DE GRUYTER DOI 10.1515/jisys-2013-0038 — Journal of Intelligent Systems 2013; 22(3): 241–251

Research Article

B. G. Nagaraja* and H. S. Jayanna

Multilingual Speaker Identification by Combining Evidence from LPR and Multitaper MFCC

Abstract: In this work, the significance of combining the evidence from multitaper mel-frequency cepstral coefficients (MFCC), linear prediction residual (LPR), and linear prediction residual phase (LPRP) features for multilingual speaker identification with the constraint of limited data condition is demonstrated. The LPR is derived from linear prediction analysis, and LPRP is obtained by dividing the LPR using its Hilbert envelope. The sine-weighted cepstrum estimators (SWCE) with six tapers are considered for multitaper MFCC feature extraction. The Gaussian mixture model–universal background model is used for modeling each speaker for different evidence. The evidence is then combined at scoring level to improve the performance. The monolingual, crosslingual, and multilingual speaker identification studies were conducted using 30 randomly selected speakers from the IITG multivariability speaker recognition database. The experimental results show that the combined evidence improves the performance by nearly 8–10% compared with individual evidence.

Keywords: Speaker identification, mel-frequency cepstral coefficients, multitaper mel-frequency cepstral coefficients, multilingual, linear prediction residual, linear prediction residual phase.

***Corresponding author: B. G. Nagaraja,** Siddaganga Institute of Technology, Department of Information Science and Engineering, Tumkur 572103, Karnataka, India,
e-mail: nagarajbg@gmail.com
H. S. Jayanna: Siddaganga Institute of Technology, Department of Information Science and Engineering, Tumkur 572103, Karnataka, India

1 Introduction

Speaker recognition is defined as a task of recognizing speakers from their voice [3]. In speaker identification, the system identifies the most likely speaker of the test speech signal. In speaker verification task, a user's speech is used to clas-

sify him as being either who he claimed to be or an impostor [10]. Depending on the mode of operation, speaker identification can be classified as text-dependent identification or text-independent identification. Text-dependent identification requires the speaker to produce speech for the same text, both during training and testing, whereas text-independent identification does not rely on a specific text being spoken [20]. Speaker identification can be performed in the monolingual mode (common language during training and testing), crosslingual mode (different language for training and testing), and multilingual mode (speaker-specific models are trained in one language and tested with multiple languages) [2].

Most state-of-the-art speaker identification systems work within a single-language environment (English/European language) using sufficient data. People have the ability to learn more than one language [6]. Many countries, including India, are multilingual. In India, more than 50 languages are officially recognized. A person in a multilingual country usually speaks more than one language. For instance, criminals often switch over to another language, especially after committing a crime [2]. Therefore, training a person's voice in one language and recognizing him in a different language (multilingual environment) is an issue in many countries. In addition, data sparseness is becoming a crucial research concern in automatic speaker recognition system. In a noncooperative scenario such as forensic investigation, speech data may last for only a few seconds and the task is to identify the speaker. Such an application should be able to validate the speaker using limited amount of speech data.

Speaker characteristics in the speech signal can be attributed to the vocal tract dimension, excitation characteristics, and the learning habits of the speakers [13]. The mel-frequency cepstral coefficient (MFCC) and linear prediction cepstral coefficient (LPCC) features can accurately characterize the vocal tract configuration of a speaker and can achieve good performance [16]. In [13, 19], it was shown that the linear prediction residual (LPR) and linear prediction residual phase (LPRP) signals contain speaker-specific information that is complementary to the MFCC features. Attempts have been made to exploit the usefulness of features extracted from excitation and vocal tract characteristics for speaker recognition [13, 19, 22]. In this direction, Murty and Yegnanarayana [13] combined the evidence from LPRP and MFCC for improving the speaker recognition performance. The speaker verification experiments on the NIST 2003 database showed that the proposed combined system yields a better equal error rate than the individual systems.

