International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 21 (2017) pp. 11088-11093

Pattern Recognition of Speech Signals Using Wavelet Transform and Artificial Intelligence

[1]Oscar Rangel, [2]Dario Amaya and [3]Olga Ramos

Virtual Applications Group-GAV, Nueva Granada Military University –UMNG, Bogota, Colombia.
Orcid Id 0000-0002-1490-4970, Orcid Id 0000-0002-5422-9025

Abstract

Feature extraction in speech processing is one of the main phases to develop speech processing applications. A large set of feature extraction methods is available to implement on speech processing approaches, however the decomposition through Wavelet packets is one of the most popular nowadays for its robustness. This paper describes the development and implementation of the WPD technique using speech samples of the utterances of /cero/ and /uno/. The characteristic coefficients that result of the WPD are entered in a pattern recognition based on neural networks to classify data and recognize between the uttered words. The results show a classification above 75%, which demonstrates the suitability of the method for recognition.

Keywords: Wavelet Packet Decomposition, Wavelet Transform, Non-Audible Murmur, Neural Networks.

INTRODUCTION

Speech processing systems have allowed to develop many applications; from speech coding, text-to-speech synthesis, speaker identification/verification, to automatic speech recognition systems from which the development of natural language processing can be achieved [1].

The fundamental basis of each of the applications listed above is the same despite of the processing result between them. The fundamental step is known as the feature extraction stage of the system, and its objective is to reduce and characterize the data of a signal in a condensed way to be treatable in further processing stages. Classical approaches for feature extraction are based on temporal and frequency representations (i.e. correlation, FFT) which are the basis of more robust approaches like Linear Predictive Coding (LPC) [2], Perceptual Linear Prediction (PLP) [3], or cepstral representations (Cepstrum) [4].

The techniques listed above have been developed through the years taking into account different models of voice production [5]. Such models, have demonstrated that the voice as the output of the vocal tract, has a non-stationary behavior which means that the vocal tract system characteristics also varies over time [6].

The non-stationary nature of the sound waves produced in the vocal tract, has motivated the researchers to use new methods for extracting speech characteristics taking into account the variations over time of the statistical information of the signals. Methodologies like the Mel Frequency Cepstral Coefficients (MFCC) [7], have used the frequency representation of the signals through the Discrete Cosine Transform, the segmentation process and the iterative filtering to characterize the oscillatory changes over time [4].

Other methodologies that include the analysis of the frequency variations over time are based on the multiresolution analysis that can be performed using the Wavelet Transform [8]. Works like the one presented in [9] shows a comparison between the MFCC method and the Wavelet Packet (WP) approach to characterize speech signals for a speaker identification application. The comparison shows the accuracy of the WP approach which shows to better than the feature extraction based on the Fourier Transform.

In [10], is shown the extraction of coefficients using the WP methodology for automatic speech recognition (ASR). The method combines a similar filtering process from MFCC and describes the advantages of the analysis of variation characteristics of speech signals.

The Wavelet Packet Decomposition is used in [11] to build an ASR system for spoken words in Malayalam, this approach uses the WP for extracting characteristics coefficients and an information cost function to estimate relevant information of the signals which used for recognition.

Taking into account the advantages of the multiresolution analysis using Wavelets, in this paper is proposed a methodology for characterization and pattern recognition using the robust characteristics extracted by the Wavelet Packet Transform and neural networks to do a recognition system.

SPEECH RECOGNITION SYSTEM

The aim of the work presented in this paper is to describe the design criteria and the implementation steps taken into account in the construction of a speech recognition system, based on a wavelet extraction system and neural networks for

identification. The processing stages for speech recognition are shown in Figure 1.

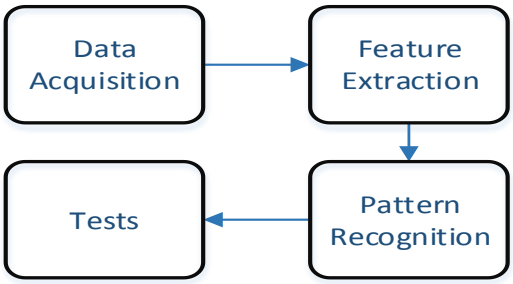

Figure 1: Sub-processing stages for speech recognition.

Figure 1 shows the four fundamental processing phases for the speech recognition. The first three correspond to the signal processing for extracting characteristics and the training of a feedforward neural network for patern recognition. The last process is the test of the system, which includes new features extracted from new speech samples to finally validate the system.

