VOL. 11, NO. 13, JULY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

FEATURE EXTRACTION AND PATTERN IDENTIFICATION OF SILENT SPEECH BY USING MFCC, DTW AND AI ALGORITHMS

Diego Alfonso Rojas[1], Olga Lucia Ramos[1]and João Mauricio Rosário[2]
[1]Department of Mechatronic Engineering, Nueva Granada Military University, Bogota, Cundinamarca, Colombia
[2]Faculty of Mechanical Engineering, University of Campinas, Campinas, Brasil
E-Mail: u3900213@unimilitar.edu.co

ABSTRACT

At present are many the ways of communication, that allow the interaction among people of the same society, a particular case of these communication ways appears the silent speech, but it is still in study and development. Silent speech, is to acquire the signals generated in the vocal apparatus before a sound occur, in order to establish a channel of information transmission in environments with a considerable amount of noise or among people whose ability to emit sounds is limited due to various pathologies. In this work the results of capturing and analyzing signals of silent speech, with the aim of identifying phonological units of Spanish language are presented. Initially the signals acquisition was performed by NAM microphone, for further processing with feature extraction techniques like MFCC (Mel Frequency Cepstral Coefficients) and DTW (Dynamic Time Warping), which provided the necessary data for training the neural network for the pattern recognition task. The classification algorithm was trained with the data of 9 test subjects, all of the male gender, with 5 samples from each of the three phonological units that want to be recognized ('Uno','Dos','Tres'), as a final result the algorithm is able to classify and identify patterns with a success rate over 85%.

Keywords: silent speech, MFCC, DTW, neural networks, pattern recognition.

1. INTRODUCTION

Communication in any society or culture, is the process of to emit signals with the intention to convey a message [1] and for to be successful the receiver and the transmitter must have the necessary skills to interpret correctly the signals that are in the message. For humans, this process comes from the neurological and psychological activity derived from the production of language, thought and psychosocial skills, therefore, it is imperative to develop interaction skills that allow establish relations with peers and its environment.

Communication of humans can be divided into two categories, verbal communication [2], it refers to transmit the information by using linguistic signs typical of a language or culture that can be written or oral, and the non-verbal [3], that is sending and receiving information through gestures or signs.

One of the most natural forms of communication is speech, that involves three stages in its production, the first is intended to generate a concept related to a particular sound, the second includes the grammatical, morpho-phonological and phonetic encoding for generating the corresponding activation in the organs that belongs to the vocal apparatus, and finally, articulation or voice output is performed. [4]

Currently, there are technological developments able to process, interpret and generate speech known as ASR (Automatic Speech Recognition) [5], that base their operation on the acoustics of the voice signals thus have certain limitations, such as susceptibility to noise in the transmission environment of the message and also depend on the amplitude of the speech, so pathologies such as dysphonia and aphonia, that affect voice production, also affect system performance.

As a possible solution to the limitations mentioned above, the use of sub-vocal speech is proposed, since this does not depend on the production of audible voice, instead of this uses signals acquired from the bio-electrical processes that control the different parts of vocal apparatus. From this concept, developments as the Silent Speech Interfaces appears [6], nevertheless, these systems are still experimental.

Some studies have been made approaches to processing and classification of these kind of signals using different devices to acquire them as is the case of [7] where used sEMG sensors, capable of recording muscle activity generated during the sound production, obtaining a recognition error in spoken words equal to 16.8% and whispered with 34.8%, in the same way [8] uses similar sensors, but with another feature extraction, Wavelet Packets, achieving a recognition rate between 74% and 76%. In other works, the behavior of organs that belong to the vocal tract, such as the glottis, have been studied by vibration analysis [9] or electromagnetism [10], where they conclude that both systems have a wide potential in the detection of sub-vocal stimulus.

Developments with ultrasonic sensors and image processing to analyze the movement of the tongue have been developed as shown in [11], where achieves a recognition rate of 60% for phonemes of the English language using PCA (Principal Component Analysis) y HMM (Hidden Markov Models).

