Source Microphone Identification Using Multitaper MFCC Features

Ömer Eskidere, Ali Karatutlu

Department of Electrical Electronics Engineering, Bursa Orhangazi University, Bursa, Turkey
omer.eskidere@bou.edu.tr, ali.karatutlu@bou.edu.tr

Abstract

In this paper, we show studies of Multitaper spectrum estimation techniques to obtain the Mel-frequency cepstral coefficients (MFCCs) for source microphone identification systems. In the previous works, MFCCs were computed commonly from a single windowed (Hamming) DFT spectrum for source microphone identification. Emphasis on Multitaper MFCC features led us to extract intrinsic source microphone features obtained from non-speech segments. Experimental results on a set of 16 microphones yield that the Multitaper MFCC features show better performance compared to the Hamming windowed MFCC features.

1. Introduction

Recently, media forensics has received considerable attention in different situations such as authenticity and integrity of image, video and audio recordings. In a criminal investigation and law enforcement, alterations of these digital contents (forged contents) are an important issue. On the other hand, the needs of methods to assess their authenticity boost the demand. This issue is strictly related with audio authenticity and tampering detection [1]. The speech signals from source microphones can be very helpful in such situations. Complementary information about these situations can also be obtained by correctly identifying the source microphone.

Kraetzer *et al*. [2] proposed the first approach to identify the source microphone. They used audio steganalysis features as feature extraction method and the Kmeans and Naive Bayes as classification techniques to determine the source microphone and its recording environments. Then, Buchhols *et al.* presented [3] a Fourier coefficient histogram of near-silence segments of the recording as features. Simple Logistic, J48 decision tree, K-nearest neighbor, Support Vector Machine (SVM) classifiers carried out microphone identification task. Then, they investigated the performance of unweighted information fusion approach, using time domain (TD) features combination with frequency domain (FD) and the MFCC features [4]. In a later study [5], the identification of 8 microphones was studied and all feature computations were done using MFCC on NIST 2006 Speaker Recognition Evaluation database. Recently, we presented a microphone identification system using MFCC, linear prediction cepstrum coefficients (LPCC) and perceptually based linear predictive coefficients (PLPC) in [6]. Our results shows, when the MFCC features were classified by a Gaussian Mixture Model (GMM), 16 different microphones were identified with an accuracy of 96.87%. A comparison of the source microphone identification experiments is given in Table 1.

In addition to the microphone identification experiments, a landline telephone handsets identification method was proposed in the study [5] and further enhanced by Panagakis, and Kotropoulos [7]. Moreover, the source computer device identification from the recorded calls was recently presented [8]. In our earlier study, we extended the device identification by identifying cell phones from speech recordings [9]. We applied MFCC for the feature extraction and the vector quantization (VQ) and the SVM classifiers on a dataset of 14 cell phones.

Table 1. Comparison of feature extraction methods for microphone identification.

Reference	No. of Microphone	Features	Correct ID rate (%)
Kraetzer *et al*. [2]	4	7 TD* 56 MFCC	75.99
Buchhols et al. [3]	7	2048 FD*	93.5
Kraetzer *et al*. [4]	4&7	7 TD 35 FD 56 MFCC	100
Kraetzer *et al*. [1]	4&7	9 TD 529 FD 52 MFCC	82.51
Garcia-Romero&Epsy-Wilson [5]	8	23 MFCC	99
Eskidere [6]	16	13 MFCC	96.87

* TD: Time domain features, FD: frequency domain features.

According to the previous studies, the MFCC features were frequently used for identification of unknown source device. These features help in capturing device specific discriminative information from speech recordings. All of these works single window (taper) MFCC features carried out. A single windowing such as Hamming help to reduce bias but large variance is still a problem, and therefore, MFCCs calculated from this technique have also high variance. To eliminate this problem, Multitaper spectral estimation method can be used. As a result, the large variance for spectrum estimation can be reduced by multiple time domain windows instead of hamming-windowed power spectrum [10-12]. This technique is based on analyzing the speech frame using a number of spectrum estimators. Each of these estimators has a different taper. Thus, the final spectrum as a weighted mean of each sub-spectrum is calculated.

