Wireless Pers Commun (2017) 97:933–950
DOI 10.1007/s11277-017-4544-1

Frame Selection for Robust Speaker Identification: A Hybrid Approach

Swati Prasad[1] · Zheng-Hua Tan[2] · Ramjee Prasad[3]

Published online: 30 May 2017

Abstract Identification of a person using voice is a challenging task under environmental noises. Important and reliable frame selection for feature extraction from the time-domain speech signal under noise can play a significant role in improving speaker identification accuracy. Therefore, this paper presents a frame selection method using hybrid technique, which combines two techniques, namely, voice activity detection (VAD) and variable frame rate (VFR) analysis. It efficiently captures the active speech part, the changes in the temporal characteristics of the speech signal, taking into account the signal-to-noise ratio, and thereby speaker-specific information. Experimental results on noisy speech, generated by artificially adding various noise signals to the clean YOHO speech at different SNRs have shown improved results for the frame selection by the hybrid technique in comparison with any one of the techniques used for the hybrid. The proposed hybrid technique outperformed both the VFR and the widely used Gaussian statistical model based VAD method for all noise scenarios at different SNRs, except for the Babble noise corrupted speech at 5 dB SNR, for which, VFR performed better. Considering the average identification accuracies of different noise scenarios, a relative improvement of 9.79% over the VFR, and 18.05% over the Gaussian statistical model based VAD method has been achieved.

Keywords Frame selection · Robust speaker identification · Biometric · Variable frame rate (VFR)

✉ Swati Prasad
in_sp@es.aau.dk; swatiprsd@gmail.com

[1] Department of Electronics and Communication Engineering, Birla Institute of Technology, Mesra, Ranchi, Jharkhand 835215, India

[2] Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark

[3] Department of Business Development and Technology, Aarhus University, Herning 7400, Denmark

1 Introduction

The process of finding out the speaker of a given speech utterance is referred to as Speaker Identification, and a system performing this task is referred to as Speaker Identification System. It is basically a Biometric system based on human voice. A speaker identification system broadly consists of three parts: (1) feature selection and extraction, (2) training/enrollment, and (3) testing/recognition. In feature selection and extraction, important and reliable speech parts from the speech signal are selected, and is then transformed into features. The features provide a compact representation of the speaker-specific information, sufficient to distinctly identify the individual from the rest of the speakers. The training/enrollment develops a model of each speaker using the features obtained from their recorded speech samples (training data). In the testing/recognition, the features of a given speech utterance from an unknown speaker is compared with each of the speaker models developed during training. The model with the highest similarity score is decided as the true speaker of the given utterance. A speaker identification system can be classified as Closed or Open. When it is known in advance that the given test utterance is spoken by a person belonging to a group/pool of N known speakers, it is called a closed system. Otherwise, it is called an open system. In an open system, one has to first determine, if the speaker belongs to the known group, and then the speaker is to be identified. A speaker identification system can be further classified as Text-independent, if the training and the test speech utterances are different; otherwise it is called as Text-dependent [1–4]. It is easy to understand that text-independent speaker identification is more difficult than the text-dependent.

In this paper, closed text-independent speaker identification system is studied. Here, identification of the speaker of the given test utterance is carried out from a group of N people (closed), and the test utterance is different from the training utterances used for speaker modeling (text-independent). For convenience, it is simply referred to as speaker identification system, in the rest of the paper.

Speaker identification system finds applications in surveillance, in crime scenes where the crank caller can be identified from a list of suspects and in automatic ID tagging. A promising usage of the proposed hybrid technique is in accessing remote devices which is commonly shared by many users. Here, speech can be used for identification and is provided from a distance through mobile phones. Once the device identify a given speech utterance as belonging to the group of authorized speakers, the person gain access of the device and a personalized service can be provided to him/her. It makes the whole process more secure and easy compared to the password based system which can be stolen/forgotten.

Achieving the best identification accuracy is the ultimate goal of any speaker identification system. One of the major challenges faced by these systems is its poor performance, when an acoustic mismatch between the training and the test conditions occur, referred to as mismatch problem. A commonly observed mismatch scenario is that, the speaker models are trained with clean training speech data, and the test data consists of environmental noises as well. To mitigate the mismatched condition problem, several approaches were tried in the recent past, and it is still a very active research area [5–7].

Robust methods have been developed at different levels of the speaker identification process. Robust features such as multitaper Mel-Frequency Cepstral Coefficients (MFCC) [8], multitaper Perceptual Linear Prediction (PLP) [9] and Mean Hilbert Envelope Coefficients (MHEC) [10] were studied. Speech enhancement techniques were applied, to make the noisy speech test data close to the clean training data [11–13]. The effects of "multicondition training" was evaluated, where multiple copies of the training data were generated by adding noises of different characteristics to the clean training data, and then using it for modeling speakers [14–16]. Recently, i-vector speaker models were used for compensation of channel/speaker variations [17]. Finally, robust methods at classifier's decision level were also explored [18].

Speech signals, being non-stationary are usually processed by first windowing it into shorter frames, where it exhibits quasi-stationary behaviour. Typically 20–30 ms long frame size is used with a fixed frame shift of approx. half the frame size [6, 19]. Capturing frames at a Fixed Frame Rate (FFR) is inefficient under environmental noise conditions, as it does not take into account the following points:

1. The speech utterance consists of both speech and non-speech regions. Generating frames in the non-speech regions, convey negligible information about the speaker. In addition to this, when environmental noises are present during testing and speaker models were trained in a clean condition, the non-speech regions depicting noise may greatly decrease the speaker identification accuracy. Therefore, it is desirable to remove the non-speech regions from the speech signal.
2. Speech signal also consists of fast changing and steady state speech regions. Fast changing speech regions, like plosives, appear for a very short duration of time and more frames are required from these regions to capture its characteristics properly. In contrast, steady state speech regions, like vowels, appear for a longer duration and fewer frames are required from these regions to avoid the unnecessary addition of the same type of speech characteristics.
3. Apart from these, a speech region may be unreliable, measured by the signal-to-noise ratio. Removal of these regions shall enhance the identification accuracy.

