MASKAY 2(1), Nov 2012
ISSN 1390-6712

Recibido (Received): 2012/05/04
Aceptado (Accepted): 2012/08/17

LPC-based Feature Coefficients for Voice Authentication Tasks

María Salomé Pérez and Enrique V. Carrera

***Abstract*— Voice authentication is a promising biometric technique based on extracting important information from the speech signal by means of computing a vector of feature coefficients. Based on that, this paper evaluates the effectiveness of linear predictive coefficients when combined with other simple metrics in voice authentication tasks. Linear predictive coefficients were chosen due to their relatively good performance and their not-so-complicated structures when compared to other similar alternatives. All the feature coefficients have been evaluated through an extensive parameter space study in order to apprehend the main limitations and potentials of voice authentication under different scenarios. For such an evaluation, a classifier based on artificial neural networks has been implemented.**

***Index Terms*— Voice authentication, linear predictive coefficients, artificial neural networks.**

I. INTRODUCTION

Nowadays, technology is been used for restricting access to our resources through user authentication. Although, there are several authentication techniques, biometric-based authentication is the most promising alternative. Biometric-based authentication measures individuals' unique physical or behavioral characteristics (*e.g.*, fingerprint verification, retinal scans, facial analysis, analysis of vein structures, voice authentication) [1]. Indeed, biometric-based authentication has some key advantages over other authentication techniques, since biometric characteristics are not easily forgotten, like a password, or lost like a key.

For the majority of biometric authentication techniques, sophisticated equipment and the physical presence of the person being authenticated is required. However, given the correct use of analytical techniques, a person's *voiceprint* can be as unique as any other biometric characteristic. Moreover, voice authentication is simple (*i.e.*, no extra hardware or software is required), is less personally intrusive, and the authentication itself can even be done remotely [2].

Considering that speech signals carry information about a speaker in various forms (*e.g.*, speaking style, context, emotional state of the speaker), developments in the voice-authentication field have generated several statistical, analytical and data processing techniques to support it properly. A convenient and well understood processing technique is to extract important information from the speech signal by means of computing a vector of feature coefficients [1]. There are many effective feature extraction algorithms available. For instance, the short-term spectrum of the speech signal is the most famous method of representing this kind of signals. Nevertheless, several approximations to the short-term spectrum, such as linear prediction coding, Mel-ceptrum coefficients and filter bank magnitudes, are also popular [3].

Based on that, this paper focuses on evaluating the usage of linear predictive coefficients (LPCs) in voice authentication tasks. We are interested in using a vector of feature coefficients based mainly on LPCs due to their relatively good performance and their not-so-complicated structures when compared to other similar alternatives. Along this evaluation, LPCs are also combined with other feature coefficients (*e.g.*, fundamental frequency, standard deviation, energy, kurtosis) in order to have an extensive parameter space study.

In our particular implementation, the vector of feature coefficients is used to authenticate people using a classifier based on artificial neural networks (ANNs) [4]. In order to apprehend the main limitations and potentials of voice authentication under different scenarios, results for the identification of groups of people (*i.e*, gender classification), identification of one person among a group of strange people, and individual authentication of people, are presented.

II. BACKGROUND

This section introduces some fundamental concepts required in tasks of voice-pattern classification.

A. The Essence of the Voice

For most people, speech is an efficient and natural form of exchanging information. Nevertheless, this condition is limited due to the variable characterization of some features depending of accent, emotional state and vocal tract modeling of the speaker. Thus, voice signals cannot be consider stationary processes, but within a sufficiently short interval, the first and second order statistical moments of voice show little variations. In other words, voice signals can be considered as wide-sense stationary processes around a period of 30 ms [5].

B. Voice Features

1) Linear Predictive Coding: Linear predictive coding models across its coefficients the future value of a signal from a linear combination of its past values. This coding is widely used as a strong tool for modeling vocal tracks. In this case, the transfer function of a vocal track can be approximated by an all-pole filter model [3]. This voice feature can also be examined through an auto-regressive algorithm, which is able to achieve higher resolutions with smaller number of samples and lower sampling frequencies when compared to the FFT (Fast Fourier Transform) approach [6].

