Feature Extraction Using LPC-Residual and Mel Frequency Cepstral Coefficients in Forensic Speaker Recognition

Jose B. Trangol Curipe and Abel Herrera Camacho

Abstract—In this paper, we investigated the form to improve the performance in the recognition, involved at the forensic area. To improve, we use Linear Predictive Coding (LPC) and it residual; it was compared with Mel Frequency Cepstral Coefficients (MFCC). The classification technique was Gaussian Mixture Model (GMM). The collection data is in Spanish language, using spontaneous speech from 37 male speaker of Mexican Spanish, we have three recordings, between each recording exist 3 week and one month of separation respectively, this allows us in real condition work, we use non contemporaneous recording and scarcity of data to training and testing the performance of the forensic recognition task. Two conclusions can be drawn from the results, the first, MFCC has better performance with long recording, LPC-residual has better performance with short recording.

***Index Terms*—Forensic speaker recognition (FSR), gaussian mixture model (GMM), linear predictive coding-residual (LPC- residual), mel frequency cepstral coefficients (MFCC).**

I. Introduction

In the last years, the interest in forensic speaker recognition is growing in the research area [1]; also the judicial necessity has grown. Due a la variability of speech, the speaker recognition is one the most difficult task in biometrics identifications. In general, the court wants to known the odds that the suspected speaker has produced the questioned recording, given the circumstances of the case and the observations made by the forensic scientist [1]. A forensic expert has to interpret evidence material in the course of a criminal investigation. In the case of questioned recording (trace), the evidence does not consist in speech itself, but in the quantified degree of similarity between speaker de- pendent features extracted from the trace, and speaker dependent features extracted from recorded speech of a suspect, represented by his/her model [1], [2]. Different methods can be applied to determine if the unknown voice of the questioned recording (trace) belongs to the suspected speaker (source). The most persistent real- world challenge in this field is the variability of speech. There is within-speaker (within-source) variability as well as between-speakers (between-sources) variability. Consequently, forensic speaker recognition methods should provide a statistical probabilistic evaluation, which attempts to give the court an indication of the strength of the evidence, given the estimated within-source variability and the between-sources variability [3].

Manuscript received September 14, 2012; revised November 28, 2012.

The authors are with the Department of Signal Processing, Engineering School, UNAM, Mexico (e-mail: abelhc@hotmail.com).

Forensic speaker recognition shows very good performance in discriminating between voices of speakers under controlled recording conditions and sufficient data. However, the conditions in which recordings are made in investigative activities (e.g., anonymous calls and wire-tapping) cannot be controlled and pose a challenge to automatic speaker recognition [4].

In the automatic speaker recognition, we used LPC-resid and MFCC like feature extraction technique of each speaker, and Gaussian Mixture Models are used to create the models characteristics of signal speech, for future applications in forensic speaker recognition.

II. Voice Database

In the Work 27 Mexican subjects with an age range of 18 to 27, all male, all are university student, all native speakers of Mexican Spanish, using spontaneous speech and read out speech from each one, and none of the speakers present any speech or voice problems, each one, was recorded in three non-contemporaneous recordings, the separation in each recordings was of two week, and one month's respectively.

III. Signal Analysis and Modelling

MFCC is a powerful coding technique [5]. MFCC imitate the ear perception behavior and give, good identification [6]. MFFCC uses a subjective results scale called the 'mel' scale. The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels [5]. The waveform, is applied a pre- emphasis and cut into a number of overlapping segments. A Hamming window is multiplied and the Fourier Transform (FFT) is computed for each frame. The power spectrum is warped according to the Mel-scale in order to adapt the frequency resolution to the properties of the human ear. Then the spectrum is segmented into a number of critical bands by means of a filter-bank typically consists of overlapping.

A. Mel Frequency Cepstral Coefficients

MFCC is a powerful coding technique [5]. MFCC imitate the ear perception behavior and give, good identification [6]. MFFCC uses a subjective results scale called the 'mel' scale. The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. As a reference point, the pitch of a 1 kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels [5].

The waveform, is applied a pre-emphasis and cut into a number of overlapping segments. A Hamming window is

DOI: 10.7763/IJCEE.2013.V5.658

multiplied and the Fourier Transform (FFT) is computed for each frame. The power spectrum is warped according to the Mel-scale in order to adapt the frequency resolution to the properties of the human ear. Then the spectrum is segmented into a number of critical bands by means of a filter-bank typically consists of overlapping triangular filters. A discrete cosine transformation (DCT) applied to the logarithm of the filter-bank outputs results in the raw MFCC vector [7] triangular filters. A discrete cosine transformation (DCT) applied to the logarithm of the filter-bank outputs results in the raw MFCC vector [7].