In this paper we study the source microphone identification issue using the speech recordings. We also compared the microphone identification performances of Hamming windows MFCCs and Multitaper MFCC feature extraction methods. It was previously proposed that use of non-speech segments of

recorded signal which help in capturing information about the microphone easier than the speech segment [13]. Though, high identification accuracy was obtained in using whole speech signal in our earlier work [6], we also tested in this study that information about the source microphone might be more pronounced in the non-speech parts of the signal. Experimental results indicate that the multitaper MFCCs method outperformed Hamming windowed MFCCs for microphone identification task.

2. Methods

Microphone identification task includes of two steps: training and identification. In the first step, discriminative features were obtained from the training speech recordings of each microphone in the dataset. A microphone model was created using these features. Using speech recordings, microphone properties such as transducer type, sensitivity, frequency response, and directionality helped us to capture intrinsic characteristics of the microphone. These properties of the source microphone were transferred to the recorded signal. In the identification step, the features were fed into a classifier to determine a specific microphone existing in our dataset.

2.1. Feature extraction

We consider each microphone has unique traces transferred to the recordings and these traces can be represented by features and signal modeling techniques. In speech processing, MFCCs are the most popular features and have been proven to be very effective as feature extraction technique [14-16]. Generally, MFCCs were carried out to date as fundamental acoustic features in microphone identification problems [1-6].

2.1.1. Multitaper MFCC

A popular approach in the spectrum estimation is the Multitaper spectrum estimation technique which has low variance. This technique has been replaced by the Hamming-windowed DFT spectrum used for single taper MFCC in many applications [17-19]. For the Multitaper spectrum estimation, the speech frame was performed from a series of spectra using FFT and then each signal was weighted and averaged in frequency domain. For $x(t)$ is a frame of signal with length L, the multitaper spectrum estimator $\hat{S}_{Mt}(f)$ is given by

$$\hat{S}_{Mt}(f) = \sum_{m=1}^{N} P(m) \left| \sum_{t=0}^{L-1} w_m(t) x(t) e^{-\frac{2i\pi f}{L}} \right|^2 \qquad (1)$$

where N is the number of the tapers, w_m is the m^{th}data taper ($m = 1,2,\dots,N$) and $P(m)$ is the weight of the m^{th} taper.

Figure 1 shows block diagrams for MFCC and Multitaper MFCC feature extraction methods. For traditional MFCC extraction, the speech sample was first decomposed into overlapping frames. Then signal was windowed using a window function which produced the more weight to the center of the signal than to its ends. Hamming window is the most popular window which can be defined as:

$$w(t) = 0.54 - 0.46cos\left(\frac{2\pi t}{L}\right) \qquad (2)$$

The power spectrum can directly be calculated using the Fast Fourier transform (FFT) from this windowed frame. In other words, as a special case of the Multitaper spectrum estimation method, the same power spectrum is obtained in case of $m = N = 1$ and $P(m) = 1$. Then, the Mel-scale is applied to warp the power spectrum and signal is filtered with a bank of triangular filters. Finally, the discrete cosine transform is applied to the logarithmic filterbank outputs to obtain MFCCs.