M. S. Pérez studies Electronics and Telecommunications Engineering at the Ecuadorian Armed Forces University, P.O. Box 17-15-231B, Sangolqui, Ecuador msperez1@espe.edu.ec

E. V. Carrera is with the Department of Electrical Engineering, Ecuadorian Armed Forces University, P.O. Box 17-15-231B, Sangolqui, Ecuador evcarrera@espe.edu.ec

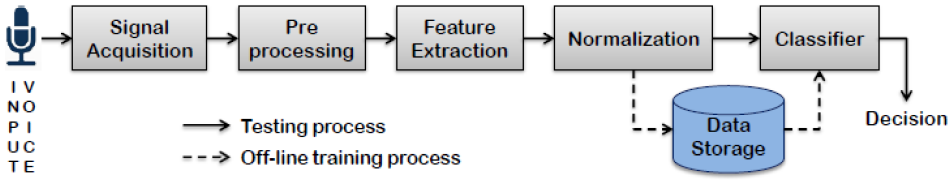

Fig. 1. Block diagram of the voice processing system.

2) Fundamental Frequency: The fundamental frequency F_0, knows also as pitch, models the main vibrations of the glottal cords that generate the sequence of quasi-periodic excitation pulses associated to each sound [7].

3) Time-Domain Features: In order to enhance LPC representation, it is usual to introduce statistical features such as the energy value (E) of the digital signal and the standard deviation (σ) of the samples contained in each frame [8]. Another possible factor for enhanced speech characterization is kurtosis (K), which represents the speech signal waveform by means of a numerical value, measuring the relative concentration (*i.e.*, flatness or peakedness) of a real-valued random variable when related to the normal distribution [5].

C. Artificial Neural Networks

ANNs are a combination of relatively simple non-linear adaptive processing elements, arranged in a structure that resemblances the processing of biological neurons. Hence, several layers of parallel processing elements are interconnected, and their connection weights are adjusted to perform some specific functions such as classification or prediction. Its main applications include non-linear partitioning of vector spaces, feature extraction, and decision-making systems [3].

The most common algorithm used for training ANNs is know as back-propagation. This algorithm looks for a local minima in the error function while adjusts the connection weights of the network. Since back-propagation is a supervised learning method, the network requires the dataset of desired outputs for each input. In this way, the difference between the desired output and the current one is the error function to minimize.

In addition, back-propagation networks require differentiable transfer functions [6]. Because of that, the back-propagation algorithm generally uses the sigmoid transfer function in each processing element.

III. IMPLEMENTATION

In order to evaluate the effectiveness of the LPC-based feature coefficients in voice authentication tasks, a voice-processing system was developed using the *Matlab*® R2012a platform[1]. In particular, the data acquisition, signal processing, and neural network Matlab's toolboxes are being used.

The block diagram of the proposed system is presented in figure 1, where the solid line shows the normal testing process, while the dashed line correspond to the initial off-line training process. These blocks are detailed in the following paragraphs:

- *Signal acquisition*. Speech signals are captured by our system using the recording Matlab function and an external omni-directional microphone. This function is configured to use a sampling frequency of 8 KHz, 16 bits per sample, and recording in the WAV audio format.
- *Pre-processing*. This block segments the whole acquired input in groups of 240 adjacent samples (*i.e.*, 30 ms) and labels each group according to some identification parameters previously set in the system's interface. The main identification parameters used in this work are the name and gender of the speaker.
- *Feature extraction*. This block computes a vector of LPCs besides other values like kurtosis, standard deviation, fundamental frequency, and energy of the digital signal. These feature coefficients are calculated considering a time frame of 30-ms. In other words, our feature extraction works over each group of 240 samples created by the pre-processing block.
- *Normalization*. Since the ANN model used as classifier requires input values in the interval $[-1, +1]$ in order to maximize its performance, the normalization of every feature coefficient is required. Note that the sigmoid transfer function generates output values in the interval $(0, 1)$, but its highest sensibility is given for input values around $(-1, +1)$. Thus, the *min-max* normalization method [9] is used in our current implementation.
- *Data storage*. At the very beginning, normalized feature coefficients and their labels (*i.e.*, the targets of the classifier) are stored in MAT data files in order to create a training dataset for the ANN. In Matlab, these values are viewed as a matrix where each row corresponds to some feature coefficient or label, and each column represents a different group of 240 speech samples.
- *Classifier*. This work uses an ANN-based classifier executing the back-propagation algorithm. The implemented neural network has only 3 layers. The size of the inner layer is always configured to 10 processing elements, while the size of the input and output layers depends on the nature of the experiment (see next section). The final decision of the network is chosen according to a simple criterion of the maximum output value, since each network's output corresponds to a unique label (or target). In this way, recognition success and mismatch are easily explored through a confusion matrix [4].