Research studies taking into account the above referred points, either individually or in combination, have been made. A spectral subtraction speech enhancement based Voice Activity Detection (VAD) method [20], and a novel likelihood ratio sign test based VAD method [21] were proposed to determine the speech and the non-speech part of a speech utterance. Jung et al. proposed a phoneme based feature frame selection method, in which minimum redundancy between selected frames but maximum relevance to the speaker model was targeted [22]. A frame selection method based on the weights assigned by two Gaussian mixture models, one from the speech and the other from the noise was presented by Fujihara et al. [23]. A Variable Frame Rate (VFR) analysis method in which the frame selection rate is varied depending upon the significance of the speech signal characteristics was proposed for the speech recognition application under noise [24, 25]. It has shown good performance. "Speech recognition" is different from "speaker identification" in the sense that, speech recognition targets to predict the spoken words of the speech utterance, whereas, speaker identification target to predict the speaker of the speech utterance.

Since, the joint study of all the above three referred points for the speaker identification application has not been done, this study focuses and tries to take into account all the three referred points during frame selection. This is expected to increase the speaker identification accuracy of the system under environmental noises. The present study first investigates the VFR analysis method [24] for a different application of speaker identification under environmental noises and then, the investigated VFR analysis method is utilized with the widely used Gaussian statistical model based VAD method [26] for proposing an effective hybrid frame selection technique.

The rest of the paper is organized as follows. The next Section presents the proposed frame selection method using hybrid technique. Section 3 discusses the database, the experiments conducted in this study, and the results obtained. Section 4 presents the future prospects of the proposed hybrid technique and finally, Sect. 5 concludes the paper with some future research directions.

2 Frame Selection Using Hybrid Technique

This section describes the proposed frame selection method using hybrid technique with an example.

2.1 Description

The method utilizes the same speech signal to select frames for feature extraction in two different ways as shown in Fig. 1. It first selects the active speech part of the signal, discarding non-speech, using the statistical model based VAD. The decision about the presence and the absence of speech is done by comparing the noisy speech feature vector with the estimated noise features in accordance with a decision rule based on the Likelihood Ratio Tests (LRT) [26]. It is explained in more details in the Sect. 2.1.1. Frames are then obtained from the selected active speech part, utilizing the conventional FFR analysis.

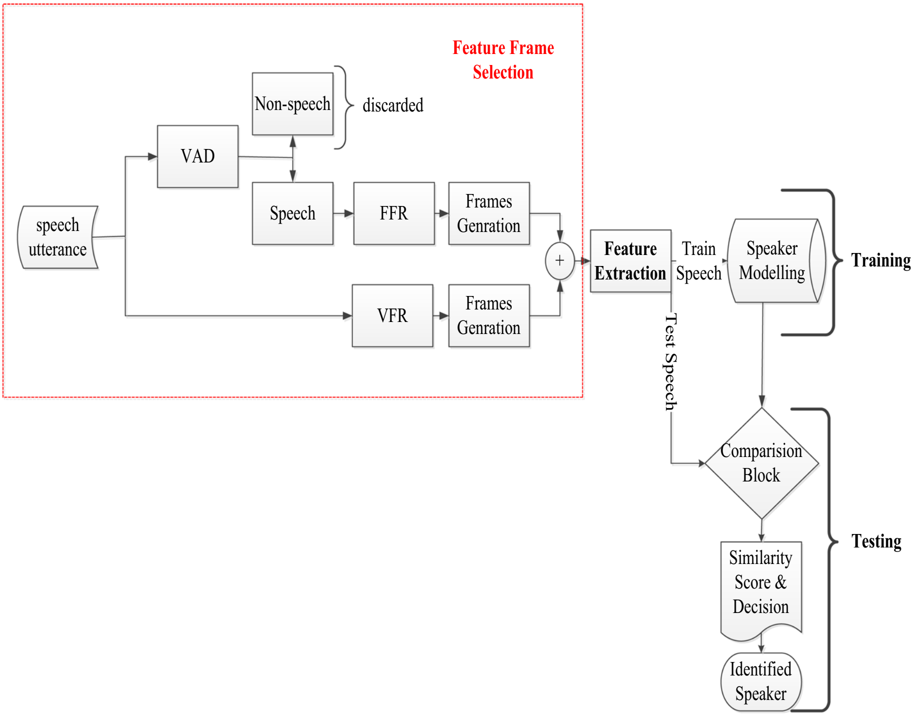


Fig. 1 Feature frame selection using hybrid technique

A frame size of 25 with 10 ms frame shift is used. This way of frame selection ensures, evenly capturing of the frames at a fixed rate, only from the active speech part of the signal.

Secondly, the same speech signal is processed again, using the VFR analysis for selecting frames according to the changes in the temporal characteristics of speech. In this, dense frames are first selected using the FFR analysis with frame size of 25 ms and frame shift as low as 1 ms. Distances between two adjacent frames are then calculated using the difference in energy. These distances are weighted by the signal-to noise ratio (SNR) value for the additional measurement of the reliability of the frame. Lastly, frames with accumulated SNR weighted energy distances above a particular threshold are selected and others are discarded. This results in dense frame selection around fast changing speech regions, sparse frame selection from the steady state regions and no frame selection from the non-speech regions [24]. It is further described below in Sect. 2.1.2.

Finally, frames selected in these two different ways from the same speech utterance are simply concatenated to yield the hybrid frames. The average number of frames selected per second (frame rate) by the hybrid method can be kept near to the 100 Hz frame rate of the conventional FFR analysis, so that more storage space may not be required. This is discussed in more details in the Sect. 3.2.

2.1.1 Statistical Model Based VAD

The widely used Gaussian statistical model based VAD is used in this study. It utilizes a generalized Likelihood Ratio Test (LRT) for decision making, which employs a decision-directed method to estimate the a priori SNR in signal [26].

