International Journal of Digital Information and Wireless Communications (IJDIWC) 8(1): 58-62
The Society of Digital Information and Wireless Communications, 2018 ISSN: 2225-658X (Online); ISSN 2412-6551 (Print)

Speaker Identification Based On Vocal Cords' Vibrations' signal: Effect Of The Window

Dany Ishac[1, 3], Antoine Abche[2] and Elie Karam[2]
[1]Dept. of Computer Engineering
[2]Dept. of Electrical Engineering
University of Balamand, Deir El Balamand,
Koura, Lebanon

Georges Nassar[3], Dorothée Callens[3]
[3]Laboratoire de l'IEMN
Département Opto-Acousto-Electronique
Université de Valenciennes et du Hainaut-Cambrésis,
Valenciennes, France

***Abstract*—A new speaker identification technique was proposed and presented. It is based on the acquisition of the individual's signal of its vocal cords' vibrations by attaching a piezoelectric transducer element to a collar and wrapping it around the neck. Having collected the signal, the Short Term Fourier Transform (STFT), in conjunction with a particular window, is applied on the signal to decompose it into its frequency components. Then, the resulting spectrogram is normalized, the noise is removed and the corresponding features are extracted for identification purposes. The performance of the developed approach in conjunction with various windows' types (the Bartlett, the Blackman, the Hamming, the Hanning and the Rectangular windows) and window's size is studied and evaluated quantitatively. The results show that the proposed approach using the Hamming window of size 64 yields the best accuracy in the identification of the desired individuals.**

Keywords*—Vocal cords vibration frequency, STFT, Barlett, Blackman, Hamming, Hanning, Rectangular, collar, transducer, speaker identification

I. Introduction

The biometric recognition technology has been lately gaining a tremendous popularity due to its importance as a robust security measure. Biometric security systems are favorable and convenient to users because the persons are not required to remember long passwords or to carry any identification cards. Furthermore, the biometric recognition consists of the extraction of a feature vector based on a physiological characteristic, which is exclusive and unique to each individual, such as the retina, the iris, the face, the voice, etc. Therefore, the identification methods provide a high degree of security and have been used in a wide variety of applications [1, 2].

The speaker recognition, particularly based on the biometric technology, has been studied by researchers for many years. Numerous network models and signal processing techniques have been developed and have been tested for recognition and identification purposes [3] such as the linear predictive coding (LPC) technique [4-6], the Mel frequency cepstral coefficient (MFCC) [7, 8], the wavelet transform [9-11]. The field of speaker recognition can be divided into two categories: speaker verification and speaker identification. The first category involves the comparison of an individual's sound with an existing sound's sample to decide if he/she is who he/she claims to be. However, speaker identification involves the matching of the input sound with known sounds stored in the database. The latter category can be divided into two branches: text-dependant identification and text-independent identification. While the text-dependant identification system has a prior knowledge of the spoken text by the user, the text-independent identification system has to recognize the user from any spoken text [12-14].

This paper describes a new text-dependant speaker identification system as well as its performance using various windows. Besides its good percentage of accuracy, its main advantage lies in the fact that the text used for recognition is only an utterance which gives the system a very high classification speed. Also, its main novel characteristic is that the signal of the vocal cords' vibrations is used for the individual's identification unlike the existing systems in the literature that use voice's signals acquired by microphones.

II. Method

The proposed approach is based on acquiring the signal of the vocal cords' vibrations and analyzing it. In this context, a piezoelectric transducer element is attached to a collar which is wrapped around the individual's neck. It generates a charge when a pressure hits the surface of the transducer and a sound wave when a voltage is applied across the element. That is, an electrical energy is transformed into a mechanical energy and vice versa. When a mechanical vibration hits the crystal's surface of the transducer, a current signal of proportional intensity and of the same frequency of the vibration is generated.

