Text-independent Speaker Identification in Emotional and Whispered Speech Environments

Naresh P. Jawarkar
Babasaheb Naik College of Engineering
Pusad (MS), India

Raghunath S. Holambe
SGGS Institute of Engineering and Technology
Nanded (MS), India

Tapan Kumar Basu
Academy of Technology
Aedconagar,
Hoogly (WB) India

ABSTRACT
This paper describes challenging task of closed set text-independent speaker identification in emotional and whispered speech environments. In the first phase of the work, speaker identification system is developed using neutral speech and tested using speech samples comprising of six basic emotions of anger, happiness, sadness, disgust, neutral and fear. The performance is analyzed using Mel frequency cepstral coefficients (MFCC), Line spectral frequencies (LSF), and temporal energy of subband cepstral coefficients (TESBCC) feature sets. The second phase of work involves the process of speaker identification system in whispered speech environment. The performance of the speaker identification system degrades drastically for whisper speech utterances. A new feature called temporal Teager energy based subband cepstral coefficients (TTESBCC) is proposed. The comparison of the performance of MFCC, TESBCC, weighted instantaneous frequency (WIF) and TTESBCC feature sets is done for this process. A novel classifiers fusion technique is developed and its performance is compared with that of the individual classifiers. Two databases with speech utterances of thirty nine speakers recorded in the six basic emotions and speech utterances of twenty five speakers in whispered speech are used for experimentation. The speech utterances for database were recorded in Indian language – Marathi. It is observed fusion of classifiers considerably enhances the speaker identification accuracy in both emotional and whispered speech environments.

General Terms
Speaker recognition, Pattern recognition

Keywords
Speaker identification, whispered speech, temporal Teager energy based subband cepstral coefficients, emotional environment, classifier fusion.

1. INTRODUCTION
Speaker recognition is the process of extracting the identity of the person speaking. Speaker recognition task can be further classified into speaker identification and speaker verification processes. The speaker identification refers to determining the person talking from a set of known voices or speakers. The speaker verification refers to process of authenticating whether a given voice sample is produced by a claimed speaker. If there is a possibility that the target speaker is none of the registered speakers the task is called open-set problem. Speaker identification can be classified into text-dependent and text-independent tasks. In the text dependent case, the utterance presented to the recognizer is known before hand where as no assumptions about the text being spoken is made in text independent case. Literature survey shows many studies on the speaker recognition in the neutral (normal) environment [1]-[3]. However, there are few studies that focus on challenging task of the speaker recognition in emotional and whispered speech environments.

Speaker recognition in emotional environment is considered one of the nascent research fields in human-computer interaction [4]. Wu et al. [5] studied the effect of emotion on the performance of GMM-UBM based speaker verification system. H. Bao et al. [6] proposed emotion compensation method called emotion attribute projection to reduce the intra-speaker variability for speaker verification on emotional speech. Li and Yang [7] proposed the approach that exploits the prosodic difference to cluster affective speech for speaker modeling for speaker recognition and evaluated it using the Mandarin affective speech corpus. Shahin I. [8]-[9] carried out studies in speaker recognition in emotional environments and shouted talker conditions, using LFPC feature and different models such as second-order circular hidden Markov model and suprasegmental Hidden Markov Models. They inferred the two reasons for worsening the performance of the system namely, mismatched emotions between the speaker models and the test utterances, and the articulating styles of certain emotions which create intense intra-speaker vocal variability. Koolagudi et al. [10] used transformation of MFCCs for improving speaker identification performance under different emotions. They used emotional database of ten speakers under eight emotions in Telugu. Jawarkar et al. [11] studied the performance of speaker identifications in emotional speech environments for thirty four speakers using four features. Hanilci et al. [12] proposed joint density GMM mapping technique for compensating the MFCC features.

Whispered speech is a natural mode of communication that is used under situations to protect the content of speech information in natural conversation. It has been reported that the speech spectra reflect significant differences between whisper and neutral speech production. Differences include a complete loss of voiced excitation structure and a shift in formant frequencies in low frequency region [13]-[16]. Secondly, the spectral slope of whispered speech is flatter than that of neutral speech. The performance of speaker identification system trained with neutral speech degrades significantly due to the major differences between whisper and normal speech in both excitation and vocal tract function [15]. Fan and Hansen [15] have used three features viz. mel-frequency cepstral coefficients (MFCC), linear-frequency cepstral coefficients (LFCC) and exponential-frequency cepstral coefficients (EFCC) for speaker identification with whispered speech. Grimaldi and Cummins [17] compared the performance of MFCC and a mean-amplitude weighted short time estimate of instantaneous frequency based on AM-FM representation of speech signal for speaker identification. Sarria-Paja et al. [18] used feature based on AM-FM signal for speaker verification and gender detection. Jawarkar et al. [19] studied the performance of three features for speaker identification within whispered speech environment. Wang,

Jia-Ching, et al. [20] proposed a speaker identification system using instantaneous frequencies of the whispered speech signal approximated probability product kernel support vector machine for an access control and compared the performance with pyknogram-based system.

There are various approaches to speaker recognition. Amongst the prominent speaker modeling techniques are the Gaussian mixture model (GMM), vector quantization, artificial neural networks, support-vector machines, polynomial classifier. Jawarkar et al. [21] have employed fuzzy neural network for speaker identification. The GMM approach is most widely used for speaker recognition.

Main objective of the of the present work is to compare the performance of the feature sets, namely MFCC, TESBCC, LSF and newly proposed feature set TTESBCC, and study the effect of fusion of classifiers for test independent speaker identification for closed-set speaker identification within emotional and whispered speech environments.

Rest of the paper is organized as follows. Feature extraction process is described in Section 2. Experimental studies of the speaker identification system using various features are mentioned in Section 3. The classifier fusion technique is also discussed in this section. Finally conclusions are drawn in Section 4.

2. FEATURE EXTRACTION

The speech signal is first passed through anti-aliasing filter with cut-off frequency of 44.1 KHz. The signal is then sampled at the sampling frequency of 22050 Hz and converted into digital signal using analog to digital converter with 16-bit resolution. The silence removal stage removes the non-voiced portion of the signal based on the energy threshold criterion. The process of Feature extraction for various features used in the study is discussed in this section.

