International Journal of Multimedia and Ubiquitous Engineering
Vol.11, No.10 (2016), pp.349-360
http://dx.doi.org/10.14257/ijmue.2016.11.10.33

# Pre-emphasis, Windowing and Spectral Estimation of Silent Speech Signals Using Embedded Systems

Leonardo A. Góngora[1], Dario Amaya[2] and Olga L. Ramos[3]

*Mechatronics Engineering Program, Faculty of Engineering, Nueva Granada Military University, Bogotá, Colombia*
*{[1]tmp.leonardo.gongora, [2]dario.amaya, [3]olga.ramos}@unimilitar.edu.co*

***Abstract***

*The feature extraction process is the fundamental stage in the development of speech processing systems. In this paper is described the methodology and implementation of an embedded algorithm for extracting characteristic information from silent speech signals. The classical approach based on frequency representations of the signals is followed as methodology for this work. First, the acquisition stage of the silent speech signals is performed using a Non-Audible Murmur microphone and the STM32F4Discovery evaluation board. Then, the digitalized signal is filtered, segmented and normalized using the pre-emphasis and windowing steps. The magnitude spectrogram is calculated from the pre-processed signal using the Fast Fourier Transform (FFT), to finally estimate characteristic data from de silent speech signal. As result of this process, the signal characteristic parameters, defined in the frequency domain are obtained and used as elements at later stages of pattern recognition, in order to build systems of Automatic Speech Recognition (ASR), Speech Coding, Speaker Recognition, among other applications.*

***Keywords****: Pre-emphasis, Windowing, Fast Fourier Transform, Magnitude Spectrum, Silent Speech Interface (SSI)*

## 1. Introduction

For a long time, researchers interested in the linguistic mechanisms involved in the voice production have been studied and analyzed speech signals. They have continually worked on the construction of techniques and devices able to recognize and understand the language automatically when the speech task is performed.

The resulted studies opened the doors to major applications, as the development of devices that could allow people to communicate with machines, which brings advantages in business, and accessibility. Benefits such as costs reduction in managing customer service staff, accessibility in automated services for natural communication, and the customization the product catalog tailored to customers [1], are examples of applications of Automatic Speech Recognition (ASR) and natural language understanding systems.

The ASR concept emerges from the idea of building a tool that would allow people to communicate and interact with machines, such as computers or similar devices, using speech as mean of communication. ASR systems enable to process and identify utterances spoken by the user through an input interface, which can be a microphone or phone. In this way, the spoken utterances are converted in treatable elements that compose a linguistically meaningful sentence for the system, and enable the human-machine interaction process through speech [2].

The development of speech recognition interfaces brings many advantages such as those mentioned above but exists a limitation in terms of applicability, which is related to using voiced speech as the only way of communication to build such kind of systems.

ISSN: 1975-0080 IJMUE

It is for this reason that the silent speech interface concept (SSI) was born. SSI systems are defined as an interaction platform that enables communication through speech, when audible acoustic signals are not available [3]. Among the applications of this kind of systems, one of the most important is to allow oral communication for people who have speech disorders. Furthermore, the improvement of the communication in noisy places to mitigate the possible distortions of orally expressed ideas by the sender in the communication process, and/or guarantee the privacy of the message if exist restrictions for public places [4].

The biggest difference between SSI and ASR systems lies in the acquisition stage of the signals for digitalizing and further processing. For the SSI systems it depends of the transductor used for acquisition, as seen in [5],while for automatic speech recognition systems only is required the speech signal captured by a microphone [6].

The Non-Audible Murmur (NAM) is one the methods used to build SSI systems. As its name suggest, NAM corresponds to the production of low-power whispers that cannot be heard by listeners nearby the speaker [7].

The NAM processing for developing silent speech interfaces is based on the theory of speech processing to build ASR systems. That is why the processing steps for ASR are applicable to NAM-based SSI systems with certain modifications.