A Non-Audible Murmur microphone [12], is the transductor used to measure the NAM signals. A STM32F4-Discovery development board is used to perform the data acquirement. Through the ADC peripheral of the MCU, the speech signals are digitalized, then using an UART interface with a PC, the data is sent to perform feature extraction processes.

The signal in discrete terms is processed using a Matlab interface. There is implemented a Wavelet based feature extraction system to estimate characteristics of the signals. The Discrete Wavelet Transform (DWT) method for calculating Wavelet Packets (WP) is utilized as feature extraction routine, and as result 13 coefficients are extracted, which represent the variation characteristics of the processed speech signals.

Finally, the 13 coefficients are entered to a multilayer perceptron neural network to identify and recognize the uttered words. The recognition word vocabulary (lexicon) in this case corresponds to isolated digits of the Spanish language (zero /cero/, one /uno/). The phoneme units for such task, are defined as complete words since the lexicon is reduced to two isolated digits.

A. Data acquisition

A NAM transductor is a modified electret microphone used to acquire Non-Audible Murmur signals. The NAM microphone was first described by Nakajima et al in [13], the primary structure of this device is shown in Figure 2.

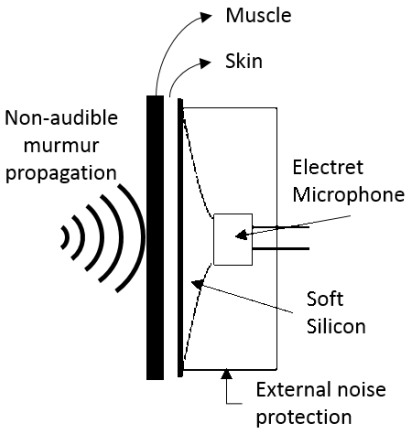

Figure 2: NAM microphone.

The NAM microphone depicted in Figure 2, is connected to the STM32F4-Discovery board to perform the acquisition stage. The STM32F407VG MCU, is configured to perform the analog-to-digital conversion at a frequency rate of 8000 Hz. The voltage resolution of the ADC in this case, is calculated from the operative voltage range of the MCU i.e. 0 V to 3 V. Eq (1) shows how to calculate the ADC resolution.

$$Q = \frac{V_{ref+} - V_{ref-}}{255} \quad (1)$$

Where Q is the resolution in volts, V_{ref+} and V_{ref-} are the upper and lower limits of operative voltage range of the MCU respectively. In this case the resolution Q is 11.76 mV, using a positive voltage of reference equal to 3 V and a negative reference of 0 V. The ADC bit resolution is 255, and was configured in this value to ease the UART data transmission.

The NAM microphone is located in the mastoid process, where has been demonstrated that is the best place for acquirement of NAM signals [14]. The mastoid process is the area located in the back part of the temporal bone, behind the ears. In this place the muscular tissue is soft, and allows the propagation of non-audible murmur waveforms when the NAM speech is performed. Figure 3 shows the NAM location in the mastoid process.

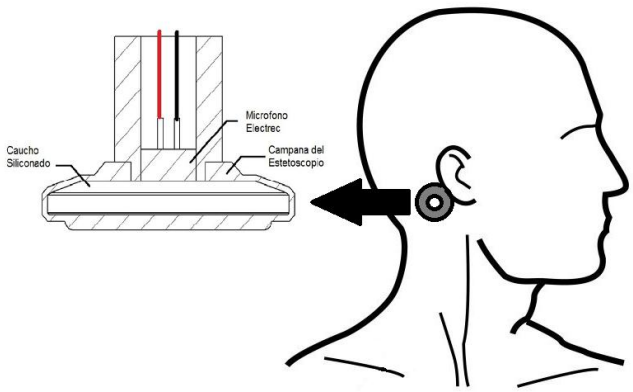

Figure 3: NAM microphone location.

The data of the digitalized signal, is transmitted to the Matlab interface using the UART protocol. The transmission baud rate was set in 9600 bits per second, ensuring a standardized rate. The summarized processing scheme for signal acquisition is depicted in Figure 4.

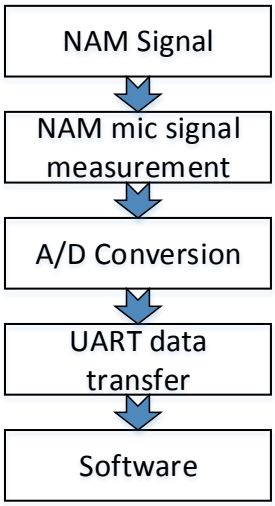

Figure 5: Acquisition and data transmission sub-processes.