2.1 Mel frequency cepstral coefficients

The voiced speech signal, after silence removal, is pre-emphasized with pre-emphasis factor of 0.97. This is followed by frame blocking with a frame length of 512 samples (23.22 ms) with 50% overlap with the neighbouring frames. Finally each frame is multiplied with Hamming window to reduce the side lobe effects. Magnitude spectrum of each frame is obtained by taking FFT. This spectrum is multiplied by the 30 mel-scale triangular filters and then log energy is computed. The log-energy filter outputs are then cosine transformed to produce the cepstral coefficients. Twenty MFCCs are extracted from each frame. The zeroth cepstral coefficient is discarded as it contains only DC term. The remaining 19 coefficients are used in feature vector.

2.2 Temporal energy subband cepstral coefficients (TESBCC)

Sen and Basu [22] have proposed a set of parallel Nyquist filters and used it for extracting TESBCC feature. The Fourier transform of the proposed Nyquist window function is given below.

$$W(e^{j\omega}) = \begin{cases} \cos^2(\gamma\omega) & -(2\pi/N) \le \omega \le (2\pi/N) \\ 0 & |\omega| \end{cases} \quad (1)$$

where $\gamma = N/4$ and N is the window length.

Steps involved in computation of TESBCC are as under.

(i) The speech signal is pre-emphasized with pre-emphasis factor of 0.97 and then passed through a bank of thirty parallel filters described above.
(ii) Log energy of the subband signal of each frame of 23.22 ms length is computed
(iii) Discrete cosine transform of log-energies in each frame is finally obtained. First 19 coefficients are used in feature vector of TESBCC.

2.3 Weighted instantaneous frequency (WIF)

The WIF computation involves following steps.

(i) Speech signal $s[n]$ is passed through bandpass filters. Analytic signal $z_i[n]$ is then computed.

$$z_i[n] = x_i[n] + jy_i[n] \quad (2)$$

where $x_i[n]$ is the bandpass filtered signal of the i^{th} filter and $y_i[n]$ is the Hilbert transform of $x_i[n]$.

(ii) The instantaneous amplitude $a_i[n]$ and frequency $f_i[n]$ are computed for each filtered signal $x_i[n]$ as under.

$$a_i[n] = \sqrt{x_i^2[n] + y_i^2[n]} \quad (3)$$

$$f_i[n] = \frac{1}{2\pi T_s}\left(\theta[n] - \theta[n-1]\right) \quad (4)$$

where

$$\theta[n] = \arctan\left(\frac{y_i[n]}{x_i[n]}\right) \quad (5)$$

and T_s is the sampling period.

(iii) Each filtered signal is divided into frames (each of M samples) with frame increment of $M/2$ samples. Weighted instantaneous frequency for i^{th} filter and k^{th} frame is computed as under.

$$Wf_{ik} = \frac{\sum_{j=1}^{M} a_{ik}^2[j] f_{ik}[j]}{\sum_{j=1}^{M} a_{ik}^2[j]} \qquad k = 1, 2, 3\ldots \quad (6)$$

Twenty five Gabor filters, with centre frequencies uniformly spaced in mel-scale and bandwidth of 106 mel, are used for the experimentation. Number of samples per frame, $M = 512$.

2.4 Temporal Teager Energy based Subband Cepstral Coefficients (TTESBCC)

Literature survey shows the applications of Teager energy in the area of speech processing. Patil and Basu [23] have used Teager energy based cepstrum for identifying phonetically similar languages. Kandali et al. [24] have employed Teager energy based wavelet packet cepstral coefficients (tWPCC) for emotion recognition.

A new feature TTESBCC, which includes computation of Teager energy has been developed. Teager energy operator is defined as:

$$\psi_d(x[n]) = x^2[n] - x[n-1]x[n+1] \quad (7)$$

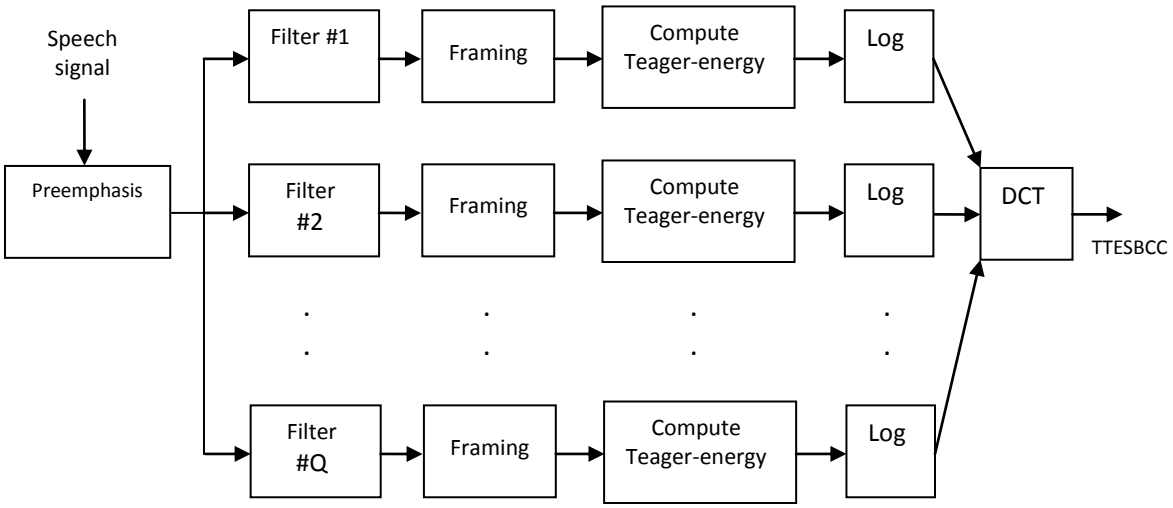


Fig.1 Block diagram for the TTESBCC computation

An important property of the Teager energy operator is that it is nearly instantaneous and has capability to capture energy fluctuations. It has time resolution that can track rapid signal energy changes within a glottal cycle [25].

The process of computation of TTESBCC involves following steps:

(i) The speech signal is pre-emphasized with pre-emphasis factor of 0.97 and then passed through a bank of thirty parallel Nyquist filters described in (1).
(ii) The Teager energy operator is applied to the output of each filtered signal to compute energy.
(iii) Discrete cosine transform of log-energies in each frame is finally obtained. First 19 coefficients are used in feature vector of TTESBCC. Fig.1 shows the block diagram for the TTESBCC computation.

