Int J Speech Technol (2011) 14:405–417
DOI 10.1007/s10772-011-9120-6

Speaker verification under degraded condition: a perceptual study

Gayadhar Pradhan · S.R. Mahadeva Prasanna

Received: 4 July 2011 / Accepted: 28 September 2011 / Published online: 13 October 2011

Abstract This study analyzes the effect of degradation on human and automatic speaker verification (SV) tasks. The perceptual test is conducted by the subjects having knowledge about speaker verification. An automatic SV system is developed using the Mel-frequency cepstral coefficients (MFCC) and Gaussian mixture model (GMM). The human and automatic speaker verification performances are compared for clean train and different degraded test conditions. Speech signals are reconstructed in clean and degraded conditions by highlighting different speaker specific information and compared through perceptual test. The perceptual cues that the human subjects used as speaker specific information are investigated and their importance in degraded condition is highlighted. The difference in the nature of human and automatic SV tasks is investigated in terms of falsely accepted and falsely rejected speech pairs. Speech signals are reconstructed in clean and degraded conditions by highlighting different speaker specific information and compared through perceptual test. A discussion on human vs automatic speaker verification is carried out and the possibility of performance improvement of automatic speaker verification under degraded condition is suggested.

Keywords Speaker information · Speaker verification · Degraded condition · Human vs automatic

G. Pradhan · S.R.M. Prasanna (✉)
Department of Electronics and Electrical Engineering, Indian Institute of Technology Guwahati, Guwahati 781039, India
e-mail: prasanna@iitg.ernet.in

G. Pradhan
e-mail: gayadhar@iitg.ernet.in

1 Introduction

Speaker recognition task aims at recognizing the speakers from their speech signal (Kinnunen and Li 2010) and can be either speaker identification or speaker verification (SV). In case of identification, since their is no claim, the objective is to identify the most likely speaker present in the test speech signal. The verification task deals with validating the identity claim with the test speech signal. The genesis of the present work is motivated from the remarkable ability of humans in recognizing speakers under degraded condition that we experience in our daily communication. For instance, even if our friend calls over mobile phone from a very degraded environment like airport, we still be able to recognize him or her by listening to the speech signal. Is there any clue that may be obtained from this for increasing the robustness of automatic speaker recognition under degraded condition? For this we may need to conduct human and automatic speaker recognition studies, simultaneously using speech data from the degraded condition. It is relatively easier for humans to perform verification than identification. From the objective of reducing the complexity of human speaker recognition task, the present work employs verification mode.

Even though intuitively their is a belief that humans are good at verifying speakers, there are still many questions. For instance, whether the accuracy of verification remains same in clean and degraded condition? What are the speaker specific cues that are mostly used to verify a person under degraded condition? Whether humans rely on a particular speaker specific cue or use different cues at different times? How does the human verification and automatic speaker verification compare under degraded condition? If we have simultaneously recorded clean and different degraded speech signals, then performing a perceptual study and automatic speaker verification task on this speech data may help in finding solutions to some of these questions. The motivation of this work is to collect such a speaker verification database and perform the perceptual and automatic speaker verification studies. The experimental results from both the tasks are then analyzed to find the solutions to some of the above mentioned questions.

There are some perceptual studies conducted earlier to compare the performance of human and machine for speaker verification task (Alexandera et al. 2004; Kreiman and Papcun 1991; Nielsen and Crystal 1998, 2000; Nielsen and Stern 1986). All these studies are conducted either for clean or channel degraded speech. To the best of our knowledge, there are no studies reported on the same set of speakers for clean and different degraded conditions. Further, for most of the perceptual studies conducted on channel degraded speech, the experimental protocol is set similar to NIST speaker recognition evaluation (NIST 2003). But, this protocol may not suit human recognition task well. This is because, humans have limited memory and may not recall the trained voice properly to compare with the test voice for verification decision, after long duration of listening. At the same time it will be time consuming for them to listen to it again and again for making the decision.

Considering the above mentioned limitations of the existing studies, in the present work, the perceptual test speech files are reconstructed by concatenating two 10 s speech files taken from the same gender. The initial 10 s (we call train speech) is taken from the clean speech and remains same for all the experiments. The second segment (we call test speech) is varied from clean to different degraded conditions. To remove the bias of the subjects towards the name of speech file and experimental condition, the speech files are coded and randomized. The perceptual test is conducted by 16 subjects having knowledge about speaker verification. The subjects were instructed to listen each speech file as many times as they want before making a decision. After completing the experiment, the subjects submitted their decision about the speech files (accept/reject) along with a mention on the perceptual cues used to take the decision. The accuracy of the perceptual test is evaluated based on the mean and majority opinion. In the second level, the speech signal is reconstructed in clean and degraded conditions by highlighting different speaker specific features. A perceptual experiment is then conducted using the reconstructed speech files as test speech to rank them depending on the level of speaker specific information perceived.

An automatic SV system is developed using the Mel-frequency cepstral coefficients (MFCC) and Gaussian mixture model (GMM) (Reynolds et al. 2000). In the first level, the verification accuracy of automatic SV system is evaluated using same speech files as the human listener (10 s training and 10 s testing) and then using a relatively larger training and testing speech (2 min training and 30 s testing). For both human and automatic SV systems, the accuracy is compared across all the experimental conditions to find the effect of degradation on verification decision. For clean and degraded conditions, the automatic SV system is compared with human listener in terms of the verification accuracy, and the falsely accepted and falsely rejected speech files. Finally, the possible causes for performance reduction in automatic speaker verification system in mismatched condition is analyzed and the possibility of performance improvement in such conditions is mentioned.

The rest of the paper is organized as follows: The human speaker verification through perceptual study is described in Sect. 2. Automatic speaker verification system is described in Sect. 3. The experimental results are presented in Sect. 4. Discussion on human vs automatic speaker verification is presented in Sect. 5. Finally, the paper is concluded in Sect. 6.

2 Human speaker verification under degraded condition

This section presents a human speaker verification (HSV) task through listening test. The main motivation of this listening test is to study the effect of degradation on humans and find the speaker specific cues that are mostly used to verify a person under degraded condition. In the second level, the listening test is aimed to investigate the speaker specific information and robustness of different features in degraded conditions.

2.1 Speaker verification database

In this work, we have used a subset of the first two phases of IITG multi-variability (MV) speaker recognition database, developed inhouse for speaker recognition studies under degraded condition (Haris et al. 2011). The first phase (Phase I) is collected in an office environment, in a setup having five different sensors, different Indian languages and two different speaking styles. The five different sensors include headphone microphone mounted close to the speaker, inbuilt tablet PC microphone, two mobile phones and one digital voice recorder. Except for the headphone microphone, all the other four sensors are placed at a distance of about 2 to 3 feet from the speaker. Speech was recorded simultaneously over these sensors. In the second phase (Phase II) of data collection, distance of the headphone microphone, inbuilt tablet PC microphone, and digital voice recorder remained same as in the first phase. Out of the two mobile phones, the speech recorded in one mobile phone was through a collar microphone attached at the waist level. The speech recorded in the second mobile phone was through the communication channel. While the subject talking to the facilitator (present outside the recording room) through the mobile phone, the data was recorded simultaneously in all the sensors. The recording was done in two different environments, namely, office/laboratory and hostel rooms. In both the phases, the data was recorded in two sessions, separated by at least one week and two speaking styles, namely, reading and conversation.

