Low Variance Non-Contemporaneous Recordings Using Multitaper Analysis in Forensic Applications

Jose Trangol and Abel Herrera

***Abstract*—In the last time, the forensic speaker recognition has focused the attention and interest of the scientific community, the voice signal present a significant challenger due to its large variability, but in the forensic science, the variability is even more complex due to multiple factors that are present in this area of forensic biometrics, e.g., short time recordings, noise environmental, channel mismatched, non-contemporaneous recordings and so on.**

In this paper we study the performance of the Mel Frequency Cepstral Coefficients, using Multitaper Spectrum Estimate with promising preliminary results. This technique reduces the variance and improves the performance in forensic applications. In this work we focus in non-contemporaneous recordings, to modelling we use Gaussian Mixture Models.

***Index Terms*—Forensic speaker recognition, GMM, MFCC, multitaper analysis.**

I. Introduction

Nowadays, the forensic speaker recognition has been obtained from the scientific community to solve the problems that arise in this biometric area [1], [2]. Some of the challenges that arise are due to the lack of consensus on methods of analysis, and the lack of understanding on the subject [3]. In forensic speaker recognition, we have a recording, which known, as offender voice, this recording is obtained through the police intelligence or from some systems recording, for the other hand, we have to a recording with a known voice, usually a suspect [4]. Some systems, especially commercial speaker recognition, work very well in lab condition (under control), but when work in forensic real conditions (don't exist control), resulting in yield reduction; we work far to the lab condition. The limited amount of speech data obtained in forensic recordings is a problem that is all recognition system. In general, forensic cases exhibit a small amount of voice data, in many occasions, far from ideal conditions [5], and have noted that a speaker can show large variation occasion to occasion, and even within a single recording session [6], the variability is more complex when used non-contemporaneous recordings (common recordings). Variability is one of the major problems that occur in speech processing and in forensic situation the variability is still more. All these factors highly degrade the speaker performance process.

In these paper, our focus is in control variability, reduced the variability in non-contemporaneous recordings and use short time recordings to create real forensic conditions. The spectral variance can be reduce, if using a method knows as multi-taper spectral estimation, this consists in replace the traditional method which used one window (traditionally Hamming), by one of multiple windows [7]. Mel Frequency Cepstral Coefficients (MFCC) is used to extract information from the speech signal. It is one of the feature extraction techniques popular and is used in the speaker recognition and speech recognition, MFCC has obtained significant returns in both tasks (speech and speaker recognition) [8]. The aim of the research, statistically speaking, is to obtain MFCC's with small variance, and on average the estimated cepstrum should be similar to the original and small bias [9]. The multitapering aim is to analyze the input signal using different windows, then estimate the resulting spectrum as an average of the individual sub-spectral [9]. The multitaper method has been used in various research areas and has shown acceptable yields improved the traditional method [7]. In this work, different tapers are evaluated, such as Thomson [10], sine [11], and multipeak [12], to reduce the variance in non-contemporaneous recording to future forensic applications.

II. Voice Database

In the experiment, 35 Mexican male speakers were recorded, from 18 to 30 years, all university students and native speakers, 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 recording was two week, and one month's respectively.

III. Forensic Speaker Recognition

Biometrics is a scientific discipline that aims to capture relevant information of the individual, in the context of law enforcement [13]. In many situations, a voice recording is a key element [14]. The speaker recognition is a process aims to identify people by their voices [15]. This biometric application, is a difficult task, and is a recognition application that lacks a complete understanding [16]. The role of forensic science is the provision of information to help answer question of importance to investigators and to court of law, The goal, is identify if the questioned recording (trace) and suspect speaker have the same source [17].

IV. Feature Extraction and Parametrization

The speech signal is a signal having a large variation, due

Manuscript received March 20, 2013; revised June 28, 2013.

