Research Journal of Applied Sciences 13 (11): 653-662, 2018
ISSN: 1815-932X

System for Acquisition and Conditioning of Non-Audible Murmur Signals

Juan Clavijo, Olga Ramos and Dario Amaya
Universidad Militar Nueva Granada, Bogota, Colombia

Abstract: To achieve recognition of silent speech, especially of the non-audible murmur this development raises as a first step, the acquisition and transmission of vibration data from the sub vocal speech. Because of the stochastic characteristics of an audio signal in the time domain this is treated and analyzed in the complex frequency domain. To achieve the frequency analysis is implemented in this development a Fast Fourier Transform (FFT). The data acquired and transformed with FFT are transmitted via a Wi-Fi network to a computer for further analysis of signal and image processing through an application developed in C # language.

Key words: Sub-vocal speech, silent speech, non-audible-murmur, fast Fourier transform, vibration data, stochastic

INTRODUCTION

The signal acquisition stage is an indispensable part of many electronic systems which interact with analog signals. As part of these kind of systems, methods for automatic speech recognition have been extensively studied in recent years. The technological advances that have been taking place, since, the last century also provided new concepts and tools in order to improve the human-machine interaction.

Silent Speech Interfaces (SSI) are defined as systems of interaction that enable oral communication when acoustical signals are not available (Denby *et al*., 2010). Thus, SSI systems would allow speech disabled persons to communicate without restrictions or provide an alternative form to have conversations in noisy places ensuring the message integrity or in the case where exists privacy limitations could be a choice to communicate without restrictions in public places (Heracleous *et al*., 2003).

One of the most important methods catalogued as a SSI system is the detection of Non-Audible Murmur (NAM). As the name implies, the NAM can be described as the production of small whispers to be heard by anyone but the person that produces it (Nakajima *et al*., 2003).

As a first step to develop a system for silent speech recognition that uses the NAM technique was proposed the design of an electronic system to acquire the NAM signals from a modified microphone for this purpose (Nakajima *et al*., 2003).

For the physical signal acquisition of sub-vocal speech, a non-audible murmur microphone is used which is build based on the research work presented in (Nakajima *et al*., 2003; Heracleous *et al*., 2003; Toda *et al*., 2009).

The processing steps of NAM signals to develop silent speech interfaces is based on the theory of voiced-speech for the construction of Automatic Speech Recognition systems (ASR). Hence, the processing steps are applicable for NAM recognition with certain modifications.

The construction of ASR systems, takes into account 3 major phases which are related to the extraction and analysis of the characteristics of the speech signals the classification and pattern recognition and the utterance verification of words recognized by the system (Douglas, 2003; Ishii *et al*., 2011).

Since, it is intended to make an approach to the development of an ASR system applied to non-audible murmur it is used as first phase implementation of algorithms based on frequency representations for the characterization of this class of signals. The research presented by Ishii *et al*. (2011) and Toda *et al*. (2012) show the importance of implementing Fourier analysis systems in sub vocal speech.

According to, the ideas presented above in this study we describe the development of a set of algorithms for extracting implicit characteristics in NAM signals to be used as recognizable elements in subsequent pattern recognition tasks for identification of specific phonological units (Rabiner and Schafer, 2011).

This document is organized as follows; In the second part, a description of the acquisition and signal pre-processing task forsubvocal speech is made. The third part, describes in detail the feature extraction task for data processing and subsequent clustering for possible applications in the construction of a silent speech ASR system. Finally, a description of the processing tool is made and the conclusions and future research are condensed.

Corresponding Author: Juan Clavijo, Universidad Militar Nueva Granada, Bogota, Colombia

MATERIALS AND METHODS

Speech processing: Speech processing systems have been studied for a long time. The methodology used for the construction of speech processing systems is based on the classical theory of signal processing which exploits the frequency analysis as the main analysis tool. Implementations related to classical speech processing, can be applied to silent speech signals as shown in (Babani *et al*., 2011; Fan and Hansen, 2011; Denby *et al*., 2010). These conventional systems represent a start point to base the development of silent speech interfaces.

The state of the art related to the implementation of voice signals processing, shows that Fourier analysis is the main tool used in this kind of systems. For this reason the development of an acquisition and processing system based on the frequency representations of signals is proposed. Figure 1 shows the general diagram for the acquisition and signal processing of the sub-vocal speech task.

Figure 1 are defined 3 important processes that conformthe signal characterization steps for silent speech. The first stage is referred to the transformation of the acoustical silent speech signal into atreatable variable in voltage terms this is performed using a, specially adapted microphone which enhance the acquisition of silent speech signals this is called NAM microphone.

In the second stage we have the process of the ADC conversion and calculation of the frequency components of the signal. To do this a dsPIC is used to sendthe processed data to a computer via. the IEEE 802.11b protocol. The frequency representation of the signal calculated by the dsPIC, enters to the third stage which is responsible for conducting the treatment of the data to finally allow the user to graphically verify the results of the feature extraction and calculation.