2.5 Line Spectral Frequencies (LSF)

The feature Line Spectral Frequencies (also called as line spectral pair) was first introduced by Itakura [25] as an alternative to linear predictive coding (LPC) spectral representation. In the present work 16^{th} order LPC are computed using autocorrelation method for each frame of 23.22 ms length. LSF are then computed from LPC.

Cepstral mean subtraction is carried out to remove the effect of channel distortion or channel mismatch before using the feature sets described above for training and testing.

3. EXPERIMENTAL WORK

The experimental work involves the development of speaker identification systems for evaluating the performances with two databases, namely, emotional speech database and whispered speech database. Emotional speech database contains speech utterances in neutral and five basic emotions: anger, fear, sadness, happiness and disgust.

3.1 Performance evaluation for emotional speech environment

In this study emotional speech database, for thirty nine speakers in the age group of 16 to 50 years, was developed. The database includes speech utterances recorded in Marathi, one of the regional languages spoken over by 8 crore population in the Maharashtra state of India in two sessions. A notebook computer with onboard sound interface set, M-audio professional mobile digital sound recorder: Microtrack-II and Microphones were used for recording process.

Each speaker was instructed to utter isolated words, digits and sentences, five times, in neutral environment. These utterances were used for training the speaker model. Then the speakers were advised to utter two types of sentences (other than that used for training), one which are biased towards emotion and other which are unbiased towards emotion, in different emotions after rehearsal. It may be noted that the data of emotional speech is a synthetic one because it is a tutored emotion (after a few rehearsals) and not a natural emotion of anger or fear or sadness. Each sentence was uttered five times. These utterances were used for testing .The samples of sentences which are biased and unbiased towards the emotion (translated in English) are shown in Table 1.

Gaussian mixture model with 16 mixtures were developed for each speaker using Expectation and Maximization (E&M) algorithm. Each speaker model was trained using a neutral speech utterance of 1 minute duration. Identification performance for each classifier was carried out for 1, 3, 5 and 10 second test utterance lengths. The test speech was first processed to evaluate the sequence feature vectors. The sequence of feature vectors was divided into overlapping segments of feature vectors at the 23.2 ms frame rate. Thus, 1-second testing utterance contains 86 feature vectors. Speaker identification accuracy (SIA) is calculated as:

$$SIA = \frac{1}{N}\sum a_i \tag{8}$$

where a_i is the speaker identification accuracy of the i^{th} speaker and is defined as under:

$$a_i = \frac{N_C}{N_T} \times 100 \tag{9}$$

where N_C is the number of correctly identified segments and N_T is the total number of test segments for the i^{th} speaker. Base line system consists of speaker models developed using 19 dimensional MFCC using GMM. Thirty triangular filters are used for computing MFCC.

Table 1 Samples of sentences used for testing (translated in English)

Sentences that are biased towards emotions		**Sentences that are unbiased towards emotions**
Emotion	**Sentence**	1. The sky is cloudy today. 2. Pusad is a small town. 3. My uncle is a farmer. 4. What can I do now? 5. Mumbai is the capital of Maharashtra. 6. Let me explain you. 7. Is your father inside? 8. He works continuously. 9. Students study very hard. 10. Boys dance in the garden.
Anger	Why did he tear my picture? He broke my pen.	
Fear	Manager will remove me from job. What is there inside the dark room?	
Sadness	Ramu's buffalo has died. My brother has failed in the examination.	
Happiness	I stood first in the examination. Congratulations for getting job.	
Disgust	India has lost the test-match again. Why is not he behaving properly?	

Table 2 Speaker identification accuracy with feature set: MFCC, TESBCC and LSF

Em	**Speaker Identification Accuracy (%)**											
	Feature set: MFCC				**Feature set: TESBCC**				**Feature set: LSF**			
	1s	***3s***	***5s***	***10s***	***1s***	***3s***	***5s***	***10s***	***1s***	***3s***	***5s***	***10s***
A	50.95	59.38	59.41	59.98	55.80	65.33	66.89	68.08	50.30	58.16	57.39	58.06
F	53.80	63.04	65.31	67.45	59.53	68.89	71.12	73.50	64.13	68.79	70.77	72.99
S	51.15	59.24	61.49	62.04	53.81	62.05	64.27	66.70	61.33	68.88	69.91	71.02
H	45.49	50.25	51.69	52.10	58.93	62.89	65.61	68.96	56.24	61.87	63.57	65.06
D	43.81	51.45	52.88	53.03	56.40	68.07	71.82	73.69	57.44	62.58	63.33	63.95
Av	49.04	56.67	58.16	58.92	56.89	65.45	67.94	70.19	57.89	64.06	64.99	66.22
N	82.24	90.18	92.62	95.67	83.86	89.53	93.81	94.71	74.37	77.14	77.25	78.16

Em: Emotion, A: Anger, F: Fear, S: sadness, H: Happiness, D: Disgust, Av: Average, N: Neutral

Table 3 Speaker identification accuracy with combination of two-feature sets (with PCA)

Em	**Speaker Identification Accuracy (%)**											
	1s			***3s***			***5s***			***10s***		
	ML	**TL**	**TM**	**ML**	**TL**	**TM**	**ML**	**TL**	**TM**	**ML**	**TL**	**TM**
A	60.93	60.36	55.13	66.44	64.53	60.48	67.49	65.31	61.13	67.37	67.09	61.55
F	68.30	64.95	59.82	75.09	70.68	65.78	76.71	72.72	68.45	78.98	76.87	72.30
S	65.61	59.46	56.41	72.44	64.19	60.40	74.58	66.27	61.46	78.16	69.36	64.70
H	61.55	59.38	55.62	68.05	65.78	62.66	69.96	66.62	64.41	71.46	69.22	67.04
D	59.33	60.96	54.34	65.17	67.30	60.12	66.31	69.45	60.99	66.59	70.22	62.59
Av	63.14	61.02	56.26	69.44	66.50	61.89	71.01	68.07	63.89	72.51	70.55	65.64
N	83.15	83.87	87.63	86.55	87.36	92.12	87.93	87.98	93.91	89.06	88.51	95.57

Em: Emotion, A: Anger, F: Fear, S: Sadness, H: Happiness, D: Disgust, Av: Average, N: Neutral, ML: MFCC+LSF, TL: TESBCC+LSF, TM: TESBCC+MFCC