In [16], the phase information is combined with the MFCC for speaker identification and verification tasks. The modified phase information extraction method that normalizes the change variation in the phase according to the frame position of the input speech was proposed. The experimental results showed that the combination of the MFCC and the phase information was efficient for noisy speech

signal. The conventional MFCC realization based on windowed (hamming) discrete Fourier transform (DFT) may not yield good performance due to the high variance of the spectrum estimation [11, 12, 17]. The window function in speech processing smooths the spectral estimate of a frame of speech data by multiplying the data frame with the window. However, windowing by a single window has the disadvantage of producing leakage effects [21]. Kinnunen et al. [11] promoted the use of multitaper MFCC features for speaker verification. Experimental results on the NIST 2002 database indicated that multitapers outperform the conventional single-window technique (MFCC). In our previous work, we attempted to identify the speaker in the multilingual context with the constraint of limited data (15 s) using sine-weighted cepstrum estimator MFCC (SWCE-MFCC) (K=6) as feature [14]. It was observed in the study that the use of the multitaper MFCC approach gives a better identification performance in all the speaker identification experiments compared with the conventional MFCC technique.

The impact of mismatch in training and testing languages on a speaker verification system using English, Hindi, and Arunachali languages was carried out in [4]. The speaker verification system was developed using 38-dimensional features and the Gaussian mixture model–universal background model (GMM–UBM) approach. A training data of 120 s and different testing data of 15, 30, and 45 s were used. Recognition performance was observed to be greatly dependent on the training and testing languages. Further, it was observed that if the system is trained with more than one language, the relative recognition performance of the system degrades compared with that of the single-language scenarios (monolingual).

The features extracted from the multitaper MFCC, LPR, and LPRP are modeled using GMM–UBM individually. The individual scores are combined to obtain the speaker identification performance. The rest of the paper is organized as follows: the database used for the study is described in Section 2. Section 3 presents the speaker identification studies. Monolingual, crosslingual, and multilingual experimental results are discussed in Section 4. Conclusions are given in Section 5.

2 Database for the Study

Speaker identification experiments are carried out on the subset of the IITG multivariability speaker recognition (IITG-MV) database, which is collected in a setup having five different sensors, two different environments, two different languages, and two different styles [8]. The recording was done in the office (controlled environment) and in hostel rooms, laboratory, corridors, etc. (uncontrolled

environments). The speech signal was sampled at 16 kHz and stored with 16-bit resolution. The recording was done in Indian English and favorite language of the speaker, which may be one of the Indian languages such as Hindi, Kannada, Tamil, Oriya, Assami, Malayalam, and so on [18]. For the present work, we randomly selected 30 (17 male and 13 female) speakers in the IITG-MV database (headphone speech data).

3 Speaker Identification Studies

3.1 Feature Extraction using LPR

Speech recordings were sampled at 8 kHz with 16-bit resolution and pre-emphasized (0.97). A frame duration of 12 ms, with 6 ms for overlapping, was used. To obtain the LPR, the vocal tract information is predicted from the speech signal by 10^{th}-order linear prediction analysis [19]. The estimated linear predictive coefficients represent the vocal tract information and are suppressed from the speech signal using an inverse filter formulation to obtain LPR [19].

3.2 Feature Extraction using LPRP

The LPRP is obtained by dividing the LPR using its Hilbert envelope [13]. Hilbert envelope is the magnitude of the analytic signal of a given real signal. The analytic signal $r_a(n)$ corresponding to the LPR $r(n)$ is given by

$$r_a(n) = r(n) + jr_h(n), \tag{1}$$

where $r_h(n)$ is the Hilbert transform of $r(n)$ and is given by

$$r_h(n) = f^{-1}[r_h(w)], \tag{2}$$

where

$$R_h(w) = \begin{cases} -jR(w); & 0 \le w < \pi \\ jR(w); & 0 > w \ge -\pi, \end{cases}$$

where $R(w)$ is the Fourier transform of the $r(n)$ and f^{-1} denotes the inverse Fourier transform. Hilbert envelope $h_e(n)$ given by

$$h_e(n) = \sqrt{r^2(n) + r_h^2(n)}, \tag{3}$$

and the cosine of the phase of the analytic signal $r_a(n)$ is given by

$$\cos(\theta(n)) = \frac{r(n)}{h_e(n)}. \tag{4}$$

3.3 Feature Extraction Using Multitaper MFCC

Let $F = [f(0)\ f(1)\ \ldots\ f(N-1)]^T$ denote one frame of speech (N samples) signal. The windowed DFT spectrum estimate is given by [11, 12]