Three major phases are defined for developing ASR systems, and are concerned with the analysis of the characteristics of the speech signals [8]; the classification, pattern recognition [9], and the verification of the utterances of words that are recognized by the system [10-11].

For recognizing words from non-audible murmur signals, we take into account the acquisition, analysis and pattern classification of the ASR methodology. For this case, the acquisition stage for NAM signals is performed using the same method for speech signals; with a microphone, but with the difference that it has to be modified to correctly capture this kind of signals [12].

In accordance with the advantages offered by NAM-type SSI systems, and since the implementation of such systems is performed similarly to the classical speech processing approaches, this paper proposes the development of a feature extraction stage to estimate the frequency-domain representation of non-audible murmur waveforms; through the Fast Fourier Transform (FFT) to estimate the magnitude spectrum, embedded on a development board.

## 2. Methodology

The aim of this paper is to describe the stages for acquisition, pre-emphasis, windowing and calculation of the magnitude spectrum of a silent speech signal. As result, the development of a silent speech processing system was achieved. It was embedded in a development board in order to build a portable device for speech processing.

Figure 1 shows an intuitive representation of the processing stages.

 

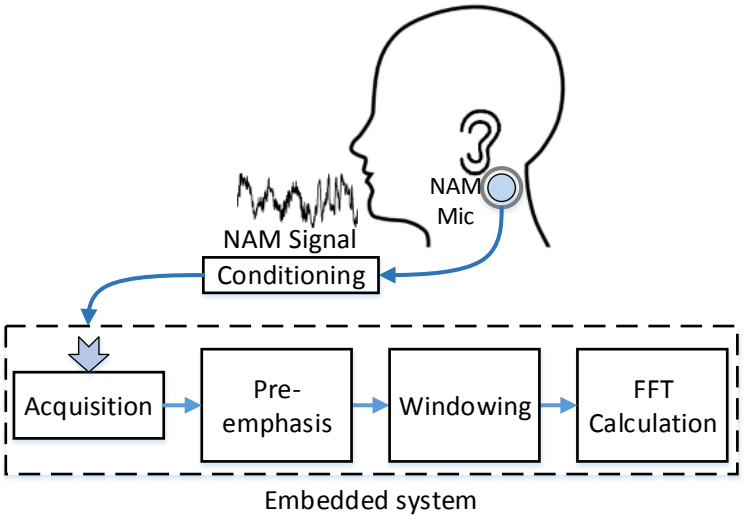


**Figure 1. Processing System Stages**

### 2.1. Acquisition of NAM Signals

The NAM signals are defined as the waveforms that are produced as result of very quietly uttered speech that cannot be perceived by anyone but the person who performs the speaking action [4]. According to this, it is necessary to build an acoustic sensor capable to detect the low-power signals of the silent speech.

A NAM microphone is the acoustic sensor used for this kind of applications. This transducer consists of a common electret condenser microphone, which is located within the cone-shaped space formed by the diaphragm of a common stethoscope. Such space is filled with soft silicone to allow propagation of the NAM signals, because of its properties similar to those of the human skin tissue [13]. Figure 2 shows the physical construction of the stethoscopic NAM microphone.

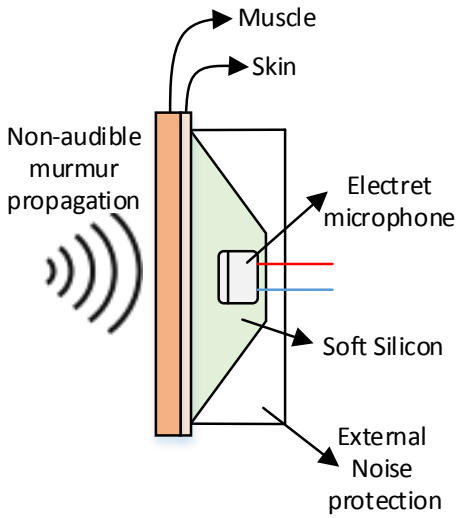


**Figure 2. NAM Microphone**

After placing the NAM microphone in the mastoid processes, which is the recommended place where a better signal acquisition of NAM signals is performed [7], it is proceeded with the signal conditioning stage. For this case, the signal conditioning

consists of a transistor configured as amplifier, which normalize the signal within the operating voltage ranges of the microcontroller of the board (0 V to 3.3 V), to then perform the digital processing.