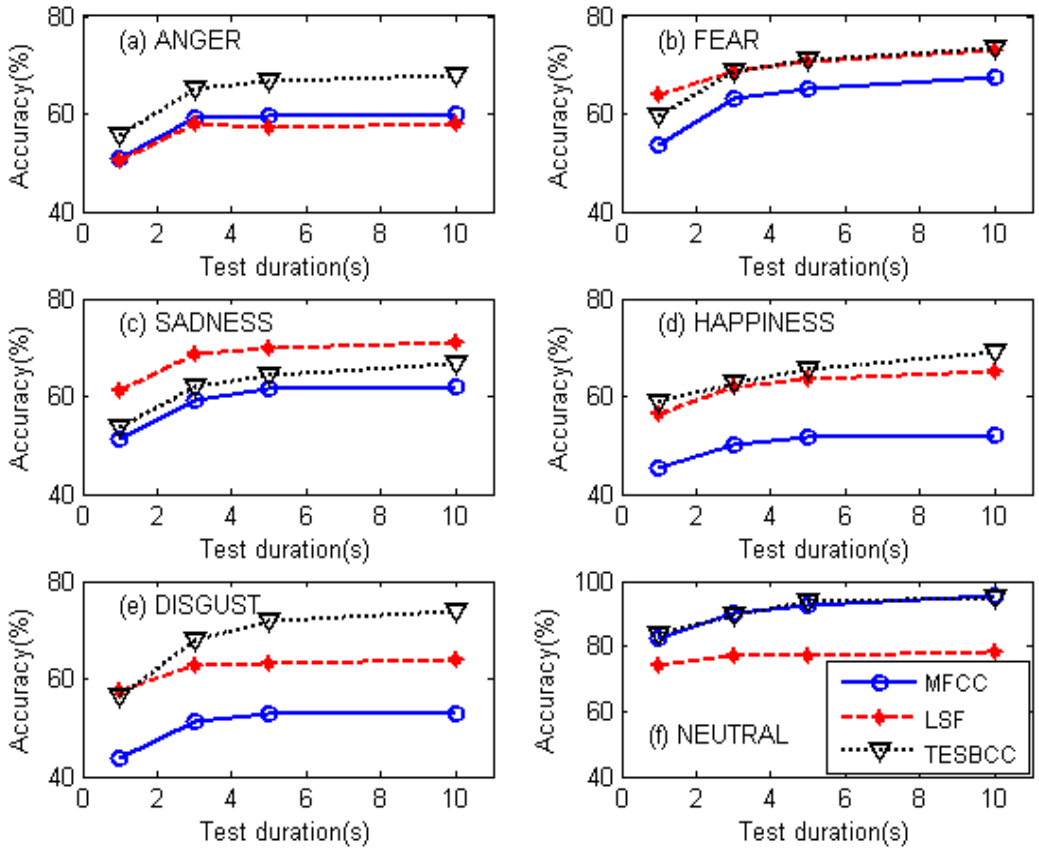


Fig. 2 Comparison of SI system performance for different emotions (a)Anger, (b)Fear, (c)Sadness, (d)Happiness, (e)Disgust and (f)Neutral

3.1.1 Performance with individual feature sets

Speaker identification accuracy for different emotions using feature sets: MFCC, TESBCC and LSF are shown in Table 2. Values in first five rows indicate SIA for different emotions. Values in sixth row indicate the average value of SIA for the five emotions for particular test duration. Values in seventh row indicate SIA under neutral environment. Comparison of the performance for different emotions is shown in Fig. 2. It is observed that MFCC outperforms the other features in neutral speech environment. However, the performance of the speaker identification system using MFCC degrades in emotional environments. In general, the speaker identification accuracy reduces in emotional environments for all the three feature sets used as compared to that in neutral environment. This may be due to intra speaker variability for different emotions. TESBCC outperforms other two features in anger, happy and disgust speech environments, whereas LSF outperforms other two feature sets in sad speech environment. TESBCC and LSF gives similar performance with speech test utterances in fear emotion. Therefore a novel classifier fusion technique is proposed.

3.1.2 Fusion of Classifier Outputs

A large number of methods have been developed for classifier fusion. It has been shown that multiple classification system can be used to enhance a number of pattern recognition applications [26]-[30]. Many fusion methods operate on classifiers which produce soft outputs. These are real values in the range [0, 1]. Mashao D. J. and Skosan M. [30] have combined the decision of two classifiers for improving the performance of a speaker recognition system. Doddington et al. [29] suggested improvement in the base line performance by a simple combination of scores obtained for different systems. Chen and Chi [28] discussed a method of combining multiple probabilistic classifiers using different feature sets extracted for speaker identification (SI) task. The new classifier fusion technique used for SI is discussed below.

Each speaker is represented by Gaussian mixture models. GMM is capable of representing a large class of sample distributions [31] and it is currently one of the principal methods for speaker identification. A speaker model which has the maximum *a posterior probability* for an unknown test utterance is declared as the winner. Thus output of each GMM classifier is the speaker number of the winning speaker. For performance evaluation of speaker identification system within emotional speech environments, three classifiers are developed namely, GMM-TESBCC, GMM-LSF and GMM-MFCC. Soft output of these classifiers is combined. Block-diagram for fusion of three classifiers is shown in Fig. 3.

The output of each GMM classifier is first computed. Final decision is made using the fusion decision logic. The decision logic can be described in following form.

```
if outputs (speaker number) of any two
classifiers are identical for a given
test utterance
then
  Apply majority rule
else
  Apply weighted sum rule
end if
```

Majority rule:
Output of the fusion system is speaker S and is determined as per the following majority rule.

$$S = S_i, \quad if \ S_i = S_j \quad ; i \neq j, \ i = 1,2,3 \ and \ j = 1,2,3 \qquad (10)$$

where S_i is the speaker class for i^{th} classifier and is computed as follows.

$$S_i = \arg \max_{1 \le k \le N} g_i^k \qquad (11)$$

where g_i^k is the normalized a posterior probability for k^{th} speaker model at the output of i^{th} GMM classifier such that $\sum_{k=1}^{N} g_i^k = 1$ and N is the number of speaker models.

Weighted sum rule:
Output class is decided by the weighted output as under.

where f_k is the weighted score of the k^{th} speaker and is given by:

$$S = \arg \max_{1 \le k \le N} f_k \qquad (12)$$

$$f_k = \sum_{i=1}^{3} w_i g_i^k \qquad (13)$$

where w_i is the weight associated with i^{th} classifier such that $\sum_{i=1}^{3} w_i = 1.0$.