The speaker was requested to utter the vowel "a". The vocal cords' mechanical vibrations were detected by the transducer and the corresponding signal is transformed to an electrical signal for processing. The latter signal is a non-stationary signal. Its properties change substantially over time and the changes are usually of primary interest for analysis purposes. The spectral analysis techniques such as the Fourier analysis provide a good description of the frequencies' contents of the waveform but not their timing. The latter information is encoded in the phase portion of the transform. However, the encoding is difficult to interpret and to recover. Therefore, a wide range of approaches have been developed to extract both the time and the frequency information from a waveform. They are known as time-frequency methods and include the Short Term Fourier Transform (STFT), the Choi-Williams Distribution (CWD) and the Wigner-Ville Distribution (WVD) [15, 16]. The STFT technique was applied to the collected signal to decompose the latter into its frequency components. It consists of slicing the waveform of interest into a number of short segments and performing the Fourier transform on each segment. A window function must be applied on the collected signal x(t) to isolate the segment of data and consequently to perform the STFT on the extracted data. Therefore, the selection of a window's type and its size can be crucial. The window's type and the window's size have a big influence on the results. While the small window's size improves the time resolution, the frequency resolution will be reduced and vice versa. Moreover, low frequencies might be lost when the size of the window is very small because they will not be included in the data segment to be analyzed [15, 16- 17]. In this context, the performance of the proposed time-frequency technique using the various windows (Bartlett, Blackman, Hamming, Hanning and Rectangular) as well as their sizes is studied and is evaluated.

Having implemented the STFT technique, further processing is performed on the acquired signal. First, the frequencies' magnitudes are normalized by dividing the latter by the highest value. Second, the low frequencies' values are considered as noise because they are mostly resulted from an outside interference and not from the vocal cords' vibrations. Subsequently, they are eliminated in order to have a better signal, well suited for the latter processing and consequently to achieve a better accuracy for identification purposes. Finally, the meaningful frequencies of the signal are extracted to identify the individual. The ranges of frequencies extracted from the spectrogram are those who have magnitudes' values greater than a threshold value. In other words, if the magnitude at a particular frequency is greater than a certain percentage of the highest frequency's amplitude, it will be kept. Otherwise, the corresponding frequency's amplitude is removed (set to zero). The extracted features are transformed into a 1-D array for classification purposes [18].

III. SPEAKER IDENTIFICATION

In this section, the classification of the collected signal of an individual is described. In other words, the desired person who has uttered the vowel "a" is identified. The identification is accomplished using the correlation as a similarity measure. The extracted 1-D array (X) is compared against a set of vectors that are already extracted in the same manner and are stored in the database for different individuals (Trained set – vector Y). The linear correlation coefficient Corr(X, Y) between two vectors X and Y is expressed as:

$$Corr(X,Y) = \frac{1}{MxN} \sum_{i=0}^{MxN-1} \frac{X_i Y_i - \mu_X \mu_Y}{\sigma_X \sigma_Y} \qquad (1)$$

Where X and Y are the vectors to be compared, μ_x and μ_y are the mean values of X and Y, respectively, σ_x and σ_y are the standard deviations of X and Y, respectively and MxN is the length of the extracted vector X (or Y). For any two vectors, the closer the coefficient's value is to 1; the higher the similarity is between the two vectors. Consequently, the closer the correlation value is to zero, the higher is the dissimilarity. Then, the highest correlation value points to the identified person [19].

As stated before, the identification process requires the existence of a database in which a template of the feature vector of each individual to be identified is stored. In this work, the database consists of feature vectors of N individuals. Actually, each person utters the vowel "a" and the corresponding signal of the vocal cords' vibrations is collected. This experiment is repeated three times for each individual. Thus, 3N signals were collected and each is processed as outlined earlier in order to obtain the feature vector. Then, one feature vector per individual is stored in the database and the remaining 2N feature vectors are used to evaluate the proposed approach.