The speech recorded in the headphone microphone (H01) is clean compared to other sensors. Speech recorded over digital voice recorder (D01), due to its high sensitivity and position, is worst affected by the environmental noise like air conditioner, fan sound, room reverberation and other surrounding noises present at the time of recording. The speech recorded through the online mobile phone (M03) is affected by the channel degradation. The speech recorded in the tablet PC microphone (T01) and offline mobile is degraded compared to H01, but relatively less degraded compared to D01. The recording was done in two languages, namely, English and favorite language of the speaker which happens to be one of the Indian languages like Hindi, Telugu, Kannada, Oriya and so on. However, the present work uses only a subset, recorded using English language in reading style from both the phases.

2.2 Perceptual studies

The perceptual studies are performed in two parts. The objective of the first part is to evaluate the performance of human speaker verification under degraded condition and also to know whether vocal tract based information is preferred over excitation based information or otherwise. The objective of the second part is to evaluate the perceptual preference of five different components extracted from the speech signal for the task of speaker verification.

2.2.1 Perceptual study for speaker verification

For this study, by fixing language as English and reading style, we considered 30 speakers (20 male and 10 female) set of IITG MV speaker recognition database. These speakers are common in Phases I and II of the database. For a human, few seconds of speech is sufficient to verify the speaker. Further, the verification task is easier, if the train and test speech are from different genders. The perceptual test speech files are constructed by concatenating two speech segments of duration 10 s taken from the same gender. The first speech segment is treated as the train speech and the second as the test speech. The train and test speech signals are taken from the initial portions of speech recorded in the first and second sessions, respectively. In Phases I and II of the database, same text is used for reading style in both the sessions. Therefore the reconstructed speech does not contain any style and text variation. For each speaker, the constructed speech is designed once as true trial (both the train and test speech are from the same speaker) and once as impostor trial (train and test speech are from different speakers). Therefore, in the present database of 30 speakers, for a particular train and test condition, there are in total 60 concatenated speech files each having duration of 20 s.

The first motivation of the perceptual study is to analyze the human speaker verification accuracy in clean and different degraded conditions. Therefore, for each constructed speech file, the train speech is taken from the sensor H01 recording (clean speech) and the test speech is taken from the sensor H01, T01, D01 and M03, respectively. The only variation from one experimental condition to other is the degradation present in the test speech. Thus including all the four experiments, we have 240 (60 × 4) concatenated speech files.

The second motivation of the perceptual study is to analyze whether perceptually vocal tract based speaker specific information is preferred over that of excitation source or otherwise. This is because, for the past few decades, the automatic speaker verification studies mainly focus on the two speaker specific information present in the speech signal, namely, the vocal tract and the excitation source. To perform this, the speech is synthesized by highlighting only either the vocal tract or excitation source information using the linear prediction (LP) analysis.

The LP analysis is performed on the second session of H01 recording to separate the speaker specific vocal tract information from the excitation source (Murty and Yegnanarayana 2006). The speech signal is processed in blocks of 20 ms with a shift of 10 ms. For each 20 ms analysis block, the sample $s(n)$ is estimated as a linear weighted sum of the past p samples (10 for the present case). The predicted sample $\hat{s}(n)$ is given by

$$\hat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k) \tag{1}$$

where p is the order of prediction, and $\{a_k\}$, $k = 1, 2, \ldots, p$ is the set of linear prediction coefficients (LPCs). The LPCs are obtained by minimizing the mean squared error between the predicted and actual sample values over the analysis frame. The error between the actual $(s(n))$ and predicted $(\hat{s}(n))$ value is given by

$$e(n) = s(n) - \hat{s}(n) = s(n) + \sum_{k=1}^{p} a_k s(n-k) \tag{2}$$

The LPCs model the vocal tract information. Hence, the LP residual signal $e(n)$ mostly contains the excitation source information.

In the first case, to highlight the vocal tract information, the speech is synthesized by exciting the residual phase of the analytic signal of the LP residual to the LPCs. Let $e_a(n)$ be the analytic signal of the residual signal $e(n)$. Then,

$$e_a(n) = e(n) + je_h(n) \tag{3}$$

where $e_h(n)$ is the Hilbert transform of $e(n)$. The Hilbert transform is computed as

$$e_h(n) = IDTFT(E_H(\omega)) \tag{4}$$

where

$$E_H(\omega) = \begin{cases} +jE(\omega), & -\pi \leq \omega < 0 \\ -jE(\omega), & 0 \leq \omega \leq \pi \end{cases} \tag{5}$$

and $E(\omega)$ is the DTFT of $e(n)$. DTFT refers to the discrete time Fourier transform and IDTFT refers to the inverse of DTFT. Let $h_e(n)$ be the Hilbert envelope (HE). It is defined as the magnitude of $e_a(n)$ i.e.,

$$h_e(n) = |e_a(n)| \tag{6}$$

$$h_e(n) = \sqrt{e^2(n) + e_h^2(n)} \tag{7}$$

Let $\phi(n)$ be the phase of $e_a(n)$. It is defined as

$$\cos(\phi(n)) = \frac{\Re(e_a(n))}{|e_a(n)|} = \frac{e(n)}{h_e(n)} \tag{8}$$

In the second case, the average of LPCs is computed across all the frames to remove the time varying nature of the vocal tract. To highlight the excitation source, the average LPCs are excited by the LP residual signal. The perceptual test files are then designed using speech recorded in the first session as train speech and synthesized speech as test speech (only for H01 train and H01 test). Thus the total number of constructed speech files for all the six experiments is 360 (60 × 6).

There are sixteen subjects participated in the present perceptual studies. All are BTech and MTech students taking speech technology course at the Indian institute of technology (IIT) Guwahati. Hence, all these subjects have knowledge about speech processing. They were also briefly explained about the goal of speaker verification task and also motivation of the perceptual study. There are in total 360 speech files for six experimental conditions. The subjects are instructed to listen to each speech file carefully as many time as they want before making the decision. The perceptual study is conducted in a common laboratory. The subjects were allowed to take break during the perceptual experiment, if they feel so. After completing the experiment, subjects submitted their decision about the speech files (accept/reject) along with the perceptual cues used to take the decision.

2.2.2 *Perceptual study for ranking speaker-specific features*

The second set of perceptual experiments are conducted on the speech signals constructed by highlighting different information components. The objective of this study is to rank the level of speaker specific information in each of the highlighted components and their robustness to different degraded conditions. The LPC, LP residual signal, residual phase, zero frequency filtered signal (ZFFS), and epoch strength derived from ZFFS are used for finding the level of speaker information present in them.

