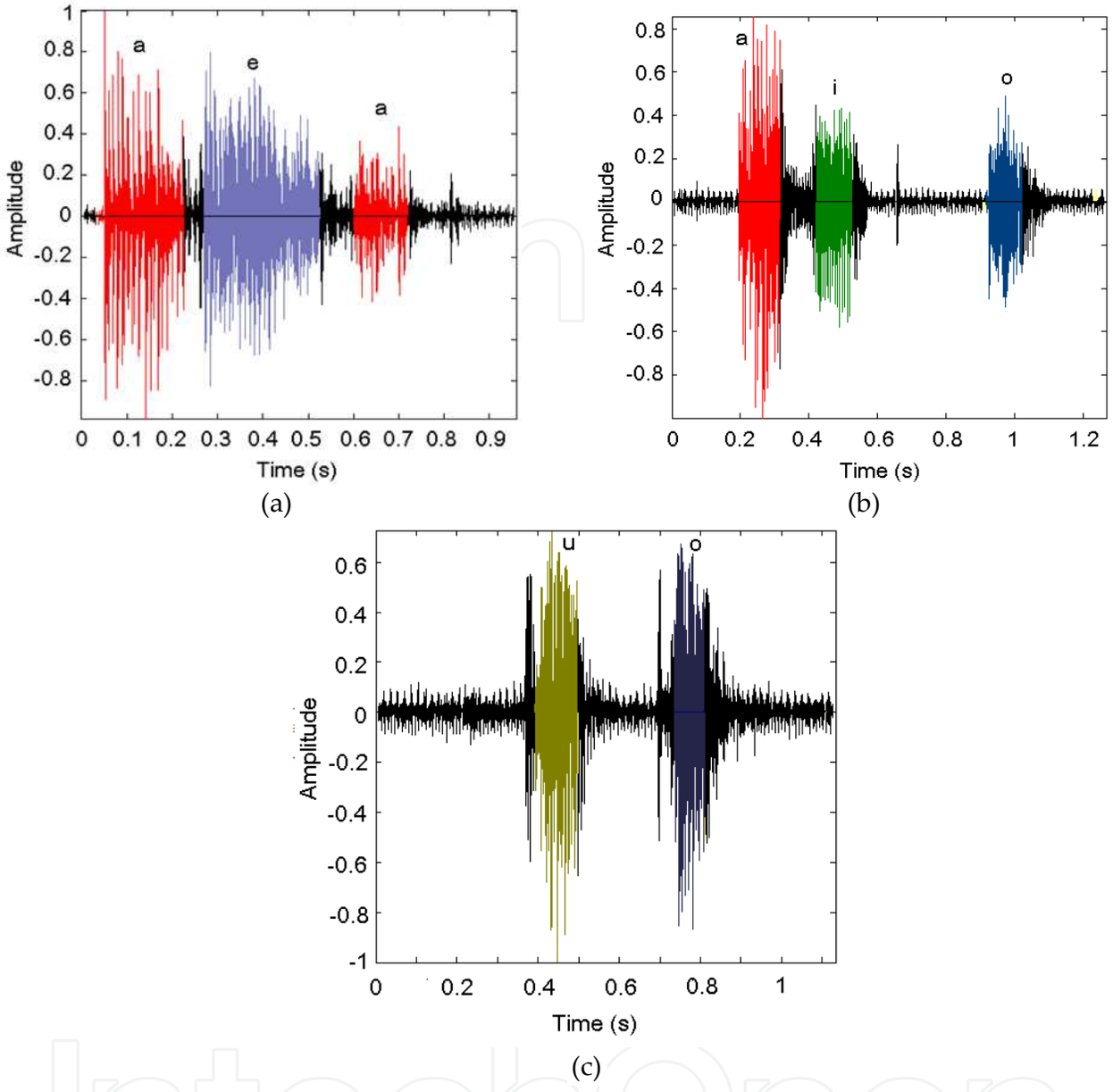

Figure 3. Detected voiced/unvoiced segments of esophageal speech signal of Spanish words "abeja" (a), "adicto" (b) and "cupo" (c).

Figure 4 shows the produced esophageal speech signal corresponding to the Spanish word "cachucha" (cap) together with the restored signal obtained using the proposed system. The corresponding spectrograms of both signals are shown in Figure 5.

To evaluate the actual performance of the proposed system, 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 as follows.