The embedded system used for this application is the STM32F4Discovery board of STMicroelectronics®, which was configured to work at a clock frequency of 8 MHz. The HAL library from ST® and the CMSIS from ARM® were used as the software tools for programming.

The main modules of the MCU involved in the development of the feature extraction system were the ADC (Analog to Digital Converter) and the DAC (Digital to Analog Converter).

The digitization of the silent speech signal was performed using the ADC module of the board, which was configured to work using 12 bits of resolution and an approximate sampling time of 78.125 microseconds. The spoken utterance chosen for processing, in this case corresponds to the pronunciation of the word 'tres' [t ɾ e s] (three in English), whose approximate duration is 360 ms.

The total duration of the signal (360 ms), corresponds to 4608 samples according to the sampling time reached by the MCU, however, if an array whose dimension is limited to the exact number of samples is defined, it could represent problems to acquire the signal if it is generated and the ADC conversion is not synchronized. That is why an input vector of 6400 samples, corresponding to 500 ms for the acquisition of the NAM signal is defined, thus the mismatch is mitigated. Figure 3 depicts the NAM signal of the utterance of [t ɾ e s].

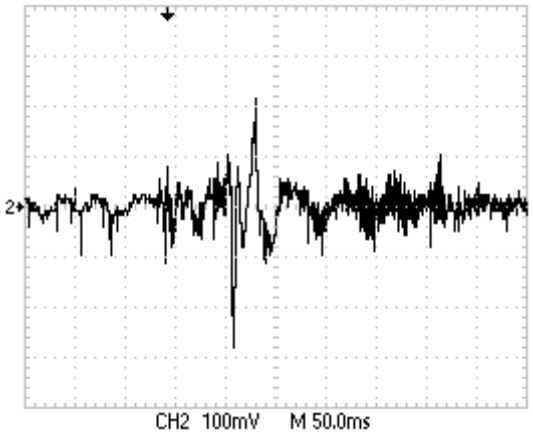


**Figure 3. NAM Signal of Utterance of 'Tres'**

### 2.2. Pre-emphasis

The second stage after the acquisition and amplification of the NAM waveforms is the pre-emphasis applied to the digitalized silent speech. The pre-emphasis phase aims to boost the energy magnitude of the high frequency components of the signal, to make the characteristic information of the signal more available and improve detection of phonological units, in further processing stages (*i.e.* pattern classification) [14].

The pre-emphasis in this work, is performed using a first order high pass filter, applied to the input signal $s[n]$. The filter is defined as shown in (1).

$$s_f[n] = s[n] - 0.97s[n-1] \quad (1)$$

Where $s_f[n]$ represents the filtered signal defined as a difference equation, and involves the present and previous components of $s[n]$.

The signal of 500 ms, was digitalized and stored in the data memory of the microcontroller, in a 6400 index data array of float32_t type. Then it was applied the first

 

order filter to the data vector that stores the digitalized signal. Such processes are represented in the block diagram of Figure 4. The new $s_f[n]$ vector represents the signal $s[n]$ when the pre-emphasis is performed. The notation in terms of $i$, is used to represent the iterative process in the algorithm.