Next, experiment was carried out by combining the two feature sets: (i) MFCC+LSF, (ii) TESBCC+LSF and (iii) TESBCC+MFCC, after applying principal component analysis (PCA). Results of speaker identification accuracy for combinations of feature sets are shown in Table 3. Finally, output of the three GMM classifiers: MFCC-GMM, TESBCC-GMM and LSF-GMM are fused as discussed in earlier in this section. The effect of variation of weights for three classifiers was first studied. The performance showing the effect of weight variation is shown in Fig. 4 and 5. It is observed that weights [W_T, W_L, W_M] = [0.5, 0.25, 0.25] gives the optimum SIA for emotional speech test utterances. There is minor change in SIA using neutral speech test utterances with variation in weights. Table 4 shows the results for classifier fusion.

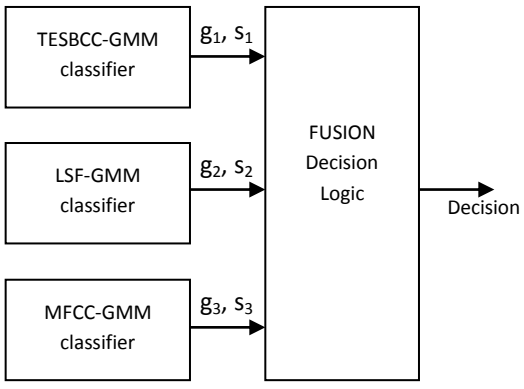


Fig. 3. Block diagram for fusion of three classifiers

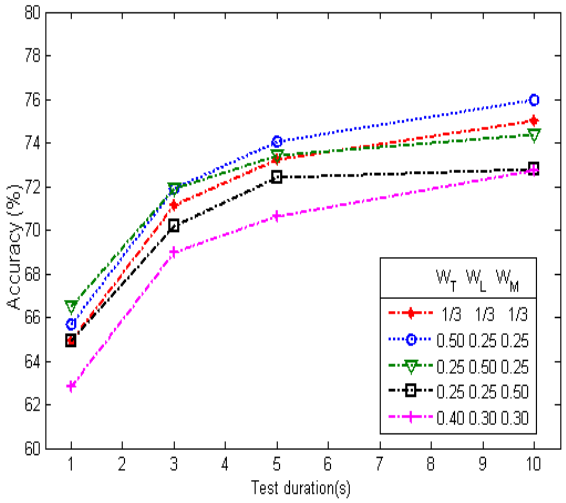


Fig. 4 Effect of variation in weights with emotional test speech

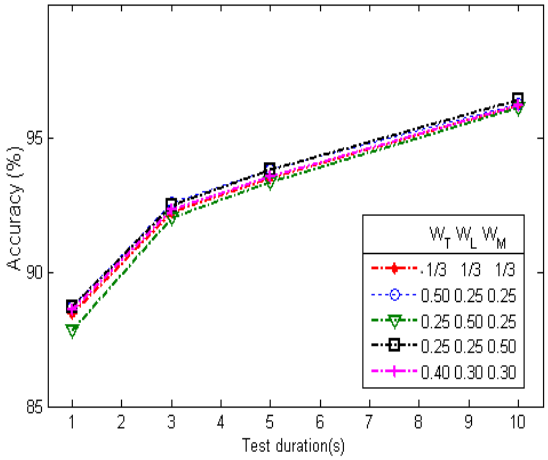
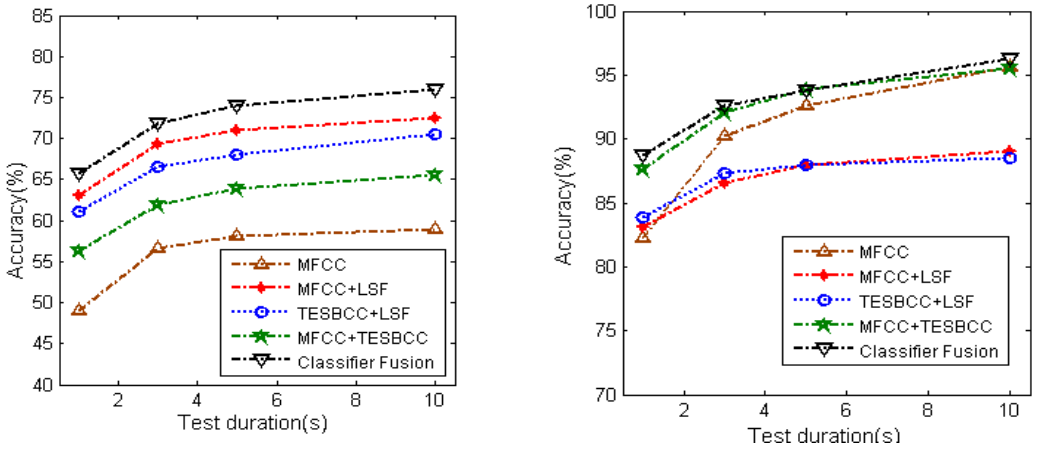


Fig. 5 Effect of variation in weights with neutral test speech

Fig. 6 shows the accuracy of speaker identification in emotional speech environment for (i) MFCC, (ii) MFCC+LSF, (iii) TESBCC+LSF, (iv) TESBCC+MFCC and (v) Classifier Fusion. It is observed that fusion of two features improves speaker identification in emotional speech environments. However, feature-fusion results in decrease in the speaker identification in neutral environment. It can be further seen that classifier fusion technique improves SIA both in emotional and neutral speech environments.

3.2 Performance evaluation within whispered speech environment

Whispered speech is a natural mode of communication that is used under situations to protect the content of speech information in natural conversation. Fig.7 shows the neutral and whispered speech signal versus time corresponding to a sentence in Marathi with meaning "Pusad is a small town". The corresponding magnitude spectra are shown in Fig. 8. It can be seen that there is drastic reduction magnitude spectra of the voiced portion of the whispered speech as compared to that with the unvoiced portion. This section evaluates the performance of speaker identification system with four different feature set within whispered speech environment.



