Traditional Method and Multi-Taper to Feature Extraction Using Mel Frequency Cepstral Coefficients

Jose Trangol and Abel Herrera

***Abstract*—This paper makes a comparison of two methods in speech processing that aims to obtain cepstral coefficients that represent characteristics of each speaker, in order to obtain a greater difference between each speaker (to build a representative model, in a next step), traditionally feature extraction is performed through a single periodic window function (generally hamming windows), we compare the traditional method with multi-taper, which has been used in verification speakers tasks and language recognition. The results obtained using 40 speakers indicate a reduction of the variance by using multi-taper and better separation between speakers; with this we have more robust coefficients. To feature extractions we use Mel Frequency Cepstral Coefficients.**

***Index Terms*—Feature extraction, mel frequency cepstral coefficients, multi-taper.**

I. Introduction

In the last decades the voice signal processing has been subject of study by different research groups in the world. When human being wants to convey a linguistic message, not only conveys the linguistic message, also conveys characteristic features from the speaker that is the result of the interaction of several physiological, psychological and environmental factors [1], [2]. These multiple influences will be simultaneously present in each speech act, and some or all of them will contain specificities of the speaker [2]. A technology, which use speech data is speaker recognition, this encompasses verification and identification systems [3]. The first verify a person's claimed identity from his voice from a particular enrolled speaker, the second, identification; associating an unknown voice with ones from a ser of enrolled speakers [4], the systems could be working in text-dependent (constraint to a specific word, the recognition phrases are fixed or prompted) or in text independent (there are no constraints on the words which the speakers are allowed to use) utterances may have completely different content, and the recognition system must take this phonetic mismatch into account [5]. These technology (This technologies), work under ideal conditions and have obtained good yields, but when working far from ideal conditions, its performance is reduced. One application of this technology is in forensic speaker recognition, but in this application the system should work in adverse conditions (short time recording, mismatched conditions, etc.).

In a recognition system, there are three key points; i) a feature extraction unit for representing speech signal in a compact manner, ii) a modeling scheme to characterize those features using statistical approach (3), iii) a classification scheme for characterizing the unknown utterance [6]. Fig. 1; categorizes the level information from speech signal [5], we use short-term spectral information.

Point of statistical modeling is dependent on the extraction of features [6], therefore, the step of the extraction of features is an important issue, so this research work seeks to find which of the two methods; Hamming windows or multi-taper method can generate better cepstral coefficientes. In the present work, we used Mel Frequency Cepstral Coefficients to extract features of each speaker; this information will allow us to differentiate between speakers.

The idea in multitapering, is to pass the analysis frame through different windows functions and form the final spectrum estimate as a weighted average of the individual sub spectra [7]. The multi-taper method has been widely used in geophysical applications, verification speaker [8], [9], speech recognition [10], and it has been show shown to outperform the windowed periodogram [8]. It has also been used in speech enhancement applications and, recently in speaker recognition with promising preliminary results [8].

We performed a comparative study between two methods the traditional and multi-taper; the traditional method use a single analysis window, usually Hamming window; this window is applied to the whole signal and they have high variance [7]. On the other side we have the multi-taper method, this method has low variance [8], and in this case we apply different windows to the whole signal, we use Thomson, sine, multi-peak in our comparative study.

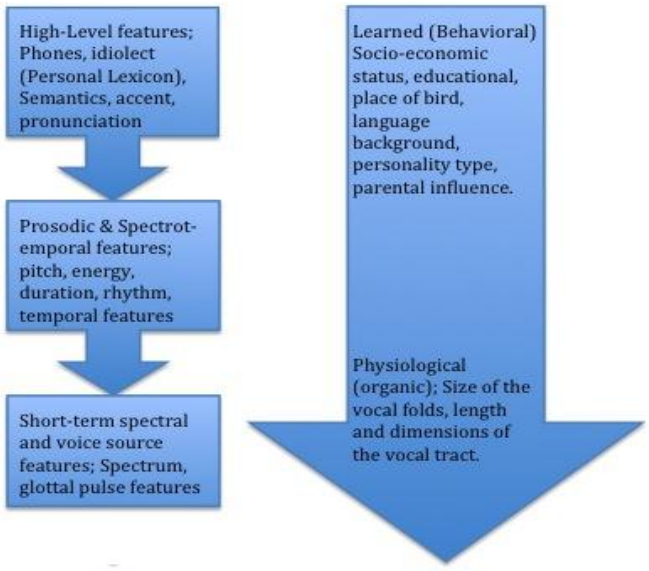


Fig. 1. Scheme of speech.