Acquisition system and pre-processing: As indicated above the first two stages summarized in Fig. 1 correspond to the acquisition and pre-processing of the signal through a NAM microphone and a dsPIC, respectively.

NAM microphone: The acquisition of silent speech is a big challenge in order to build ASR systems for unvoiced speech. The NAM microphones were developed as an alternative to acquire accurately silent signals. Hereby, NAM microphones are the result of multiple research works (Heracleous *et al*., 2003; Nakajima *et al*., 2003; Toda *et al*., 2009). These works defined a modified structure of a stethoscope using an electret microphone and its diaphragm. In order to make a system for acquiring and processing silent signals, we proposed the development of a prototype of the NAM microphone. To do this a pediatric stethoscope and an electret microphone were used. The configuration and modification of the stethoscope are depicted in Fig. 2.

After construction of the microphone, a signal conditioning stage was made to standardize the output voltage values of the microphone within the operating range of the ADC module used by the dsPIC. This conditioning step was performed using a simple amplification, constructed from a transistor configured as an amplifier mode.

FFT calculation: The first processing stage of the silent signal acquired through the NAM microphone is performed using a Digital Signal Controller (DSC). The device used is a dsPIC30F4013 that has a 2 kB RAM memory and a 48 kB program memory. Its maximum operating speed is 128 MHz with a maximum of 32×10^6 instructions per second.

The DSC used is employed for digital conversion of the analog signals from the NAM microphone. A 12 bit resolution and a sampling frequency of 2 kHz is reached. The sampled data are storedin the RAM of the controller to then calculate a 128 samples FFT. This means that the FFT is performed every 64 msec to generate a FFT at a

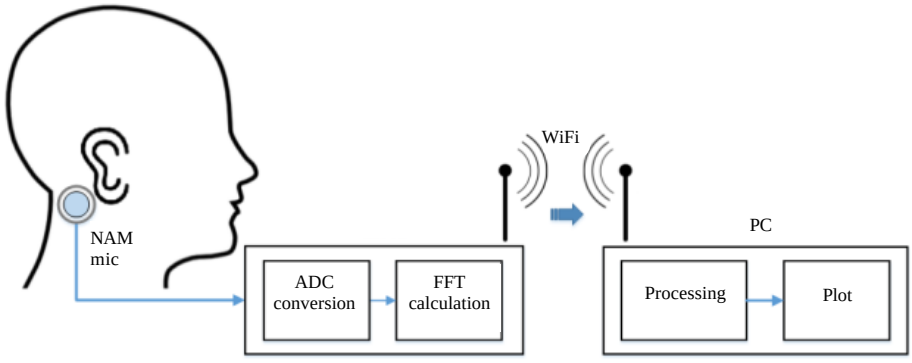

Fig. 1: General diagram for acquisition and processing

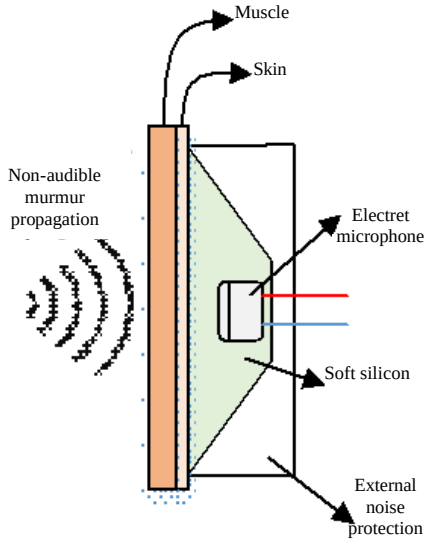

Fig. 2: NAM microphone structure

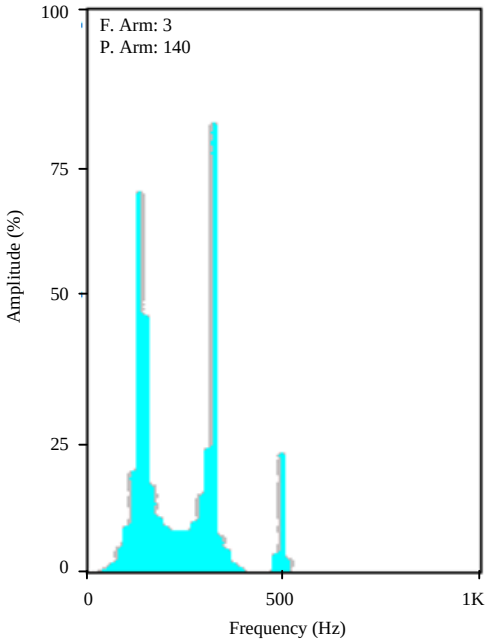

Fig. 3: Magnitude spectrum (FFT)

frequency of 15.625 Hz (Proakis and Manolakis, 1996). Calculating the fast Fourier transform is performed using the Cooley-Tukey algorithm and is embedded in the dsPIC using the definition summarized in Eq. 1 and the optimization concepts presented in:

$$X(j) = \frac{1}{N}\sum_{k=0}^{N-1} x(k) e^{\frac{i2\pi jk}{N}} \quad (1)$$

The transformation of the 128 samples, generates a bandwidth between 0 and 1 kHz with a total of 64 bands. This relationship indicates that each value calculated by the FFT is separated each 15.625 Hz in the bandwidth making this value the minimum measurable frequency for the FFT.