B. Linear Predictive Coding

Linear predictive analysis is one of the most powerful and widely used speech analysis techniques. The importance of this method lies both in its ability to provide accurate estimates of the speech parameters and in its relative speed of computation [8]. LPC analysis is based on the assumption that the speech signal can be characterized by a predictor model which looks at past values of the output alone; hence it is an all pole model in the Z transform domain [9].

C. LPC-Residual

The prediction residual signal, according to the LPC model.

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum \alpha_k s(n-k) \quad (1)$$

where α_k, are the LPC predictor coefficients, $s(n)$, are the samples of the speech signal. It is evident that $e(n)$, might contain information which has not been captured by the LPC coefficients and which can be useful for the speaker recognition task [10].

IV. Automatic Speaker Recognition

Over the past several years, Gaussian Mixture Models (GMMs) have become the dominant approach for modeling in text independent speaker recognition applications [11]. For text independent speaker recognition, where is no prior knowledge of what the speaker will say, the most successful likelihood function has been Gaussian Mixture Models [11]. The parametric modeling capabilities of the GMM allow it to model any arbitrarily shaped probability density function (pdf) with a weighted sum of M component Gaussian densities [1].

For a D-dimensional feature vector;

$$p(x/\lambda) = \sum_{i=1}^{M} w_i p_i \quad (2)$$

$w_i = 1,2,3,..,M$, are the mixture weights and $p_i = 1,2,3...,M$, are the component density. Each component density is a D-variate Gaussian function of the form:

$$p_i(x) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} exp\left\{-\frac{1}{2}(\vec{x}-\vec{u}_i)\sum_i^{-1}(\vec{x}-\vec{u}_i)\right\} \quad (3)$$

where u_i, is the mean, $\sum_i$, is the covariance matrix, the mixture weights satisfy the constraint that $\sum_{i=1}^{M} w_i = 1$, where $w_i \geq 1$, are the mixture weights. Therefore a GMM consisting of M Gaussian, and can be specified by:

$$\lambda = \{p_i, u_i, \sum_i\} i = 1,2,3...M \quad (4)$$

For speaker identification, each speaker is represented by a GMM and is referred to by his/her model λ [12].

V. Procedure

Speech data was collected from 47 young male speakers, they were aged between 18 and 27, they are university student, they were recorded three times, with the second a third recording sessions being approximately three week and one months after the first session. Each recordings has approximately a duration of 140 seconds, in all recordings, is extracted noise background, in the place where not exist speech signal, The resulting recordings were edited by hand to eliminate speech portions where the structure was unclear, the first and the second recording, is used to create the training models of each speaker in the base data, the third recording, is used to testing the recognition performance. To work in real conditions, was used short time to training, and create the models from each speaker in the automatic speaker recognition, we use, 15, 20 and 25 seconds, to create more realistic models. This can be considered as a plausible scenario in forensic casework regarding bandwidth limiting and signal duration [13].

The first aim in this work, was compare the performance, using parameterization techniques, in this paper was used MFCC front-end and LPC-residual. Each vector was created using short time of training, to create real conditions forensic work. The second aim is work using vowels of each recordings, with this we can obtain important information about the vowels, and which of them is useful to forensic applications, for this point, each vowel was extracted in each one of the recordings to create training and testing models, to know which of vowels is the best in forensic applications. 13 coefficients is used in both, LPC-residual and MFCC, to create the vector. Determining the number of component *M* in a mixture needed to model a speaker adequately is an important by difficult problem [12]. The number of mixtures was *M* = 128 to this work to this base data.

VI. Results

Order to assess MFCC and LPC-residual, in the feature extraction; we use short time in training, to evaluate the future applications in forensic speaker recognition. In the Table 1 we can see, the results in a comparison using the first and the second recording to create training models, and the third to test the automatic speaker recognition.

TABLE I: Recognition Using MFCC.

MFCC	Testing	Testing	Testing
Training	10 s	20 s	30 s
10s	57,69	69,23	76,92
15s	61,53	73,07	80,76
20s	80,76	84,61	92,30

TABLE II: RECOGNITION USING LPC-RESIDUAL.