The LPCs model the vocal tract information. The other four features mostly contain the excitation source information. The literature shows that the cepstral features derived from the speech signal are severely affected by degradation (Pelecanos and Sridharan 2001). This is mainly due to the modification of speech spectrum in the presence of degradation. The smoothed speech spectrum generally represents the speaker specific vocal tract characteristics. The literature also shows that the features derived from the excitation source information is relatively less affected by degradation like noise (Wang et al. 2011). Therefore, for the present perceptual study, more emphasis is given to the features representing the excitation source information.

In zero frequency filtering (ZFF) method, speech is passed through a resonator located at the zero frequency that preserves the signal energy around the impulse present at zero frequency and removes all other information, mainly due to the vocal tract resonances. The trend in the output of the zero frequency resonator is removed further by considering a window of length one to two pitch periods and the trend removed signal is termed as the ZFFS (Murty and Yegnanarayana 2008). The positive zero crossings of the ZFFS give the location of epochs.

The algorithmic steps to estimate the epochs in speech by ZFF method are as follows (Murty and Yegnanarayana 2008):

- Difference input speech signal $s(n)$

$$x(n) = s(n) - s(n-1) \tag{9}$$

- Compute the output of cascade of two ideal digital resonators at 0 Hz

$$y(n) = -\sum_{k=1}^{4} a_k y(n-k) + x(n) \tag{10}$$

where $a_1 = 4$, $a_2 = -6$, $a_3 = 4$, $a_4 = -1$.

- Remove the trend i.e.,

$$\hat{y}(n) = y(n) - \bar{y}(n) \tag{11}$$

where $\bar{y}(n) = \frac{1}{(2N+1)} \sum_{n=-N}^{N} y(n)$ and $2N + 1$ corresponds to the average pitch period computed over a longer segment of speech.

- The trend removed signal $\hat{y}(n)$ is termed as ZFFS.

The zero crossings give the location of epochs. The slope around the zero crossings contains the strength information (Murty et al. 2009). Thus in the present work, first order difference of ZFFS is computed to find the strength of excitation.

To highlight the vocal tract information, the speech files are reconstructed by exciting residual phase to the LPCs. Alternatively, in the other four cases, the average LPCs is excited by the excitation source features. For all the experimental conditions, the speech files are reconstructed using the second session data. For a particular experimental condition, for each speaker a speech folder is generated which contains the training speech and five reconstructed speech files. The speech recorded in the first session over sensor H01 is used as training speech. To remove the bias of subjects towards speaker and speech feature, in each case the speech folders and reconstructed speech files are coded and randomized.

The perceptual experiments are conducted for clean and different degraded conditions. The subjects were briefly explained about the goal and motivation of the present listening test. The subjects were instructed to rank the coded speech files depending on the level of speaker specific information. The highest speaker information bearing speech file gets rank 1 and the lowest gets rank 5. If a subject feels the level of the speaker specific information is same in more than one speech file, then all these files are assigned same rank. Finally, the ranks are normalized across all the subjects and testing speakers as follows: (1) For a particular speaker folder, each reconstructed speech file is ranked depending on the majority opinion of the subjects. (2) Each speaker specific feature is then ranked depending on its rank for majority number of testing speakers.

3 Automatic speaker verification under degraded condition

This section presents an automatic speaker verification (ASV) system. The first objective is to evaluate the performance of ASV system for limited data under different degraded conditions and for a relatively larger training and testing data. This part of the experiment helps to understand the effect of data duration on the speaker verification performance under different degraded conditions. The second objective of this experiment is to systematically investigate the ASV system by comparing the performance across different experimental conditions and with the HSV to address some issues for robustness of the system.

3.1 Speaker verification database

The performance of statistical modeling based speaker verification system like GMM-UBM depends largely on the duration of training and test speech. In the first level, the experiment is conducted to study the automatic speaker verification performance for limited training and testing speech under different degraded conditions. The speech data and experimental conditions remained same as explained in Sect. 2.2. In the second level, keeping the experimental conditions same, the initial 2 min of speech recorded in the first session is used for training and the initial 30 s of speech recorded in the second session is used for testing.

3.2 Feature extraction

During the training and testing process, the speech signal is processed in frames of 20 ms at 10 ms frame rate. For each 20 ms Hamming windowed frame, MFCCs are computed using 22 logarithmically spaced filters (Davis and Mermelstein 1980). The first 13 coefficients excluding zeroth coefficient value are used as a feature vector. Delta (Δ) and delta-delta ($\Delta\Delta$) of MFCC are also computed using two preceding and two succeeding feature vectors from the current feature vector. Thus the feature vector will be of 39 dimension with 13 MFCC, 13 ΔMFCC and 13 $\Delta\Delta$MFCC. The feature vectors corresponding to the speech regions are identified using an energy based voice activity detector based on 0.06 times the average energy as the threshold. The energy threshold is based on several speaker verification experiments on NIST-2003 speaker recognition database with different thresholds and using the one that gives best performance (Prasanna and Pradhan 2011 in press).

3.3 Parameter normalization

The blind deconvolution like cepstral mean subtraction (CMS) reduces the performance when there is not much variability in the recording sensor and environment, and it improves the performance when there is variation (Reynolds 1995). In the present experimental setup, except one sensor match experiment (H01 train, H01 test), there is variation either in sensor, channel or degradation effect. Further, the models are built by adapting a sensor mix universal background model (UBM). In the present experimental setup, the feature vectors are therefore normalized to fit a zero mean and unit variance distribution.

3.4 Speaker modeling and testing

The main motivation of this work is to analyze the effect of degradation on automatic speaker verification system by comparing the results to the perceptual test. For automatic speaker verification, GMM-UBM based speaker modeling is employed (Reynolds et al. 2000). The UBM is a large GMM which represents the speaker independent distribution of features. The UBM is generally built using large population speech. The UBM is the core part of GMM-UBM speaker verification system. The UBM should balance with respect to male and female speakers, and the speech should come from every possible sensor which will be encountered at the time of speaker verification. The UBM is represented by a weighted sum of C component densities as $U = \{\mu_c, \Sigma_c, \eta_c\}, c = 1, 2, \ldots, C$, where μ_c, Σ_c and η_c are the mean vector, covariance matrix and weight associated with mixture c, respectively. The speaker dependent models are built by adapting the components of UBM with the speakers training speech using maximum a posteriori (MAP) algorithm (Reynolds et al. 2000). During the testing stage, the log likelihood scores are calculated between the claimed model and UBM.

In this experimental setup, six hours of UBM speech were selected from 17 male and 17 female speakers those who are not belonging to the present 30 speakers set. This six hours of speech contains three hours of male speech and three hours of female speech. For each speaker, the UBM speech is distributed equally among the four sensors H01, T01, D01 and M03. Using the sensor mixed data, two gender dependent 512 mixture size GMM are built, one for the male and other for the female speech. Finally, a 1024 mixture size gender independent UBM is built by pooling the two models and normalizing the weights (Reynolds et al. 2000).