IV. RESULTS AND ANALYSIS

As already stated earlier, the proposed time-frequency based approach requires the implementation of STFT using a particular window, namely, the Bartlett, the Blackman, the Hamming, the Hanning and the Rectangular. The selection of the window type and its size affect the precision of the developed approach and the accuracy in the identification of the desired individuals. In this context, the various windows are presented before proceeding to the quantitative analysis.

The Bartlett window is defined as follows [20]:

$$W_b(n) = \begin{Bmatrix} \frac{2n}{N-1}, & 0 \le n \le \frac{N-1}{2} \\ 2 - \frac{2n}{N-1}, & \frac{N-1}{2} < n \le N-1 \\ 0, & otherwise \end{Bmatrix} \qquad (2)$$

The Blackman window is given by [20, 21]:

$$W_{bl}(n) =$$

$$\begin{Bmatrix} 0.42 - 0.5\cos\left(\frac{2\pi n}{N-1}\right) + 0.08\cos\left(\frac{4\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0, \qquad otherwise \end{Bmatrix} \qquad (3)$$

The Hamming window is defined as follows [20, 21]:

$$W_h(n) = \begin{Bmatrix} 0.54 - 0.46\cos\left(\frac{2\pi n}{N}\right), -\left(\frac{N-1}{2}\right) \le n \le \left(\frac{N-1}{2}\right) \\ 0, \qquad otherwise \end{Bmatrix} \qquad (4)$$

The Hanning window is expressed by [20, 21]:

$$W_c(n) = \begin{Bmatrix} 0.5 - 0.5\cos\left(\frac{2\pi n}{N} - 1\right), 0 \le n \le N-1 \\ 0, \qquad otherwise \end{Bmatrix} \qquad (5)$$

The Rectangular window can be considered as the simplest window. It is represented by the following weighted function [20]:

$$W_r(n) = \begin{Bmatrix} 1, & \frac{-(N-1)}{2} \le n \le \frac{N-1}{2} \\ 0, & otherwise \end{Bmatrix} \qquad (6)$$

TABLE I. COMPARISON OF WINDOWS' PARAMETERS [20]

Window Type	Approximate amplitude of the peak side lobe	Approximate main lobe's width	Peak estimation error (dB)
Bartlett	-25	$\frac{8\pi}{N}$	-25
Blackman	-57	$\frac{12\pi}{N}$	-74
Hamming	-41	$\frac{4\pi}{N}$	-53
Hanning	-31	$\frac{8\pi}{N}$	-44
Rectangular	-13	$\frac{4\pi}{N+1}$	21

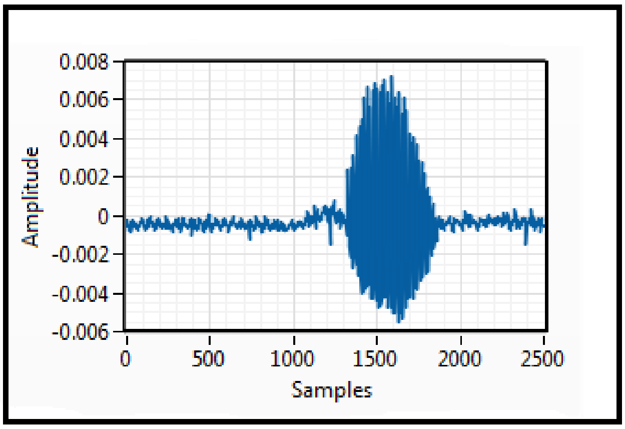

Fig. 1. Acquired signal for an individual

Table I shows a summary of the main differences between the various implemented windows in terms of the main lobe's width, the amplitude of the peak side lobe with respect to the main lobe and the error associated with the peak's estimation [20].

As it is stated earlier, the measurements were performed by attaching a piezoelectric transducer on the skin of the individual's neck using a collar that is wrapped around the individual's neck. The transducer element is the Ferroperm Piezoceramic Pz26. It has a length (L) =2.2cm, a width (W) = 0.4cm, a thickness (Th) = 0.1 cm and a coupling factor K of 33% [22, 23]. The transducer element was connected to the input port of a NI Elvis II+ board (16-bit resolution) and the resulting electrical signal was acquired and analyzed using labview. The sampling frequency was selected to be 2500 Hz.