B. Pre-processing

The pre-processing stage of the recognition system is implemented to isolate speech signals. It works using a calculation of the overall energy of a uttered sample, then a threshold is set to reject the parts of the signal where no variation is detected. The result of the isolating process is shown in Figure 6.

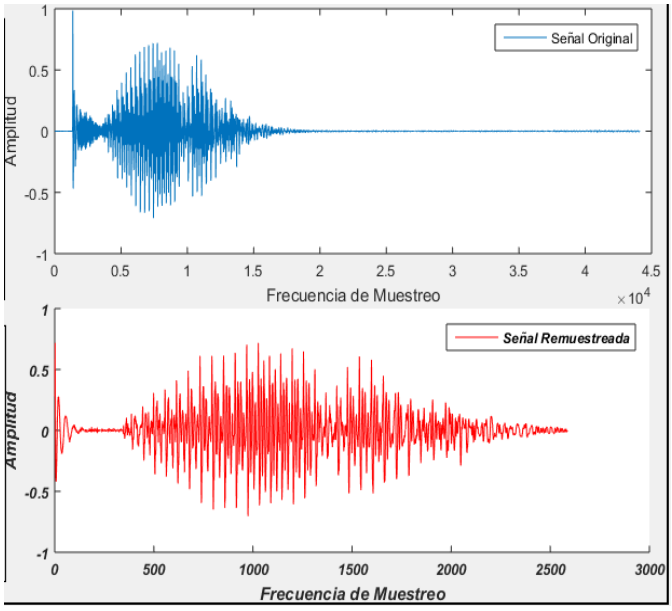

Figure 6: Original signal vs. Isolated signal.

WAVELET BASED FEATURE EXTRACTION SYSTEM

In general, a feature extraction system for speech processing involves several phases to estimate unique characteristics of the analyzed signals. In this case, the feature extraction phase is carried out with a Wavelet approach. The temporal and frequency versatility of Wavelet analysis is the main characteristic to consider it as an important method for signal analysis.

The multiresolution analysis of Wavelets is initiated with the Wavelet transform (WT) (Eq. (2)). The Wavelet transform is defined as the sum in time of the multiplication of the scaled version with the shifted version of the original signal. The mapping between time and frequency is an advantage over the Fourier analysis [15].

$$S(\tau, a) = \int_{-\infty}^{+\infty} S(t) \frac{1}{\sqrt{a}} \varphi * \left(\frac{t-\tau}{a}\right) dt \qquad (2)$$

Where φ is the mother Wavelet conjugate, which is scaled and shifted point to point to estimate the comparation levels with the analyzed signal $S(t)$. The value of $a = \frac{f}{f_0}$ from the dilatation of the Wavelet, using f_0 as the fundamental frequency and τ as the time shift [16].

A Wavelet Packet (WP) decomposition is used for feature extraction. The WP method is summarized in Figure 7, and is based on the WT definition.

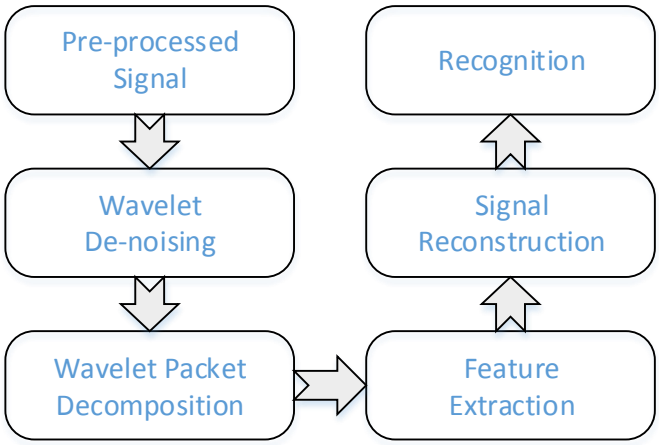

Figure 7: Wavelet Feature Extraction phases.

C. Denoising and Wavelet Packet Decomposition:

Denoising of a signal is the rejection of disturbance components that are present on a waveform. Such perturbations (or noise), can produce misunderstandings about the behavior of a signal which is a critical problem when signal analysis is performed. Using the filtering characteristics of the Wavelet method, is possible to reject the noise of a signal according to the environmental source and the frequency ranges of the speech (0, 8000 Hz).