4 Experimental results

This sections presents the performance of HSV and ASV systems for different experimental conditions and address some of the issues to improve the performance under degraded conditions. This section mainly focuses on different threshold finding methods and their significance under degraded conditions.

4.1 Performance measure

Speaker verification system validates the identity claim of a person (Campbell 1997). A perfect speaker verification system should accept all the true claims and reject all the false claims. In practical applications, some true trials may be rejected and some false trials may be accepted. Hence, the speaker verification performance is measured in terms of false rejection rate (FRR) and false acceptance rate (FAR). In the present study, the speaker verification performance is evaluated in terms of average error rate (AER). The AER is computed by taking the average of FRR and FAR.

4.2 Human speaker verification

It is a well known fact that the speaker discrimination ability varies from person to person. Hence, the performance of human speaker verification system is measured in term of mean and majority opinion of the subjects participated in the listening test. In the mean opinion measure, performance is first evaluated based on each individual opinion and then mean of the performance is calculated with equal weight to all the subjects. In the majority opinion measure, for each speech file the performance is evaluated based on the majority opinion.

4.2.1 Based on mean opinion

The performance of human speaker verification system for different experimental conditions based on mean opinion measure is given in Table 1(a). The table also contains the maximum and minimum AER across all the subjects. By comparing the second and third column of the table, it can be observed that the AER of the best speaker discriminating subject varies significantly compared to the least speaker discriminating subject. This experiment also shows that no subject is able to verify the speaker with 100% accuracy, even for clean and sensor matched speech. This may be due to the gender matching of the training and testing speech. By comparing the results for different experimental conditions, it can be seen that, as the level of degradation increases, the accuracy of human speaker verification decreases for all the subjects. This shows that the human speaker verification accuracy is also affected by degradation like sensor, noise and communication channel.

4.2.2 Based on majority opinion

The performance of perceptual experiments based on majority opinion is given in Table 1(b). The significance of the differences in the pairs of the scores is tested using hypothesis testing (Hogg and Ledolter 1987). The level of confidence (LC) for the observed differences in the sample means is obtained using the sample variances and values of *student-t* distribution. For the present case, the LC is calculated taking the mean of the opinion (8 for the present 16 subject set) as the reference. The LC for each experimental condition is given in Table 1(b). The performance is improved significantly compared to the mean opinion case. The level of confidence is also high (>99.5%) in all cases. This indicates that the number of opinions used to take the decision is significantly more. The table also contains the FAR and FRR for each experimental condition. Except M03 (channel degraded) test speech, the error is mainly attributed due to the false acceptance. For channel degraded test speech, the false rejection is significantly large compared to the false acceptance. This shows that the human listeners are unable to

Table 1 Performance of human speaker verification based on mean and majority opinion

(a) Mean opinion

Train vs test speech	Error rate (%)		
	Max	Min	Avg
H01 vs H01	16.66	5	9.79
H01 vs T01	26.66	8.33	16.35
H01 vs D01	23.33	5	15.93
H01 vs M03	35	11.66	20.52
H01 vs Vocal tract	30	10	16.45
H01 vs Exitation source	33.33	16.66	24.58

(b) Majority opinion

Train vs test speech	Error rate (%)		
	FAR	FRR	AER (LC)
H01 vs H01	3.33	0	1.66 (>99.5%)
H01 vs T01	6.66	0	3.33 (>99.5%)
H01 vs D01	16.66	3.33	10 (>99.5%)
H01 vs M03	3.33	20.00	11.66 (>99.5%)
H01 vs Vocal tract	6.66	0	3.33 (>99.5%)
H01 vs Exitation source	33.33	0	16.66 (>99.5%)

match properly two voices when one is wideband and the other is narrowband. Tables 1(a) and (b) show that human verification accuracy decreases as the level of degradation increases.

4.3 Ranking for speaker specific information

The rank for speaker specific information for different experimental conditions is summarized in Table 2. The ZFFS and strength of excitation contain better speaker information compared to the residual phase. The speaker information present in the ZFFS is comparable to the LP residual. By comparing the experimental results given in Table 1 and the ranks given in Table 2, it can be observed that the vocal tract contains better speaker information compared to the excitation source. This is a well known fact and reported in many literatures (Murty and Yegnanarayana 2006; Yegnanarayana et al. 2005). But through this perceptual experiment we have tried to investigate the ground truth about the robustness of vocal tract and excitation source information to different degradations. The study shows that the excitation source information is robust to the degradations, but the speaker information is still less compared to vocal tract. The important thing observed from this study is that if the excitation strength is removed from the speech signal, human verification accuracy decreases even for clean speech. Hence, the excitation source information may contain robust and different speaker specific information compared to the vocal tract.

Table 2 Rank of the speaker specific features depending on the level of speaker information. The abbreviations VT, Res, ResPh, ZFFS, and EpoStr refer to vocal tract information, LP residual, LP residual phase, zero frequency filtered signal and epoch strength

Train vs test speech	Rank (best = 1)				
	VT	Res	ResPh	ZFFS	EpoStr
H01 vs H01	1	2	5	2	3
H01 vs T01	1	2	5	2	3
H01 vs D01	1	2	5	2	2
H01 vs M03	1	2	5	3	3

4.4 Automatic speaker verification

Unlike the HSV, to accept or reject a claim in ASV, score should be matched to the claimed model above certain threshold. In this work the effect of degradation on the test speech varies from one experimental condition to other. The verification score and the threshold will also vary from one experimental condition to other. Therefore, to make a hard decision (accept/reject), a set of five cohorts is used. The cohort speakers are selected using clean train and clean test (H01 train and H01 test) condition. In the present experimental set up the train speech remained same for all the experiments. For a particular degraded condition, it is assumed that all the test speech files are affected by similar degradation. Since, all the test files transformed from clean to a particular degradation, it is expected that the cohort speaker set will remain same (Wu et al. 2007). Hence, the same cohort set is used for all the experiments. Two measures are used to validate the identity claim, namely, winning over cohort set and winning to cohort mean score with threshold.

4.4.1 Winning over cohort set

In this measure, for each test speech, the score is computed against the claimed model and its five nearest cohorts set. If the score obtained by the claimed model is best among the six scores (one claimed model score and five cohort scores), then the claim is accepted else rejected. The LC is obtained using the model scores and the nearest cohort scores. The performance of ASV system based on this measure for different experimental conditions is given in Table 3(a). The table also contains the FAR and FRR for each experimental condition. By comparing the results for clean and different degraded conditions, it can be seen that as the level of degradation increases, the FAR and FRR increases. But the increase in FRR is significantly more compared to the FAR. This may be due to the close competition among the nearest cohorts. For any modification of speaker information in degraded condition, the test speech deviated from the claimed model and matches to the cohort model.