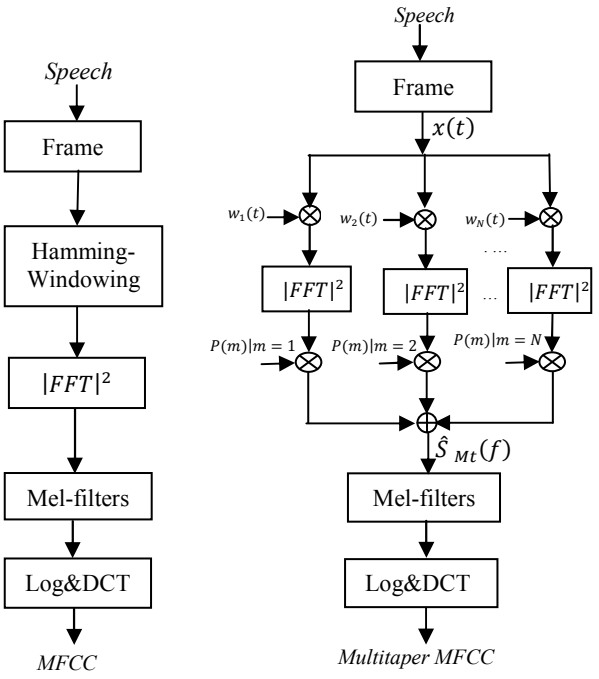

Fig. 1. Block diagrams for MFCC and Multitaper MFCC feature extraction

In Multitaper MFCC extraction, the Multitaper MFCC was replaced by the single taper FFT spectrum of MFCC. Firstly, a frame of signal was multiplied by a family of tapers such as Thomson [20], multipeak [21] and sinusoidal weighted cepstrum estimator (SWCE) tapers [22]. Fig. 2 denotes the Thomson, the multipeak and the sine tapers were used for Multitaper spectrum estimation. Then, the power spectrum was calculated for the each signal and the logarithmic filter bank outputs obtained of weighted and averaged signal. Lastly, Multitaper MFCC features were obtained from the discrete cosine transform. Detailed information about these techniques can be found in [17, 23].

2.2. Classification

In the experiments, the GMM classifier was used and tested to validate the applicability of the proposed method with speech and non-speech parts of the whole utterance. For our microphone identification system, each microphone is represented by such a GMM and this model λ is defined as

$$\lambda = (k_j, \mu_j, \Sigma_j) \quad j = 1, 2, \dots, M. \qquad (3)$$

where M is the number of mixture components, μ_j is the mean vectors, Σ_j is the covariance matrices and $k_j \in [0,1]$ is the mixture weights.

In the training phase of the GMM, model parameters λ are computed from the training data using the expectation maximization (EM) algorithm which converges to the maximum likelihood estimate of the mixture parameters locally [25-26]. In the identification phase of the GMM, the calculation of a set of likelihood functions using the test signal is applied. In our earlier study [6], 2, 4, 8, 16, 32, 64, and 128 Gaussian component densities were studied and $M = 64$ was found to give the highest accuracy for MFCCs. For this reason we use the GMM with 64 component Gaussian densities.

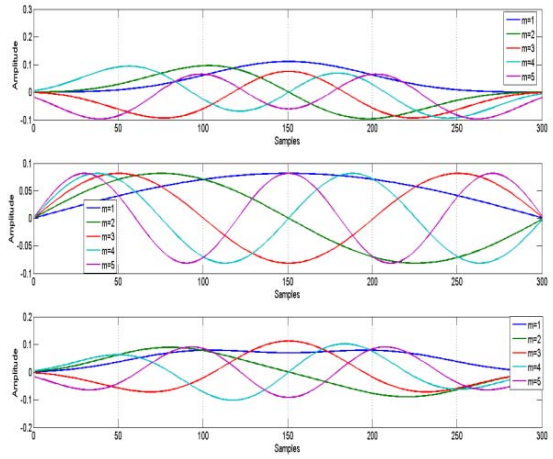

Fig. 2. Plot of three types of widely used tapers for Multitaper spectrum estimation. (a) The Thomson tapers (b) the multipeak tapers, and (c) the *sine* tapers. Window length is 300, m is the taper number.

3. Experiments

The microphone identification experiments were conducted on the same subset of the TIMIT Database as described previously [6]. This subset consisted of speech recordings from 40 speakers (20 males and 20 females) acquired by 16 microphone. In the experiments, 120 phonetically diverse sentences of these 40 speakers were played and recorded by each microphone in the same room. The brands and models of microphones are shown in Table 2.

Table 2. The brands and models of the microphones used in the experiments.