[1] http://www.mathworks.com/

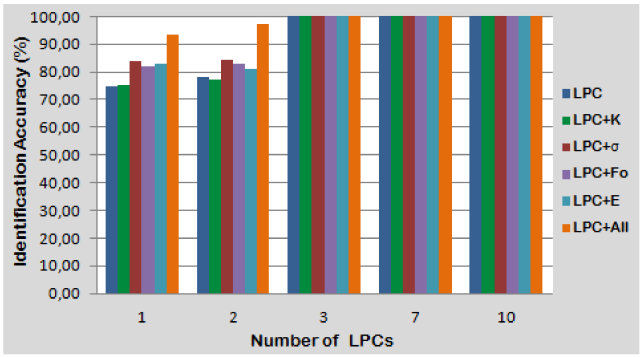

Fig. 2. Gender identification using samples from 2 people.

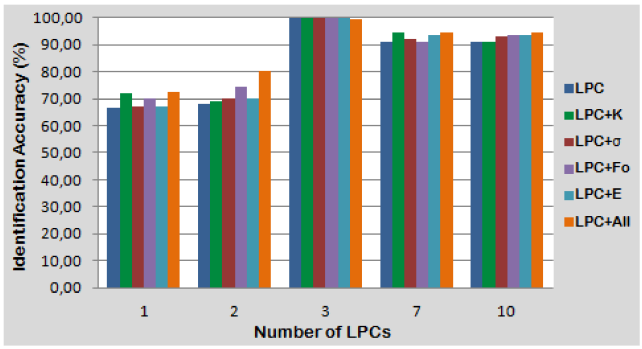

Fig. 3. Gender identification using samples from 8 people.

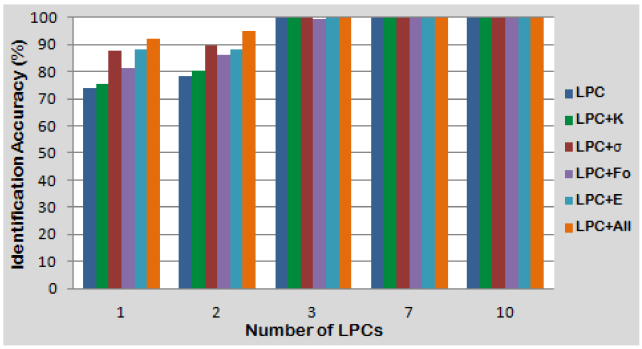

Fig. 4. One person identification considering two strangers.

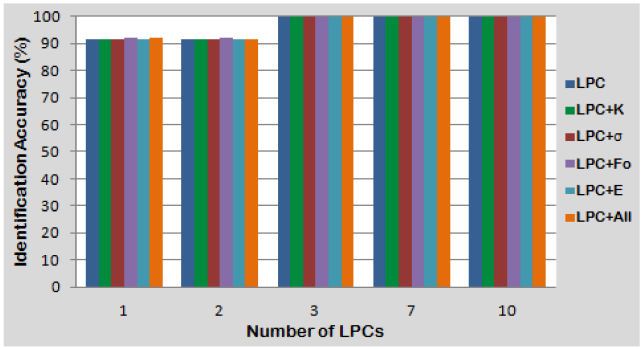

Fig. 5. One person identification considering eight strangers.