1. Assuming additive noise, a binary hypothesis may be formulated as:

$$\begin{aligned} H0 &: Y(t) = N(t) \qquad \text{(Speech absence)} \\ H1 &: Y(t) = S(t) + N(t) \qquad \text{(Speech presence)} \end{aligned} \tag{1}$$

where, $Y(t)$, $N(t)$ and $S(t)$ represents the noisy speech, noise, and speech, respectively at frame t, given by the k-dimensional Discrete Fourier Transform (DFT) coefficients :

$$\begin{aligned} Y(t) &= [Y_0(t), Y_1(t), \ldots, Y_{k-1}(t)]^T \\ N(t) &= [N_0(t), N_1(t), \ldots, N_{k-1}(t)]^T and \\ S(t) &= [S_0(t), S_1(t), \ldots, S_{k-1}(t)]^T \end{aligned} \tag{2}$$

2. Considering $Y(t)$, $N(t)$, and $S(t)$ as asymptotically independent Gaussian random variables, the probability density functions conditioned on $H0$ and $H1$ are given by:

$$p(Y(t) \mid H0) = \prod_{j=0}^{k-1} \frac{1}{\pi\lambda_{n,j}} \exp\left(- \frac{| Y_j(t) |^2}{\lambda_{n,j}} \right) \tag{3}$$

$$p(Y(t) \mid H1) = \prod_{j=0}^{k-1} \frac{1}{\pi(\lambda_{n,j} + \lambda_{s,j})} \exp\left(- \frac{| Y_j(t) |^2}{(\lambda_{n,j} + \lambda_{s,j})} \right) \tag{4}$$

where, $\lambda_{n,j}$ and $\lambda_{s,j}$ represents the variances of N_j and S_j, respectively.

3. The likelihood ratio for the jth frequency bin is defined as:

$$\wedge_j \equiv \frac{p(Y_j(t) \mid H1)}{p(Y_j(t) \mid H0)} = \frac{1}{1+\xi_j} \exp\left(\frac{\gamma_j \xi_j}{1+\xi_j}\right) \tag{5}$$

where, $\xi_j \equiv \frac{\lambda_{s,j}}{\lambda_{n,j}}$, and $\gamma_j \equiv \frac{|Y_j(t)|^2}{\lambda_{n,j}}$ represent the a priori and a posteriori signal-to-noise ratio, respectively. ξ_j is estimated by the decision directed method [26].

4. The final decision rule is given by the geometric mean of the individual frequency bands,

$$\log \wedge = \frac{1}{k} \sum_{j=0}^{k-1} \log \wedge_j \gtrless_{H0}^{H1} \eta \tag{6}$$

where, η represents a preset threshold.

2.1.2 VFR Analysis Method

The VFR analysis method used is based on the "a posteriori" SNR weighted energy distance" [24].

1. Dense frames of the speech utterance are selected at a fixed rate with 25 ms frame size and frame shift as low as 1 ms.
2. A posteriori SNR weighted energy distance between two adjacent frames is calculated as:

$$D_{SNR}(t) = |\log E(t) - \log E(t-1)| \times SNR_{post}(t) \tag{7}$$

where, $E(t)$ is the energy of frame t, and $SNR_{post}(t)$ is the estimated a posteriori SNR value of frame t.

3. For the calculation of the threshold T for frame selection, a variant sigmoid function of $\log E_{noise}$ is used. This function is used to set a smaller threshold for clean speech, so that more frames may be generated from it. The variant sigmoid function is expressed as:

$$f(\log E_{noise}) = A + \frac{B}{1 - e^{-2(\log E_{noise} - 13)}} \tag{8}$$

Here, E_{noise} is the estimated noise energy of the utterance, parameters A and B decide the average frame rate (number of frames per second) and is discussed again in Sect. 3.2. The value 13 is chosen to make the turning point of the sigmoid at an a posteriori SNR between 15 and 20 dB.

Finally, threshold T is given by:

$$T = \overline{D_{SNR}(t)} \times f(\log E_{noise}) \tag{9}$$

where, $\overline{D_{SNR}(t)}$ is the average of the distances $D_{SNR}(t)$ taken over the entire utterance.

4. Frame selection is performed in this step. The weighted distances $D_{SNR}(i)$ are accumulated as: $Acc(i) = Acc(i-1) + D_{SNR}(i)$ from frame $i = 1, 2, 3, \ldots$ and $Acc(0) = 0$.

Whenever, $Acc(i) > T$ at i = n, frame n is selected and the Acc value is reset. The accumulation process is restarted from i = n+1. This is continued till the end of the frames.

The $SNR_{post}(t)$ in (7) is calculated as:

$$SNR_{post}(t) = \log \frac{E(t)}{E_{noise}} \quad (10)$$

E_{noise} in (8) and (10) is estimated by taking the average energy of the initial 10 frames that are assumed to be non-speech only, corresponding to approximately 34 ms long signal.

2.2 An Example

Figures 2 to 4 illustrates how frames are selected using the VAD, VFR and the proposed hybrid technique for the speech utterance "73" i.e. Seventy Three. For comparison, frames selected using the FFR analysis has also been shown. In these Figs., Panel-1 shows the time-domain waveform of the speech utterance together with the VAD decision (pulsed waveform, in which speech part is shown above 0 level and the non-speech part is shown as the 0 level). Panel-2 shows the wideband spectrogram of the utterance. Panel-3 shows the frame selection by the conventional FFR analysis with 25 ms frame size and 10 ms frame shift. Each bar indicates that a frame has been selected. Panel-4 shows the frame selection using FFR analysis as in Panel-3, but only from the active speech part selected by the VAD (indicated in Panel-1). Panel-5 shows the VFR frame selection, and Panel-6 shows the proposed frame selection using the hybrid technique, referred to as Proposed.