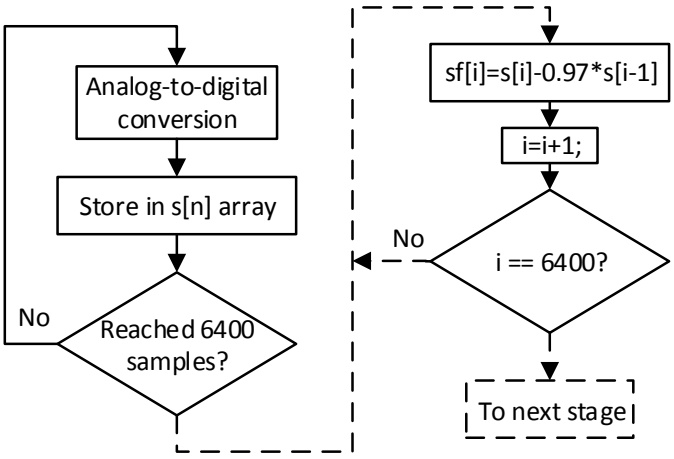


**Figure 4. Block Diagram of the Embedded Pre-emphasis Stage**

### 2.3. Signal Windowing

Speech signals (both voiced and silent), are part of the group of waveforms of non-stationary type, in short, this means that frequency components are not static and vary over time, as well as its statistical characteristics. This represents problems when implementing classical signal analysis (*i.e*. Fourier or temporal analysis), but exists the advantage that these frequency variations over time, have a very slow rate of variation due to the properties of the vocal tract system [1].

Accordingly, if segmentation of the signal is performed in frames short enough to be considered stationary, and long enough to find important features in the signal, it is possible to continue the classic signal processing without drawbacks due to the non-stationarity of these waveforms[15].

Some important parameters should be noted in the windowing process of the signal. The first is the frame width ($f_w$), which in this case is is defined in 40 ms and corresponds to the total duration of the frame, the frame shift ($f_s$) is 10 ms, and marks the dead-time between the beginning of one of the frames, with the next, and the final parameter is the shape of the window. Figure 5 depicts the process of segmentation of speech signals, and are based on the selected parameters.

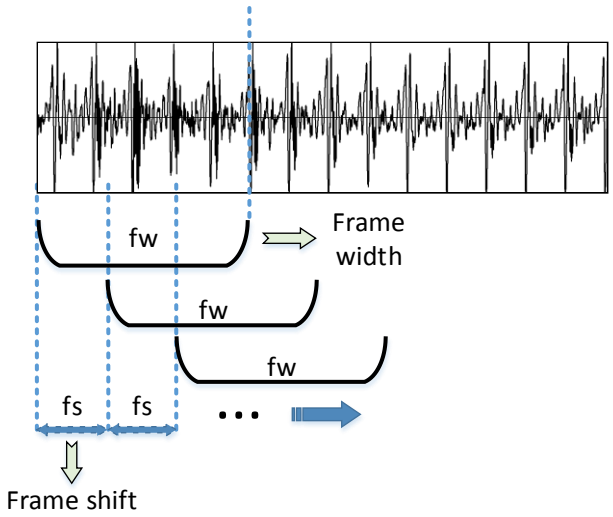


**Figure 5. Signal Windowing**

The window shape in this case, is defined as a Hamming window. The window is represented as a Gaussian type function, and is multiplied term by term with each of the resultant frames of the segmentation. The window multiplication process, is used to avoid discontinuities when the FFT calculation is performed, limiting the boundary magnitudes of the frame vector to zero.

The window width is selected to 40 ms, since this value is within the recommended range for this kind of applications (20 ms – 40 ms), and according to the sampling time this window width corresponds to 512 data of the filtered signal $s_f[n]$, which is a power of 2 and a requirement for the FFT calculation. Finally, the frame shift between the samples is set in 10 ms, based on the recommendation criteria summarized in [16].

To embed the windowing process on the development board, a two-dimensional matrix arrangement as data container is defined. This arrangement was defined in relation to the number of windows obtained from the $s_f[n]$ signal, which as will be seen later this value is 33 segments. Thus, the resulted segments were stored in a float32_t type data matrix ($M_w[m,n]$) of size 33 x 512.