The resulted spectrum is then transmitted through one of the UART ports of the dsPIC at a rate of 115200 bps. The information of each FFT calculation is packaged in blocks of 66 bytes and then transmitted. Wireless transmission of samples is implemented with a Wi-Fi connection to a computer using a S6 XBee node. The XBee module is configured with a specific port and IP address to manage the Wi-Fi network. The XBee module has a maximum range of 30 m which represents portability and convenience in the acquisition task.

Signal post-processing: Following the methodology for developingthe silent speech processing system which is summarized in Fig. 1 is proceeded with the construction of the interface for visualization of the spectrum data supplied by the dsPIC. This is solved designing auser interface application in C # language. The resulted post-processing tool allows to read and store data transmitted by the DSC, allocating them in memory spaces in order to achieve a real-time processing of information displaying the result of the magnitude of the FFT and the corresponding spectrogram. A first view of the FFT calculation is depicted in Fig. 3.

Figure 3 illustrates an instant sample of the calculated FFT, however, the depicted data is not clear about the temporal information associated to the frequency, making it necessary to complement with a sequence of readings of the FFT. A possible approximation to resolve this issue is the use of the concept of the Short-Time Fourier Transform (STFT) also known as the spectrogram of a signal. The STFT is a graphical way to show the spectral trends of a signal (in this case) of sub-vocal speech. The spectrogram is a 3 dimensional graph in which the x-axis is the time axis, the y-axis represents the frequency spectrum and in its last axis is condensed the power of each spectral component, represented by color intensity in grayscale (Oppenheim, 1970). The graph shown in Fig. 4 shows the spectrogram calculated using the user interface for the phrase “Military University” in terms of silent speech.

As a complementary form of study, the software performs other calculations and estimations of parameters for creating features to be used in a posterior stage for identification that may be applied to speech coding, speaker recognition, automatic speech recognition, among other kind of speech applications.

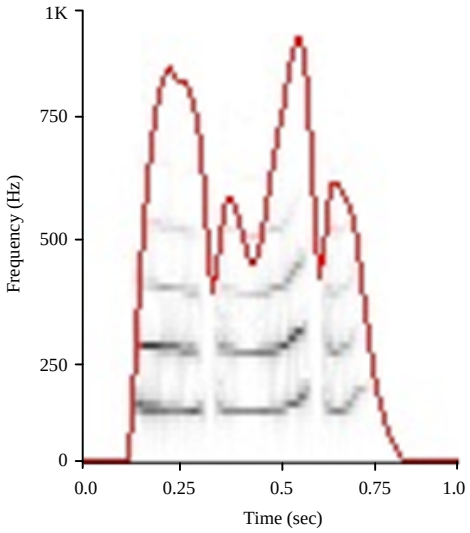

Fig. 4: Spectrogram

The smoothed line in Fig. 4 represents the spectrum envelope of the signal in the frequency domain. The envelope is the average value of the magnitudes of the spectrum, condensed through the FFT either for each instant of time or for each of the samples.

RESULTS AND DISCUSSION

Parameters calculation: One of the most notable difficulties of identifying patterns is to achieve that each pattern be comparable repeatedly. In the case of speech this is an important feature to take into account given that the pronunciation of the letters, syllables or phonemes generally vary from performance time, power or volume. This the reason to implement techniques to try to standardize and simplify the patterns in as many features as possible.

In order to enhance the feature extraction process for frequency representations of speech the software implements the following strategies:

- FFT windowing
- FFT average
- FFT derivative
- Harmonics detection
- Harmonics normalization
- Spectrogram gradient
- Spectrum binarization
- Temporal normalization

To implement suitable algorithms to perform the calculation of the parameters listed above is used an initial sample spectra set of values, allocated by a sample vector which comprises 64 instantaneous values of power of the FFT send by the dsPIC. Each of the values of the power spectrumis distanced from each other 15,625 Hz. The container vector of such information is defined as shown in Eq. 2:

$$fft^*[kT]=\begin{bmatrix} p^*_{1k} \\ p^*_{2k} \\ . \\ p^*_{bk} \\ . \\ p^*_{64k} \end{bmatrix} \quad (2)$$

Where:

fft*[kt] = Vector, groups each of the FFT samples
p* = Represents the power value of the b band on the k sample that is acquired at aT time

FFT windowing: The characteristics of the Fourier transform, make the first components related to the lower frequencies of the signals to have large values with respect to high frequencies. This particularity is due to the representation of the zero levels in the signal when no dynamic behavior exists on this (Jurafsky and Martin, 2000). The condition explained above, requires to find a method to minimize the large components of the spectrum that are not part of the dynamic behavior of the speech signals.