Table 3 Performance of automatic speaker verification system

(a) Winning over cohort set

Train vs test speech	Error rate (%)					
	10 s train vs 10 s test			2 min train vs 30 s test		
	FAR	FRR	AER (LC)	FAR	FRR	AER (LC)
H01 vs H01	6.66	0	3.33 (>99.5%)	0	0	0 (>99.5%)
H01 vs T01	6.66	3.33	5 (>99.5%)	3.33	3.33	3.33 (>99.5%)
H01 vs D01	26.66	36.66	31.66 (>97.5%)	10	36.66	23.33 (>97.5%)
H01 vs M03	16.66	40	28.33 (>97.5%)	6.66	26.66	16.66 (>99.5%)
H01 vs Vocal tract	10	13.33	11.66 (>99.5%)	10	10	10 (>99.5%)
H01 vs Exitation source	10	63.33	36.66 (<90%)	6.66	63.33	35 (<90%)

(b) Winning over cohort mean score with threshold

Train vs test speech	Error rate (%)					
	10 s train vs 10 s test			2 min train vs 30 s test		
	FAR	FRR	AER (LC)	FAR	FRR	AER (LC)
H01 vs H01	6.66	0	3.33 (>99.5%)	0	0	0 (>99.5%)
H01 vs T01	13.33	0	6.66 (>95%)	3.33	0	1.66 (>97.5%)
H01 vs D01	30	30	30 (<90%)	10	13.33	11.66 (<90%)
H01 vs M03	23.33	33.33	28.33 (>95%)	13.33	20	16.66 (>99.5%)
H01 vs Vocal tract	10	3.33	6.66 (>90%)	13.33	0	6.66 (>95%)
H01 vs Exitation source	16.66	56.66	36.66 (<90%)	16.66	60	38.33 (<90%)

4.4.2 Winning over cohort mean score with threshold

The speech files in the IITG MV database are degraded by the real environment degradations. For a particular degraded condition, the degradation effect varies within the speech file and from one speech file to other. In the experimental process, it is observed that the scores obtained by the cohort set significantly varying among themselves. This may be due to different bias of test speech towards the cohort models. Therefore, to nullify the cohort score variation on the verification decision a method is proposed. In this method, mean and standard deviation of cohort scores are used to impose a threshold dynamically. If the claimed model score exceeds the cohort mean score by the standard deviation, the claim is accepted else rejected. The merit of this measure lies in the fact that depending on the quality of test speech signal, the threshold value is changed automatically. Hence, if one cohort gains significantly over the other cohort scores due to some bias, it can be suppressed to certain extent. The LC is obtained using the model scores and the threshold scores (mean of the cohort scores + threshold). The LC is high for all the sensor recordings, except the sensor D01. As explained earlier the speech recorded in sensor D01 is worst affected by the degradation. Due to sever degradation effect, the verification score seems to be more random in nature. The verification performance based on this measure is given in the Table 3(b). By comparing FAR and FRR given in the Tables 3(a) and 3(b), it can be observed that by using this measure, the FRR is reduced significantly. Although, in some cases the FAR is increased, the rate of increase in FAR is very less compared to the rate of decrease in FRR. Hence, the AER is reduced significantly.

5 Human vs automatic speaker verification: a discussion

The performance of HSV, especially under degraded condition, is significantly better compared to ASV under limited data condition (10 s training data and 10 s testing data). This indicates that humans are good at verifying speakers even with limited data. HSV performance also decreases with increase in degradation. However, the performance degradation is significantly less compared to ASV. This demonstrates the relative robustness of HSV for degradation. The performance of ASV increases with the increase in the amount of training and testing data. However, the improved performance is still poor compared to HSV under degraded condition. This reinforces the intuitive feeling that humans are good at verifying speakers compared to machine. FAR seem to be relatively high compared to FRR in HSV. This may be due to the confusability created due to degradation.

Table 4 Number of common speech pairs falsely accepted and falsely rejected by human subjects and automatic speaker verification system

Train vs test speech	No of common speech pairs	
	False accepted	False rejected
H01 vs H01	0	0
H01 vs T01	0	0
H01 vs D01	1	0
H01 vs M03	0	1
H01 vs Vocal tract	0	0
H01 vs Exitation source	3	0

Alternatively, both FAR and FRR are high in ASV task. This infers that in case of ASV, the effect of degradation seem to be random.

To find similarity between HSV and ASV in the error domain, the speech pairs falsely accepted and falsely rejected in both cases are compared. The number of such pairs common in both are given in Table 4. The speech pairs falsely accepted and falsely rejected are significantly different for HSV and ASV systems. This indicates that the speaker information and approach employed by humans and machines seem to be different. Based on the report submitted by the subjects, it seems that the human subjects rely more on higher level features like accent, pronunciation of specific words, pitch contour, speaking rate, stress at particular word. Especially under degraded condition, most of the subjects rely on the pronunciation of specific words and speaking rate. The human subjects for most of the cases used different cues at a time to verify the claim. For difficult speech files where the amount of degradation is too much or the two speakers seem to have same speaker quality, they compared similarity as well as difference between the two voices to make the verification decision. Also, the speaker information exploited seem to be unique for each person. For instance, the way of pronouncing a particular word was observed to be unique and wherever that person speech signal comes, the subjects used this cue for verifying the same. For ASV, we have tried to find the similarity and difference by matching the test speech with the claimed model and a set of cohort speakers, and also a threshold depending on the cohort scores. However, ASV performance is poor compared to HSV. In case of ASV we use segmental MFCC features. The main challenge is therefore automatic extraction of higher level features from a conversational speech, especially under degraded condition and using them for speaker modeling and hence increasing robustness of ASV under degraded condition.

After analyzing the experimental results as above, now we will go back to looking into the speech signals in the time and frequency domains to get feel about the nature of degradation and suggesting future direction for addressing robustness in case of ASV. The signals given in Fig. 1 are segments of speech taken from the IITG MV speaker recognition database. Figures 1(a)–(d) show the speech recorded over H01, T01, D01 and M03, respectively. Figures 1(e) and (f) show the speech signals reconstructed by highlighting the vocal tract and excitation source information. The speech signals recorded in different sensors is affected by different levels of degradation. In such a condition, most of the existing speaker verification systems focus to reduce the mismatch between the training model and testing features either by removing the degradation effect from both training and testing speech signals (Auckenthaler et al. 2000; Boll 1979; Pelecanos and Sridharan 2001; Reynolds et al. 2000) or by biasing the parameters of the speaker model towards the testing environment (Ming et al. 2007; Teunen et al. 2000). These methods rely more on the estimation of degradation or use of *a priori* knowledge of degradation for normalization of degradation effect. For the present study we have also followed some of these approaches, like CMS and cepstral variance normalization (CVN) for normalization in the feature domain, building sensor mixed model instead of using only sensor H01 speech and in the score level using five nearest cohorts. Although, these methods provide some advantage under degraded condition, they may not solve the issue completely and hence poor performance of ASV.