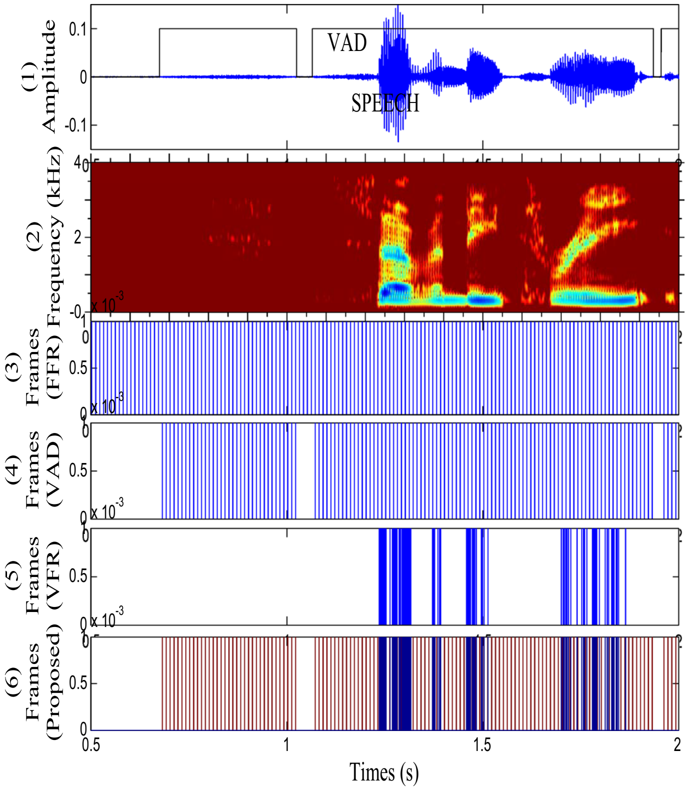


Fig. 2 Frame selection for the clean speech utterance

Figure 2 presents the frame selection by the above mentioned methods for the clean utterance. Figure 3 depicts the frame selection for the Babble noise corrupted speech utterance at an SNR of 15 dB (Fig. 3a) and 5 dB (Fig. 3b), respectively. Similarly, Fig. 4 shows the frame selection for the Car noise corrupted speech utterance at an SNR of 15 dB (Fig. 4a) and 5 dB (Fig. 4b), respectively.

From Figs. 2 to 4, the followings can be observed:

- Compared to the FFR analysis, VAD tries to captures frames at a fixed rate only from the active speech part. VFR analysis method selects more frames at the transient regions, fewer frames at steady state regions and no frames from the non-speech regions.
- In the presence of Babble noise (Fig. 3), VFR performed better, compared to the VAD in rejecting the non-speech part, which is before the start of the utterance.
- But in the presence of Car noise (Fig. 4), the VAD method captured the speech part more efficiently than the VFR.

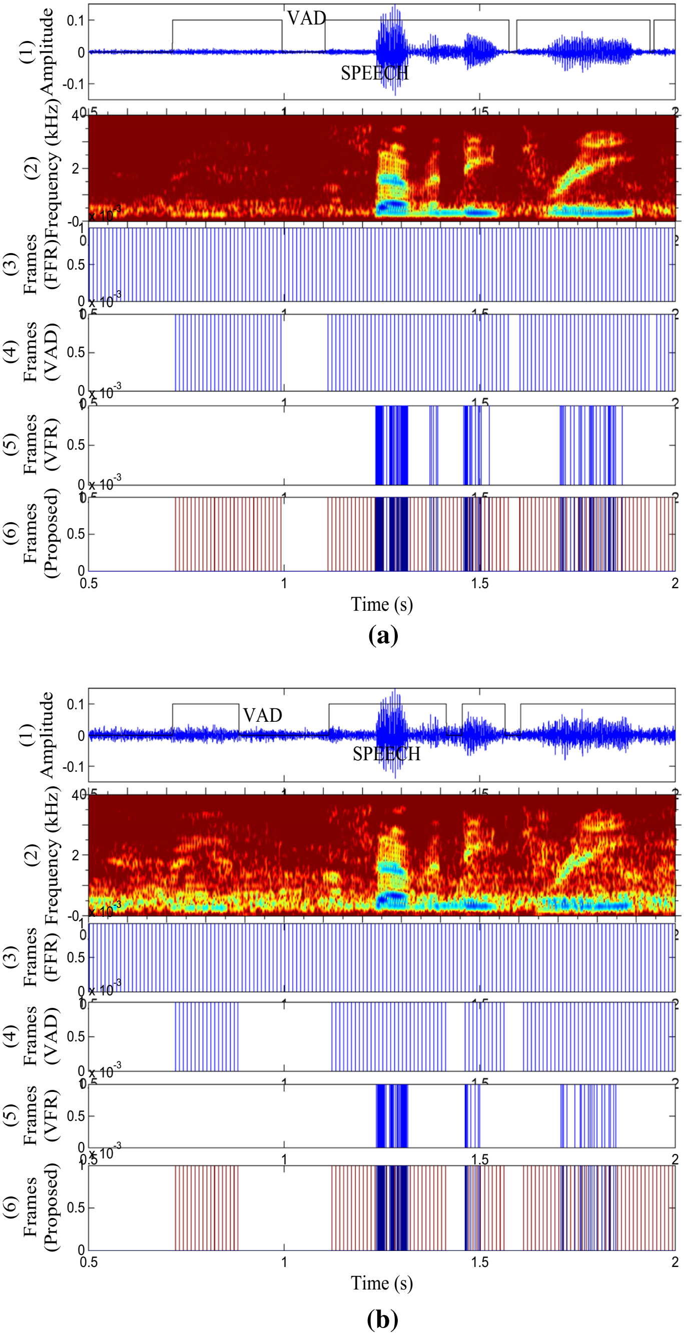
This gives an indication that the use of the hybrid technique for frame selection may prove beneficial as it will be adding up the different and complementary characteristics of the individual methods used for the hybrid.

3 Experiments and Results

This section is organized as follows. Section 3.1 describes the speech database used for the experiments. Section 3.2 investigates the different parameter values for the threshold function of the VFR method for selecting the average frame rate, and its effect on the identification accuracy is studied. Section 3.3 describes the different speaker identification experiments conducted for evaluating the performance of the proposed frame selection method using hybrid technique. Finally, Sect. 3.4 presents the results obtained and discussions.