Windowing is the operation used to normalize the high frequency data that is found in the signal. For the application presented in this study, a linear window was used and is defined as presented in Eq. 3:

$$w[b]=\begin{bmatrix} c_1 \\ c_2 \\ . \\ c_b \\ . \\ c_{64} \end{bmatrix} \quad (3)$$

Where:

$c_i = 3\ (i-1)/20$, for $i \leq 7$
$c_i = 1$, for $i>7$

The window defined in Eq. 3 mitigates inversely proportional the first seven spectrum bands. The implementation of the window is performed by multiplying term by term the window coefficients and each of the values of the FFT samples. This operation raises a new definition of normalized vector as shown in Eq. 4:

$$fft[kT]=\begin{bmatrix} c_1 p^*_{1k} \\ c_2 p^*_{2k} \\ . \\ c_b p^*_{bk} \\ . \\ c_{64} p^*_{64k} \end{bmatrix} \quad (4)$$

The spectral differences between a windowed FFT and the FFT without treatment is depicted in Fig. 5.

Spectrum average: The data of each new sample FFT is averaged to generate a new statistical and pattern feature. This new data is annexed to the vector as one raw element. The definition of the vector is shown in Eq. 5:

$$fft[kT]=\begin{bmatrix} p_1 \\ p_2 \\ p_{bk} \\ . \\ p_{64k} \\ pr_k=\left(\sum_{i=1}^{64} p_{ik}\right)/64 \end{bmatrix} \quad (5)$$

Sub Vocal Speech Pattern (PHSV): The Sub-Vocal Speech Pattern (PHSV) is formed by a finite number of FFT samples which are set up as a matrix of 65 rows and L columns, corresponding to the length of the pattern data:

$$PHVS=\begin{bmatrix} p_{1k} & p_{1(k-1)} & \cdots & p_{1(k-L)} \\ p_{2k} & p_{2(k-1)} & \cdots & p_{2(k-L)} \\ \vdots & \vdots & \ddots & \vdots \\ pr_k & pr_{(k-1)} & \cdots & pr_{(k-L)} \end{bmatrix} \quad (6)$$

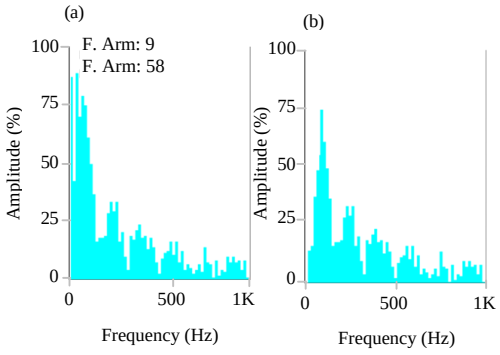

Fig. 5: Magnitude spectrum; a) FFT without treatment and b) Windowed FFT

Under the PHSV concept is possible to represent the silent speech signals features as an array with the windowed spectral components and the phoneme duration. The graphical representation of a specific example with a pattern is depicted in Fig. 4. Where the spectrogram represents a pattern of the silent speech.

Derivative of the FFT: The derivative of the FFT is calculated by a simple method that determines the algebraic difference between samples this concept is described by the following relation Eq. 7:

$$\frac{d(fft[kT])}{T}=\begin{bmatrix} p_{1k}-p_{2k} \\ p_{2k}-p_{3k} \\ p_{bk}-p_{(b+1)k} \\ . \\ . \\ . \end{bmatrix} \quad (7)$$

For practical purposes in the calculation task, the value of the period (T) of the signal has the value of 1. This assumption does not alter the proportions of the derivative.

Graphically, the derivative of the FFT is depicted in Fig. 6 wherein the yellow curve represents the derivative of the FFT.

The b points defined in Eq. 7 correspond to the values wherein the derivative has zero crossing, causing a change from a positive value to a negative value and

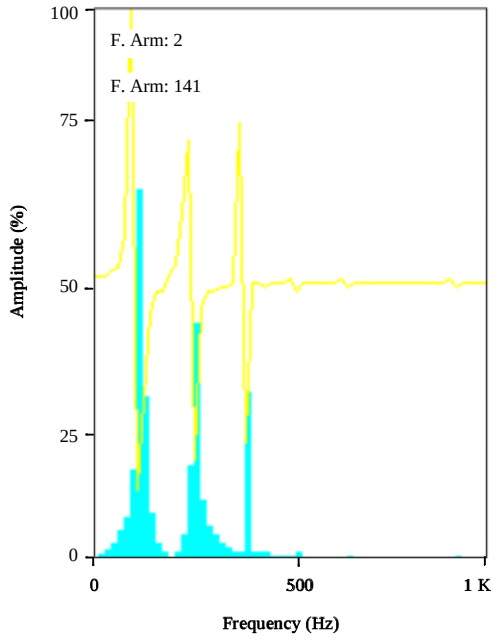

Fig. 6: Derivative of the FFT

determining harmonic components in the signal. Conveniently small changes in the results of the derivative, allow to identify false harmonic frequencies whose amplitudes are lower than the true harmonic magnitudes. To avoid this undesirable situation, the algorithm makes an inspection of the calculated derivative and the harmonic values to rejectthe false data located far from the bands with high magnitude. This process is performed iteratively on the calculated vector, evaluating harmonic frequencies neighborhoods for each spectral point.