IV. RESULTS

With the purpose of evaluating the effectiveness reached by our authentication system, voice samples from eight adult people (4 women and 4 men) were collected. Each person pronounced a codeword of 4 Spanish terms spaced in time, generating more than a thousand sets of feature coefficients and labels. In each execution, a random set that includes 70% of the samples is used for training the ANN, while the remaining 30% of the samples are used for testing purposes. All the presented results correspond to the average of a hundred executions, since the ANN training process is not completely deterministic.

The following subsections analyze the effectiveness of the system in different authentication tasks according to their complexity.

A. Gender Identification

Gender identification uses only two labels for training the ANN outputs: male and female. The input feature coefficients correspond to groups of 2, 4, 6 and 8 people, maintaining the number of men equal to the number of women. In addition, the number of feature coefficients used as inputs to the ANN is also varied. Figures 2 and 3 show the identification accuracy for different numbers of LPCs and their combination with other broad metrics. In the figures, 'Fo' stands for fundamental frequency, 'σ' for standard deviation, 'E' for energy, 'K' for kurtosis, and 'All' for the combination of all these 4 extra metrics.

According to the results, 3 LPCs are enough to reach a successful gender identification among a small number of people (*e.g.*, 2 or 4 people). However, if the number of people increases, 3 LPCs plus energy or kurtosis produces identification accuracy rates above 92%.

B. One Person Identification

The main goal of this evaluation is distinguishing a selected person from the rest of the group. In particular, a female person has been selected for identification and a varying number of strangers is added to the group. Figures 4 and 5, summarize the results for different numbers of LPCs and their combination with other already described metrics. Note that in this case, the ANN has only two possible outputs: the person is recognized or rejected.

Again, 3 LPCs seem to be enough to distinguish one person from the rest of the group. Regarding the task of identifying one person against a group of strangers, it can be considered as an specialization of the experiment evaluated in the previous subsection. Instead of having two separate large groups of people, one of the groups contains exactly one person. In any case, the maximum level of identification accuracy in both experiments is reached using 3 LPCs plus the signal energy as inputs to the classifier.

C. People Authentication

The goal of this authentication experiment is to identify every single person within a group of people. In similar form

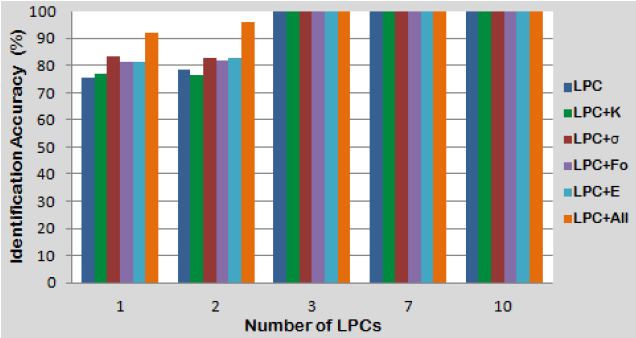

Fig. 6. People authentication considering a universe of 2 people.

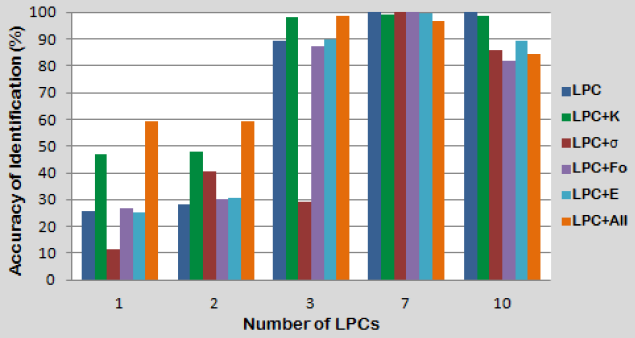

Fig. 7. People authentication considering a universe of 8 people.

to previous experiments, the size of the group is varied from 2 to 8 people. The average identification accuracy for groups of 2 and 8 people is shown in figures 6 and 7, respectively. Be aware that in this case, the ANN has a separate output for each person to be identified.