3.1 Database

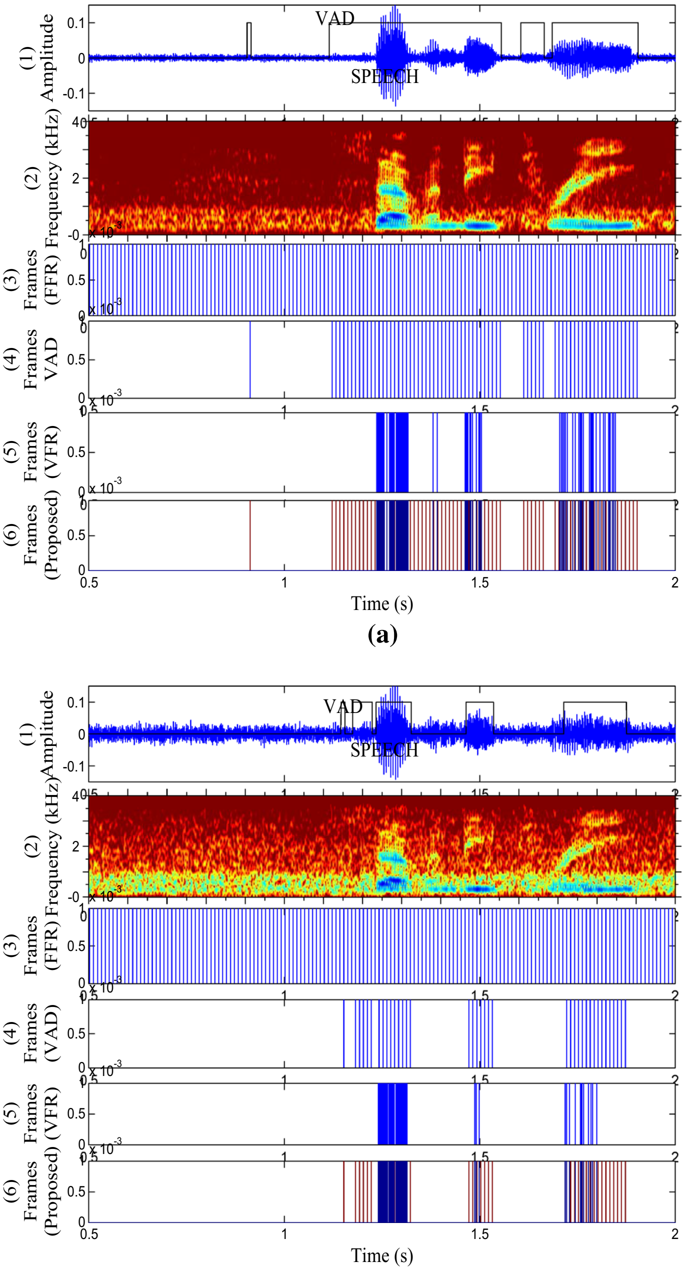
Noisy YOHO database were used for conducting the experiments. For this, eight different types of noise signals from the Aurora II database [27], representing environmental noises were taken. These noise signals were artificially added to the clean YOHO speech [28] to generate the noisy YOHO database. YOHO database consists of 138 speakers (106 males and 32 females) with connected number utterances like “76-39-57”. It was recorded in a three months time period from a quite office environment. Although some office noise was present, in this work, it was considered as clean speech. The training data were collected separately from the testing data from 4 recording sessions collecting 24 utterances (approximately 5s long) from each. Therefore, a total of 96 utterances (480s of speech) per speaker were used for training. For testing, 10 recording sessions were conducted collecting 4 utterances (approximately 5s long) from each. Therefore, a total of 40 utterances per speaker were used, resulting in 5520 tests to evaluate the system. To generate the noisy YOHO speech database, eight different noises, namely, Babble, Exhibition, Restaurant, Airport, Car, Street, Subway and Train from the Aurora II database were artificially added to the clean YOHO speech at four different SNRs of 5, 10, 15 and 20 dB. For noise addition, an equal speech length noise signal is randomly cut from the noise signal (assumed to be very long compared to the speech utterances) and is then added to the speech



(a)

(b)

Fig. 3 Frame selection for the Babble noise corrupted utterance at SNR of 15 and 5 dB. **a** 15 dB SNR. **b** 5 dB SNR



(a)

(b)

Fig. 4 Frame selection for the car noise corrupted utterance at SNR of 15 and 5 dB. **a** 15 dB SNR. **b** 5 dB SNR

at the desired SNR. To train the speaker models, only clean YOHO train data were utilized, and for testing, both clean and noisy YOHO test data were used.

3.2 Average Frame Rate and Speaker Identification Accuracy

The average frame rate i.e., the number of frames selected per second by the VFR analysis method varies with the values of the parameter A and B of the sigmoid function (8), used in the calculation of the threshold (9) for frame selection. For the speech recognition application [24], in which, what is spoken i.e., the content, plays an important role, it is found that, using the parameter values of A = 9 and B = 2.5, gave an average frame rate of 100 Hz and it showed good speech recognition performance. Here, choosing a frame rate which was not near to the frame rate value of 100 Hz, resulted in a mismatch with the front-end processing and significant decrease in recognition accuracy was observed.

The present study deals with the speaker identification, where what is spoken is not important, instead who is speaking i.e., the person plays an important role. Therefore, the effect of different values of parameters A and B on the speaker identification accuracy is investigated in this Subsection. Different A and B values were chosen to select the average frame rate in the range between 50 and 100 Hz, and the same is tabulated in Table 1.

Speaker identification experiments with VFR analysis method were carried out by using these values of A and B corresponding to frame rates between 50 and 100 Hz. These experiments were conducted on a smaller YOHO database, consisting of initial 21 speakers, and the noisy test data corrupted with only Babble and Car noises were included. The results obtained were tabulated in Table 2.

From Table 2, it can be observed that changing the parameters A and B that results in different frame selection rate did not result in any significant change in the average speaker identification accuracy. An improvement in average speaker identification accuracy is seen, when the value of the parameters were A = 9 and B = 2.5 corresponding to a frame selection rate of 60 Hz. As the present study aimed for obtaining the highest identification accuracy, the average frame rate of 60 Hz with the parameter values of A = 9, and B = 2.5 were chosen for all further experiments involving VFR analysis method, either individually or in hybrid.

The proposed frame selection method using hybrid technique simply concatenates the frames selected by the VAD and the VFR method. Using the values A = 9 and B = 2.5 in the threshold function, corresponding to 60 Hz average frame rate of VFR analysis method, the average frame rate of the proposed frame selection using the hybrid technique is found to be 110 Hz. This is considered optimal as it will use almost the same storage space, which is used in the conventional FFR analysis with 100 Hz frame rate.