Id	**Brand (Model)**
M1	Sennheiser (PC-31)
M2	A4Tech (HS-5P)
M3	Creative (HS-350)
M4	Enzatec (HS-903)
M5	Genius (HS-500X)
M6	Hama (CS-408)
M7	Hama (NB-402)
M8	Logitech (PC-120)
M9	Microsoft (LX-2000)
M10	Philips (SHM-1900)
M11	Philips (SHM-1000)
M12	Sony (DR-115DP)
M13	Sony (ECM-DS70P)
M14	Teac (HP-5S)
M15	Trust (MC-1200)
M16	Trust (Clip-17358)

For our experiments, MFCCs and Multitaper MFCC features were extracted on 25 ms frames, with 10 ms frame shift and the speech recordings were normalized prior to feature extraction. MFCCs were calculated from the Hamming windowed spectrum estimates. For Thomson, multipeak and SWCE methods, the steps were treated as the same, except that the spectrum was estimated using Eq. (1). These features extraction steps demonstrated in Fig. 1 were employed. For the Thomson multitaper method a weighted average, adaptive weights computed from the eigenvalues, was performed instead of simply averaging the individual estimates. As a result, we obtained 13 MFCCs or 13 Multitaper MFCC features (excluding the DC coefficient c_0). Non-speech parts of the signal were determined using a silence detector described in ref. [24]. Following the experimental setup used in ref. [6], dataset was divided into 2 parts, as training and testing datasets.

4. Results and Discussion

In the experiments, suitable approaches of the baseline Hamming window system and Multitaper systems were investigated on microphone identification performance. Thus, Thomson, multipeak, and SWCE tapers was tested. Additionally, the effects of amount of microphone-specific information in speech and non-speech parts were compared for the source microphone identification. Figure 3 shows the microphone identification results using whole speech recordings for the baseline Hamming window system and multitaper systems.

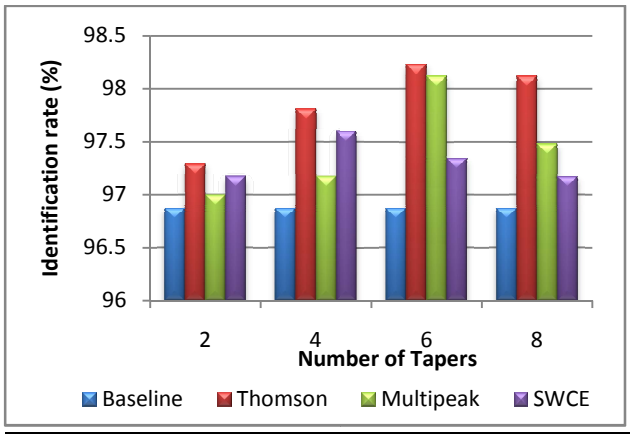

Fig. 3. Microphone identification results (%) using whole speech recordings for the baseline Hamming window system and multitaper systems.

Next, we evaluated the impact of use non-speech segments only for the proposed feature extraction methods. In this case, features obtained from the non-speech parts were employed. Identification results using these features are given in Figure 4.

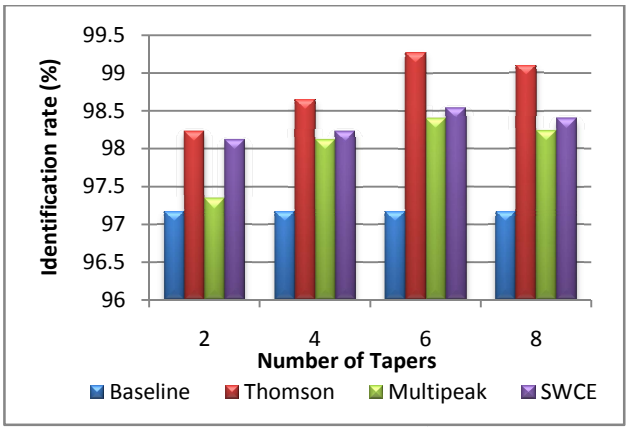

Fig. 4. Microphone identification results (%) using non-speech segments for the baseline Hamming window system and multitaper systems.

In the last experiment, Microphone identification results using the speech parts only for the baseline Hamming window system and multitaper systems are given in Figure 5.