(a)Testing with emotional speech **(b) Testing with neutral speech**

Fig. 6. Performance with feature fusion and Classifier fusion

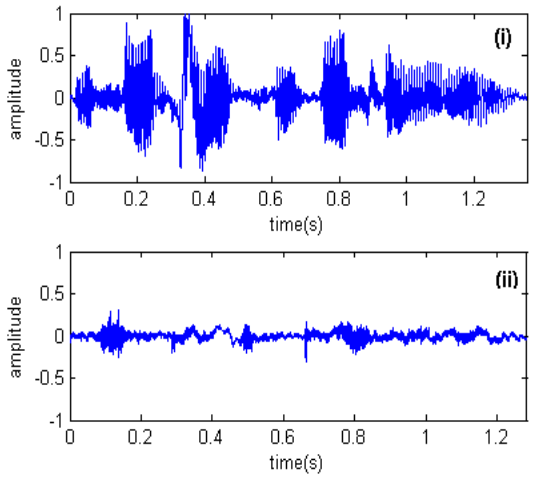


Fig. 7. Speech signal (i)Normal, (ii) Whispered

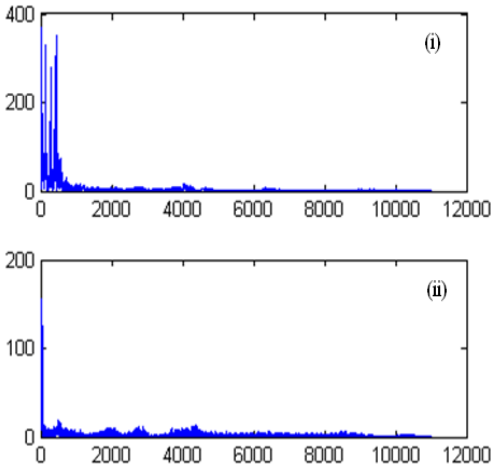


Fig. 8. Magnitude spectra (i) Normal and (ii) Whispered speech

Table 4 Speaker identification accuracy classifier fusion (weights: W_T = 0.5, W_L = 0.25 and W_M = 0.25)

Emotion	Speaker Identification Accuracy (%)			
	1s	*3s*	*5s*	*10s*
Anger	63.33	69.53	70.96	73.33
Fear	70.96	77.00	78.67	79.51
Sadness	65.87	71.79	74.33	77.63
Happiness	64.91	70.53	72.50	74.69
Disgust	63.36	70.74	73.89	74.64
Average	65.68	71.98	74.07	75.96
Neutral	88.70	92.61	93.81	96.28

3.2.1 Baseline System

Baseline system consists of speaker models developed using 19 dimensional MFCC using GMM. Table 5 shows the performance of the baseline system. It can be seen that there is a drastic reduction in the speaker identification accuracy (SIA) for the system tested with whisper speech utterances. This is mainly due to the major differences in excitation and vocal tract function of whispered and neutral speech.

Table 5 Results with MFCC-GMM Base line system

Testing speech mode	Speaker Identification Accuracy (%)			
	1 sec.	*3 sec.*	*5 sec.*	*10 sec.*
Whisper	31.12	34.38	35.73	37.87
Neutral	93.77	96.82	97.79	98.48

3.2.2 Performance using other features

Three systems TESBCC-GMM, WIF-GMM and TTESBCC-GMM were developed to study the speaker identification performance. Systems were trained using normal speech. Table 6 shows the performance of the speaker identification system. It can be seen that TTESBCC outperforms the other features when tested with whisper speech. This may be because of the capability of the *Teager* energy operator to capture energy fluctuations in whispered speech. However, there is a slight reduction in the accuracy of the system with TTESBCC when tested using neutral speech as compared to that of TESBCC.

3.2.3 Feature Separability Analysis

The accuracy of the recognition system is closely related to the structure of the feature space in which the speaker is being modelled. The most commonly used separability measures

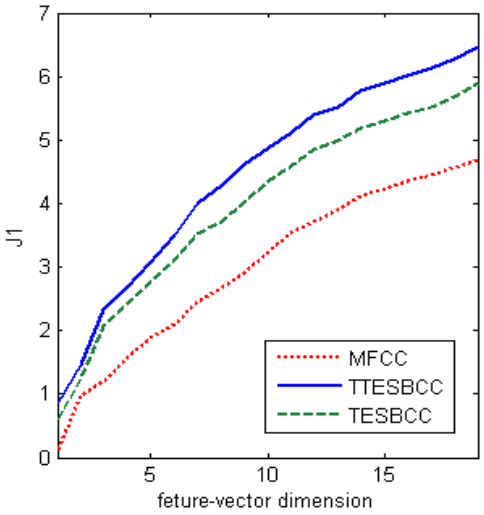


Fig. 9 Similarity measures J1 for MFCC, TTESBCC and TESBCC features for whispered speech

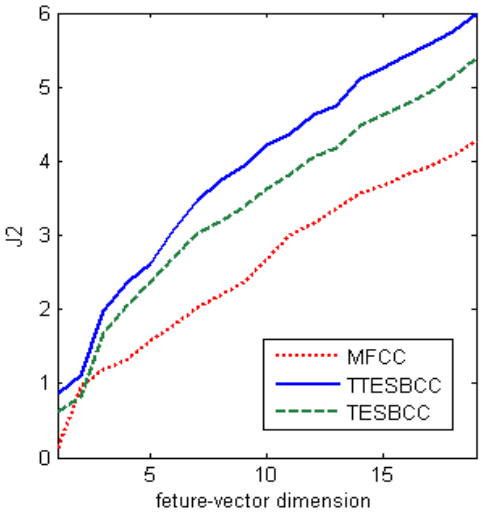


Fig. 10 Similarity measures J2 for MFCC, TTESBCC and TESBCC features for whispered speech

Table 6 Comparison of Results with Different Features

Feature	Testing speech mode	Speaker Identification Accuracy (%)			
		1 s	*3 s*	*5 s*	*10 s*
TESBCC	Whisper	37.31	46.46	49.52	51.29
	Neutral	94.55	98.18	98.78	98.81
TTESBCC	Whisper	38.39	49.59	52.41	55.80
	Neutral	94.46	98.12	98.50	98.62
WIF	Whisper	28.96	36.93	38.12	38.63
	Neutral	92.49	96.31	96.59	97.14

Table 7 System Performance with Classifier Fusion

Testing speech mode	Speaker Identification Accuracy (%)			
	1 s	*3 s*	*5 s*	*10 s*
Whisper	49.33	54.64	56.35	58.71
Neutral	98.47	99.54	99.58	99.73

used in speech/speaker recognition are the measures like F-ratio, and Chernoff and Bhattacharya bound [32]-[33]. The F-ratio only measures the separability of a single coefficient or dimension of the feature vector. J-measures are an extension to the F-ratio. J-measures are used to evaluate the discrimination of an entire feature set. Two of these measures that were used for the separability analysis in this work are:

$$J_1 = tr(W^{-1}B) \qquad 14$$

$$J_2 = \sum_{i=1}^{D} \frac{b_{ii}}{w_{ii}} \qquad 15$$

Where matrix B is the between class covariance, or covariance of class means, and measures how close the speech classes are separated from each another. Matrix W is the within class covariance matrix. This indicates how large the speaker classes are. b_{ii} and w_{ii} are the i^{th} diagonal elements of matrices B and W, respectively. D is number of diagonal elements.