Manuscript received March 21, 2014; revised June 24, 2014.

Jose Trangol and Abel Herrera are with Department of Signal Processing, Engineering School, Universidad Nacional Autonoma de Mexico, Mexico (e-mail: abelherrerac1@gmail.com).

II. Data

For our aim, in this study we use a corpus creates in the Speech Laboratory from Universidad Nacional Autonoma de Mexico (UNAM). This database contains 40 speakers; Spanish language, all male, university students 18-27 years old, they were recorded two times, the second recording have three week after the first session.

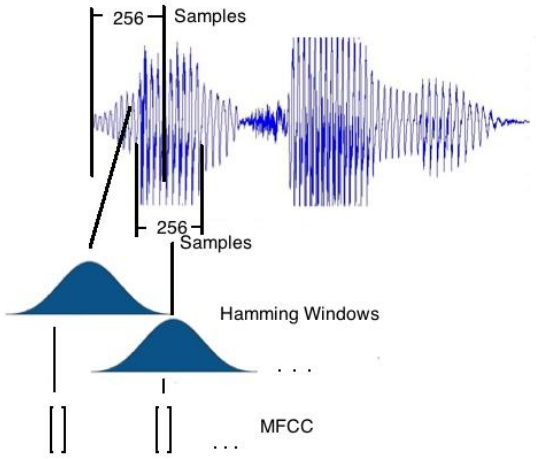


Fig. 2. Traditional method.

III. Mel Frequency Cepstral Coefficients

To perform spectral analysis, signal must show stationary properties [4]. The speech signal continuously changes due to articulator's movements [4]. The analysis of the speech signal is done locally by the application of windows whose duration in time is shorter than the whole signal, see Fig. 2, and for the parameterization of the speech signal we use Mel Frequency Cepstral Coefficients (MFCC).

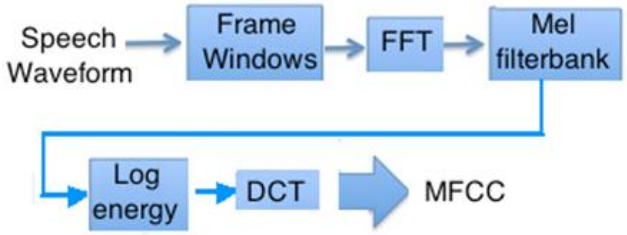


Fig. 3. Mel Frequency Cepstral Coefficients.

MFCC is a most popular technique in feature extraction [10], in part to the robustness in the process. Using the traditional method, we use a single periodic window, usually the hamming windows [4], in short frames of about 20–30 ms in duration, 50% overlapping [7]. Within this interval, the signal is assumed to remain stationary and a spectral feature vector is extracted from each frame [5], see Fig. 2.

MFCC imitate the ear perception behavior and give, good identification [12]. 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 [13]. 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, see the Fig. 3.

IV. Multi-Taper Method

A multi-taper spectral estimator is an average of windowed periodograms using different orthogonal windows (aka tapers), e.g. the Thomson, the sine, and the peak matched multi-tapers. The multi-taper spectrum estimator has low variance [8].

A Hamming-windowed DFT spectrum is the most often used power spectrum estimation method for speech processing applications.

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

Where, $f \in \{0,1,...,N-1\}$, denotes the discrete frequency index and $W = [w(0),.....,w(N-1)]^T$, is a time-domain window function which usually is symmetric and decreases towards the frame boundaries (e.g. Hamming) [10]. With this method the variance is high [14]. One way to reduce the variance of the MFCC estimator is to replace a windowed periodogram estimate by a so-called multi-taper spectrum estimate [9], [14].

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

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 $\hat{S}_{MT}(.)$, which is also called the pth eigenspectrum, M denotes the number of tapers and $\lambda(p)$ is the weight corresponding to the pth taper, see Fig. 4.

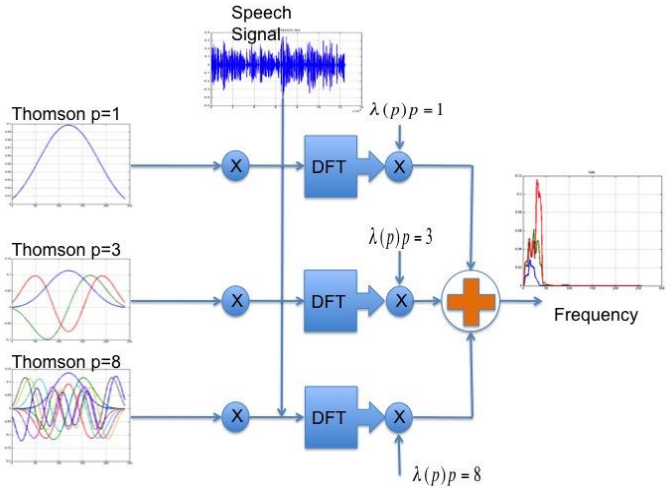


Fig. 4. Multi-taper method.