The authors are with Departamento de Procesamiento de Señales, Facultad de Ingeniería, UNAM(e-mail: jtrangol@yahoo.com, abelhc@hotmail.com)

DOI: 10.7763/IJMO.2013.V3.312

to the articulatory movements; therefore, the signal must be analyzed using short time windows, usually 20-30 ms duration with 50% of overlapping [18]. By using this window interval, the signal is called pseudo stationary and a spectral feature vector is extracted from each frame [19]. Speech parametrization consists in transforming the speech signal to a set of feature vectors [20]. To parameterization the signal we used MFCC.

A. Mel Frequency Cepstral Coefficients (MFCC)

MFCC is a powerful coding technique [22]. MFCC imitate the ear perception behavior and give, good identification [23], MFCC uses a subjective results scale called the mel scale [22]. The mel-frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz [22]. The waveform, is applied a pre-emphasis and cut into a number of overlapping segments. 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) is applied to the logarithm of the filter-bank [24]. This DCT results in the raw MFCC vector [24].

B. Multitaper Analysis

Typically, the spectrum for MFCC is estimated using the tradicional method using a single periodogram window. The bias can be reduced by windowing the time series with, for example, a Hamming window [25], [26]. The windowed periodogram has low bias in general, but it still suffers from high variance [25], [26]. A multitaper spectral estimator is an average of windowed periodograms using different orthogonal windows (aka *tapers)*, e.g. the *Thomson* [10], the *sine* [11], and the *multipeak* multitapers [12]. The Fig. 1 shows the different tapers in Multitaper method. The multitaper spectrum estimator has low variance [27]. The multitaper MFCC estimator lacks researching, especially for its statistical properties [27].

C. Spectrum Estimate MFCC Using Multitapering

The power spectrum estimation method for speech processing applications widely used is a windowed [8]:

$$\hat{S}_d(m,k)=\left|\sum_{j=0}^{N-1} w(j)s(m,j)e^{-\frac{2i\pi j}{N}}\right|^2 \quad (1)$$

where $k \in \{0,1,....,K-1\}$, denotes the discrete frequency index and $w(j)$ is a time-domain window function [9]. The Hamming window reduces the bias of the spectrum, but still has large variance [9]. To reduce the variance of the MFCC estimator is using the multitaper spectrum estimate [8, 25].

$$\hat{S}_{MT}(m,k)=\sum_{p=1}^{M}\lambda(p)\left|\sum_{j=0}^{N-1} w_p(j)s(m,j)e^{-\frac{2i\pi j}{N}}\right|^2 \quad (2)$$

The equation (2) is the multi-taper spectrum estimator, where N is the frame length, w_p is the *pth* data taper used for the spectral estimate, M denotes the number of tapers and $\lambda(p)$ is the weight corresponding to the *pth* taper.

$$\sum_j w_p(j)w_q(j)=\delta_{pq} \quad (3)$$

The multitaper spectrum estimate is therefore obtained as the weighted average of M individual sub-spectra [8].

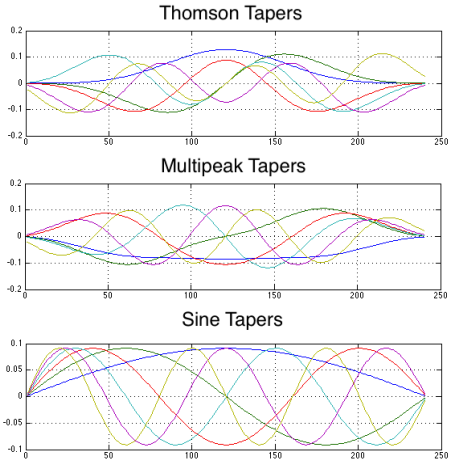

Fig. 1. Different tapers, multitaper method, *M=6.*

V. Modeling

A speaker recognition system should create a model λ_i for speaker's using speech signal from a determined speaker [14]. That is, given the feature representation obtained through the short-term acoustic analysis, a speaker model is required in order to perform the pattern matching and to make some statistical measure as to the likelihood of the observed features being produced by claimed identity. The basis of Gaussian Mixture Models (GMM) is the feature model by a mixture of Gaussian densities [26].