The similarity measures, J_1 and J_2 versus feature vector dimension for the MFCC, TESBCC and TTESBCC features for the whispered speech for 25 speakers are shown in Fig. 9 and Fig. 10, respectively. It can be seen that TESBCC provides better feature separability than the other two features.

3.2.4 Effect of variation of number of speakers

Three separate systems using TESBCC & GMM are developed for 8, 16 and 25 speakers and tested with 10 seconds of speech utterance. Effect of variation of number of speakers on the performance of system is shown in Fig. 11. It can be seen that there is very small variation in the accuracy of the system tested with normal speech as number of speakers is increased from 8 to 25, whereas, the accuracy reduces from 71.90% to 55.80% for the same system tested with whispered speech.

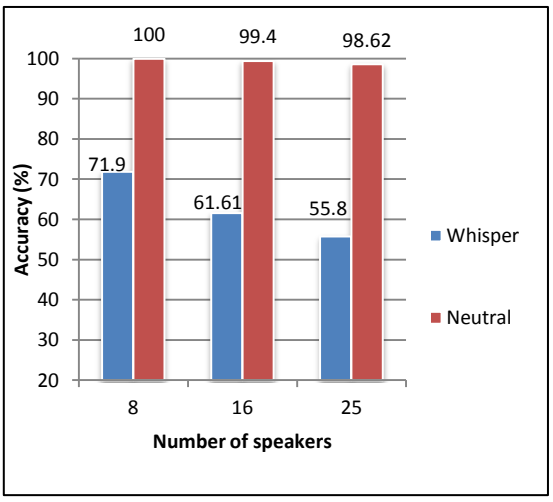


Fig. 11 Effect of variation in number of speakers

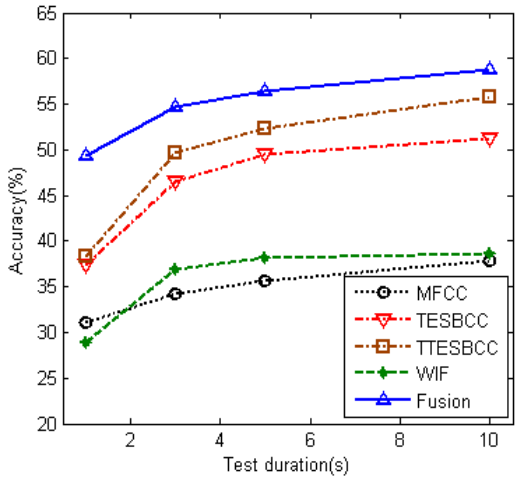


Fig. 12 Performance comparison with whispered speech

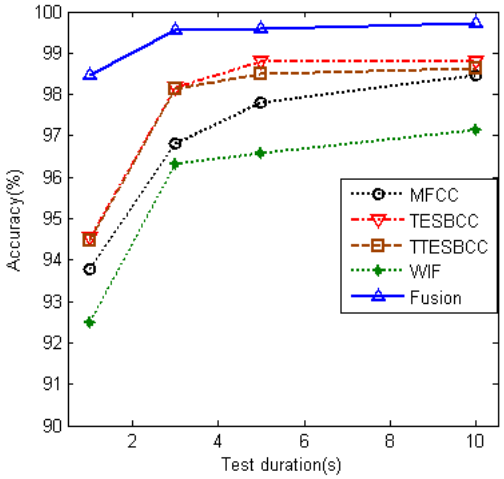


Fig. 13 Performance comparison with neutral speech

3.2.5 Classifier Fusion

In the present study output of the three systems MFCC-GMM, TTESBCC-GMM and WIF-GMM were combined and final decision is made using the fusion decision logic. Performance of the system with classifier fusion is shown in Table 7. It is observed that there is improvement in accuracy of speaker identification in both neutral and whispered environments. Comparison of the performance of various methods is shown in Fig. 12 and Fig.13. Fig.12 shows the accuracy for different test durations for whispered speech and Fig.13 shows the same for the normal speech.

4. CONCLUSIONS

A text-independent speaker identification system in emotional and whispered speech environments is presented. First, the comparative study of the performance of the MFCC, LSF and TESBCC feature sets is carried out for speaker identification within emotional speech environment. MFCC in general gives better performance with test utterance in the neutral environment. However, the performance of the system with MFCC deteriorates in the emotional speech environment. The combinations of features MFCC+LSF and TESBCC+LSF improve the accuracy of identification in the emotional speech environments; however their performances degrade in neutral environment. The fusion of three classifiers GMM-MFCC, GMM-LSF and GMM-TESBCC improve the speaker identification performance in both emotional and neutral environments.

Next, comparative study of the performance of the MFCC, TESBCC, WIF and TTESBCC feature sets was carried out for speaker identification in whispered speech environment. Speaker identification accuracy of the GMM-MFCC system for the test utterance in the whispered speech is lowest. The newly proposed feature TTESBCC outperforms the other features for testing using whispered speech. Fusion of GMM-MFCC, GMM-WIF and GMM-TTESBCC classifiers improve the speaker identification performance in both whispered and neutral environments.

The net improvement in the average accuracy of speaker identification in emotional and whispered speech environments using classifier fusion method over that with GMM–MFCC classifier for 10 second test-speech utterance is 27.04% and 20.84%, respectively.

In the present study is based on the use of GMM based classifier. Future work includes the use of multiple classifiers such as Support Vector Machine, Artificial Neural Network, etc. to enhance the system performance.

5. REFERENCES

[1] Furui, S. 1997. Recent advances in speaker recognition. *Pattern Recognition Letters*, vol. 18, No. 9, pp. 859–872.