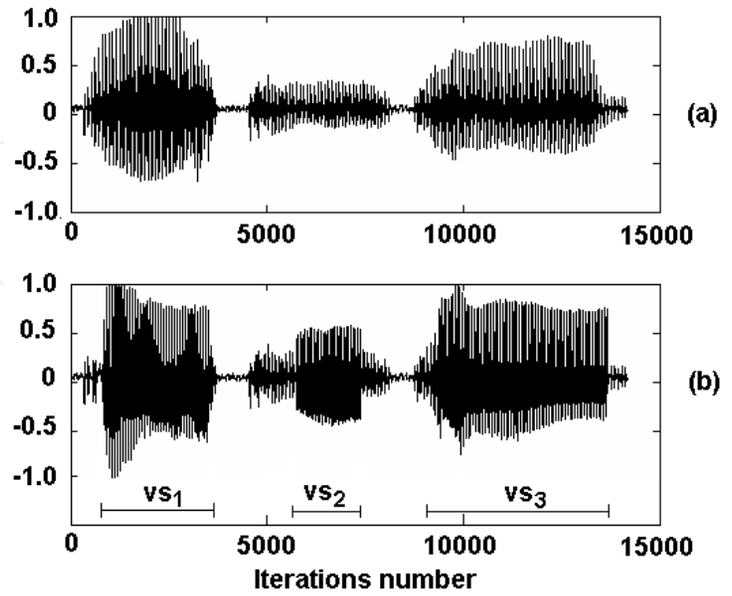

Figure 4. Waveforms trace corresponding to the Spanish word, "Cachucha", (Cap). a) produced Esophageal speech, b) restored speech.

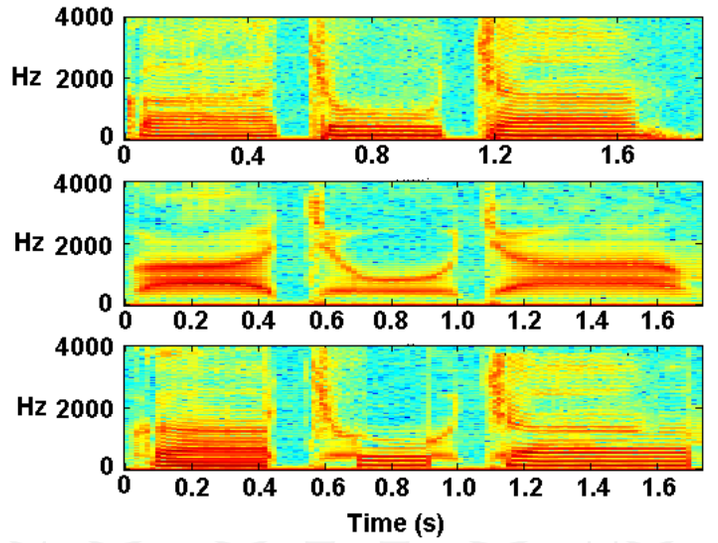

Figure 5. Spectrograms trace corresponding to the Spanish word, "Cachucha" (Cap). a) Normal speech, b) Produced Esophageal Speech, c) Restored speech.

$$MBSD = \frac{\sum_{k=1}^{N}\sum_{i=1}^{M}\left[L_n(k,i) - L_p(k,i)\right]^2}{\sum_{k=1}^{N}\sum_{i=1}^{M}L_n^2(k,i)} \tag{26}$$

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 6 and 7

show the Bark spectral trace of both, the esophageal speech 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. The distortion decreases during the voiced periods as suggested by (26). Evaluation results using the Bark spectral distortion measures show that a good enhancement can be achieved using the proposed method.

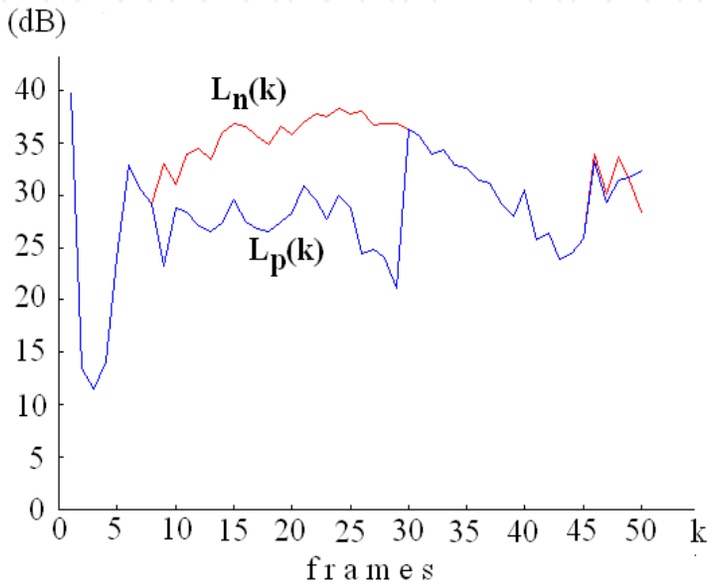

Figure 6. Bark spectral trace of normal, $L_n(k)$, and enhanced, $L_P(k)$, speech signals of the Spanish word "hola".