In past years, GMM's have become the baseline for a speaker recognition, using a text independent system [27]. 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 [28].

For a D-dimensional feature vector;

$$p(x/\lambda)=\sum_{i=1}^{M} w_i p_i \quad (4)$$

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

$$p_i(x)=\frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}}\exp\left\{-\frac{1}{2}(\vec{x}-\vec{u}_i)\Sigma_i^{-1}(\vec{x}-\vec{u}_i)\right\} \quad (5)$$

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

$$\lambda=\{p_i,u_i,\Sigma_i\} \qquad i=1,2,3...M \quad (6)$$

For speaker identification, each speaker is represented by a GMM and is referred to its model λ [28].

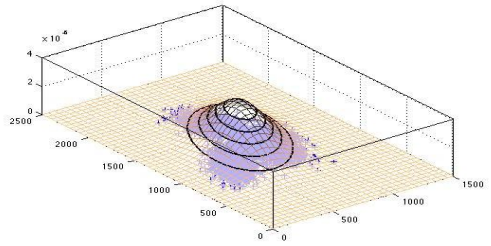

Fig. 2. Example of the GMM for the word 'auto', M=128.

VI. Procedure

Speech data was collected from 35 young male speakers from 18 to 27 years, 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 recording has approximately a duration of 140 seconds it 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 are used to create the training models of each speaker in the base data, the third recording are used to testing the performance. To work in real conditions, short time at training was used. To create the models from each speaker in the speaker recognition system, we use 15 and 30 seconds, to create more realistic models. This can be considered as a forensic scenario regarding limiting and signal duration.

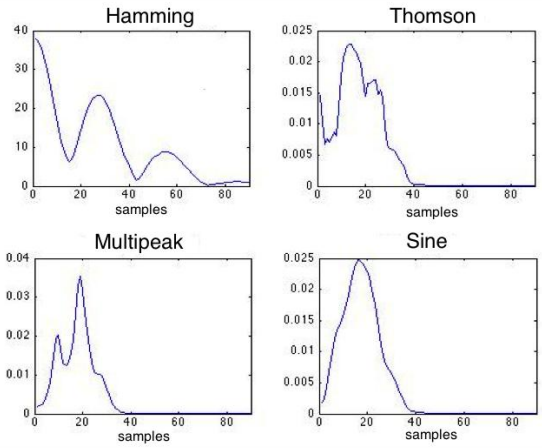

Fig. 3. Hamming analysis and multi-tapers, applied to one portion signal.

The first aim in this work was correctly model the data to train the system, for this we use different number of Gaussians, 128, 256 and 512 was use in this basedata. Determining the number of component M in a mixture is an important and difficult problem [28]. The second aim is work using cepstral coefficients using the traditional method and the multi-tapers method to compare the performance, in the traditional method we use Hamming windows periodogram; in the multi-taper case, we use Thomson, sine and multi-peak. To know which one has the best performance in forensic situations, 13 coefficients are used in MFCC using multi-taper analysis.

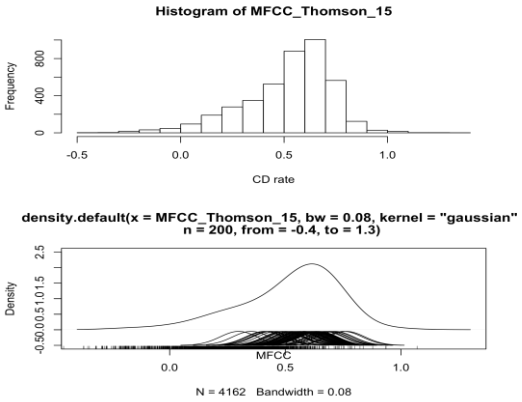

Fig. 4. MFCC Distribution and Kernel Density Estimation, Thomson 15s test, k=8.