The effect of mismatch between the training and testing speech on the verification scores is analyzed in Fig. 2. Figure 2(a) shows the verification scores for clean train and test conditions (H01 vs H01). For the true trials (test number 1–30), the scores are significantly above the threshold, and for the false trials (test number 31–60), the scores are significantly below the threshold. A system threshold around 0.3 can easily separate the true claims from the false claims. But if this threshold is set for other mismatched experiments, all the true trials get rejected. As the level of degradation is increased, the difference between the claimed model score and the cohort mean score reduces. The training model, UBM and cohort models are same for all the experimental conditions. The only variation present from H01 test to other experimental conditions is the degradation present in the test speech. The deviation of claimed model score towards the cohort model may be due to the failure of speech frame selection algorithm resulting in selecting significantly degraded frames and hence overriding of degradation on the speaker information in the extracted speaker specific feature vectors.

In the present work for the detection of speech frames, an energy based method is used and the threshold varies in an adapting manner depending on the energy of the speech signal. This may fail to select the speech regions for highly corrupted speech file like D01 recording (Fig. 1(c)). But, this speech detection method is sufficient for H01, T01 and M03 recordings. If we compare the speaker verification performance from Table 3(b), the performance for H01, T01

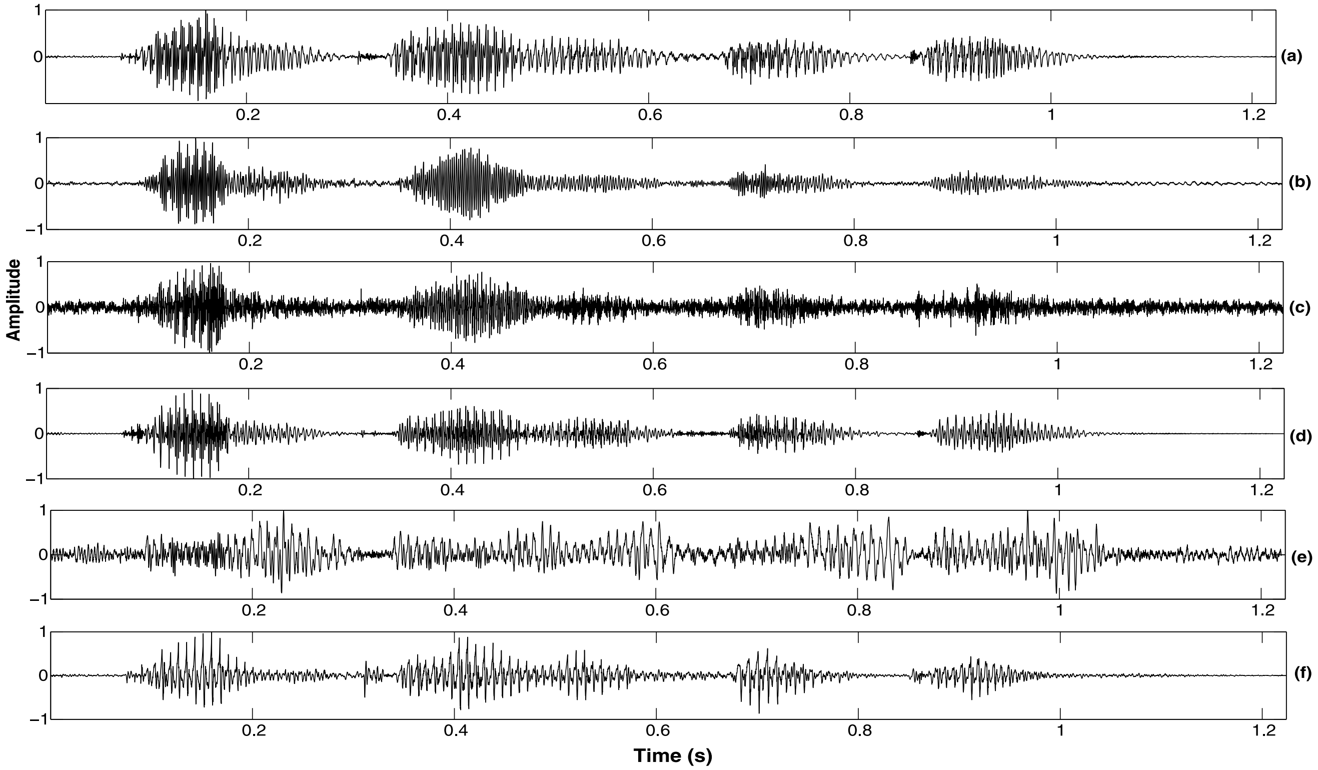

Fig. 1 Effect of degradation on the speech signal recored in IITG MV speaker recognition database, (**a**)–(**d**) speech recorded in sensor H01, T01, D01 and M03, respectively. (**e**) and (**f**) speech reconstructed by highlighting the vocal tract and excitation source information, respectively

and M03 test speech is significantly different. This shows that even if the speech selection is perfect, all speech regions may not be speaker specific in degraded conditions. From Fig. 1, it can be seen that as the level of degradation increases, the low signal to noise ratio (SNR) portion of the speech is progressively merged into the nonspeech regions (degradation). As a human being we have the ability to separate the speech regions from the background degradation and focus more on the speaker specific regions. Figure 3 shows the Hamming windowed 30 ms speech for the vowel /a/ and the corresponding magnitude and log magnitude spectra for clean and different degraded conditions. For same speaker and speech files, 30 ms speech for the low SNR voiced consonant /d/ and corresponding magnitude and log magnitude spectra are shown in Fig. 4. These two figures show that the spectral mismatch between the clean and degraded speech for the vowel is significantly less compared to the low SNR consonant. Hence, under degraded condition mismatch between the training model and the test feature vectors may be reduced by using the high SNR speech regions that may be vowel, semivowel and diphthong sound units. As discussed earlier robustness can be further provided by combing excitation source and suprasegmental features with vocal tract features derived from these regions. The merit of this approach is we rely more on the less degradation affected speaker specific features rather on the degradation effect and this approach does not require *a priori* knowledge of degradation.

6 Summary and conclusions

In this work, perceptual studies are conducted for understanding the robustness offered by HSV for degraded condition. The performance of human subjects is compared with that of the ASV developed using MFCCs and GMM-UBM. For limited data and degraded conditions, HSV provides better performance in higher level degradation. The human subjects seem to rely more on higher level features like accent, pronunciation of specific words, pitch contour, speaking rate and stress of particular word. The human subjects used different cues to verify the claim. The perceptual study conducted on the reconstructed speech files to compare the level of speaker specific information and effect of degradation on vocal tract and excitation source features reveals that both the features are affected by degradation. Perceptually, vocal tract information is preferred by the subjects.