### 2.4. FFT Calculation

The result in the segmentation of the signal is the division into equal length windows of the vector data that represents the NAM signal in discrete-time. This allows to perform specific processing tasks, such as feature extraction from the frequency representation of the signal. The next step in the feature extraction process is the calculation of the Fourier transform to find the frequency representation of the signal to characterize the analyzed waveforms. Equation (2) represents the Fourier transform in discrete terms.

$$S[j] = \frac{1}{N}\sum_{n=0}^{N-1} s[n]e^{-\frac{i2\pi jn}{N}} \quad (2)$$

The DSP-CMSIS library was used to calculate the FFT of the discrete-time frames in order to implement an algorithm for feature extraction using the embedded system. To properly calculate this information, the estimation of the magnitude of the Fourier transform of a real sequence was performed for each frame stored from the previous stage. The diagram in Figure 6 shows the procedure for this task.

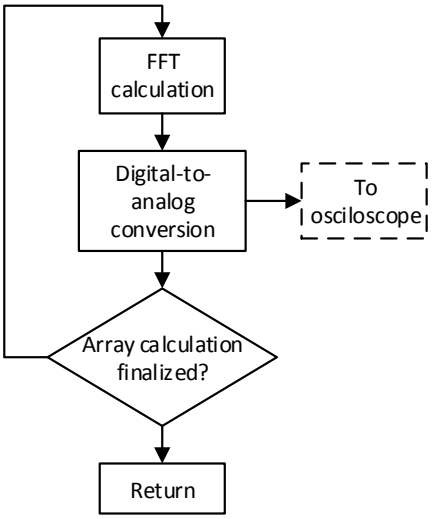


**Figure 6. FFT Calculation**

As shown in Figure 6, after the calculation of the FFT, the data is displayed on an oscilloscope as a test method for verification. This is accomplished using the DAC module of the embedded system, which was configured for this task.

## 3. Results

The proposed methodology summarized in

Figure 1, was successfully applied to the silent speech signal for the utterance [t ɾ e s], whose temporal representation is depicted in Figure 3. Thus, with the methodology application, it was possible to obtain the magnitude spectrum of each of the signal frames.

Figure 3 shows the time domain representation of the analyzed signal. In this figure which is taken from an oscilloscope, is shown that the content of relevant information in the signal begins at approximately 70 ms. The frames out of this time threshold were discarded.

According to this the first signal segment starts at the index 896 of the signal vector (equivalent to 70 ms). The last segment is located in the index 4992 of the data array that is 390 ms from the start of the signal. By defining these time limits and using equation (3) is possible to calculate the number of segments to be analyzed in 360 ms of total duration of the signal.

$$N_f = \frac{N_s - w_f}{s_f} + 1 \tag{3}$$

Where $N_f$, is the number of frames contained in the signal, $N_s$ corresponds to the number of samples of the signal, $w_f$ is the width of the frame in terms of number of samples (512), and $s_f$ is the frame shift in samples (128).

The number of frames of the silent speech signal, calculated from (3), correspond to 33. This means that 360 ms of the signal, contain 33 frames of 40 ms width with a shift of 10 ms. For explanation purposes, in this case only the spectrum of the first four windows are shown.

Figure 7 shows the magnitude spectrum of the first window which lies between the time limits 70 ms and 110 ms, which corresponds to a segment width equal to 40 ms (512 samples). It can be clearly noted in the spectrum that due to the effects of normalization through the pre-emphasis phase, the amplitude peaks present at high frequencies are dominant compared to the values at low frequencies. However, a max value occurs at approximately 5000 Hz in the signal, and its magnitude can be used as a unique characteristic of the signal.