Table 1 Parameters A, B and average frame rate

Parameters		Average frame rate (Hz)
A	B	
12	2.5	50
9	2.5	60
7	2	70
5	2	80
4	1.5	90
3	1.5	100

Table 2 Identification accuracies (%) for different average frame rates in VFR method

Noise	SNR (dB)	Average frame rate (Hz)					
		100Hz A = 3, B = 1.5	90 Hz A = 4, B = 1.5	80 Hz A = 5, B = 2	70 Hz A = 7, B = 2	60 Hz A = 9, B = 2.5	50 Hz A = 12, B = 2.5
Clean	—	97.98	98.21	98.33	97.86	97.98	98.33
Babble	20	96.31	95.83	96.31	96.55	96.67	95.95
	15	92.86	92.74	93.21	93.21	93.45	92.98
	10	80.36	81.55	81.79	82.86	81.43	82.50
	5	55.36	56.19	52.62	57.98	59.40	56.55
Car	20	94.52	94.17	94.52	94.17	93.69	93.81
	15	84.88	85.60	85.83	85.24	85.95	84.17
	10	63.69	64.05	63.45	63.93	65.36	64.88
	5	40.95	41.43	37.36	41.07	41.79	38.10
Average		78.55	78.86	78.14	79.20	79.52	78.59

3.3 Speaker Identification Experiments

A comparison of the main methods used in the VFR analysis has been carried out by Guarasa et al. [29] for speech recognition application. The energy weighting and the distance accumulation approach [30] showed better performance over the other methods. Recently, "a posteriori" SNR weighted energy distance based VFR analysis method has shown good performance for the speech recognition over these and other methods [24, 25]. Therefore, in the present study for speaker identification under various noise conditions, "a posteriori" SNR weighted energy distance based VFR analysis method has been selected for investigation and as one of the baseline. Let us call this method as Bsln-VFR.

The standard Gaussian statistical model based VAD has been selected as the second baseline for the study. Let us call this method as Bsln-VAD. In addition to these two, the conventional FFR approach, employing no robustness method has also been included for performance comparison. Let us call this method as Bsln-no robustness. Speaker identification experiments using the complete YOHO database consisting of 138 speakers were conducted for the Bsln-no robustness, Bsln-VAD, Bsln-VFR and the proposed frame selection method using hybrid technique, referred to as Proposed HVV (Hybrid VAD and VFR).

Speech features for speaker modeling were extracted from each of the selected frames. Twelve Mel-Frequency Cepstral Coefficients (MFCC) excluding the 0th coefficient were used for feature extraction. Speakers were modeled by a 64 component Gaussian mixture model utilizing the expectation-maximization algorithm. The speaker whose model maximizes the likelihood of the test utterance was decided as the correct speaker.

3.4 Results and Discussions

The identification accuracies (no. of correctly identified utterances/total no. of utterances tested × 100 %) for the different speaker identification experiments conducted in this study under clean and noisy test data are shown in Table 3. It can be observed that, the Proposed

Table 3 Identification accuracies (%) for the different methods across various noise scenarios

Noise	SNR (dB)	Bsln-no robustness	Bsln-VAD	Bsln-VFR	Proposed HVV
Babble	20	67.71	88.81	88.25	91.91
	15	46.63	73.51	81.91	86
	10	22.79	46.86	65.05	70.61
	5	6.35	16.13	36.82	36.66
Average		35.87	56.33	68.01	71.29
Exhibition	20	37.13	78.25	75.91	84.16
	15	17.71	54.19	55.64	65.25
	10	7.78	26.65	26.63	32.72
	5	3.05	9.48	7.81	10.71
Average		16.41	42.14	41.5	48.26
Restaurant	20	70.59	88.74	88.51	91.79
	15	50.38	72.34	82.12	85.97
	10	24.6	45.45	63.67	69.76
	5	6.76	16.35	29.35	33.14
Average		38.08	55.72	65.93	70.17
Airport	20	60.66	88.74	88.67	91.82
	15	39.49	75.69	82.04	86.64
	10	18.91	52.79	67.31	72.44
	5	6.64	22.68	37.13	41.13
Average		31.43	59.98	68.88	73.01
Car	20	44.53	87.27	84.3	89.14
	15	26.11	73.06	70.07	78.57
	10	11.71	49.77	46.37	56.11
	5	3.44	22.62	21.08	26.76
Average		21.45	58.18	55.46	62.65
Street	20	46.12	82.23	79.08	86.46
	15	27.5	63.71	64.96	73.93
	10	12.78	39.99	40.48	51.16
	5	5.02	18.33	16.01	22.46
Average		22.86	51.07	50.13	58.5
Subway	20	29.15	74.6	73.77	80.98
	15	11.86	48.95	53.77	60.77
	10	4.21	20.74	26.63	30.82
	5	1.12	7.32	8.56	9.92
Average		11.78	37.9	40.69	45.63
Train	20	52.96	86.55	86.68	90.52
	15	30.75	73.48	77.27	82.81
	10	12.69	52.62	57.43	64.3
	5	3.83	26.42	27.36	32.83
Average		25.07	59.77	62.19	67.62
Total average		25.34	52.64	56.6	62.14
Clean		91.32	96.74	91.78	94.98

HVV method has shown the highest identification accuracies compared to the baseline methods for all noise scenarios at all SNRs considered, except for the Babble noise corrupted speech at 5 dB SNR in which Bsln-VFR performed better. When an average of the identification accuracies of different noise scenarios has been considered, Proposed HVV method achieved an absolute improvement of 36.80% and a relative improvement of 145.22% from the Bsln-no robustness method. From the Bsln-VAD, it achieved an absolute improvement of 9.50% and a relative improvement of 18.05%. It has also shown an absolute improvement of 5.54% and a relative improvement of 9.79% from the Bsln-VFR method.

It can also be observed that, the Proposed HVV has achieved a good performance for the Babble, Restaurant and Airport noise scenarios at SNR of 5 dB compared to the Bsln-VAD. It has achieved an absolute improvement of 20.53% and a relative improvement of 127.28% for the Babble noise, an absolute improvement of 16.79% and a relative improvement of 102.69% for the Restaurant noise and an absolute improvement of 18.45% and a relative improvement of 81.35% for the Airport noise over the Bsln-VAD method.

Compared with the Bsln-VFR, it has achieved an absolute improvement of 6.45% and a relative improvement of 40.29% for the Street noise and an absolute improvement of 5.47% and a relative improvement of 20% for the Train noise at SNR of 5 dB. It is seen that the performance of the Bsln-VAD method is better than the Proposed HVV method for clean speech. It is also seen that the performance of the Proposed HVV method is better than the Bsln-VFR for clean speech. As Proposed HVV combines VAD and VFR method, it is concluded that for clean speech, the effects of VFR is more on the Proposed HVV method. However in case of noisy conditions, Proposed HVV provides better results in comparison to both VAD and VFR methods.