$$S(f) = \left| \sum_{n=0}^{N-1} w[n] f[n] e^{-i2\pi fn/N} \right|^2, \tag{5}$$

where $W = [w(0)\ w(1)\ \ldots\ w(N-1)]^T$ is the time-domain window (Hamming) function. The basic idea in multitapering is to pass the analysis frame through several window functions and then approximate the weighted mean of individual subspectra to obtain the final resultant spectrum [12]. The multitaper spectrum estimation is given by [11]

$$\hat{S}(f) = \sum_{j=1}^{K} \lambda(j) \left| \sum_{n=0}^{N-1} w_j[n] f[n] e^{-i2\pi fn/N} \right|^2, \tag{6}$$

where K represents the number of multitapers used, $W_j = [w_j(0)\ w_j(1)\ \ldots\ w_j(N-1)]^T$ is the multitaper weights, and $j = 1,2, \ldots, K$ are used with corresponding weights $\lambda(j)$. The sine tapers are defined as [1]

$$w_j(n) = \sqrt{\frac{2}{N+1}} \sin\left(\frac{\pi j(n+1)}{N+1} \right);\ n = 0, \ldots, N-1. \tag{7}$$

The weight used in the SWCE method is given by [1]

$$\lambda(j) = \frac{\cos\left(\frac{2\pi(j-1)}{M/2} \right) + 1}{\sum_{j=1}^{M} \left(\cos\left(\frac{2\pi(j-1)}{M/2} \right) + 1 \right)}. \tag{8}$$

In this work, the SWCE multitaper is used with $K = 6$ windows. A mel-warping is then performed using 22 triangular band-pass filters followed by a discrete cosine transform. Figure 1 shows the block diagram representation of the multitaper MFCC method. A 13-dimensional (excluding 0^{th} coefficient) MFCC feature vector is finally obtained. Figure 2 shows the Hamming and SWCE multitaper representation in the frequency domain.

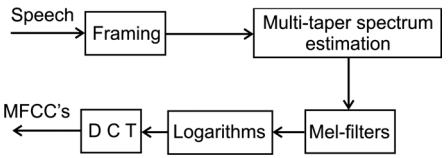

Figure 1. Block Diagram of Multitaper MFCC Technique.

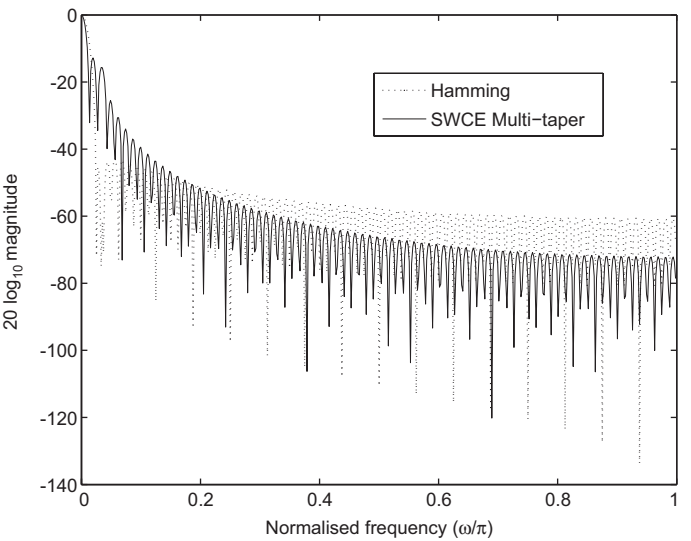

Figure 2. Frequency Domain for a Hamming Window and SWCE Multitaper Window.

3.4 Speaker Modeling Using GMM–UBM

To build the UBM, we have used 1 h of speech data from all the 138 speakers of the YOHO database. The speaker-specific models were created by adapting only the mean vectors of the UBM using the maximum a posteriori adaptation algorithm [7]. The parameters of the GMM models (mean vector, covariance matrix, and mixture weights) were estimated using the expectation maximization algorithm. We have modeled speakers using GMMs with 8, 16, 32, 64, and 128 mixtures.

3.5 Speaker Testing

During testing, the frame scores of each speaker are stored as a confidence score (C_i). The average confidence score for a given test signal is computed as

$C=1/S\sum_{i=1}^{M}C_i$, where S is the sum of the individual frame scores of a speaker and M is the total number of speakers. The confidence scores C_1 and C_2 are obtained using the multitaper MFCCs and LPRs/LPRPs, respectively. For each speaker, C_1 and C_2 are combined using the linear weighted sum [13], given by

$$C_c=\alpha^{\star}C_1+\beta^{\star}C_2, \tag{9}$$

where $\beta=(1-\alpha)$. To find the optimal weights for the linear combination, we performed a simple search of the best weightings ($\alpha=0.75$).