VII. Results

With the experiments in this paper, it is remarkable, the difference between the traditional method and the multitaper spectrum estimate. In the Fig. 3, we show multitaper and traditional analysis, in the top left, hamming window, top right Thomson taper, bottom left multi-peak taper, and bottom right sine taper, is evident the difference.

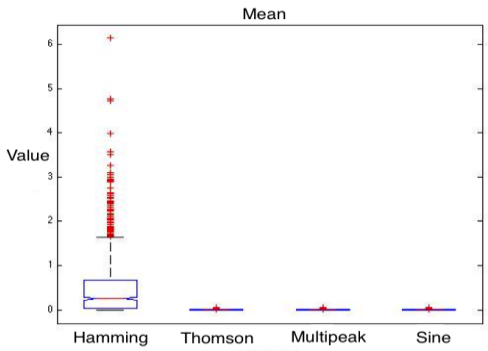

Fig. 5. Variance from recordings, 15 seconds, from left to right, Hamming, Thomson, sine, multi-peak.

The Fig. 4 shows the kernel density estimation, using 15 second of recording from Thomson multitaper, in the traditional case the distribution data is more large in comparison with the multitaper method. In the Fig. 5, we see the large variance in the data's from Hamming, comparing it with the different multitapers. Multitaper reduces the variance in compare with Hamming windows. The Table I shows the result from an evaluation using 35 peoples from Spanish language, this experiment was used a text-independent system. Multitaper method outperforms the traditional method.

VIII. Conclusion

In this paper, the multi-taper method using MFCC obtains betters results than classical windowing, for forensic applications. Thomson taper is the type with better performance; the necessary number of tapers is seven to ten.

When used more time in the recording, we have low variance using multitaper analysis, if we have less variance, we can create more accurate GMM models. Due to the base

data and the experiment without likelihood ratio, to evaluate forensic situation, more tests should be performed.

TABLE I: RECOGNITION HAMMING VS. MULTITAPER.

Train test	15 sec Hamming	Thomson	Multipeak	sine
30 sec	77.14 %	88.57 %	82.85 %	82.85 %
45 sec	80 %	91.57 %	85.71 %	88.57 %

ACKNOWLEDGMENT

The report described is an on-going research project supported by UNAM-DGAPA-PAPIME (PE105311), the authors are deeply grateful for this support.