The feature extraction phase in this project is carried out using a generalization of the WT called the Wavelet Packet decomposition. The WP method acts as a series of high-pass and low-pass filtering steps, which in each level produce packet vectors that characterize a signal through an information cost function. Figure 8, shows the Wavelet Packet decomposition scheme.

The schematic diagram depicted in Figure 8, shows the signal decomposition through high-pass and low-pass filters. For each *scale index,* a corresponding number of filters which varies in a power of two basis (2^j) are defined. The original signal $s[n]$, is splitted in two filtered parts (A1 and A2), which are filtered again to extract new Wavelet Packet vectors, which in turn are filtered until reach a final scale ($j = m$). In Figure 8, the functions $H_{j,i}$ and $G_{j,i}$ denote the high-pass and low-pass filters respectively, and produce the new WP vectors for the next scale. Each vector in each depth scale depending of the filtering process is known as an Approximation or Detail (A, D) set of coefficients.

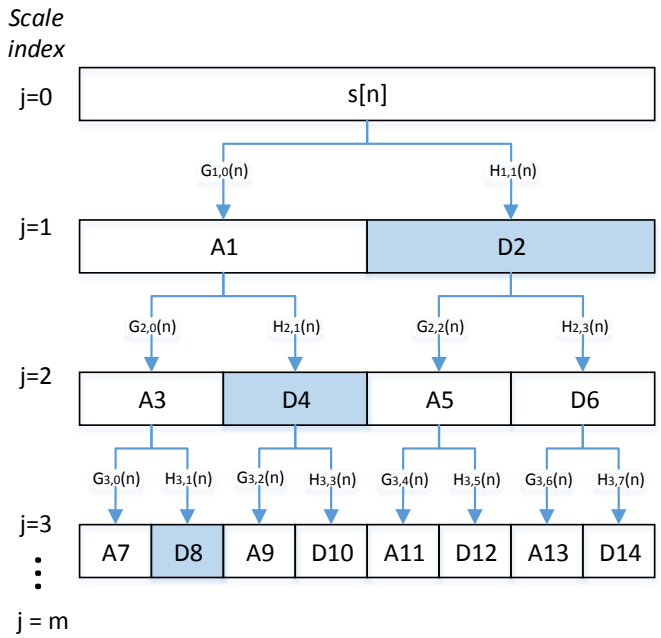

Figure 8: Diagram of Wavelet Decomposition.

As Figure 8 shows, at each scale level the WP vectors are reduced in size at a power of two rate, this leads to extract specific characteristics in specific time intervals, which is similar to the multiresolution analysis performed using the discrete Wavelet transform. The WP approach differs from the DWT analysis in the way of filtering the interest data from the original signal. For the case of Wavelet Packets, in each scale level, the previous WP vector is decomposed using the filters, however for the multiresolution analysis, the coefficients extraction is performed only for the detail coefficients of the predecessor approximation vector at the previous depth level as shown in Figure 8 for the colored tiles.

The advantage of Wavelet approaches over classical frequency methods is the analysis in both temporal and frequency domains [17]. The concept of multiresolution for exploring frequency characteristics in a not uniform variation of time intervals, allows to perform robust analysis of the frequency variation over time of speech signals; taking into account the fundamental tenet of speech processing that states that voice signals are of the non-stationary type [6].

The resulted coefficients of the filtering phase at the final depth scale reached by the WP decomposition, is often too large to be used as is in further recognition steps. To resolve this issue, an information cost function is implemented to measure and retain important features in a small set of coefficients as possible.

D. Feature Extraction:

The feature extraction process for extracting unique data using the WP approach is performed by implementing an information cost function. For this case, such data retention is carried out by the Shannon Entropy measure, which is defined for a discrete sequence s_i as shown in Eq. 3.

$$E(p) = -\sum_i p_i^2 \log(p_i^2) \qquad i = 0,1,\dots,N-1 \tag{3}$$

Where $E(s)$ is the resulted Shannon Entropy, p_i^2 is the set of normalized energies from the discrete Wavelet packet analyzed and N is the length of the WP vector.

The Shannon entropy for the detail WP and the las approximation set is estimated for the speech signals in the process described in this paper. The information measure is the maximum limit of compression of a signal without loss of information [18], which is a suitable method to acquire feature information of the signal. The final coefficients represented using the Shannon entropy for 6 speech samples are depicted in Figure 9.