4 Experimental Results

In this section, monolingual, crosslingual, and multilingual speaker identification results are presented. In all our experiments, the speaker set (30 speakers) and amount of speech data (15 s) are kept constant to make a relative comparison of the performance of speaker identification using different techniques. (Note: X/Y indicates training with language "X" and testing with language "Y"; "Multilanguage" includes Kannada, Telugu, Tamil, Assami, Bengali, Malayalam, and Oriya.)

4.1 Monolingual Speaker Identification

The performance of LPR, LPRP, MFCC, and multitaper MFCC and the different combined evidence for monolingual speaker identification are given in Table 1. The speaker identification system trained and tested in English (E/E) gives the highest performance of 73.33% for the combined multitaper MFCC and LPR system. The performance of the speaker identification trained and tested in Hindi (H/H) is 66.66% for the combined multitaper MFCC and LPRP system.

4.2 Crosslingual Speaker Identification

The performance of LPR, LPRP, MFCC, and multitaper MFCC and the different combined evidence for crosslingual speaker identification are given in Table 2. The combined evidence from the multitaper MFCC and LPR system and the multitaper MFCC and LPRP system gives the highest recognition performance of 60% for training in English and testing in Hindi (E/H). Similarly, the combined multitaper MFCC and LPRP system gives the highest recognition performance of 53.33% for a system trained in Hindi and tested in English (H/E).

Table 1. The Monolingual Speaker Identification Performance using LPR, LPRP, MFCC, and Multitaper MFCC-Based Individual and Combined System. P_i Represents the Maximum Identification Performance among the Number of Gaussian Mixtures.

Train/Test	Features	Gaussian Mixtures					P_i
		8	16	32	64	128	
E/E	LPR	16.66	20.00	26.66	30.00	26.66	30.00
	LPRP	20.00	20.00	26.66	30.00	30.00	30.00
	MFCC	43.33	50.00	56.66	53.33	56.66	56.66
	Multitaper MFCC	50.00	66.66	66.66	70.00	60.00	70.00
	MFCC and LPR	46.66	60.00	56.66	56.66	53.33	60.00
	MFCC and LPRP	50.00	60.00	50.00	60.00	56.66	60.00
	Multitaper MFCC and LPR	46.66	73.33	70.00	70.00	60.00	73.33
	Multitaper MFCC and LPRP	60.00	63.33	66.66	70.00	70.00	70.00
H/H	LPR	10.00	20.00	23.33	26.66	30.00	30.00
	LPRP	13.33	16.66	20.00	26.66	30.00	30.00
	MFCC	23.33	30.00	46.66	46.66	50.00	50.00
	Multitaper MFCC	26.66	43.33	46.66	56.66	50.00	56.66
	MFCC and LPR	33.33	36.66	53.33	56.66	56.66	56.66
	MFCC and LPRP	40.00	36.66	50.00	50.00	56.66	56.66
	Multitaper MFCC and LPR	43.33	46.66	53.33	60.00	63.33	63.33
	Multitaper MFCC and LPRP	40.00	46.66	56.66	60.00	66.66	66.66

Table 2. The Crosslingual Speaker Identification Performance using LPR, LPRP, MFCC, and Multitaper MFCC-Based Individual and Combined System. P_i Represents the Maximum Identification Performance among the Number of Gaussian Mixtures.

Train/Test	Features	Gaussian Mixtures					P_i
		8	16	32	64	128	
E/H	LPR	10.00	16.66	16.66	23.33	26.66	26.66
	LPRP	20.00	20.00	16.66	23.33	23.33	23.33
	MFCC	23.33	30.00	46.66	46.66	50.00	50.00
	Multitaper MFCC	36.66	46.66	46.66	50.00	56.66	56.66
	MFCC and LPR	16.66	33.33	26.66	53.33	53.33	53.33
	MFCC and LPRP	20.00	30.00	30.00	43.33	50.00	50.00
	Multitaper MFCC and LPR	33.33	36.66	50.00	60.00	56.66	60.00
	Multitaper MFCC and LPRP	30.00	40.00	46.66	56.66	60.00	60.00
H/E	LPR	10.00	13.33	20.00	23.33	23.33	23.33
	LPRP	10.00	16.66	20.00	23.33	26.66	26.66
	MFCC	26.66	23.33	33.33	36.66	40.00	40.00
	Multitaper MFCC	36.66	43.33	40.00	43.33	43.33	43.33
	MFCC and LPR	20.00	30.00	43.33	36.66	40.00	43.33
	MFCC and LPRP	20.00	33.33	40.00	46.66	46.66	46.66
	Multitaper MFCC and LPR	46.66	40.00	43.33	50.00	40.00	50.00
	Multitaper MFCC and LPRP	50.00	46.66	50.00	53.33	50.00	53.33