Another acquisition device widely used for recording signals of this type is the NAM microphone, with this many studies have been developed, in addition, significant results were obtained as it can be seen [12] and [13], also using techniques like LAD (Linear Discriminant Analysis) y GMM (Gaussian Mixture Models) was obtained a identification rate over 60%, as shows [14], finally in [15], a review of the various devices mentioned above was made, concluding that the NAM microphone has certain advantages over other acquisition devices,

because it is easy-to-use tool, for a commercial use and is close to the speaker independence.

This paper presents the results of a recognition algorithm based on neural networks, capable of recognizing patterns from the silent speech of the words 'One', 'Two' and 'Three' of Spanish language, using feature extraction techniques as Mel frequency cepstral coefficients (MFCC) and Dynamic Time Warping (DTW), obtaining as best result a classification and identification average rate of 86% for a neural network with 13 neurons in the hidden layer, trained with data obtained from 9 subjects.

The second part of this document has the description of the materials and methods used for the development of this research, explaining the individual and combined importance of each one. In the third section are summarizes and analyzes the results obtained from the artificial intelligence algorithm trained with the data generated by the feature extraction techniques. Finally, the conclusions reached in the development of this work are discussed, suggesting a future perspective for further progress in the investigation of this kind of communication systems.

2. MATERIALS AND METHODS

2.1 Materials

2.1.1 NAM Microphone

It is a capture device principally designed for silent speech applications, is composed by an electret microphone covered with a soft silicone layer in order to emulate the mechanical properties of the skin surface and to acquire signals produced by vibration of the vocal cords, just before any sound occurs in the speech apparatus [12].

The best place for the location of this sensor is below the ear, right at the bottom of the mastoid bone as is depicted in Figure-1, because this position is convenient to record the various signals produced by the use of normal or silent speech.

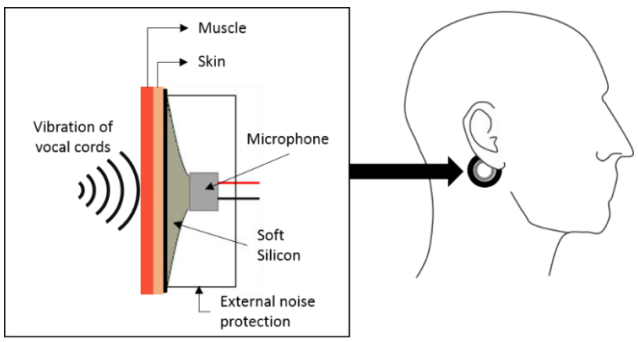

Figure-1. Structure and placement of NAM microphone.

2.1.2 Software

To implement the feature extraction techniques and training neural networks, the software MATLAB from MathWorks, in its version 8.3 was used.

2.2 Methods

2.2.1 Acquisition

5 samples of each of the three phonological units that wanted to be identified, in a population of 9 test subjects of male gender were taken, giving a total of 135 samples. The sensor for capturing the silent speech signals was a NAM microphone designed for that purpose, as well as an acquisition card for recording audio signals. Figure-2 shows one of the signals acquired from the murmur of 'Dos'.

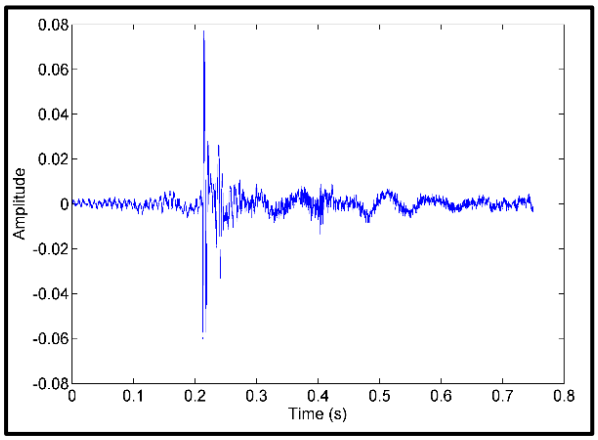

Figure-2. Silent signal acquired from the pronunciation of 'Dos'.