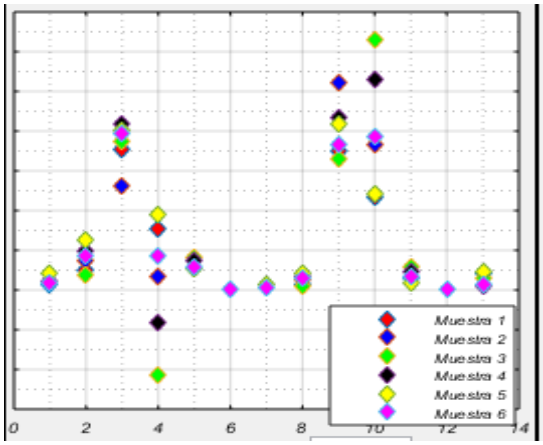

Figure 9: Coefficient representation of 6 speech signals of the utterance of /uno/ (one).

Figure 9, shows 13 coefficients of 6 different signals for the utterance of the word /one/. A correspondence between coefficients is observed and represent unique data of the signals.

Taking into account the Eq. 4, the reconstruction of the signal for performed to estimate how well the extracted coefficients represent the signal. The success of this stage determined the suitability of the feature extraction method to proceed with pattern recognition phases.

$$\tilde{x} = \tilde{a}_j + \tilde{d}_{j-1} + \cdots + \tilde{d}_1 \quad (4)$$

Where j is the decomposition level and $\tilde{a}$, $\tilde{d}$ are the approximation and detail coefficients respectively.

PATTERN RECOGNITION

Pattern recognition is a derivation of the methodologies used in machine learning, whose objective is to identify regular behavior in data series. These techniques search for unique patterns to classify certain events or categories in order to recognize information [19]. In the development of a recognition system of speech signals for utterances of the "uno" and "cero" digits, the classifier is based on neural networks, and the categories are defined as the digits to recognize with a defined set of characteristics extracted from the WP method.

A *feed-forward* neural network was used as the classifier. Such network is composed by three layers, the first layer receives the WP characteristic vectors as inputs, the second layer processes the data, and the last layer adjusts the processed values to be used as outputs. A generic representation of the neural network is shown in Figure 10.

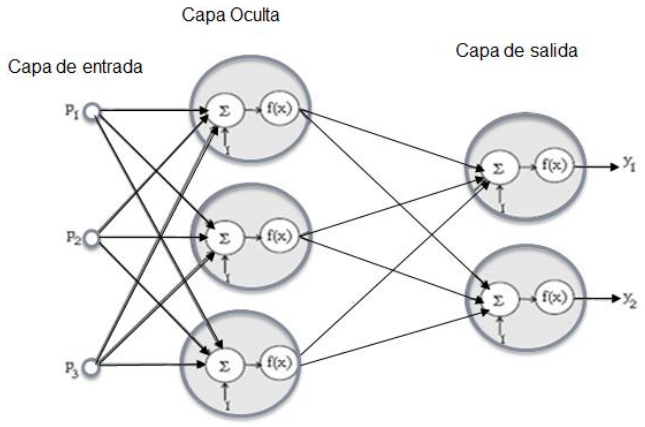

Figure 10: Generic representation of a multilayer neural network.

The activation function for the hidden layer is a sigmoidal function, which is represented by Eq. (5).

$$f(n) = \frac{1}{1 + e^{-n}} \quad (5)$$

The activation function for the output layer is based on a *softmax* function, as represented by Eq. (6).

$$g(n) = \frac{e^n}{\sum_{i=1}^{k} e^n} \quad (6)$$

RESULTS

The feature extraction phase was performed using 6 speech samples of one test subject for each of the categories to recognize; the phoneme units treated as words, and the categories as /cero/ and /uno/ utterances.

To validate the speech recognition system, 10 samples were recorded for the utterances of /uno/ and /cero/, the results of recognition are condensed in the Tables 1 and 2.

Table 1: Recognition results for the first five samples.

Utterance	***Test 1***	***Test 2***	***Test 3***	***Test 4***	***Test 5***
/cero/	81.15	81.96	68.91	72.10	71.28
/uno/	76.95	75.90	74.24	79.79	75.16

Table 2: Recognition results for the last five samples.

Utterance	***Test 6***	***Test 7***	***Test 8***	***Test 9***	***Test 10***
/cero/	71.13	74.99	78.18	74.18	70.86
/uno/	77.59	75.84	76.85	72.84	79.26

From the Table 1 and 2, it is shown that the percentage of recognition has an average of 75.45 %, which is sufficient taking into account the amount of samples used for training and the number of categories that the neural network has to recognize.