Table 3. The Multilingual Speaker Identification Performance using LPR, LPRP, MFCC, and Multitaper MFCC-Based Individual and Combined System. P_i Represents the Maximum Identification Performance Among the Number of Gaussian Mixtures.

Train/Test	Features	Gaussian Mixtures					P_i
		8	16	32	64	128	
E/multi	LPR	10.00	10.00	16.66	20.00	23.33	23.33
	LPRP	13.33	20.00	16.66	20.00	23.33	23.33
	MFCC	20.00	40.00	36.66	40.00	50.00	50.00
	Multitaper MFCC	30.00	36.66	50.00	46.66	60.00	60.00
	MFCC and LPR	30.00	40.00	46.66	43.33	56.66	56.66
	MFCC and LPRP	26.66	36.66	40.00	43.33	53.33	53.33
	Multitaper MFCC and LPR	30.00	43.33	50.00	46.66	63.33	63.33
	Multitaper MFCC and LPRP	36.66	40.00	46.66	56.66	60.00	60.00
H/multi	LPR	6.66	13.33	13.33	20.00	16.66	20.00
	LPRP	10.00	16.66	16.66	23.33	20.00	23.33
	MFCC	26.66	30.00	33.33	33.33	36.66	36.66
	Multitaper MFCC	23.33	26.66	30.00	36.66	36.66	36.66
	MFCC and LPR	30.00	30.00	33.33	40.00	36.66	40.00
	MFCC and LPRP	30.00	33.33	40.00	43.33	43.33	43.33
	Multitaper MFCC and LPR	23.33	30.00	33.33	36.66	50.00	50.00
	Multitaper MFCC and LPRP	30.00	36.66	43.33	50.00	53.33	53.33

4.3 Multilingual Speaker Identification

The performance of LPR, LPRP, MFCC, and multitaper MFCC and the different combined evidence for multilingual speaker identification are given in Table 3. The speaker identification system trained in English and tested with the multilanguages (E/multi) gives the highest performance of 63.33% for the combined multitaper MFCC and LPR system. The performance of the speaker identification system trained in Hindi and tested with the multilanguages (H/multi) gives 53.33% for the combined multitaper MFCC and LPRP system.

Some of the observations we made from the monolingual, crosslingual, and multilingual results are as follows:

- The proposed combined multitaper MFCC and LPR system and multitaper MFCC and LPRP system yields good recognition in all the speaker identification experiments. The performance is higher than the individual system. The improvement in performance may be due to different speaker-specific information (excitation and vocal tract) provided by each feature [9, 13, 15, 16].
- The multitaper (SWCE) MFCC performs better than the usual MFCC method. This may be due to the use of multiple windows (multitapers) that reduce the variance of the MFCC features, thus making the spectrum less sensitive

to noise compared with the conventional single-window (hamming) method [11, 12].
- The weights used to combine the two systems ($\alpha = 0.75$ and $\beta = 0.25$) suggest that the multitaper MFCC-based system is more reliable than the LPR-based one.
- It was observed that the results are better for monolingual experiments than for the crosslingual and multilingual experiments. This may be due to the variation in fluency and word stress when the same speaker speaks different languages and also due to different phonetic and prosodic patterns of the languages [5].

5 Conclusions

The main objective of the work was to increase the performance of the multilingual speaker identification system in limited data condition by combining the evidence from LPR, LPRP, and multitaper MFCC features. It was demonstrated by conducting speaker identification experiments on 30 randomly selected speakers from the IITG-MV database. The results showed that the information captured by the multitaper MFCC in combination with LPR/LPRP provides good speaker identification performance.

Acknowledgment: This work was supported by Visvesvaraya Technological University (VTU), Belgaum, Karnataka, India.