2.2.2 Mel Frequency Cepstral Coefficients (MFCCs)

The MFCCs are a representation of audio signals, specifically the sounds that human hearing can perceive, and are relevant due to the need of ASR systems of having a robust feature extraction technique for not to be affected by factors like background noise and emotions of the speaker, since the emotions affects the pitch and volume of voice sounds. Commonly are calculated having into account the processes shown in the Figure-3.

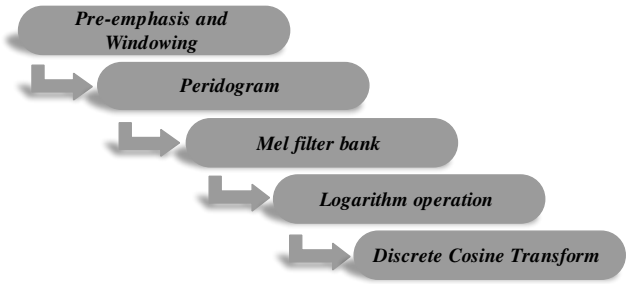

Figure-3. Processes involved in the computation of MFCC.

The first step to implement the MFCCs algorithm, are the pre-emphasis and windowing. The pre-emphasis improves the magnitude of high frequency components for turning this parameter into relevant information for the next stages of processing, and the windowing operation allows a detailed analysis by

segmenting the signal in different sections of equal size, with a width of 25 ms and a displacement of 10 ms.

After windowing operation, the periodogram was realized for finding the energy in each frequency band to obtain a data vector with energy values. With the filter bank, the vector information was debugged, then was applied a logarithm operation and the Discrete Cosine Transform to obtain the final value of the coefficients.

As a final result of this stage, 39 MFCCs for each window were obtained which were used to perform DTW technique.

2.2.3 Dynamic time warping (DTW)

DTW is a technique for analyzing time series to quantify the difference among two or more signals that may or may not vary in length and speed. This technique is useful because in practice two voice signals never have the same speed (different pronunciation) and usually have different lengths.

The DTW algorithm is performed in three stages, first executes a two-dimensional mapping of the Euclidean distances from point to point of both signals, in the second, the cost matrix is computed in terms of the distances obtained previously, and finally performs the search of the optimal path that gives the shortest distance between the two signals and allows comparison. An example of algorithm is depicted in Figure-3.

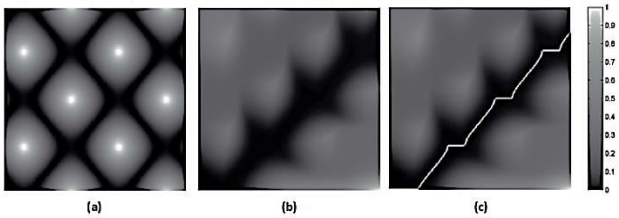

Figure-4. Dynamic time warping algorithm (a) Signal distances matrix, (b) Cost matrix (c) Optimal warp path.

In Figure-4 is depicted another representation of DTW results, this consists on to plot the distances of the optimal path calculated by the algorithm.

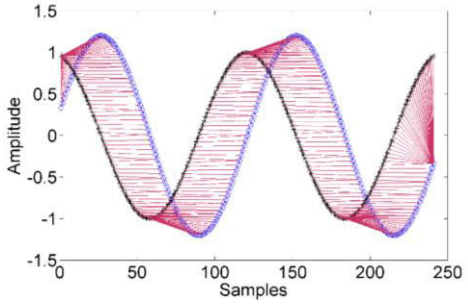

Figure-5. Warp visualization.

2.2.4 Pattern recognition

Pattern recognition is a concept from the theory related to artificial intelligence and machine learning, which is responsible for classifying and identifying common behaviors in a data sets. This kind of systems, discriminate characteristics and patterns extracted from events, objects, people, or as in this case signals of silent speech.

The recognition algorithm used in this paper is made with neural networks, for which was necessary a set with all patterns to identify properly labeled. Two factors that have effect on the neural network performance are the topology and the stop criteria for training process.

For the pattern recognition system was used a Feed-forward setup, it is made by three layers, the first one is the input layer, the second one is the hidden layer that has in charge all the data processing, and finally the output layer adjusts the values according to the labels or groups previously defined. The chosen function for the hidden layer was the sigmoid tangent, and for the output layer was the normalized exponential or softmax, which is expressed in the equation 1.