CONCLUSIONS

In this paper was described the development of a speech recognition system using Wavelet methods. The use of frequency and temporal representations jointly with neural networks, demonstrated to be a suitable strategy for pattern recognition.

The decomposition approach based on the implementation of high-pass and low-pass filters in each of the depth levels of the multiresolution analysis, allows to perform analysis in time and frequency to extract variation characteristics of the speech

signals, and the behavior that they have between the oscillatory changes over time.

The pattern recognition phase of the system, build from the design of neural networks is a suitable strategy which brings high recognition rates, that jointly with a robust feature extraction process can result in high recognition percentages as shown in Tables 1 and 2.

REFERENCES

[1] K. Brown, L. R. Rabiner, and B.-H. Juang, *Speech Recognition: Statistical Methods*. Elsevier, 2006.

[2] J.-D. Wu and B.-F. Lin, "Speaker identification based on the frame linear predictive coding spectrum technique," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 8056–8063, May 2009.

[3] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP features for speaker verification using i-vectors," *Speech Commun.*, vol. 55, no. 2, pp. 237–251, Feb. 2013.

[4] D. Jurafsky and J. H. Martin, *Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition*, 2nd ed. Prentice Hall, 2009.

[5] T. Drugman, B. Bozkurt, and T. Dutoit, "Causal–anticausal decomposition of speech using complex cepstrum for glottal source estimation," *Speech Commun.*, vol. 53, no. 6, pp. 855–866, Jul. 2011.

[6] L. R. Rabiner and R. W. Schafer, *Theory and Applications of Digital Speech Processing*, 1st ed. Pearson, 2011.

[7] M. A. Hossan, S. Memon, and M. A. Gregory, "A novel approach for MFCC feature extraction," in *2010 4th International Conference on Signal Processing and Communication Systems*, 2010, pp. 1–5.

[8] F. L. Sanchez, S. Barbon, L. S. Vieira, R. C. Guido, E. S. Fonseca, P. R. Scalassara, C. D. Maciel, J. C. Pereira, and S.-H. Chen, "Wavelet-based cepstrum calculation," *J. Comput. Appl. Math.*, vol. 227, no. 2, pp. 288–293, May 2009.

[9] C. Turner and A. Joseph, "A Wavelet Packet and Mel-Frequency Cepstral Coefficients-Based Feature Extraction Method for Speaker Identification," *Procedia Comput. Sci.*, vol. 61, pp. 416–421, 2015.

[10] E. Pavez and J. F. Silva, "Analysis and design of Wavelet-Packet Cepstral coefficients for automatic speech recognition," *Speech Commun.*, vol. 54, no. 6, pp. 814–835, Jul. 2012.

[11] S. Sunny, D. P. S., and K. P. Jacob, "Feature Extraction Methods Based on Linear Predictive Coding and Wavelet Packet Decomposition for Recognizing Spoken Words in Malayalam," in *2012 International Conference on Advances in Computing and Communications*, 2012, pp. 27–30.

[12] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Remodeling of the sensor for non-audible murmur (NAM)," in *Interspeech 2005*, 2005, pp. 389–392.

[13] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, "Non-audible murmur (nam) speech recognition using a stethoscopic nam microphone," *Word J. Int. Linguist. Assoc.*

[14] T. Toda, "Statistical approaches to enhancement of body-conducted speech detected with non-audible murmur microphone," in *2012 ICME International Conference on Complex Medical Engineering (CME)*, 2012, pp. 623–628.

[15] D. Y. Loni and S. Subbaraman, "Formant estimation of speech and singing voice by combining wavelet with LPC and Cepstrum techniques," in *2014 9th International Conference on Industrial and Information Systems (ICIIS)*, 2014, pp. 1–7.

[16] Doubert Geovanny Sánchez Marín, "Segmentación y Realce de Señales de Voz Usando la Transformada Wavelet y DSP's," University of Quindío, 2004.

[17] D. O'Shaughnessy, "Acoustic Analysis for Automatic Speech Recognition," *Proc. IEEE*, vol. 101, no. 5, pp. 1038–1053, May 2013.

[18] D. M. Ballesteros, "Aplicación de la transformada wavelet discreta en el filtrado de señales bioeléctricas," *Umbral Científico*, vol. 5, pp. 92–98, 2004.

[19] O. L. Ramos, L. A. Góngora, and D. A. Rojas, "Reconocimiento de patrones vocálicos utilizando MFCC y redes neuronales," in *XI Congreso Internacional de Electrónica, Control y Telecomunicaciones*, 2015, pp. 56–63.