We can clearly see that the accuracy of the classifier decreases as the number of people to authenticate increases. This can be explained because the decision made by the network uses a simple maximum output value criterion. Thus, increasing the number of decision outputs also increment the possible fragmentation of the output vector.

In general, these results show that an increment in the number of LPCs still has advantage in the interval from 3 to 10 LPCs. Note also that the common relation found in previous experiments for gender and one person identification is also maintained here: 3 LPCs plus kurtosis or energy of the signal provide acceptable identification accuracy levels.

V. RELATED WORK

Linear predictive models are used in a wide range of signal processing applications, such as data forecasting, speech and video coding, speech recognition, model-based spectral analysis, signal restoration and noise reduction [10]. Indeed, modern digital mobile phones employ voice coders based on linear prediction modeling of speech for efficient coding [3]. There are two main motivations for using predictors in applications of signal processing [2]: the first one is forecasting the form of a signal, and the second one is removing the predictable part of signals in order to avoid transmitting those parameters (saving time, bandwidth, power and storage).

In addition, many voice authentication technologies available are geared towards telephony making it easily possible to use remote voice authentication [11].

VI. CONCLUSIONS

This paper shows the effectiveness of LPC-based feature coefficients in tasks of voice authentication. We can conclude that 3 LPCs are enough to distinguish people gender and one person from the rest of a group. Even better identification accuracy can be obtained adding the energy of the signal to the 3 previous LPCs. In the case of several people authentication, identification accuracy rates of the classifier increases with the number of LPCs and adding the energy of the signal frame is also a good booster.

We are planning to extend this work evaluating more deterministic classifiers, like Bayesian belief networks, and include other static and dynamic feature coefficients (*e.g.*, Mel-cepstrum coefficients).

REFERENCES

[1] W. Thanhikam, Y. Satirasombat, and C. Charoenlarpnopparut, "Voice authentication system: Lpc and mel-ceptrum bases with vector quantization." Sirindhorn International Institute of Technology. Thammasat University, Thailand, Tech. Rep., 2011.

[2] S. Bengio and J. Mariéthoz, "A statistical significance test for person authentication." IDIAP, Switzerland, Tech. Rep., 2010.

[3] S. V. Vaseghi, *Advanced Digital Signal Processing and Noise Reduction.*, 3rd ed. John Wiley & Sons, Ltd, 2006.

[4] G. L. Torres, H. G. Martins, C. R. Santos, R. A. Carminati, and W. S. Vieira, "Speech recognition with industrial purposes: An approach using intelligent systems." in *Acoustics, Speech and Signal Processing (ICASSP)*, vol. 2. Instituto de Sistemas Eléctricos e Ingeniería. Universidad de Taubaté (UNITAU), Brasil, 2008.

[5] S. Perez and E. V. Carrera, "Simple speech recognition using kurtosis." *Science and Technology Magazine*, vol. 7, pp. 62–69, 2012.

[6] U. C. Hyungseob Han, Sangjin Cho, "Fault diagnosis system using lpc coefficients and neural network," *Strategic Technology (IFOST), Department of Computer Engineering and Information Technology. University of Ulsan, Korea*, pp. 87–90, 2010.

[7] L. R. Rabiner and R. W. Schafer, *Digital Processing of Speech Signals.* Prentice - Hall, Signal Processing Series, 1978.

[8] A. M. Peinado and J. C. Segura, *Speech Recognition over digital channels- Robustness and Standards.* John Wiley & Sons, Ltd, 2006.

[9] J. Han, M. Kamber, and J. Per, *Data Mining - Concepts and Techniques.*, 3rd ed. Morgan Kafmann Publishers is an imprint of Elsevier, 2012.

[10] S. Cho and C. J. Kuo, "Current developments and future trends in audio authentication." in *Multimedia in Forensics, Security and Intelligence. University of Southern California*, 2012.

[11] E. Bocchieri, D. Caseiro, and D. Dimitriadis, "Speech recognition modeling advances for mobile voice search." in *IEEE International Conference Acoustics, Speech and Signal Processing*, 2011.