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

Where,
n =Pattern value
i =Pattern index
k =Total number of patterns

The algorithm used for training the net was the backpropagation, and the Cross Entropy Error (CEE) was chosen as stop criteria. The formula of the CEE is shown in the equation 2.

$$F(a,b) = -\sum_{i}^{N} a(i) \log b(i) \quad (2)$$

Where,
a =Desired output
b =Obtained output
i =Pattern index
N =Total of patterns

3. RESULTS

For the study case presented in this paper, an algorithm for the recognition of the Spanish language phonological units, more specifically 'One', 'Two' and 'Three' was developed. To achieve this, data of 9 test subjects were used, all of the male gender. For each subject were recorded 5 samples of each word through a NAM microphone.

The acquired data group was used as the main parameter for the processing stage, in which MFFCs and calculating distances through DTW were used. Finally, the data obtained from the extraction and processing stage were considered for training the neural network that will be in charge of identifying the patterns found.

As a first result the MFCCs were obtained, as it can be seen in the Figure-6.

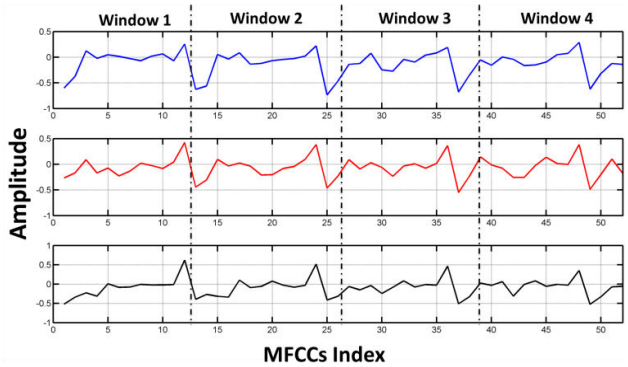

Figure-6. MFCCs for the first four windows of the words *'Uno'*, *'Dos'* and *'Tres'*.

As mentioned previously, each window provided 39 coefficients, of which 13 are known as static, 13 as velocity coefficients and 13 as coefficients of acceleration. In Figure-4 are only shown the statics coefficients for the first four windows, and at first sight can be noticed variations among different signals. But due to this method includes features from the timing of the signals, it is affected by the displacements of these, so the use of another method for comparing signals with different length and speed as the DTW is necessary. Table-1 has the data related to the computation of DTW for the three signals and its templates.

Table-1. Computation of DTW among three signals and its respective templates.

MFCCs Templates		DWT Distances		
		UNO'	DOS'	TRES'
Static	'UNO'	84.1195	74.3330	81.1985
	'DOS'	93.3664	73.8425	86.9987
	'TRES'	97.1829	82.8945	79.5458
Velocity	'UNO'	512.1436	567.2193	585.5249
	'DOS'	540.8464	514.9834	582.5394
	'TRES'	609.3000	584.7396	573.5895
Acceleration	'UNO'	599.1794	687.9006	658.8242
	'DOS'	633.6743	671.0908	694.3006
	'TRES'	647.4758	692.1127	667.0850

The data shown in Table-1 were used for training the neural network because described better the behavior of silent speech signals, regarding the stored templates and also reduces the dimensionality of the data to be used in the training.

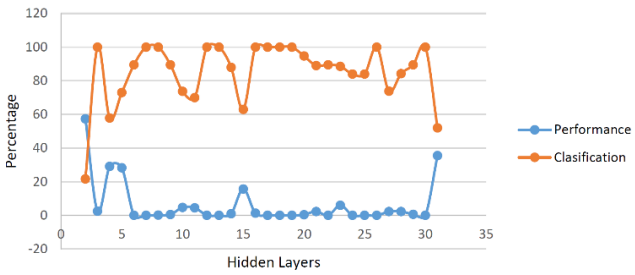

Figure-7. Neural Network error in terms of the number of hidden layers.

According to the data shown in Figure-7, there are three values that have a low training error and a significant percentage of classification; these are 7, 13 and 26 neurons. Tests were conducted with the three networks, obtaining the values specified in Table-2.