[2] Faundez-Zanuy, M., and Monte-Moreno, E. 2005. State – of – the – art in speaker recognition. *IEEE Aerospace & Electronic Systems Magazine*, vol. 20, No. 5, pp. 7–12.

[3] Kinnunen, T., and Haizhou, L. 2009. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, vol. 52, pp.12–20.

[4] Picard, R. W. 1995. Affective computing, MIT Media Lab Perceptual Computing Section Tech. Rep. 321.

[5] Wu, W., Zheng, T. F., Xu, M. X., Bao, H. J. 2006. Study on speaker verification on emotional speech. *INTERSPEECH 2006.* Pittsburgh, Pennsylvania, USA, pp. 2102–2105.

[6] Bao, H., Xu, M. and Zheng, T. F. 2007. Emotional attribute projection for speaker recognition on emotional Speech. *INTERSPEECH 2007*, Antwerp, Belgium, pp. 758–761.

[7] Li D. and Yang Y. 2009. Emotional Speech Clustering based Robust Speaker Recognition System. *CISP09*, Tianjin, China, pp.1–5.

[8] Shahin, I. 2009. Speaker identification in emotional environments, *Iranian Journal of Electrical and Computer Engineering*, vol. 8, pp. 41–46.

[9] Shahin, I. 2013. Speaker identification in emotional talking environments based on CSPHMM2s. *Engineering Applications of Artificial Intelligence*, vol. 26, pp.1652–1659.

[10] Koolagudi S. G., Fatima S. E., Rao, K. S. 2012. Speaker recognition in the case of emotional using transformation of speech features, *Proceedings of CUBE International Information Technology Conference 2012*, Pune, India, pp.118–123.

[11] Jawarkar, N., Holambe, R., & Basu, T. 2012. Text-Independent Speaker Identification in Emotional

Environments: A Classifier Fusion Approach. *Advances in Intelligent and Soft Computing*, 133, pp.569–576.

[12] Hanilçi, C. 2013. Speaker identification from shouted speech: analysis and compensation. *ICASSP 2013*, Vancouver, Canada, pp. 8027–8031.

[13] Morris, R. W., and Clements, M. A. 2002. Reconstruction of speech from whispers. *Medical Engg. Physics*, vol. 24, no. 7–8, pp.515–520.

[14] Ito, T., Takeda, K., and Itakura, F. 2005. Analysis and recognition of whispered speech. *Speech Communicatin*, vol. 45, no. 2, 139–152.

[15] Fan, X., and Hansen, J. H. L. 2011. Speaker Identification within Whispered Speech Audio Streams. *IEEE Trans. Audio, Speech and Language Processing*, vol. 19, No. 5, pp.1408–1421.

[16] Matsuda, M. and Kasuya, H. 1999. Acoustic nature of the whisper. In *Proceeding of Eurospeech*. Budapest, Hungary, pp.133–136.

[17] Grimaldi, M., and Cummins, F. 2008. Speaker Identification Using Instantaneous Frequencies. *IEEE Trans. Audio, Speech and Language Processing*, vol. 16, no. 6, pp.1097–1111.

[18] Sarria-Paja, M., Falk, T. H., O'Shaughnessy, D. 2013. Whispered speaker verification and gender detection using weighted instantaneous frequencies. *ICASSP 2013*. Vancouver, Canada, pp. 7209–7213.

[19] Jawarkar, N. P., Holambe, R. S., and Basu, T. K. 2013. Speaker Identification using Whispered Speech. *IEEE Conf. CSNT 2013*, Gwalior, India, pp. 778–781.

[20] Wang, Jia-Ching, et al. 2015. Speaker identification with whispered speech for the access control system. *IEEE Transactions on Automation Science and Engineering*, vol.12, no.4, pp. 1191-1199.

[21] Jawarkar, N. P., Holambe, R. S., and Basu, T. K., 2011. Use of Fuzzy Min-Max Neural Network for Speaker Identification. *IEEE ICRTIT–2011*, Chennai, India, pp. 178–182.

[22] Sen, N., and Basu, T. K. 2011. Temporal Energy and Correlation Features from Nyquist Filter Bank for Text-Independent Speaker Identification. *Proceeding of IEEE Students Technology Symposium*, IIT Kharagpur, India, pp. 166–170.

[23] Patil, H. A., and Basu, T. K. 2008. Identifying perceptually similar languages using Teager energy based cepstrum, *Engineering Letters*, vol. 16 No. 1, pp.151–159.

[24] Kandali, A. B., Routray, A., Basu, T. K. 2009. Vocal emotion recognition in five native languages of Assam using new wavelet features. *Int. J. Speech Tech.*, pp.1–13.

[25] Kaiser, Z. F. 1990. On Teagers energy algorithm and its generalization to continuous signals. Proceeding of 4th IEEE digital signal processing workshop, MOHONK, New Paltz, NY.

[26] Itakura, F. 1995. Line spectrum representation of linear predictive coefficients of speech signals. *J. Acoust. Soc. Am*., vol. 53, pp.537A

[27] Ho, T.H., Hull, J.J., Srihari, S. N. 1994. Decision combination in multiple classifier system. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, pp.66–75.

[28] Kittler, J., Hatef, M., Duin, R., Mataz, J. 1998. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 20, pp.226–239.

[29] Chen, K., Chi, H. 1998. A method of combining multiple probabilistic classifiers through soft competition on different feature sets. *Neurocomputing*, vol. 20, pp. 227–252.

[30] Doddington, G., Przybocki, M., Martin, A., Reynolds, D. 2000. The Nist speaker recognition evaluation overview, methodology, systems, results, perspective. *Speech Communication*, pp.225–254.

[31] Mashao, D. J., Skosan, M. 2006. Combining classifier decisions for robust speaker identification. *Pattern Recognition*, vol. 39, pp.147 –155.

[32] Reynolds, D. A., Rose, R. C. 1995. Robust text-independent speaker identification using Gaussian mixture models, *IEEE Trans. on Speech & Audio Processing*, vol. 3, pp.72–83.

[33] Nicholson S., B. Milner and S. Cox 1997. Evaluating feature set performance using the F-ratio and J-measures, *Proc. of Eurospeech Conf. Speech Communication and Technology EUROSPEECH 1997*, pp. 413-416.

[34] Fukunaga K. 1990. Introduction to statistical pattern recognition. Academic Press, Boston, MA.