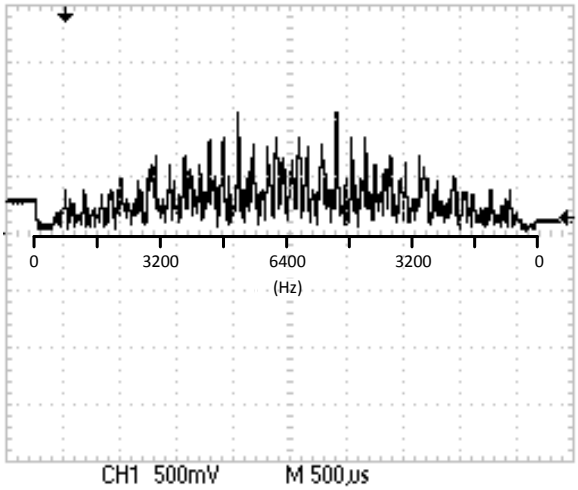


**Figure 7. Magnitude Spectrum of the First Frame**

In the case of the second frame, in the calculated magnitude spectrum (Figure 8), a variation from the previous spectrum is observed. This spectrum shows a significant low frequency component, located approximately at 730 Hz, and is constant for the frequency spectrums of the next frames. The low frequency component found in this frame, is estimated according to the time limits of this segment within the interval 80 $ms$ to 120 $ms$. This time interval presents more frequency information due to the shift process, which spans the spectral estimation process.

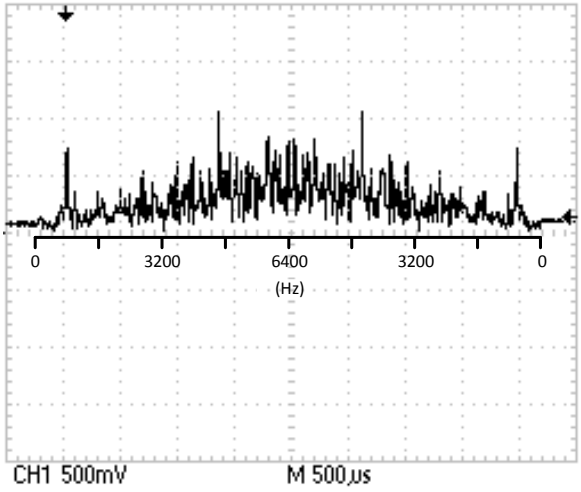


**Figure 8. Magnitude Spectrum of the Second Frame**

The third window magnitude spectrum is represented in Figure 9. There can be found better results with respect to high frequencies of the signal, while maintaining the low frequency value of 730 Hz, as mentioned above. The prominent peak at high frequencies, ranks like the first two windows at a frequency of about 5000 Hz, this value becomes in a key element in the frequency representation of the signal.

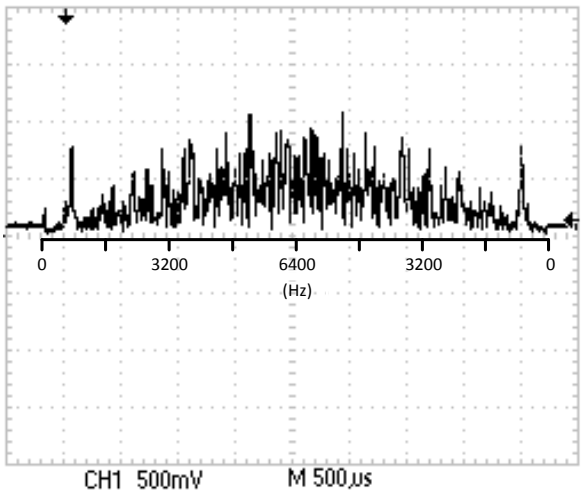


**Figure 9. Magnitude Spectrum of the Third Frame**

While the number of frames increases, it also does the range in which are defined, and therefore is found more information defining the behavior of the signal in frequency-domain. This behavior can be clearly noted in Figure 10, where the window defined in the interval (100 ms to 140 ms), maintains a low frequency with a high amount of energy in the 730 Hz value and also high frequency components in different points; 5000 Hz, 5500 Hz and about 6100 Hz.