Harmonics detection: The harmonic detection task, aims to identify the fundamental frequency and remove similar information of the spectrum, since, during the execution of the speech, the response of the FFT results in 64 bands but the number of harmonics may range from 3-6 as observed in the experimentation.

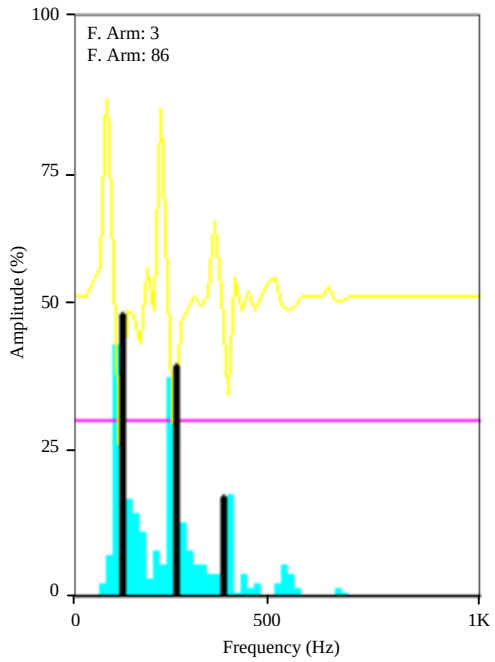

Fig. 7: Detection of harmonics

To identify the harmonics in each signal sample, the FFT is assumed as a sequence of values that are part of the signal. This condition makes the search of the harmonics becomes the search for maximum values of this signal. Given the basic principles of optimization that allow to find the maximum and minimum points of a function through its derivative, the search of the highest value is performed by taking the derivative of the signal and identifying the zero crossing of the derivative. This mathematical action becomes an equivalent form of the derivative equal to zero.

Through the optimization calculations mentioned above the algorithm can detect the specific points or bands where the harmonics are present. This feature extraction method allows the harmonics to be used as features in future identification architecture. A graphical example with the identification of the harmonics of a signal is shown in Fig. 7.

In Fig. 7, the harmonics are highlighted with black bars on the spectrum. The harmonics with lower amplitudes canbe discarded by a threshold decision parameter set by the user in the interface. In the same case of study, summarized by Fig. 7 the average power of all bands is 86 and represents a characteristic of the signal. The value is represented in the Fig. 7 as the horizontal purple line.

The calculations and features mentioned at this point are continuously performed every 64 msec in order to create the sequence of spectrums to construct patterns. A continuous acquisition of sub vocal speech is displayed in the spectrogram depicted in Fig. 8.

Based on the results presented in Fig. 8 the sub-vocal speech patterns can be represented as grayscale images. For a future comparison of patterns and its identification it is necessary that the characteristics of the patterns to be similar, i.e., whose dimensions are similar in the same manner that its contrast.

Spectrogram gradient: The interface allows to manipulate the directional derivation on the pattern image. This is achieved by calculating the gradient vector at each of the pixels in the spectrogram image for each of the signals.

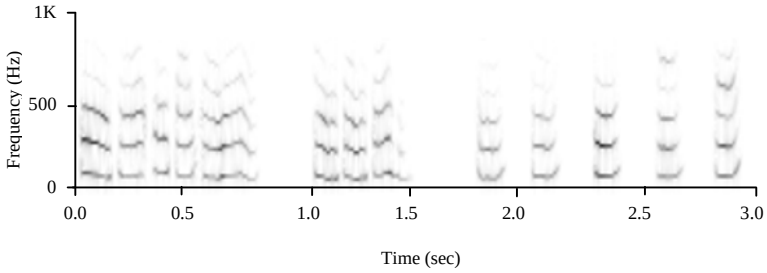

Fig. 8: Continuous spectrum

The implementation of this method allows to highlight variations in a pattern independently of its size or other characteristics. Any way to complement the gradient operation on the images, the patterns are normalized using contrast correction. In Eq. 8 the implemented mathematical structure to calculate the gradient is shown:

$$\nabla phsv(k,b)=\begin{bmatrix}\frac{\partial phsv}{\partial k}\\ \frac{\partial phsv}{\partial b}\end{bmatrix}=\begin{bmatrix}Gk\\ Gb\end{bmatrix}$$

$$G=\sqrt{Gk^2+Gb^2}$$

$$Gk=(z_1+2z_4+z_7)-(z_3+2z_6+z_9)$$

$$Gb=(z_7+2z_8+z_9)-(z_1+2z_2+z_3)$$

Donde,

$$z_1=phsv(k-1,b-1)$$
$$z_2=phsv(k,b-1)$$
$$z_3=phsv(k+1,b-1)$$
$$z_4=phsv(k-1,b)$$
$$z_6=phsv(k+1,b)$$
$$z_7=phsv(k-1,b+1)$$
$$z_8=phsv(k,b+1)$$
$$z_9=phsv(k+1,b+1) \quad (8)$$

Significantly, the new gradient image is implemented with the magnitude of the gradient vector in each of the pixels (Gonzalez and Woods, 2007). Figure 9 shows the visual representation of the result of the gradient operation on the spectrogram of the "Military University" phrase.