REFERENCES

[1] P. Rose, "Forensic speaker discrimination with Australian English vowel acoustics," *Saarbrucken*, pp. 6-10, August, 2007.
[2] M. J. Saks and J. J. Koehler, "The coming paradigm shift in forensic identification science," *Science* 309/5736, pp. 892-895.
[3] P. Rose, "Technical forensic speaker recognition: Evaluation, types and testing of evidence," *Computer Speech and Language*, vol. 20, pp. 159-191, 2006.
[4] P. Rose, "Forensic speaker recognition at the beginning of the twenty-first century- an overview and demonstration," *Australian Journal of Forensic Sciences*, vol. 37, pp. 49-72, 2005.
[5] D. Ramos and J. Gonz ález-Rodr íguez, "Addressing database mismatch in forensic speaker recognition with Ahumada III: a public real-casework database in Spanish," presented at Inter speech, Special Session: Forensic Speaker Recognition – Traditional and Automatic Approaches, Brisbane, Queensland, Australia, 2008.
[6] Y. Kinoshita, S. Ishihara, and P. Rose, "Beyond the Long-term Mean: Exploring the Potential of F0 Distribution Parameters in Traditional Forensic Speaker Recognition," presented at Odyssey: The Speaker and Language Recognition Workshop Stellenbosch, South Africa, pp. 21-24, 2008.
[7] Md. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multitaper MFCC and PLP festures for speaker verification using i-vectors," *Speech communication*, vol. 55, pp. 237-251, 2013.
[8] Md. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, and D. O'Shaughnessy, "Multi-taper MFCC feature for speaker verification using I-vectors," *Automatic Speech Recognition and Understanding IEEE Workshop*, pp. 11-15, 2011.
[9] T. Kinnunen, R. Saeidi, J. Sandberg, and M. Hansson-Sandsten. What else is new than the Hamming window? Robust MFCC for Speaker Recognition via Multitapering. (2010). *CiteSeer*. [Online]. Available: http://citeseerx.ist.psu /viewdoc/similar?=10.1.1.178.5674&type=ab
[10] D. J. Thomson, "Spectrum estimation and harmonic analysis," in *Proc. the IEEE*, vol. 70 no. 9, 1055-1096, 1982.
[11] K. S. Riedel and A. Sidorenko, "Minimum bias multiple taper spectral estimation," *IEEE, Trans. On Signal Proc.*, vol. 43, no. 1, pp. 188-195, 1995.
[12] M. Hansson and G. Salomonsson, "A multiple window method for estimation of peaked spectra," *IEEE T. on Sign. Proc.*, vol. 45, no. 3, 778-781, 1997.
[13] J. S. Dunn and F. Podio. (2007). Biometrics Consortium. [Online]. Available: http://biometrics.org
[14] *Springen Handbook of Speech Processing*, Springer-Verlag Berlin Heidelberg 2008, ch. 36, pp. 737.
[15] J. Ortega-Garcia, J. Gonz ález-Rodr íguez, and S. Cruz, "Speech variability in automatic speaker recognition system for commercial and forensic purposes," *IEEE AES System Magazine*, November, 2000.
[16] P. Rose, "Forensic speaker recognition at the beginning of the twenty-first century- an overview and demonstration," *Australian Journal of Forensic Sciences*, vol. 37, pp. 49-72, 2005.
[17] A. Drygajlo, "Forensic Speaker Recognition, Law Enforcement and Counter-Terrorism," in *Automatic speaker recognition for forensic case assessment and interpretation*, A. Neustein, H. A. Patil, eds., 2012.
[18] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*,vol. 3, no. 1, pp. 72-83, 1995.
[19] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: form feature to supervectors," *Speech Communication*, vol. 52, pp. 12-40, 2010.
[20] F. Bimbot, J. F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garc á, D. Petrovska-Delacr étaz, and D. Reynolds, "A tutorial on text-independent speaker verification," *EURASIP Journal on Applied Signal Processing*, pp. 430-451, 2004.
[21] P. Bansal, A. Dev, and C. B. Jain, "Automatic speaker Identification using Vector Quantization," *Asian Journal of Information Technology*, pp. 938-942, 2007.
[22] A. Zulfiqar, A. Muhammad, and M. Enriquez, "A Speaker Identification system using MFCC Features with VQ Technique," in *Proc. third international Symposium on Intelligent Information Technology Application*, pp. 115-118, 2009.
[23] S. Molau, M. Pitz, R. Schluter, and H. Ney, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," in *Proc. IEEE int. Conference on Acoustics, Speech and Signal Processing*, pp. 73-76, 2001.
[24] J. Sandberg, M. Hansson-Sandsten, T. Kinnunen, R. Saeidi, P. Flandrin, and P. Borgnat, "Multitaper Estimation of Frequency-Warped Cepstra with Application to Speaker Verification," *IEEE Signal Processing Letters*, vol. 17, no. 4, pp. 343-346, 2010.
[25] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of Audio Signal using AANN and GMM," *Applied Soft Computing*, vol. 11, pp. 716-723, 2011.
[26] 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.
[27] S. D. Meuwly and D. Drygajlo, "Forensic Speaker Recognition based on a Bayesian Framework and Gaussian Mixture Modeling (GMM)," in *Proc. The International Symposium on Computer Arquitecture*, The Speaker Recognition Workshop, pp. 145-150, 2001.
[28] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using Gaussian Mixture Speaker Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

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

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