MFCC	Testing	Testing	Testing
Training	10 s	20 s	30 s
10s	61,53	69,23	76,92
15s	69,23	76,92	80,76
20s	84,61	84,61	88,46

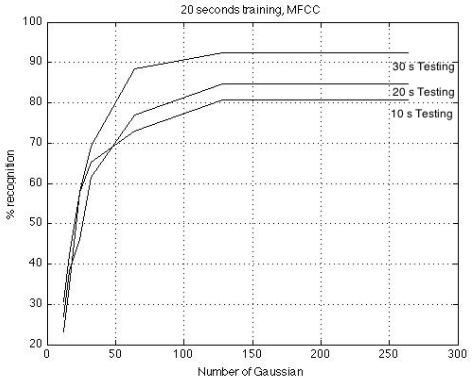

Fig. 1. MFCC using 20 s of training, and 10, 20 and 30 seconds of testing.

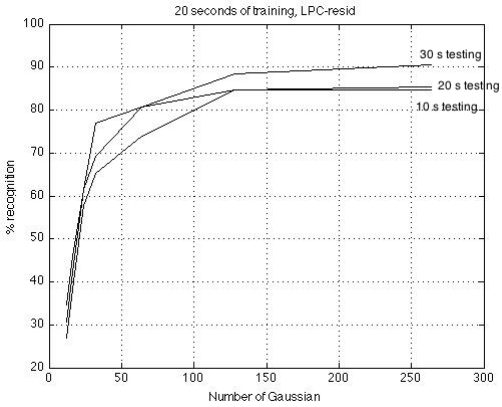

Fig. 2. LPC-residual, using 20 s of training, and 10, 20 and 30 seconds of testing.

In the Fig. 1 and 2, we can see that when we use 128 Gaussian obtain maximum recognition for this database.

The Fig. 3 shows the average frequency of occurrence of vowels in the recordings, the vowel /a/ is the highest frequency of occurrence, and Fig. 4 shows that the maximum recognition is obtained when using 128 Gaussian.

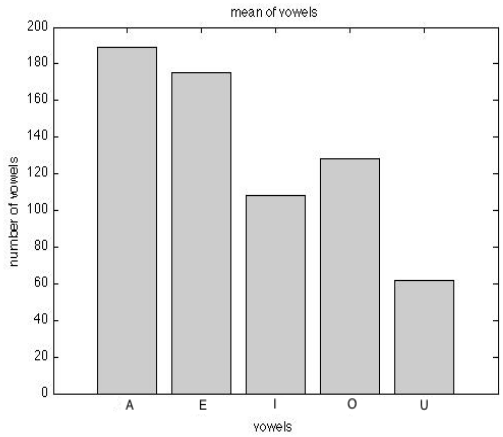

Fig. 3. Mean value of vowels

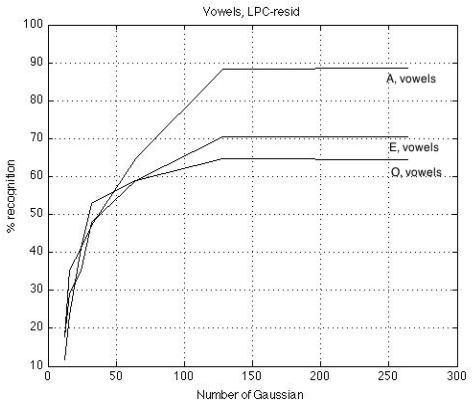

Fig. 4. LPC-residual, vowels recognition.

VII. CONCLUSION

These results allow us to continue this research project, and considering the LPC-residual technique alone for extracting characteristics of the speaker and be part of a forensic recognition system. The scientist forensic could design a speaker recognition system that can extract information characteristic of the speaker and provide an objective result in the identification of a speaker in a recording forensic. In the same time, is important see the improvement of MFCC when is used more long time to training and create the statistical models, MFCC overcoming LPC-residual. For other hand, LPC-residual overcoming MFCC, when short time is used, to train and create the statistical models from each one of speakers. This is a very important point, due to the scenario in forensic casework. In the same time this methodology presents discrimination significant in different speaker recognition.

A preliminary result, using the vowels in each recording, is obtained, and due to the great amount working, 17 speaker is using to evaluate the vowels, the vowel /a/, present the best performance, the vowels /u/ has been exclude because of its low frequency of occurrence in Spanish [14], [15], and the vowels /i/, is exclude, because has less in comparison with the others vowels, but in a future work will be used.