The individual was asked to utter the vowel "a". Then, the collected signal of the vocal cords' vibrations is processed using the approach presented in section III. That is, each individual's signal is analyzed using the STFT technique in conjunction with each window. The corresponding database for each window is generated as it is described earlier. Then, the remaining collected signals are processed to evaluate the performance of the respective STFT approach in identifying the desired person.

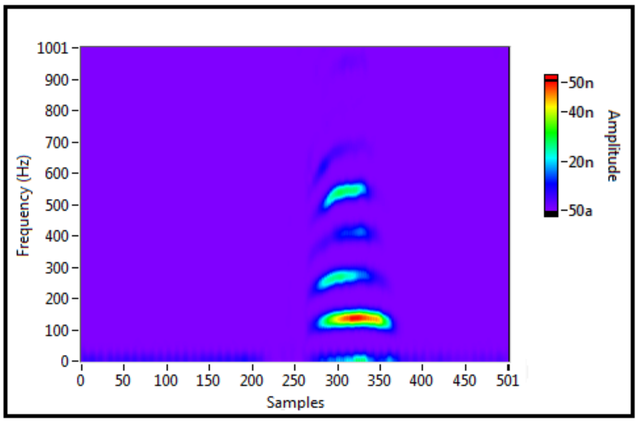

Fig. 2. Corresponding spectrogram after applying STFT

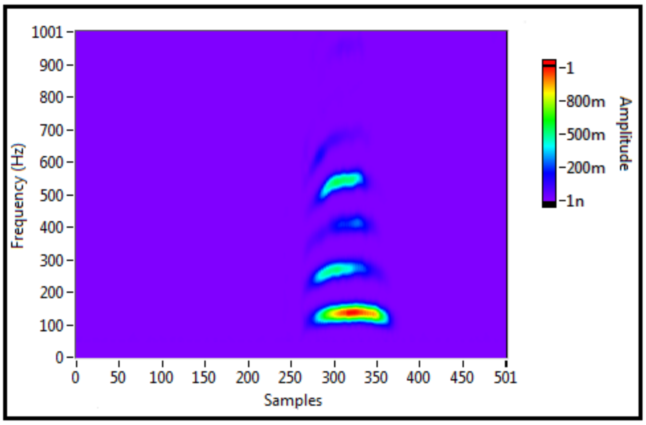

Fig. 3. Corresponding spectrogram after Normalization and noise removal

Fig. 1 shows the collected signal from an individual uttering the vowel "a". The corresponding spectrogram that is obtained using the STFT technique, in conjunction with the Hamming window of size 64 and a time step of 5 is shown in Fig.2. The amplitudes of the displayed frequencies are represented by different colors in the graph. The amplitudes' range is from 50 atto (1 atto= 10^{-18}) to 50 nano.

Fig.3 shows the resulting spectrogram after the normalization and the noise's removal procedures are performed. The corresponding extracted features are shown in Fig.4. The extraction is performed by selecting a range from the 2-D array (shown in Fig. 3) bounded by two vertical lines. While the first line corresponds to the first sample's index associated with the frequency components having a magnitude value greater than the threshold value, the second line corresponds to the last sample's index associated with the frequency components having a magnitude value greater than the threshold value. Thus, all the meaningful frequencies in the spectrogram will be inside the selected range. These features were stored and were used for identification purposes.

Table II illustrates the accuracy of the STFT technique in conjunction with the different windows' types, namely, the Barlett, the Blackman, the Hamming, the Hanning and the Rectangular. Also it displays the performance of the STFT using various windows' sizes, namely, 32, 64, 128, 256, and 512.