Table-2. Percentage of recognition regarding to the numbers of hidden layers.

Number of hidden layers	Percentage of Recognition		
	UNO	DOS	TRES
7	90	80	70
13	90	95	75
26	90	90	75

Considering the data provided in Table-2, the number of neurons chosen was 13, since for this value the average recognition exceeds 85%, which is higher than the other two values that were taken into consideration.

4. CONCLUSIONS AND FUTURE PERSPECTIVES

The MFCCs are a feature extraction technique that provides representative information of the silent speech signals, also demonstrated its ability to analyze common and measurable changes between different signals of this type. Although its implementation was limited to 13 coefficients per window, these data were enough to achieve the identification of patterns related to silent speech from the words 'Uno', 'Dos' y 'Tres'.

Using a complementary technique for Mel coefficients as DTW proved to be a determining factor for the correct identification of the desired words, since this algorithm considers the temporal changes in the signals, however, it requires templates that represent accurately the group or groups that want to identify.

The performance of neural networks in classification tasks depends largely on the methods used to extract patterns, this is because if the extraction delimits enough the features of the signals, the performance of the network will be more effective in terms of processing and recognition.

Identifying patterns related to silent speech is complicated due to the nature of the signals as they do not have the same amplitude that signals produced by normal

speech, also because of the construction of the sensor, external noise affected the measures of signals.

As future perspective, the improvement of the sensor is proposed, as well as the methodology used in the processing stages by including an embedded system that allows algorithm portability.

ACKNOWLEDGEMENTS

Special thanks to the Research Vice Chancellorship of the "Universidad Militar Nueva Granada", for financing the project ING/INV 1762 titled "Dispositivoreproductor de voz del lenguajeespañol a través de hablasubvocal e interfazcerebro-computador" project, 2015 year.

REFERENCES

[1] H. D. Lasswell. 1948. The structure and function of communication in society. Commun. Ideas. (1948): 37-52.

[2] J. Knox. 2007. Visual-verbal communication on online newspaper home pages. Vis. Commun. 6(1): 19-53.

[3] M. Knapp, J. Hall and T. Horgan. 2013. Nonverbal Communication in Human Interaction.

[4] W. J. M. Levelt. 1999. Models of word production. Trends Cogn. Sci. 3(6): 223-232.

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

[6] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert and J. S. Brumberg. 2010. Silent speech interfaces. Speech Commun. 52(4): 270-287.

[7] M. Wand, M. Janke and T. Schultz. 2014. Tackling Speaking Mode Varieties in EMG-Based Speech Recognition. Biomed. Eng. IEEE.

[8] L. Mendoza and J. Peña. 2013. Procesamiento de Señales Provenientes del Habla Subvocal usando Wavelet Packet y Redes Neuronales.

[9] S. A. Patil and J. H. L. Hansen. 2010. The physiological microphone (PMIC): A competitive alternative for speaker assessment in stress detection and speaker verification. Speech Commun. 52(4): 327-340.

[10] T. F. Quatieri, K. Brady, D. Messing, J. P. Campbell, W. M. Campbell, M. S. Brandstein, C. J. Weinstein, J. D. Tardelli, and P. D. Gatewood. 2006. Exploiting Nonacoustic Sensors for Speech Encoding. IEEE Trans. Audio, Speech Lang. Process. 14(2): 533-544.

[11] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone. 2000. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. Speech Commun. 52(4): 288-300.

[12] Y. Nakajima, H. Kashioka, K. Shikano and N. Campbell. 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. In 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 5: V-708-11.

[13] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima and K. Shikano. 2010. Silent-speech enhancement using body-conducted vocal-tract resonance signals. Speech Commun. 52(4): 301-313.

[14] V.-A. Tran, G. Bailly, H. Lœvenbruck and T. Toda. 2010. Improvement to a NAM-captured whisper-to-speech system. Speech Commun. 52(4): 314-326.

[15] O. Sandoval, E. Melo, and D. Hurtado. 2015. Revisión de las tecnologías y aplicaciones del habla sub-vocal. Ingeniería.