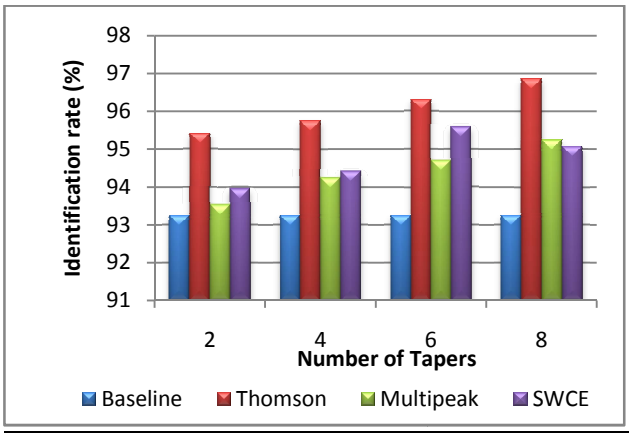

Fig. 5. Microphone identification results (%) using speech segments for the baseline Hamming window system and multitaper systems.

As shown in Figure 3, 4 and 5, the microphone identification performances are significantly affected by the number of tapers for features obtained from the speech-only parts, the non-speech parts and whole utterance. Also, as can be deduced from Figure 4, using non-speech parts only, the best identification rate of 99.27% for Thomson taper was achieved when the number of tapers was equal to 6. Thomson taper usually has superior performance compared to SWCE, multipeak and single taper (Hamming). All three Multitapers outperformed Hamming in nearly all of cases because of the variance reduction. In addition, we see that the features obtained from the non-speech parts identification performance were obviously enhanced relative to those obtained from the whole utterance or the speech-only parts.

The results show that the Multitaper MFCCs provided a high identification rate compared to the standard MFCCs. In general, the highest identification rates were obtained when number of tapers was 6.

5. Conclusions

In this work, the Multitaper MFCC features and traditional MFCCs were compared to determine the appropriate features for the microphone identification task. The experiments conducted on TIMIT dataset using sixteen microphones show that the Multitaper MFCCs are promising as a viable candidate for replacing the conventional MFCCs for identifying the recording microphone. In particular, the highest identification rates were obtained using number of tapers was six in the most cases and the Thomson taper was the best performer while estimating the power spectrum. Furthermore, the highest microphone identification rates were observed only for the conditions involving non-speech segments. Overall, the Multitaper MFCC features can effectively be employed using non-speech segments in the source microphone identification problem.

6. References

[1] C. Kraetzer, K. Qian, M. Schott, J. Dittmann, “A context model for microphone forensics and its application in evaluations”, Proc. of Media Watermarking, Security, and Forensics XIII, Electronic Imaging Conference 7880, 2011.

[2] C. Kraetzer, A. Oermann, J. Dittmann, A. Lang, “Digital audio forensics: A first practical evaluation on microphone and environment classification”, In: 9th Workshop on Multimedia & Security, pp. 63-74, 2007.

[3] R. Buchholz, C. Kraetzer, J. Dittmann, “Microphone classification using Fourier coefficients”, In Proceedings of 11th Information Hiding Darmstadt, LNCS 5806, Springer-Verlag Berlin Heidelberg, 2009.

[4] C. Kraetzer, M. Schott, J. Dittmann, “Unweighted fusion in microphone forensics using a decision tree and linear logistic regression models”, In Proceedings of the 11th workshop on Multimedia and security, Princeton, New Jersey, USA: ACM Press, pp. 49–56, 2009.

[5] D.G. Romero, C.Y.E. Wilson, “Automatic acquisition device identification from speech recordings”, In Proceeding of IEEE International Conference on Acoustics Speech and Signal Processing, Texas, US, pp. 1806–1809, 2010.

[6] Ö. Eskidere, “Source microphone identification from speech recordings based on Gaussian mixture model”, Turkish Journal of Electrical Engineering & Computer Sciences, Vol. 22, pp. 754 – 767, 2014.