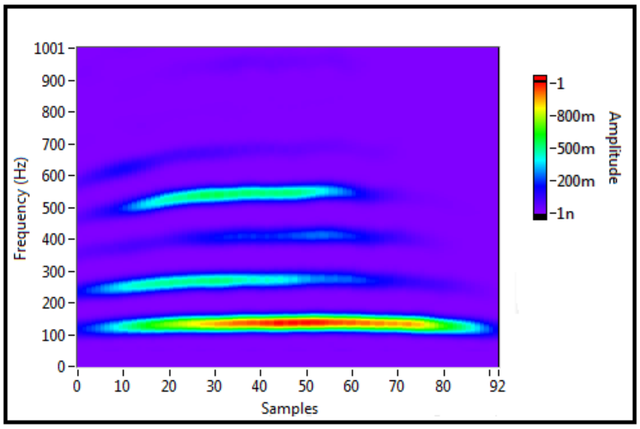

Fig. 4. Corresponding extracted features

The quantitative evaluation between the collected features of each signal and the template features of the various individuals (stored in the database) is performed using the correlation as a similarity measure. The results show:

i) For a given window, the accuracy of the proposed approach in the individual's identification decreases as the size of the window increases (for window's size ≥ 64). For example, the percentage decreases from 91% to 79%, 69% and 64% when the size of the window is increased from 64 to 128, 256 and 512, respectively.

ii) The decrease of the proposed technique that is observed in (i) might be due to the fact that as the size of the window is increased, the varying nature of the collected signal might be affected in the frequency domain.

iii) For a window's size of 32, the precision of the approach using the various windows is not high. This might be due to the fact that the small window's size does not contain enough information that will lead to a higher percentage of identification.

iv)The implementation of the STFT approach in conjunction with a Hamming window of size 64 yields the best performance i.e. 91% accuracy is achieved in identifying the desired person.

V. CONCLUSION

A new approach for speaker identification was presented. Unlike the current identification approaches, the proposed approach is based on the acquisition of the signal of vocal cords' vibrations (not the voice) using a piezoelectric transducer element attached on the skin of the neck using a collar. In this work, the collected signal is analyzed using a time-frequency approach, namely the STFT. Different windows' types and window's sizes were incorporated with the STFT. The results were acceptable for all windows with a size of 64 and 128. However, the proposed technique, in conjunction with the Hamming window of size 64, has yielded the best precision for identification purposes. The future work will be concentrated on the exploration of new signal processing algorithms as well as new similarity measures in order to improve the accuracy of persons' identification.

TABLE II. PERCENTAGE OF ACCURACY OF THE PROPOSED TECHNIQUE IN IDENTIFYING THE DESIRED INDIVIDUAL FOR DIFFERENT WINDOW'S TYPES AND WINDOW'S SIZES

Window's Type	Window's Size	32	64	128	256	512
Barlett		76	88	79	70	66
Blackman		62	85	85	72	69
Hamming		77	91	79	69	64
Hanning		73	87	81	70	66
Rectangular		88	86	74	67	65

REFERENCES

[1] E. Avci, "A new optimum feature extraction and classification method for speaker recognition: GWPNN," Expert Systems with Applications, vol. 32, no. 2, pp. 485-498, Feb. 2007. (Pubitemid 44647815)

[2] J. D. Wu, Y. J. Tsai, C. W. Chuang, L. H. Fang, & D. E. Song, "Speaker identification based on voice signal using Wigner-Ville distribution and neural network," *International Conference on Control, Automation and Robotics (CAR),*Singapore, 2012, pp. 40-45.

[3] D. Avci, "An expert system for speaker identification using adaptive wavelet sure entropy", *Expert Systems with Applications*, vol. 36, no. 3, pp. 6295-6300, 2009.

[4] A.G. Adami, D.A.C. Barone, A speaker identification system using a model of artificial neural networks for an elevator application, Information Sciences 138, 2001, pp.1-5. (Pubitemid 32698485)

[5] A. Haydar, M. Demirekler, and M.K. Yurtseven, "speaker identidication through use of features selected using generic algorithm," Electron. Lett., vol. 34, no. 1, pp. 39-40, 1998.