Figure 5 also shows the comparison of the Proposed HVV method with the different baseline methods for various noise scenarios. The graph shows the identification accuracies calculated by averaging the accuracy values at the four SNR values of a noise scenario. It

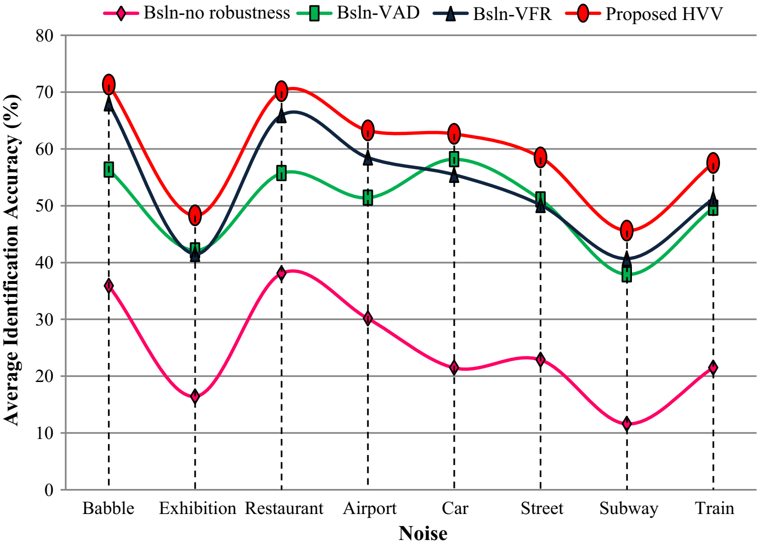


Fig. 5 Comparing average identification accuracies of the proposed and baseline methods for various noise scenarios

also confirms the better performance of the Proposed HVV method over the baseline methods.

Figure 6 compares the average identification accuracies of the proposed and the baseline methods at four SNR values and is also summarized in Table 4. The average identification accuracy is calculated by averaging the identification accuracies of the eight noise scenarios at a particular SNR value. It also shows the better performance of the Proposed HVV over other baseline methods.

4 Future Prospectus

A promising usage of the proposed hybrid method is in person identification through voice from a distance. It is used for gaining access to a shared application from a distance. Since, telephony system is already in use, the voice of the person for identification can be easily provided from a distance through a mobile phone. Once the person is identified as an authorized person from a group of N people, the person gain access to the application and a customized service can also be provided to the person. Another usage of this technique is in smart television. Imagine an old person giving command to the smart TV through voice, the smart TV, on the other hand, will first recognize the voice as one of the authorized person from a group, say, one of the family members, and then play the channels according to the person preferences. Smart TV is one such example, others can be smart air conditioner, smart door lock and smart washing machine. The success of the speaker identification has the power to completely revolutionize the consumer electronics market. Through this, more secure, compared to the password based method, which can be forgotten or stolen, and easy interaction with the electronic device can be achieved. This will be particularly liked by the elderly members of the society.

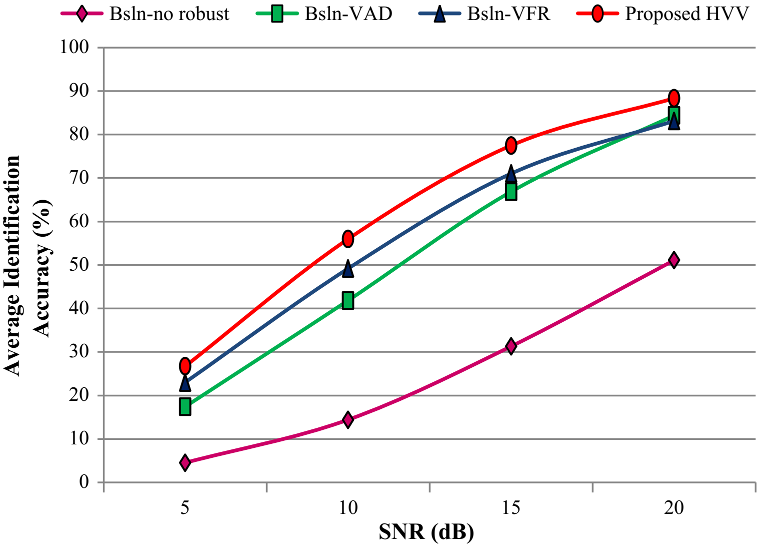


Fig. 6 Comparing average identification accuracies of the proposed and the baseline methods at different SNR values

Table 4 Average identification accuracies (%) for the proposed and the baseline methods at different SNR values

SNR (dB)	Bsln-no robustness	Bsln-VAD	Bsln-VFR	Proposed HVV
20	51.11	84.4	83.15	88.35
15	31.3	66.87	71.03	77.49
10	14.43	41.86	49.2	55.99
5	4.53	17.42	23.02	26.72

5 Conclusions

This paper presents a frame selection method using hybrid technique which combines Voice Activity Detection (VAD) and Variable Frame Rate (VFR) analysis for robust speaker identification. The method efficiently captures the speaker specific information from the time-domain speech signal, under various noise scenarios compared to the conventional fixed frame rate analysis method. It also provides the flexibility to adjust the average frame rate of the method, so that an optimum performance according to the application may be attained.

Experimental results on noisy YOHO database has shown that, the proposed feature frame selection method using hybrid technique outperforms the “a posteriori” SNR weighted energy distance based VFR analysis method and the widely used Gaussian statistical model based VAD method for all eight noise scenarios. Future studies will involve the use of speech enhancement techniques together with the hybrid technique to further improve the identification accuracy of the system.

Acknowledgements This study was supported by the Erasmus Mundus Mobility for Life Project funded by the European Commission. It is carried out at the Center for TeleInFrastruktur, Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark.