V. Procedure

Speech data was collected from 40 male speakers, they were aged between 18 and 27, university student, they were recorded two times, the second session recording is approximately three weeks after the first session. The first

recording has a duration of 40 seconds, in the second recording was extracted only a phrase that is repeated 5 times in each of the recordings for each speaker. In all recordings, is extracted background noise, in the place where not exist speech signal; the resulting recordings were edited by hand to eliminate speech portions where the structure was unclear (first session recording).

In the Fig. 4, we can see the block diagram for the MFCC coefficients, from two procedures; a) the traditional method, b) the multi-taper method

VI. Results

For our comparative study, we create the multitaper method using the different type of taper to processing the speech signal with the purpose to extract the MFCC; Thomson, sine and multipeak, was used. Hamming windows is used in the traditional method, in the case of multi-taper we used eight different analysis windows per frame (240 samples, 50% overlap, 14 MFCC coefficients, the same values was used with the traditional method), see Fig. 5.

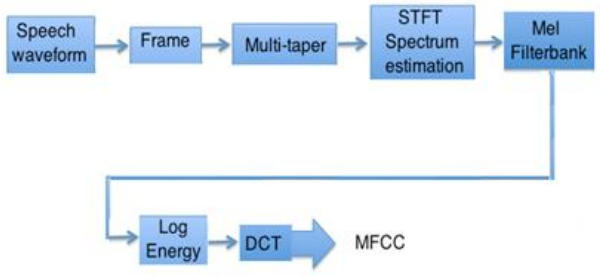


Fig. 5. Block diagram traditional and multi-taper.

TABLE I: Thomson Multi-Taper, Variance Distance between Speakers

	Thomson					
	SP1	SP2	SP3	SP4	SP5	SP6
SP1	0	0.0248	0.0045	0.0083	0.0499	0.0473
SP2	0.0248	0	0.0203	0.0331	0.0251	0.0225
SP3	0.0045	0.0203	0	0.0128	0.0454	0.0428
SP4	0.0083	0.0331	0.0128	0	0.0582	0.0556
SP5	0.0499	0.0251	0.0454	0.0582	0	0.0026
SP6	0.0473	0.0225	0.0428	0.0556	0.0026	0

Extensive experiments have been conducted to examine the speech, and robustness of the procedures, each phrase have duration of 1-second approx. to measure the relationship between each of the processing windows we use a Euclidean distance.

In the Table I, we can see the relationship using only the Thomson windows. In the Table I and the next, only show 6 speakers variance values; tables with 40 speakers are very long to show.

TABLE II: Multi-Peak Multi-Taper, Variance Distance between Speakers

	Multi-peak					
	SP1	SP2	SP3	SP4	SP5	SP6
SP1	0	0.0442	0.0092	0.2023	0.0512	0.0493
SP2	0.0442	0	0.0534	0.2465	0.007	0.0051
SP3	0.0092	0.0534	0	0.1931	0.0604	0.0585
SP4	0.2023	0.2465	0.1931	0	0.2535	0.2516
SP5	0.0512	0.007	0.0604	0.2535	0	0.0019
SP6	0.0493	0.0051	0.0585	0.2516	0.0019	0

TABLE III: Multi-Peak Multi-Taper, Variance Distance between Speakers

	Sine					
	SP1	SP2	SP3	SP4	SP5	SP6
SP1	0	0.0122	0.0111	0.0097	0.012	0.0285
SP2	0.0122	0	0.0233	0.0025	0.0242	0.0163
SP3	0.0111	0.0233	0	0.0208	0.0009	0.0396
SP4	0.0097	0.0025	0.0208	0	0.0217	0.0188
SP5	0.012	0.0242	0.0009	0.0217	0	0.0405
SP6	0.0285	0.0163	0.0396	0.0188	0.0405	0

TABLE IV: Hamming Window, Variance Distance between Speakers

	Hamming					
	SP1	SP2	SP3	SP4	SP5	SP6
SP1	0	0.0262	0.0176	0.1231	0.0264	0.0262
SP2	0.0262	0	0.0086	0.1493	0	0
SP3	0.0176	0.0086	0	0.1407	0.0088	0.0086
SP4	0.1231	0.1493	0.1407	0	0.1495	0.1493
SP5	0.0264	0.0001	0.0088	0.1495	0	0
SP6	0.0262	0	0.0086	0.1493	0	0

In the comparative study, in the Fig. 1, we can see the separation between each speaker; in this table we show the performance of Thomson using 5 repetitions of the same sentence. The Tables II, III and IV show the same procedure, but in that table we can see that the multi-taper method discriminates better each speakers compared to the traditional method using Hamming windows. In the Table IV, notice the performance of the traditional method and we can see that in some cases (e.g. SP2-SP5 and SP5-SP6), there is no difference.