[6] C. Wutiwiwatchai, V. Achariyakulporn, C. Tanprasert, "Text-dependent speaker identification using LPC and DTW for Thai language", *TENCON 99. Proceedings of the IEEE Region 10 Conference*, vol. 1, pp. 674-677, 1999.

[7] DJ. Mashao, M. Skosan, "Combining Classifier Decisions for Robust Speaker Identification", Pattern Recognition, vol.39, pp. 147-155, 2006.

[8] J. J. Sroka and L. D. Braida, "Human and machine consonant recognition", Speech Commun., vol. 45, no. 4, pp. 401-423, 2005. (Pubitemid 40423287)

[9] S. Y. Lung, "Wavelet feature selection based neural networks with application to the text independent speaker identification," Pattern Recognition, vol. 39, pp. 1518-1521, 2006.

[10] J. D. Wu, B. F. Lin, "Speaker identification using discrete wavelet packet transform technique with irregular decomposition", *Expert Systems with Applications*, vol. 36, pp. 3136-3143, 2009.

[11] J. D. Wu, S. H. Ye, "Driver identification based on voice signal using continuous wavelet transform and artificial network techniques", *Expert Systems with Applications*, vol. 36, pp. 1061-1069, 2009.

[12] X. Zhao, Y. Wang, D.L. Wang, "Robust speaker identification in noisy and reverberant conditions", *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 4, pp. 836-845, 2014.

[13] W. Jian-Da, L. Bing-Fu, "Speaker identification based on the frame linear predictive coding spectrum technique", *Expert Systems With Applications*, vol. 36, no. 4, pp. 8056-8063, May 2009.

[14] E. Avcci, D. Avci, "The speaker identification by using genetic wavelet adaptive network based fuzzy inference system", *Expert Systems with Applications*, vol. 36, no. 6, pp. 9928-9940, August 2009.

[15] L. Cohen, "Time-frequency distributionA review", *Proc. IEEE*, vol. 77, pp. 941-981, July 1989.

[16] R. A. Brown, M. L. Lauzon & R. Frayne, "Developments in Time-Frequency Analysis of Biomedical Signals and Images Using a Generalized Fourier Synthesis", Recent Advances in Biomedical Engineering, pp. 191-210, 2009.

[17] R. Schafer, L. Rabiner, "Design and simulation of a speech analysis-synthesis system based on short-time Fourier analysis", *IEEE Trans. Audio Electroacoust.*, vol. 21, no. 3, pp. 165-174, 1973.

[18] D. Ishac, A. Abche, G. Nassar, E. Karam, & D. Callens, “A text-dependant speaker recognition system”, in press.

[19] D. Ishac, G. Yammine & A. Abche, "Face recognition using a fourier polar based approach," *2015 International Conference on Systems, Signals and Image Processing (IWSSIP)*, London, 2015, pp. 200-203.

[20] S. Dogra, N. Sharma, “Comparison of Different Techniques to Design of Filter”, *International Journal of Computer Applications*, vol. 97, no. 1, July 2014.

[21] P. Podder, T.Z. Khan, M.H. Khan, M.M. Rahman, "Comparative Performance Analysis of Hamming Hanning and Blackman Window", *International Journal of Computer Applications*, vol. 96, no. 18, pp. 1-7, 2014.

[22] W. W. Wolny, Piezoceramic thick films- Technology and application. State of the art in Europe, Proc. 2000 12th IEEE Int. Symp. on Applications of Ferroelectrics (edited by. S. K. Streiffer, B. J. Gibbons, T. Tsurumi), Honolulu, (2000), 257-262.

[23] Ferroperm Piezoceramics. High Quality Components and Materials for the Electronic Industry. Available online: http://www.ferroperm-piezo.com/files/files/Ferroperm%20Catalogue.pdf (accessed on 5 October 2016).