[7] Panagakis, Y., C. Kotropoulos, “Telephone handset identifiation by feature selection and sparse representations”, Presented at the Proceedings of the WIFS, Tenerife, 2012.

[8] M. Jahanirad, A. W. A. Wahab, N. B. Anuar, M. Y. I. Idris, and M. N. Ayub, “Blind source mobile device identification based on recorded call,” Eng. Appl. Artif. Intell., vol. 36, pp. 320–331, Nov. 2014.

[9] C. Hanilçi, F. Ertaş, T. Ertaş, Ö. Eskidere, “Recognition of brand and models of cell-phones from recorded speech signals”, IEEE Transactions on Information Forensics and Security, Vol. 7 (2), pp. 625-634, 2012.

[10] Sandberg, J., Hansson-Sandsten, M., Kinnunen, T., Saeidi, R., Flandrin, P., Borgnat, P., “Multitaper estimation of frequency-warped cepstra with application to speaker verification”, *IEEE Signal Process. Lett.* 17 (4):343–346, 2010.

[11] Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M., Li, H., “Low-variance multitaper

MFCC features: A case study in robust speaker verification", *IEEE Transactions on Audio, Speech and Language Processing* 20 (7):1990–2001, 2012.

[12] Kinnunen, T., Saeidi, R., Sandberg, J., Hansson-Sandsten, M., What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering. In: Proc. Interspeech. 2734–2737, 2010.

[13] C. Hanilçi, Kinnunen, T. "Source Cell-Phone Recognition from Recorded Speech Using Non-Speech Segments", Digital Signal Processing, Vol.35, pp. 75-85, 2014.

[14] M. Slaney, "Auditory Toolbox: A MATLAB Toolbox for Auditory Modeling", Work Technical Report, Interval Research Corporation, pp. 29-32, 1998.

[15] P. Melmerstein, S. Davis, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences", IEEE Transactions on Acoustics, Speech and Signal Processing, Vol. 28, pp. 336-357, 1980.

[16] R. Mammone, X. Zhang, R. Ramachandran, "Robust speaker recognition: a feature-based approach", IEEE Signal Processing Magazine, Vol. 13, pp. 58-71, 1996.

[17] Wieczorek, M. A., Simons, F. J. Minimum-Variance Multitaper Spectral Estimation on the Sphere. *Journal of Fourier Analysis and Applications* 13(6):665-692, 2007.

[18] Alam, M. J., Kenny, P., O'Shaughnessy, D., A study of low-variance multi-taper features for distributed speech recognition. In: Proc. NOLISP. LNAI. 7015: 239–245, 2011.

[19] Attabi, Y., Alam, M. J., Dumouchel, P., Kenny, P. Douglas O'Shaughnessy, D. Multiple windowed spectral features for emotion recognition. ICASSP. 7527-7531, 2013.

[20] Thomson, D.J. Spectrum estimation and harmonic analysis. Proc. IEEE 70(9):1055–1096. 1982.

[21] Hansson, M., Salomonsson, G., A multiple window method for estimation of peaked spectra. *IEEE Trans. Signal Process.* 45(3):778–781, 1997.

[22] Riedel, K. S., Sidorenko, A., Minimum bias multiple taper spectral estimation. *IEEE Trans. Signal Process.* 43(1):188–195, 1995.

[23] Alam, M. J., Kinnunen, T., P., Kenny, P., Ouellet, D., O'Shaughnessy, D. Multitaper MFCC and PLP features for speaker verification using i-vectors. *Speech Communication*. 55(2):237-251, 2013.

[24] Aliaa, A. Y., A. S. Ebada, W. H. El Behaidy. Development of Automatic Speaker Identification System. *st* 21 National Radio Science Conference. 2004.

[25] D.A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transactions on Speech Audio Processing, Vol. 3, pp. 72-83, 1995.

[26] Reddy, Chandan K., Hsiao-Dong Chiang, and Bala Rajaratnam. "Trust-tech-based expectation maximization for learning finite mixture models", *Pattern Analysis and Machine Intelligence, IEEE Transactions on* , vol 30 (7), pp. 1146-1157, 2008.