Received May 15, 2013; previously published online June 29, 2013.

Bibliography

[1] M. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet and D. D. O'Shaughnessy, Multitaper MFCC and PLP features for speaker verification using i-vectors, *Speech Commun*. **55** (2013), 237–251.
[2] P. H. Arjun, *Speaker Recognition in Indian Languages: A Feature Based Approach*, PhD thesis, Indian Institute of Technology, Kharagpur, India, 2005.
[3] B. S. Atal, Automatic recognition of speakers from their voices, *Proc. IEEE* **64** (1976), 460–475.
[4] U. Bhattacharjee and K. Sarmah, A multilingual speech database for speaker recognition, *Proc. IEEE ISPCC* (2012), 15–17.
[5] G. Durou, Multilingual text-independent speaker identification, in: *Proceedings of Multilingual Interoperability in Speech Technology (MIST)*, pp. 115–118, Leusden, Netherlands, 1999.

[6] U. Halsband, Bilingual and multilingual language processing, *J. Physiol. Paris* **99** (2006), 355–369.
[7] B. C. Haris and R. Sinha, Exploring sparse representation classification for speaker verification in realistic environment, in: *Centenary Conference – Electrical Engineering*, pp. 1–4, Indian Institute of Science, Bangalore, 2011.
[8] B. C. Haris, G. Pradhan, A. Misra, S. Shukla, R. Sinha and S. R. M. Prasanna, Multivariability speech database for robust speaker recognition, *Proc. IEEE Commun. (NCC)* (2011), 1–5.
[9] H. S. Jayanna, *Limited Data Speaker Recognition*, PhD thesis, Indian Institute of Technology, Guwahati, India, 2009.
[10] T. Kinnunen, E. Karpov and P. Fränti, Real-time speaker identification and verification, *IEEE Trans. Audio Speech Lang. Process.* **14** (2006), 277–288.
[11] T. Kinnunen, R. Saeidi, J. Sandberg and M. H. Sandsten, What else is new than the Hamming window? Robust MFCCs for speaker recognition via multitapering, *Proc. Interspeech* (2010), 2734–2737.
[12] T. Kinnunen, R. Saeidi, F. Sedlak, K. A. Lee, J. Sandberg, M. H. Sandsten and H. Li, Low-variance multitaper MFCC features: a case study in robust speaker verification, *IEEE Trans. Audio Speech Lang. Process.* **20** (2012), 1990–2001.
[13] K. S. R. Murty and B. Yegnanarayana, Combining evidence from residual phase and MFCC features for speaker recognition, *IEEE Signal Process Lett.* **13** (2006), 52–55.
[14] B. G. Nagaraja and H. S. Jayanna, Multilingual speaker identification with the constraint of limited data using multi-taper MFCC, *Proc. SNDS (Springer)* **335** (2012), 127–134.
[15] B. G. Nagaraja and H. S. Jayanna, Multilingual speaker identification by combining evidences from LP residual and multi-taper MFCC, *Proc. ICCVSP* (2013), 1–4.
[16] S. Nakagawa, L. Wang and S. Ohtsuka, Speaker identification and verification by combining MFCC and phase information, *IEEE Trans. Audio Speech Lang. Process.* **20** (2012), 1085–1095.
[17] D. B. Percival and A. T. Walden, *Spectral Analysis for Physical Applications*, Cambridge University Press, Cambridge, MA, 1993.
[18] G. Pradhan and S. R. M. Prasanna, Significance of vowel onset point information for speaker verification, *Int. J. Comput. Commun. Technol.* **2** (2011), 56–61.
[19] S. R. M. Prasanna, C. S. Gupta and B. Yegnanarayana, Extraction of speaker-specific excitation information from linear prediction residual of speech, *Speech Commun.* **48** (2006), 1243–1261.
[20] D. A. Reynolds and R. C. Rose, Robust text-independent speaker identification using Gaussian mixture speaker models, *IEEE Trans. Speech Audio Process.* **3** (1995), 72–83.
[21] L. P. Ricotti, Multitapering and a wavelet variant of MFCC in speech recognition, *Proc. IEEE Vis. Image Signal Process.* **152** (2005), 29–35.
[22] B. Yegnanarayana, K. S. Reddy and S. P. Kishore, Source and system features for speaker-recognition using AANN models, *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* **1** (2001), 409–412.