Image contrast: The contrast of an image is a parameter that determines what proportion of the color depth is taking advantage. In the specific case of this development every pixel on the spectrogram is encoded in a gray scale with possible values from 0-255, where 0 represents absolute black and 255 absolute white color. An image with high contrast or maximum contrast has content in their pixels values distributed in all possible values of gray while a low-contrast image occupies only a portion of levels, discarding the other possibilities (Gonzalez and Woods, 2007).

One same pattern can be represented by 2 images with different contrasts. This difference of contrasts makes difficult to correctly identify the features this can be explained as when we calculate similar vectors but with different magnitude values.

Figure 10 presents 2 patterns representing the phrase in sub-vocal speech, "Military University" with different contrasts, the audio level have a correspondence with the speaker volume.

The manipulation of the image contrast of the spectrogram becomes a normalization of the volume of the features this modification allows the patterns to be similar in terms of this variable.

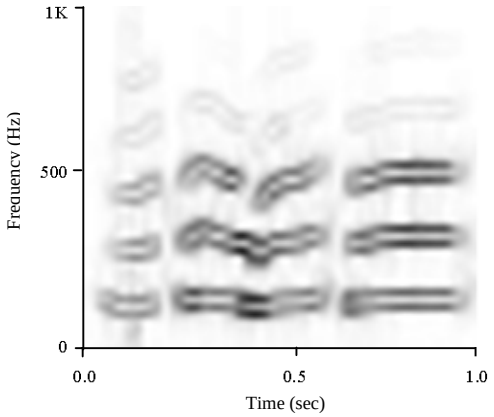

Fig. 9: Gradient of a pattern

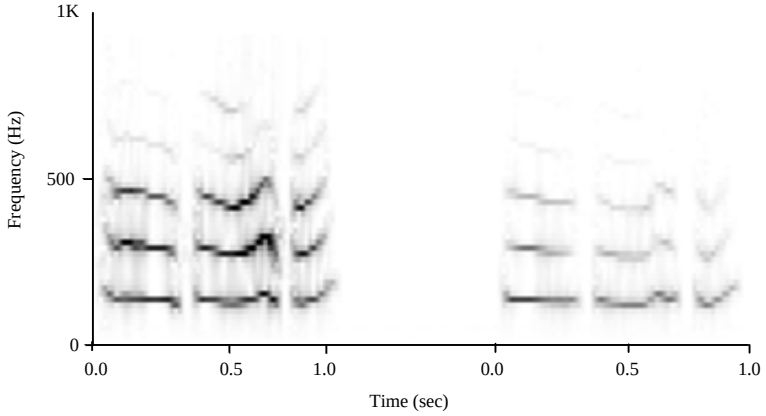

Fig. 10: Contrast comparison

In Eq. 9 is presented the mathematical relationship that can correct the contrast according to the image recorded in the matrix, thus, the new array is the normalized volume pattern matrix.

$$phsv^{*}(k,b)=\left(\frac{255}{phsv_{max}-phsv_{min}}\right)\left(phsv(k,b)-phsv_{min}\right) \quad (9)$$

In Eq. 9 the phsv max parameter is the highest value of the pattern and phsv min the corresponding minimum value (Gonzalez and Woods, 2007).

Harmonics normalization: One of the most important characteristics of the speech patterns is the fundamental frequency. This is defined as the first harmonic value in frequency of the spectrum. In speech signals, the position of the first harmonic in the spectrum varies depending on various characteristics such as age of the person, gender, mood, state of stress or relaxation, etc. However, the differences between harmonics tend to be the same, meaning that a normalization of the patterns according to the first harmonic is useful to increase the similarity between patterns.

To achieve this purpose to normalize patterns in this development, the position corresponding to the first harmonic is assumed as reference and the other frequency bands are moved respect to the first harmonic. This process is performed using the Eq. 10:

$$phsv^{**}(k,b)=phsv(k,b-b_0) \quad (10)$$

where, phsv **represents the normalization of the first harmonic or fundamental frequency pattern. In the mathematical relationship it can be seen that this new pattern is merely a displacement which is calculated using the high amplitude values in the search of the maxims (Gonzalez and Woods, 2007).

Figure 11 shows the results of the normalized sample (Fig. 11b) and the original sample (Fig. 11a), taking as parameter the first harmonic of the phrase "Military University".

Binarization of the spectrum: In order to simplify the number of features to calculate the user can apply the operation of binarization of the spectrum. This option sets a threshold within the range of the possible values of the image generated from the spectrum when a pixel value exceeds the threshold magnitude the algorithm sets the color value of the pixel to the absolute black and otherwise to the absolute white. Figure 12 depicts the main differences between a binarized spectrum (Fig. 12b) and an original spectrum (Fig. 12a).