TABLE V: Mean Using 40 Second of Recording

	Thomson	Multi-peak	Sine	Hamming
SP1	0.5741	0.5746	0.5744	0.5700
SP2	0.5725	0.563	0.5701	0.5378
SP3	0.5703	0.5743	0.5643	0.5745
SP4	0.5649	0.5722	0.5722	0.5511
SP5	0.5829	0.5741	0.5751	0.5636
SP6	0.5156	0.5745	0.5746	0.5678

In the Table V, we show the mean value from the MFCC coefficient, in this case we cannot observe clearly what methodology represents a better result may be due to the value used, the mean of a MFCC matrix.

VII. Conclusion

This study conducted a comparative research work of two methodologies that help extract information representative of each speaker, in a first view of the use variance; we observe that the performance of multi-taper method exceeds the traditional method using Hamming windows, but when we use the average of the MFCC coefficients, no processing tool improves performance noticeably.

A second point is the database, for a mayor study and more deep, we need work with a different database, using non-contemporaneous recording.

Acknowledgment

The report described is an on-going research project supported by UNAM-FI-II ("Reconocimiento del hablante para audio forense"), the authors are deeply grateful for this support

REFERENCES

[1] A. Alexander, "Forensic automatic speaker recognition using Bayesian interpretation and statistical compensation for mismatched conditions," PhD thesis, École Politechnique Fédérale de Lausanne, Laussane, Switzerland, 2005.

[2] J. Gonzalez-Rodriguez, D. T. Toledano, and J Ortega-Garcia, *Voice Biometrics*, Springer, 2007.

[3] J. P. Campbell, "Speaker recognition: A tutorial," in *Proc. the IEEE*, vol. 85, no. 9, pp. 1437-1462, Sep. 1997.

[4] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. M. Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska-Delacretaz, and D. A. Reynolds, "A Tutorial on text-independent speaker verification," *Journal on Applied Signal Processing*, pp. 430-451, 2004.

[5] T. Kinnunen and H. Li, "An Overview of text-independent speaker recognition from features to supervectors," *Speech Communication*, vol. 52, pp. 12–40, 2010

[6] M. Sahidullah and G. Saha, "Design, Analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition," *Speech Communication*, vol. 54, pp. 543–565, 2012.

[7] 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.

[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 (ASRU) IEEE Workshop*, pp. 11-15, 2011.

[9] T. Kinnunen *et al.*, "Low-Variance Multitaper MFCC Features: A Case Study in Robust Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 7, pp. 1990, 2001, Sept. 2012

[10] M. J. Alam, P. Kenny, and D. O'Shaughnessy, "A study of low-variance multi-taper features for Distributed speech recognition, springer," *Lecture Notes in Computer Science*, vol. 7015, pp. 239-245, 2011.

[11] 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*, vol. 1, pp. 73-76, 2001.

[12] M. Sahidullah and G. Saha, "Design, analysis and experimental evaluation of block based transformation in MFCC computation for Speaker Recogition," *Speech Communication*, vol. 54 pp. 543-565, 2012.

[13] 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.

[14] P. Bansal, A. Dev, and C. B. Jain, "Automatic speaker Identification using Vector Quantization," *Asian Journal of Information Technology*, pp. 938-942, 2007.

[15] M. Jahangir Alam, P. Kenny, D. O'Shaughnessy, "A Study of lo-variance multi-taper features for distributed speech recognition," *Springer*, vol. 7015, pp. 239-245, 2011.

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



Abel Herrera C. graduated in 1979 as a mecanical-electrical engineer, aquiring the master in electrical engenieering in 1985 and his Ph.D. in 2001, all at the University of Mexico (UNAM), the Ph.D. with help from the University of California-Davis. He carried out a post-doctoral internship at Carnegie Mellon University in 2001, as well as a research internship at the University of Southern California. Author of over 50 papers in speech codification, recognition and synthesis. He is the head of the Specch Processing lab, and has taught at the University since 1979.