References

1. Atal, B. S. (1976). Automatic recognition of speakers from their voices. *Proceedings of the IEEE*, *64*, 460–475.
2. Doddington, G. R. (1985). Speaker recognition—identifying people by their voices. *Proceedings of the IEEE*, *73*, 1651–1664.
3. Campbel, J. P, Jr. (1997). Speaker recognition: A tutorial. *Proceedings of the IEEE*, *85*(9), 1437–1462.
4. Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communications*, *52*(1), 12–40.
5. Mammone, R. J., Zhang, X., & Ramachandran, R. P. (1996). Robust speaker recognition–a feature based approach. *IEEE Signal Processing Magazine*, *13*, 5871.
6. Togneri, R., & Pullela, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *IEEE Circuits Systems Magazine*, *11*(2), 23–61.
7. Zhao, X., Wang, Y., & Wang, D. L. (2014). Robust speaker identification in noisy and reverberant conditions. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22*(4), 836–845.
8. Kinnunen, T., Saeidi, R., Sedlak, F., Lee, K. A., Sandberg, J., Hansson-Sandsten, M., et al. (2012). Low-variance multitaper MFCC features: A case study in robust speaker verification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *20*(7), 1990–2001.
9. Alam, M. J., Kinnunen, T., Kenny, P., Ouellet, P., & O’Shaughnessy, D. (2013). Multitaper MFCC and PLP features for speaker verification using i-vectors. *Speech Communications*, *55*, 237–251.

10. Sadjadi, S. O., Hasan, T., & Hansen, J. H. L. (2012). Mean hilbert envelope coefficients (MHEC) for robust speaker recognition. In *Proceedings of Interspeech* (pp. 1696–1699).
11. Ephraim, Y., & Van Trees, H. (1995). A signal subspace approach for speech enhancement. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *3*(6), 251–266.
12. Brajevic, Z., & Petosic, A. (2012). Signal denoising using STFT with Bayes prediction and Ephraim–Malah estimation. In *Proceedings of the 54th international symposium ELMAR* (pp. 183–186).
13. Govindan, S. M., Duraisamy, P., & Yuan, X. (2014). Adaptive wavelet shrinkage for noise robust speaker recognition. *Digital Signal Processing*, *33*, 180–190.
14. Kim, K., & Kim, M. Y. (2010). Robust speaker recognition against background noise in an enhanced multicondition domain. *IEEE Transactions on Consumer Electronics*, *56*(3), 1684–1688.
15. Zao, L., & Coelho, R. (2011). Colored noise based multicondition training for robust speaker identification. *IEEE Signal Processing Letters*, *18*(11), 675–678.
16. Venturini, A., Zao, L., & Coelho, R. (2014). On speech features fusion, integration Gaussian modeling and multi-style training for noise robust speaker classification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *22*(12), 1951–1964.
17. Dehak, N., kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *19*(4), 788–798.
18. Mashao, D. J., & Skosan, M. (2006). Combining classifier decisions for robust speaker identification. *Pattern Recognition*, *39*, 147–155.
19. Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture models. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *3*(1), 72–83.
20. Mak, M.-W., & Yu, H.-B. (2014). A study of voice activity detection techniques for NIST speaker recognition evaluations. *Computer Speech and Language*, *28*, 295–313.
21. Deng, S., & Han, J. (2012). Likelihood ratio sign test for voice activity detection. *IET Signal Processing*, *6*(4), 306–312.
22. Jung, C.-S., Kim, M. Y., & Kang, H.-G. (2010). Selecting feature frames for automatic speaker recognition using mutual information. *IEEE/ACM Transactions on Audio, Speech and Language Processing*, *18*(6), 1332–1340.
23. Fujihara, H., Kitahara, T., Goto, M., Komatani, K., Ogata, T. & Okuno, H. G. (2006). Speaker identification under noisy environment by using harmonic structure extraction and reliable frame weighting. In *Proceedings of interspeech* (pp. 1459–1462).
24. Tan, Z.-H., & Lindberg, B. (2010). Low complexity frame rate analysis for speech recognition and voice activity detection. *IEEE Journal of Selected Topics in Signal Processing*, *4*(5), 798–807.
25. Tan, Z.-H., & Kraljevski, I. (2014). Joint variable frame rate and length analysis for speech recognition under adverse conditions. *Computers and Electrical Engineering*, *40*, 2139–2149.
26. Sohn, J., Kim, N. S., & Sung, W. (1999). A statistical model based voice activity detection. *IEEE Signal Processing Letters*, *6*(1), 1–3.
27. Hirsch, H. G. & Pearce, D. (2000). The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proceedings of ISCA ITRW ASR*.
28. Campbel, J. P. Jr. (1995). Testing with YOHO cd-rom verification corpus. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing* (pp. 341–344).
29. M-Guarasa, J., Ordonez, J., Montero, J. M., Ferreiros, J., Cordoba, R., & Haro, L. F. D. (2003). Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition. In *Proceedings of Eurospeech*.
30. Zhu, Q. & Alwan, A. (2000). On the use of variable frame rate analysis in speech recognition. In *Proceedings of IEEE international conference on acoustics, speech, and signal processing*.

Swati Prasad received her M.E degree in Electrical and Electronics Engineering from Birla Institute of Technology, Mesra, Ranchi, India in 2007, and is working as Asst. Professor at the Department of Electronics and Communications since then. She also holds a B-level certificate in the National Mathematics Olympiad Contest (1995) and is a recipient of the Erasmus Mundus scholarships for pursuing her Ph.D. degree at Denmark. Her research interest lies in Speaker Identification, Brain Computer Interface (BCI) and Digital Electronics. She has also reviewed research papers for Computer Speech and Language, and Wireless Personal Communications Journal.

Zheng-Hua Tan received the B.Sc. and M.Sc. degrees in electrical engineering from Hunan University, Changsha, China, in 1990 and 1996, respectively, and the Ph.D. degree in electronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 1999. He has been an Associate Professor with the Department of Electronic Systems, Aalborg University, Aalborg, Denmark, since 2001. He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA; an Associate Professor with the Department of Electronic Engineering, Shanghai Jiao Tong University; and a Post-Doctoral Fellow with the Department of Computer Science, Korea Advanced Institute of Science and Technology, Daejeon, Korea. His current research interests include speech and speaker recognition, noise-robust speech processing, multimedia signal and information processing, human-robot interaction, and machine learning.

Ramjee Prasad is currently Professor of Future Technologies for Business Ecosystem Innovation (FT4BI) in the Department of Business Development and Technology, Aarhus University, Denmark. He is the Founder President of the CTIF Global Capsule (CGC). He is also the Founder Chairman of the Global ICT Standardisation Forum for India, established in 2009. He has published over 1000 technical papers, more than 15 patents, contributed to several books and has authored, coauthored, and edited over 30 books.