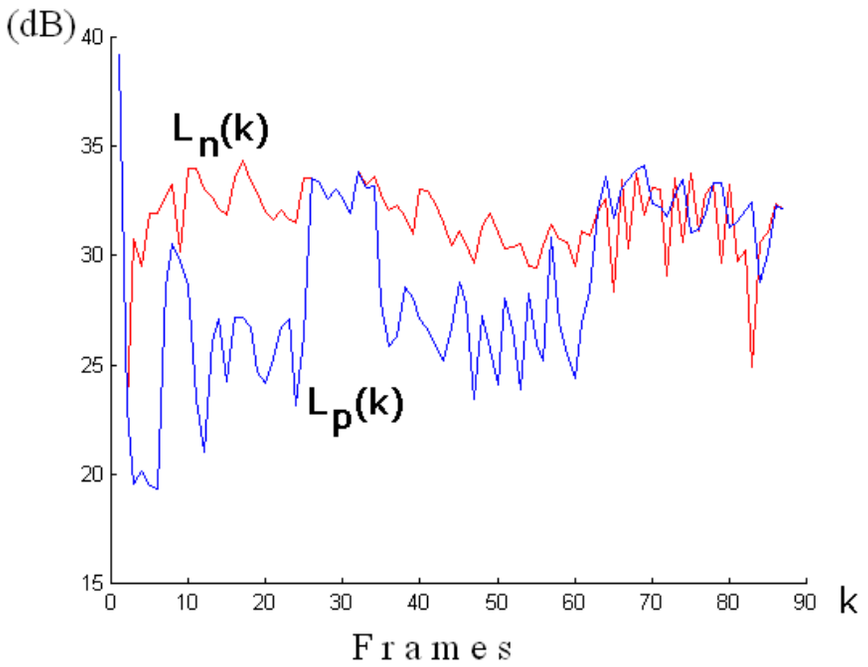

Figure 7. Bark spectral trace of normal, $L_n(k)$, and enhanced, $L_P(k)$, speech signals of the Spanish word "mochila".

A subjective evaluation was also performed using the Mean Opinion Scoring (MOS) in which the proposed system 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 where 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 proposed system 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 proposed system is expected to have a quite good acceptance among the alaryngeal speakers, because the proposed system allows synthesizing several kinds of male and female speech signals.

Finally, about 95% of alaryngeal persons participating in the subjective evaluation preferred the use of the proposed system during conversation. Subjective evaluation shows that quite a good performance enhancement can be obtained using the proposed system.

	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. 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 Hidden Markov Models (HMM). The comparison results are given in Table 3.

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 based on wavelet transform.

The behavior of proposed feature extraction method was compared with the performance of several other wavelet functions for evaluation purposes. Comparison results are shown in Table 4 which show that proposed method has better performance than other wavelet based feature extraction methods.

	Proposed method	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 chapter proposed an alaryngeal speech restoration system, suitable for esophageal and ALT produced speech, based on a pattern recognition approach where the voiced segments are replaced by equivalent segments of normal speech contained in a codebook. Evaluation results show a correct detection of voiced segment by comparison between their spectrograms to those spectrograms of normal speech signal. Objective and subjective evaluation results show that the proposed system provides a good improvement in the intelligibility and quality of esophageal produced speech signals. These results show that proposed system is an attractive alternative to enhance the alaryngeal speech signals. This chapter also presents a flexible structure that allows the use of the proposed system to enhance esophageal and artificial laryinx produced speech signals without further modifications. The proposed system could be used to enhance alaryngeal speech in several practical situations such as telephone and teleconference systems, thus improving the voice and quality life of alaryngeal people.

Author details

Alfredo Victor Mantilla Caeiros
Tecnológico de Monterrey, Campus Ciudad de Mexico, México

Hector Manuel Pérez Meana
Instituto Politécnico Nacional, México

5. References

[1] H. K. Barney, H. L. Hawork, F. E., and Dunn, (1959), "An experimental transitorized artifcial larynx",. Bell System Technical Journal, 38, 1337-1356..

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

[3] D. Cole, S. Sridharan and M. Geva, (1997), "Application of noise reduction techniques for alaryngeal speech enhancement", IEEE TECON, Speech and Image Processing for Computing and Telecommunications, pp. 491-494.
[4] H. David, et.al. (2001) "Acoustics and psychoacoustics", Ed: Focal Press. Second Edition.
[5] L. Rabiner, B. Juang, (1993), "Fundamentals of Speech Recognition", Prentice Hall, Piscataway, USA.
[6] L. R. Rabiner, B. H. Juang and C. H. Lee, (1996), 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, pp. 1-30, Norwell MA.
[7] D.G. Childers, (2000), "Speech Processing and syntesis toolboxes", Wiley & Sons, inc.
[8] A. Mantilla-Caeiros, M. Nakano-Miyatake, H. Perez-Meana, (2007), "A New Wavelet Function for Audio and Speech Processing", 50th MWSCAS, pp. 101-104.
[9] X. Zhang, M. Heinz, I. Bruce and L. Carney, (2001), "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, pp 648-670.
[10] R. M. Rao, A. S. Bopardikar (1998) ,"Wavelets Transforms, Introduction to Theory and Applications", Addison Wesley, New York.
[11] M. R. Schroeder, (1979) "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.