Temporal normalization: The pronunciation of phonemes when the speech is performed is not the same all the times and the duration of the phonological units is short. The length of a sentence can change in a matter of fractions of a second but in a digital system for processing data this condition is critical, since, processors cannot differentiate durations. To overcome this draw back and to model all patterns in a similar way the length of the patterns is fixed in a number of samples for this case 64 samples were chosen in order establish a square matrix of 64×64 for the image as normalized standard pattern. If the pattern is longer than 64 samples, the data is compressed and in the contrary case is expanded. The changes between expansion and compression are performed by a rule of 3.

User interface: The processing options described above are integrated in a unique application software developed in C # language. The resulted interface allows to connect the XBee device through a UDP network to obtain the FFT samples.

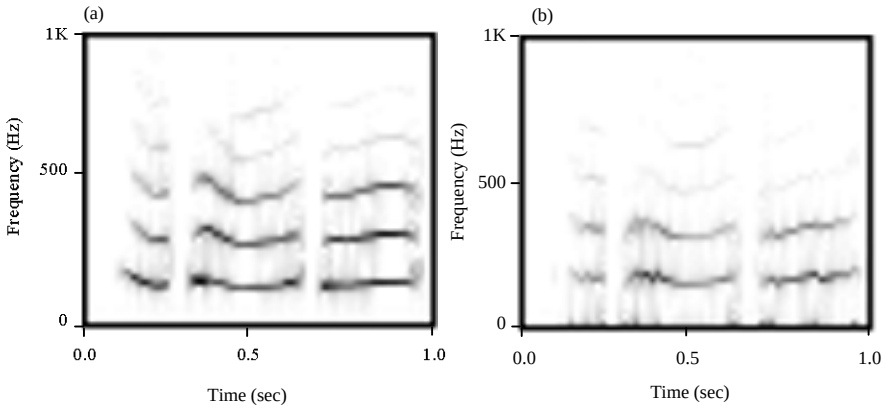

Fig. 11: Harmonics normalization; a) Original spectrogram and b) Normalized spectrogram

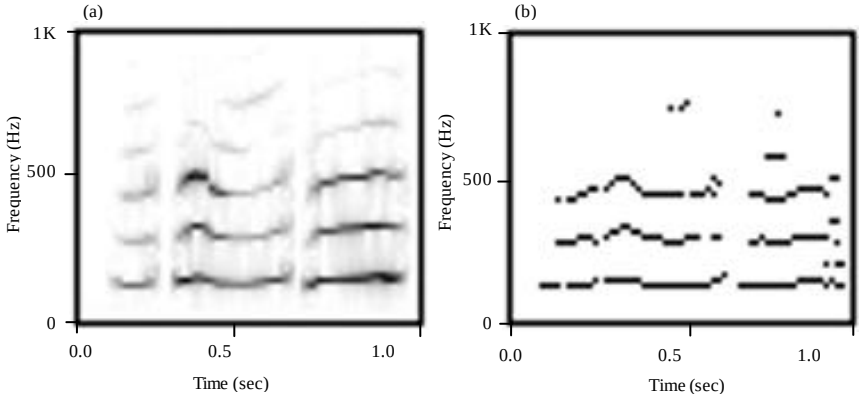

Fig. 12: Binarization of patterns

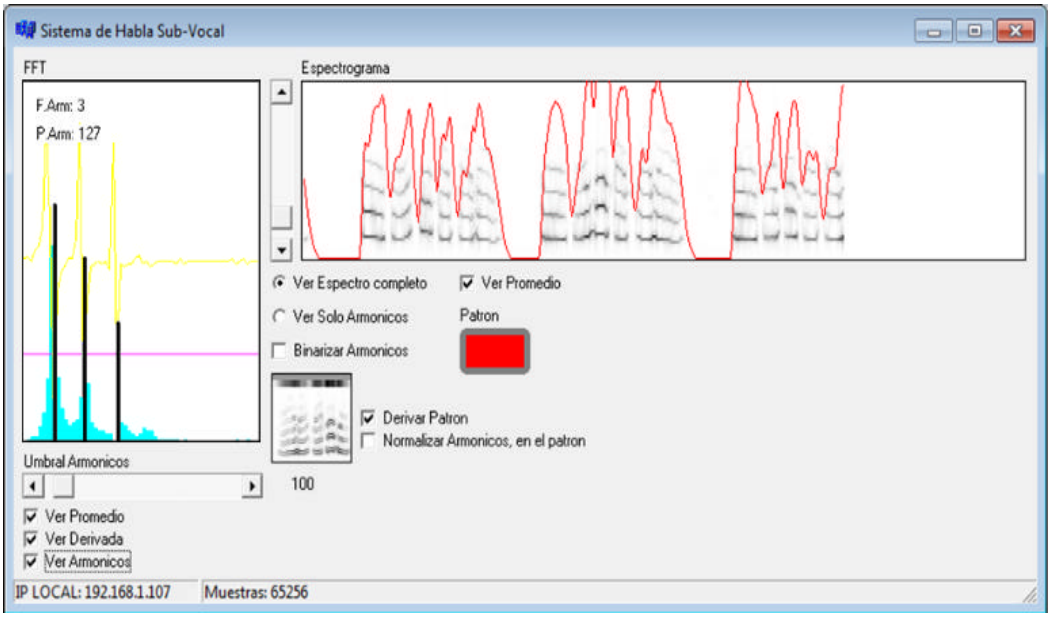

Fig. 13: User interface

The interface shows the FFT samples acquired from the dsPIC, also allows to perform the calculation of the spectrogram and the features mentioned through this study, providing to the user the control of the variables and the processing characteristics of the algorithm. The algorithm is a robust tool to analyze and represent the principal frequency characteristics of the silent speech signals. A general illustration of the interface is depicted in Fig. 13.