ACKNOWLEDGMENT

The report described is an on-going research project supported by UNAM-DGAPA-PAPIIT (IT116811-2), the authors are deeply grateful for this support.

REFERENCES

[1] S. D. Meuwly and A. Drygajlo, "Forensic Speaker Recognition Based on a Bayesian Framework and Gaussian Mixture Modeling (GMM)," *The International Symposium on Computer Arquitecture, ISCA, The Speaker Recognition Workshop*, pp. 145-150, 2001.

[2] S. R. Lewis, "Philosophy of Speaker Identification, presented at Police Applications of Speech and Tape Recording Analysis," in *Proceedings of the Institute of Acoustics*, 1984.C. Y. Lin, M. Wu, J. A. Bloom, I. J. Cox, and M. Miller, "Rotation, scale, and translation resilient public watermarking for images," *IEEE Trans. Image Process.*, vol. 10, no. 5, pp. 767-782, May 2001.

[3] A. Drygajlo, D. Meuwly, and A. Alexander, "Statistical Methods and Bayesian Interpretation of Evidence in Forensic Automatic Speaker Recognitio," *European Conference on Speech Communication and Technology Eurospeech*, pp. 689-692, 2003.

[4] A. Alexander, "Forensic Automatic Speaker Recognition using Bayesian Interpretation and Statistical Compensation for Mismatched

Conditions," Ph.D. thesis, Swiss Federal Institute of Technology at Lausanne, 2005.
[5] P. Bansal, A. Dev, and C. B. Jain, "Automatic Speaker Identification Using Vector Quantization," *Asian Journal of Information Technology*, pp. 938-942, 2007.
[6] A. Zulfiqar, A. Muhammad, and M. Enriquez A. M., "A Speaker Identification System Using MFCC Features with VQ Technique," *Third International Symposium on Intelligent Information Technology Application*, pp. 115-118, 2009.
[7] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, pp. 73-76, 2001.
[8] L. R. Rabiner and R. W. Schafer, "Introduction to Digital Speech Processing," *Oundations and Trends in Signal Processing*, vol. 1, no. 1–2, 2007.
[9] J. L. Ostrander, T. D. Hopmann, and E. J. Delp, "Speech Recognition using LPC Analysis," *Robot System Division, College of Ingineering, The University of Michigan*, January, 1982.
[10] K. P. Markov and S. Nakagawa, "Text-independent speaker recognition using multiple information sources," *International Conference on Spoken Language Processing ICSLP-1998*, pp. 0744, 1998.
[11] D. A. Reynolds, T. F. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1-3, pp.19-41, 2000.
[12] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," *Speech and Audio Processing, IEEE Transactions on* , vol.3, no.1, 1995, 72-83.
[13] T. Becker, M. Jessen, and C. Grigolas, "Forensic Speaker Verification Using Formant Features and Gaussian Mixture Models," Interspeech Special Session: Forensic Speaker Recognition, Traditional and Automatic Approaches, *Annual Conference of the International Speech Communication Association, interspeech* 2008.
[14] G. Rojo, "Frecuencia de fonemas del español actual, in Brea, M. Fernandez Rei. Santiago de compostela: universidad de santiago Compostela," *Servicio de publicacion e intercambio cientifico*, pp. 451-467, 1991.
[15] V. Marrero, E. Battaner, J. Gil, J. Llisterri, M. Machuca, M. Marquina, C. de la Mota, and A. Rios, "Identifying speaker-dependent acoustic parameters in Spanish vowels," in *Proceedings of Acoustics Acoustical Society of America, European Acoustics Association*, pp. 5673-5677, 2008.

Trangol C. Jose Benito was born in Santiago, Chile at December 11, 1975, He received M.S. degree in Electricity Engineering 2008, and actually is PhD Student at the Universidad Nacional Autonoma de Mexico. He is currently working in forensic speaker recognition and speech processing.

Herrera C. Abel received degrees in Mechanical-Electrical of Engineering, M. S. Electronic Engineering, and the Ph.D. Engineering, from Universidad Nacional Autonoma de Mexico (UNAM), Mexico, in 1979, 1985, and 2001, respectively, the Ph.D. was with support of the University of California in Davis. He did a postdoctoral research in 2001 at Carnegie Mellon University, and a sabbatical research at USC. He is author from more than 50 scientific papers on codification, recognition, and synthesis. He joined the department of Engineering of UNAM and is professor from 1979. He currently is the director of speech laboratory in the faculty of Engineering of UNAM.