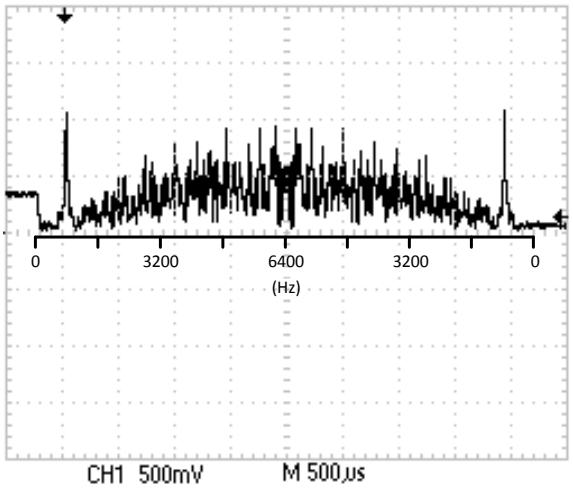


**Figure 10. Magnitude Spectrum of the Fourth Frame**

As reference data, in Table 1, are condensed some parameters related to memory consumption and processing time of the MCU. These parameters are taken into account as data to guide future work applications related to silent speech processing.

**Table 1. MCU Performance Parameters**

| Parameter | Value |
|---|---|
| Program size | 23496 bytes |
| RAM size | 127.06 kB |
| ROM size | 121.88 kB |
| Processing time | 900 ms |

## 4. Conclusions

The pre-emphasis and windowing stages that were implemented as the digital conditioning steps in this work, are the fundamental basis of development for feature extraction systems. The magnitude boosting of the high frequencies in the signal as result of the pre-emphasis phase, allows the identification of specific high frequency characteristics as unique elements of distinction in further recognition steps. Furthermore, the segmentation of the signal and normalization in defined time intervals from the total duration of the signal, reduces significantly the dynamic behavior of the signal for analysis purposes, which brings a feature extraction process by parts taking into account both temporal and frequency data.

The results show a greater contribution of noise in the temporal representation of the silent signal, this is due to the nature of the unvoiced speech which for the vocal tract is represented as a noise input. Despite of the high amount of noise, the proposed methodology is suitable for characterization, finding unique frequencies in the signal acquired as the results show.

The described system, shows clearly that there is no distinction itself related to the type of signal processing in the embedded system, if it is true that this case is aimed to NAM signal processing, if in the acquisition stage, the NAM microphone is replaced by common microphone, processing could be carried out without modification in terms of programming. This would allow the user to use this development in processing applications of voiced speech, with some modifications in the hardware, providing flexibility and versatility when working with this kind of developments.

As seen in the results, the performance parameters in this application shows that the memory consumption of the random access memory of the MCU, is in its capacity limit, which is due to the large number of bytes used for the windowing process, this issue could represent major problems if large amount of data would be processed. In comparison the ROM capacity and program size, are within the range that not exceeds the memory capacity. In the other hand, the processing time for these applications is suitable for the application described, however optimization in time for posterior developments could be accomplished if the clock frequency is augmented.

As future work to enhance the feature extraction system, the implementation of robust techniques as MFCC, PLP, LPC or RASTA is proposed as validation and comparing for the actual method.

## Acknowledgments

The authors would like to thank the Nueva Granada Military University research center for financing this work (research project INV-ING-1762, 2015).

## References

[1] L. R. Rabiner and R. W. Schafer, “Theory and Applications of Digital Speech Processing”, 1st ed. Pearson, **(2011)**.
[2] C. Vimala and V. Radha, “A Review on Speech Recognition Challenges and Approaches”, World Comput. Sci. Inf. Technol. J., vol. 2, no. 1, **(2012)**, pp. 1-7.
[3] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, “Silent speech interfaces”, Speech Commun., vol. 52, no. 4, **(2010)**, pp. 270-287.
[4] P. Heracleous, Y. Nakajima, A. Lee, H. Saruwatari, and K. Shikano, “Accurate hidden Markov models for non-audible murmur (NAM) recognition based on iterative supervised adaptation”, IEEE Workshop on Automatic Speech Recognition and Understanding, St. Thomas, Virgin Islands, **(2003)**.
[5] J. A. G. Calderón, E. N. G. Melo, D. A. Hurtado, and O. F. A. Sánchez, “Desarrollo de interfaces para la detección del habla sub-vocal”, Rev. Tecnura, vol. 17, no. 37, **(2013)**, pp. 138-152.
[6] D. O’Shaughnessy, “Acoustic Analysis for Automatic Speech Recognition”, Proceeding IEEE, vol. 101, no. 5, **(2013)**, pp. 1038-1053.