This study using degraded speech indicates that the amount of data may not be very crucial for speaker verification under degraded condition. Approaches need to be developed to identify high SNR regions and higher level features, and use them for ASV task. This may result in increasing the

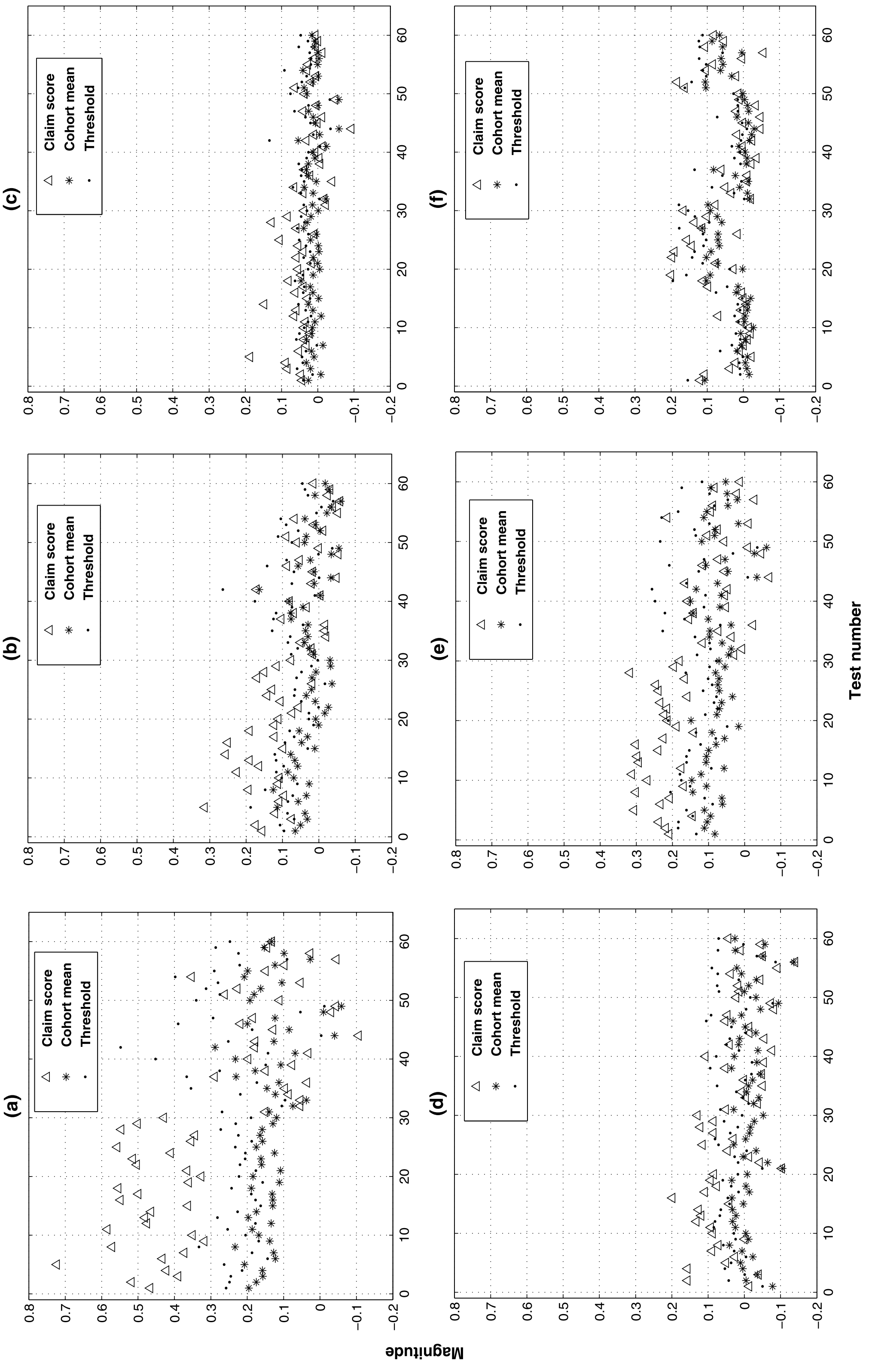

Fig. 2 Speaker verification scores of the automatic speaker verification system for clean, degraded and reconstructed speech files

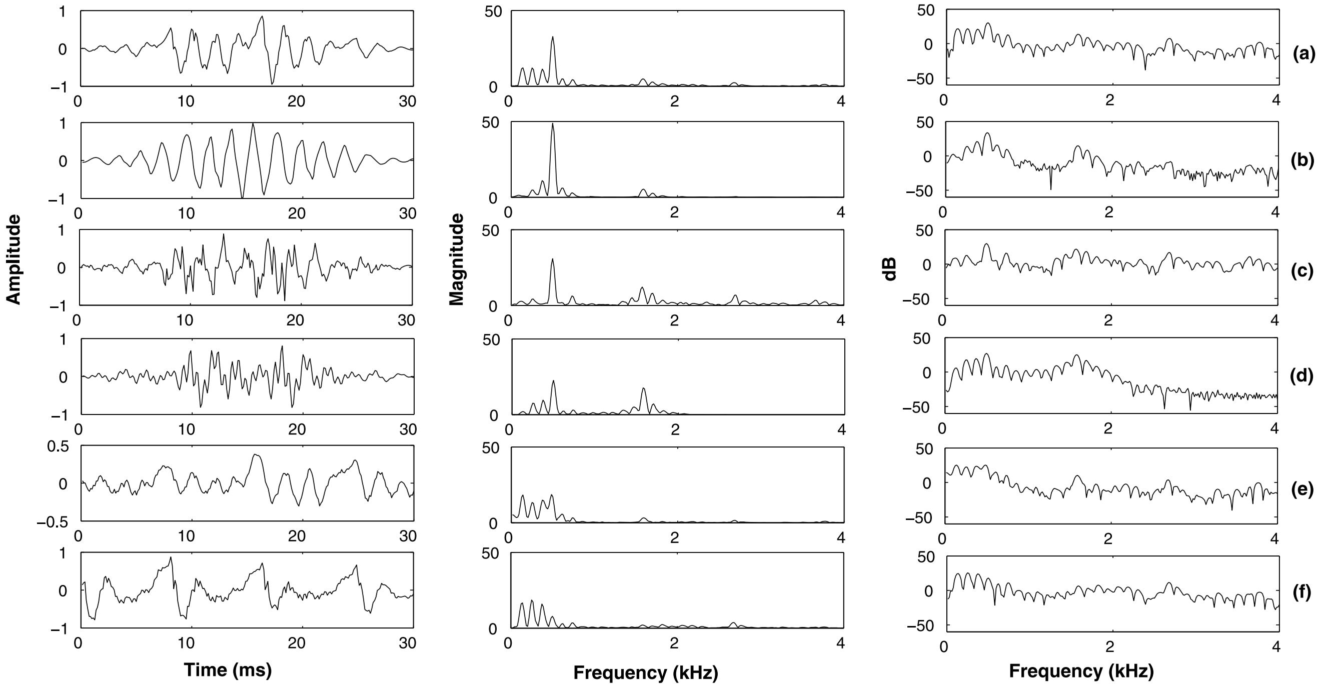

Fig. 3 Magnitude and log magnitude spectra for clean and degraded speech for the vowel /a/