CONCLUSION

Throughout this research, a detailed description of a set of different algorithmic tools for calculating signal features for silent speech was performed. Accordingly, it is made clear that the purpose of processing this kind of information is oriented as an aid in the task of pre-processing for use in posteriors complex systems such as the type of ASR, speech coding, speech and audio watermarking, speaker identification, among others.

The features observed in the sub vocal speech patterns have slightly noticeable differences, an example of this situation can be seen with the vocal patterns where the outcome is highly similar. To achieve differentiation in the patterns are recommended the use of phrases or words.

Aiming to improve the characteristics of patterns is possible to implement the calculation of a high resolution FFT this condition would improve the detail acquired in

each sample but in turn would require the use of a DSC or DSP with higher speed and capacity, since, the dsPIC used in this development is at its maximum capacity.

The relationships observed between the images of spectrogram and audio features can be highlighted that the contrast of the image is proportional to the volume of the audio the audio features can be normalized using gradient image and expand the contrast of the image.

The images of the spectrograms, denote curved shapes and geometric figures that can be analyzed in subsequent tasks in order to perform pattern recognition with the aid of the algorithmic tools discussed in this research.

REFERENCES

Babani, D., T. Toda, H. Saruwatari and K. Shikano, 2011. Acoustic model training for non-audible murmur recognition using transformed normal speech data. Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 22-27, 2011, IEEE, Prague, Czech Republic, ISBN:978-1-4577-0538-0, pp: 5224-5227.

Denby, B., T. Schultz, K. Honda, T. Hueber and J.M. Gilbert *et al*., 2010. Silent speech interfaces. Speech Commun., 52: 270-287.

Douglas, O., 2003. Interacting With computers by voice: Automatic speech recognition and synthesis. Proc. IEEE, 9: 1272-1305.

Fan, X. and J.H. Hansen, 2011. Speaker identification within whispered speech audio streams. IEEE. Trans. Audio Speech Lang. Proc., 19: 1408-1421.

Gonzalez, R.C. and R.E. Woods, 2007. Digital Image Processing 3rd Edn., Prentice Hall, New York, ISBN-13: 9780131687288.

Heracleous, P., Y. Nakajima, A. Lee, H. Saruwatari and K. Shikano, 2003. Accurate hidden Markov models for Non-Audible Murmur (NAM) recognition based on iterative supervised adaptation. Proceedings of the 2003 IEEE Workshop on Automatic Speech Recognition and Understanding, November 30-December 4, 2003, IEEE, St Thomas, VI, USA., ISBN:0-7803-7980-2, pp: 73-76.

Ishii, S., T. Toda, H. Saruwatari, S. Sakti and S. Nakamura, 2011. Blind noise suppression for Non-Audible murmur recognition with stereo signal processing. Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 11-15, 2011, IEEE, Waikoloa, HI, USA., ISBN: 978-1-4673-0365-1, pp: 494-499.

Jurafsky, D. and J.H. Martin, 2000. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. Prentice-Hall, New Jersey.

Nakajima, Y., H. Kashioka, K. Shikano and N. Campbell, 2003. Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin. Proceedings of the 2003 IEEE International Conference on Acoustics, Speech and Signal (ICASSP'03) Vol. 5, April 6-10, 2003, IEEE, Hong Kong, China, ISBN:0-7803-7663-3, V708-V711.

Oppenheim, A.V., 1970. Speech spectrograms using the fast fourier transform. IEEE. Spectr., 7: 57-62.

Proakis, J.G. and D.G. Manolakis, 1996. Digital Signal Processing. 3rd Edn., Upper Saddle River, New Jersey, USA.,.

Rabiner, L.R. and R.W. Schafer, 2011. Theory and Applications of Digital Speech Processing. Pearson Education, Hoboken, New Jersey, ISBN: 9780137050857, Pages: 1056.

Toda, T., K. Nakamura, H. Sekimoto and K. Shikano, 2009. Voice conversion for various types of body transmitted speech. Proceedings of the IEEE International Conference on Acoustics Speech and Signal ICASSP, April 19-24, 2009, IEEE, Taipei, Taiwan, ISBN:978-1-4244-2353-8, pp: 3601-3604.

Toda, T., M. Nakagiri and K. Shikano, 2012. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. IEEE. Trans. Audio Speech Lang. Proc., 20: 2505-2517.