 

[7] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Non-audible murmur recognition input interface using stethoscopic microphone attached to the skin", International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, China, **(2003)**.
[8] H. Veisi and H. Sameti, "Speech enhancement using hidden Markov models in Mel-frequency domain," Speech Commun., vol. 55, no. 2, **(2013)**, pp. 205-220.
[9] Q. B. Nguyen, T. T. Vu, and C. M. Luong, "Improving acoustic model for English ASR System using deep neural network", in International Conference on Computing & Communication Technologies - Research, Innovation, and Vision for Future (RIVF), Vietnam, **(2015)**.
[10] M. M. Azmi and H. Tolba, "Syllable-based automatic Arabic speech recognition in different conditions of noise", in 9th International Conference on Signal Processing, Beijing, China, **(2008)**.
[11] D. O'Shaughnessy, "Interacting with computers by voice: automatic speech recognition and synthesis", Proceeding IEEE, vol. 91, no. 9, **(2003)**, pp. 1272–1305.
[12] P. Heracleous, Y. Nakajima, H. Saruwatari, and K. Shikano, "A tissue-conductive acoustic sensor applied in speech recognition for privacy", in Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-Aware Services: Usages And Technologies, Grenoble, France, **(2005)**.
[13] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell, "Remodeling of the sensor for non-audible murmur (NAM)", in Interspeech 2005, Lisboa, Portugal, **(2005)**.
[14] Y. Zhang, C. He, Y. Luo, K. Chen, and W. Xing, "Improved perceptually non-uniform spectral compression for robust speech recognition", J. China Univ. Posts Telecommun., vol. 20, no. 4, **(2013)**, pp. 122–126.
[15] D. Jurafsky and J. H. Martin, "Speech and language processing : An introduction to natural language processing, computational linguistics, and speech recognition", Prentice Hall, **(2009)**.
[16] X. Huang, A. Acero, and R. Hon, Hsiao-Wuen/Foreword By-Reddy, "Spoken Language Processing: A Guide to Theory, Algorithm, and System Development", Prentice Hall, **(2001)**.

## Authors

**Leonardo Andrés Góngora Velandia**, was born in Bogotá, Colombia. He received the B.Eng degree in Mechatronics Engineering from the Universidad Piloto de Colombia in 2015. Currently he is working as research assistant in the Virtual Applications Research Group (GAV), at the Nueva Granada Military University (UMNG). His research interests include Signal Processing, Artificial Intelligence, Human-Computer Interaction and Virtual Reality.

**Olga Lucia Ramos Sandoval**, is originally from Bogotá, Colombia. She was educated at UAN, Bogotá, Colombia receiving the B.Sc. degree in Electronics Engineering in 1998. She got her specialization certified in Electronic Instrumentation in 2000 by UAN and the M.Sc. degree in Teleinformatic in 2007 by the Faculty of Engineering at the Francisco José de Caldas District University, UFJC in Bogotá, Colombia. Currently she is completing the Ph.D. degree in Engineering at UFCJ. Right now she is working as a Teacher at UMNG and Research in the Research Group GAV in different Mechatronics fields like System Control and Industrial Automation.

**Dario Amaya Hurtado**, was educated at UAN, Bogotá, Colombia receiving the B Sc. degree in Electronics Engineering in 1995 and the M.Sc. degree in Teleinformatic in 2007 by the Faculty of Engineering at the Francisco José de Caldas District University, UFJC in Bogotá, Colombia. He was awarded the Ph.D. degree in 2011 in Mechanical Engineering at Campinas State University, São Paulo, Brazil, working on hybrid control – He has worked as a professor and researcher at the Military University, Colombia since 2007 been involved in Robotics, Mechatronics and Automation areas.