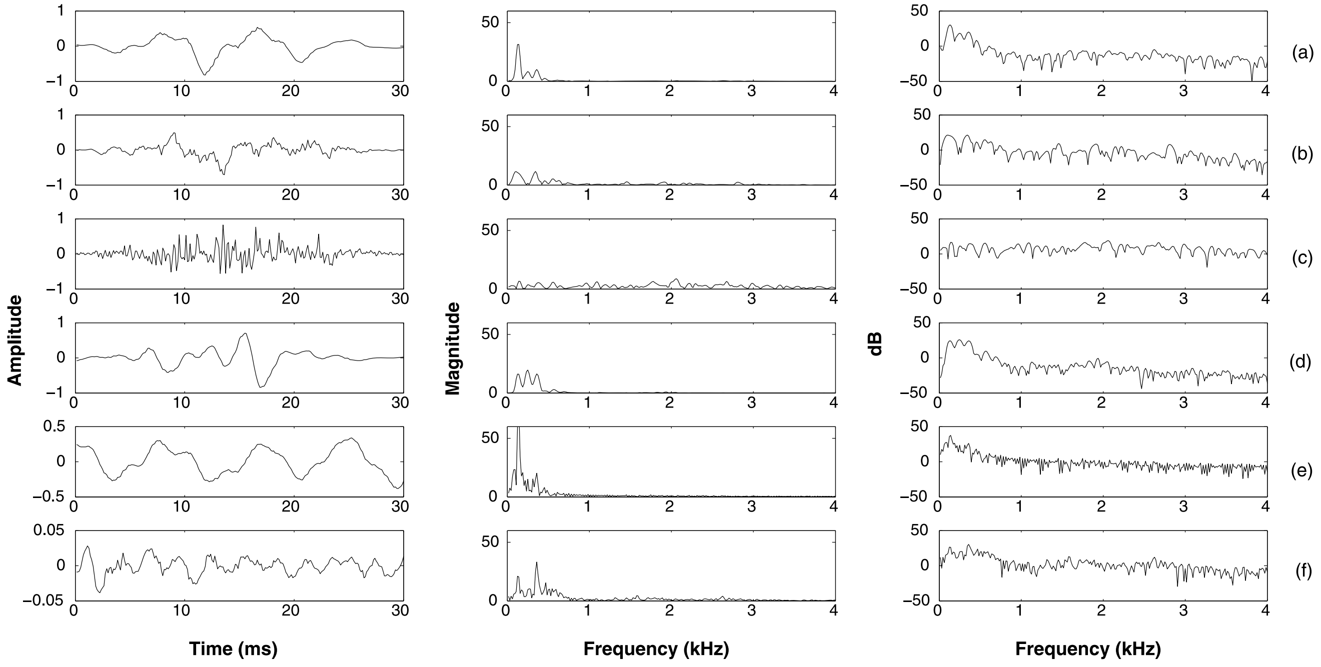

Fig. 4 Magnitude and log magnitude spectra for clean and degraded speech for the voiced consonant /d/

robustness of ASV system. The future work may focus on developing methods for the selection of high SNR regions like the vowel, semivowel and diphthong sound units during training and testing and also combing excitation source and suprasegmental features with vocal tract features derived from these regions.

References

Alexandera, A., Bottib, F., Dessimozb, D., & Drygajlo, A. (2004). The effect of mismatched recording conditions on human and automatic speaker recognition in forensic applications. In *Forensic Science International* (pp. 95–99).

Auckenthaler, R., Carey, M., & Thomas, H. L. (2000). Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, *10*(1), 42–54.

Boll, S. F. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-27*, 113–120.

Campbell, J. P. (1997). Speaker recognition: a tutorial. *Proceedings of the IEEE*, *85*(9), 1437–1462.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing ASSP-28*(4), 357–366.

Haris, B. C., Pradhan, G., Misra, A., Shukla, S., Sinha, R., & Prasanna, S. R. M. (2011). Multi-variability speech database for robust speaker recognition. In *National conf. on communication (NCC)*, Bangalore, India (pp. 1–5).

Hogg, R. V., & Ledolter, J. (1987). *Engineering statistics*. New York: Macmillan.

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, *52*, 12–40.

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication*, *10*, 265–275.

Ming, J., Hazen, T. J., Glass, J. R., & Reynolds, D. A. (2007). Robust speaker recognition in noisy conditions. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(5), 1711–1723.

Murty, K. S. R., & Yegnanarayana, B. (2006). Combining evidence from residual phase and mfcc features for speaker recognition. *IEEE Signal Processing Letters 13*(1), 52–55.

Murty, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, and Language Processing*, *16*, 1602–1613.

Murty, K. S. R., Yegnanarayana, B., & Joseph, M. A. (2009). Characterization of glottal activity from speech signals. *IEEE Signal Processing Letters*, *16*(6), 469–472.

Nielsen, A. S., & Crystal, T. H. (1998). Human vs. machine speaker identification with telephone speech. In *Inter. conf. on spoken language processing*, Sydney, Australia (pp. 221–224).

Nielsen, A. S., & Crystal, T. H. (2000). Speaker verification by human listeners: Experiments comparing human and machine performance using the NIST 1998 speaker evaluation data. Digital Signal Processing, 249–266.

Nielsen, A. S., & Stern, K. R. (1986). Recognition of previously unfamiliar speakers as a function of narrowband processing and speaker selection. *The Journal of the Acoustical Society of America*, *79*, 1174–1177.

NIST (2003). NIST-speaker recognition evaluations. In [Online], Available: http://www.nist.gov/speech/tests/spk.

Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification. In *Speaker Odessy: the speaker recognition workshop* (pp. 213–218).

Prasanna, S. R. M., & Pradhan, G. (2011 in press). Significance of vowel-like regions for speaker verification under degraded condition. IEEE Transactions on Audio, Speech, and Language Processing.

Reynolds, D. A. (1995). Speaker identification and verification using Gaussian mixture speaker models. *Speech Communication*, *17*, 91–108.

Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing*, *10*, 19–41.

Teunen, R., Shahshahani, B., & Heck, L. P. (2000). A model-based transformation approach to robust speaker recognition. In *Proc. int. conf. on spoken language processing*. Beijing, China (Vol. 2, pp. 495–498).

Wang, N., Ching, P. C., Zheng, N., & Lee, T. (2011). Robust speaker recognition using denoised vocal source and vocal tract feature. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(1), 196–205.

Wu, W., Zheng, T. F., Xu, M., & Soong, F. K. (2007). A cohort-based speaker model synthesis for mismatched channels in speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *15*(6), 1893–1903.

Yegnanarayana, B., Prasanna, S. R. M., Zachariah, J. M., & Gupta, S. (2005). Combining evidence from source suprasegmental and spectral features for a fixed-text speaker verification system. *IEEE Transactions on Speech and Audio Processing*, *13*(4), 575–582.