Int J Speech Technol (2017) 20:839–850
DOI 10.1007/s10772-017-9451-z

Processing degraded speech for text dependent speaker verification

Banriskhem K. Khonglah[1] · Ramesh K. Bhukya[1] · S. R. Mahadeva Prasanna[1]

Received: 18 November 2016 / Accepted: 7 August 2017 / Published online: 24 August 2017

Abstract This work explores the use of speech enhancement for enhancing degraded speech which may be useful for text dependent speaker verification system. The degradation may be due to noise or background speech. The text dependent speaker verification is based on the dynamic time warping (DTW) method. Hence there is a necessity of the end point detection. The end point detection can be performed easily if the speech is clean. However the presence of degradation tends to give errors in the estimation of the end points and this error propagates into the overall accuracy of the speaker verification system. Temporal and spectral enhancement is performed on the degraded speech so that ideally the nature of the enhanced speech will be similar to the clean speech. Results show that the temporal and spectral processing methods do contribute to the task by eliminating the degradation and improved accuracy is obtained for the text dependent speaker verification system using DTW.

Keywords End-point detection · Temporal enhancement · Spectral enhancement · Text dependent

✉ Ramesh K. Bhukya
r.bhukya@iitg.ernet.in

Banriskhem K. Khonglah
banriskhem@iitg.ernet.in

S. R. Mahadeva Prasanna
prasanna@iitg.ernet.in

[1] Electro Medical and Speech Technology Laboratory, Department of Electronics and Communication Engineering, Indian Institute of Technology Guwahati, Guwahati, Assam 781039, India

1 Introduction

Speaker recognition is the task of recognizing speakers using their speech signal Furui (1981). Depending on the mode of operation, speaker recognition can be either identification or verification. In case of identification, the most likely speaker of the test speech signal is identified by comparing among the enrolled speakers. Speaker verification involves identity claim and test speech signal. The objective of speaker verification is therefore to validate the identity claim. Accordingly, speaker verification is one to one matching whereas speaker identification is one to many matching task. Depending on the constraint on the lexicon used during enrollment and testing, speaker recognition is also classified into text-independent and text-dependent modes Marinov (2003). In case of text-independent mode, there is no restriction on the lexicon used for enrollment and testing. Alternatively in case of text-dependent mode, the lexicon of the test speech signal is a subset of the lexicon used during enrollment Hébert (2008).

From the practical usability point of view, the user is free to provide test speech with no constraints on duration, quality, recording condition, channel and lexical content. Accordingly, the performance of the speaker verification system is influenced by many of these possible variabilities. Among these variabilities, lexical content and channel variations are the most detrimental. Compared to channel variability which is due to uncontrolled environmental factors, lexical variability can be manageable. The performance of text-dependent speaker verification (TDSV) system is expected to be better compared to that of the text-independent case. Also matching of short duration phrases during training and testing with high accuracy makes it an attractive option for commercial speech based person authentication system Subhadeep Dey et al. (2014).

Speech is commonly used in biometric security technologies because of the user acceptance, ease of use, low cost, high accuracy and ease of implementation. In the present day situation the most common natural biometric information source is telephone based remote access control applications Chakrabarty et al. (2013); Piyare and Tazil (2011); Onukwugha and Asagba (2013); Das et al. (2009); Shahriyar et al. (2008). In real time applications, the performance of the system may degrade due to various reasons such as background noise, sensor mismatches, reverberation, channel mismatch, background speech and many more. The speech is also affected by illness, amount of stress, aging and health conditions. Depending on the above mentioned conditions, the accuracy of the TDSV system tends to degrade. In order to overcome all the above mentioned conditions in TDSV system, detection of accurate begin and end points is necessary Prasanna et al. (2003); Yegnanarayana et al. (2005); Rabiner and Juang (1993a). There are algorithms present which are based on the signal energy and have been proposed in Savoji (1989); Tsao and Gray (1984). Recently, a method using vowel onset point (VOP) was proposed to detect begin and end points accurately Prasanna et al. (2003); Yegnanarayana et al. (2005). Energy based algorithms and the VOP based method perform well in clean speech conditions. However as seen in Fig. 1b, the energy based end point detection algorithm fails for the case when there is background noise and background speech. The VOP based method also fails in the presence of background noise and background speech case as shown in Fig. 1c. The background noise is present in the region around 3.2 s and the background speech is present in the region around 3.5 s in Fig. 1.

Even though text-independent speaker verification system is widely explored, the goal of TISV system is different. TISV system involves general modeling of speaker and essentially captures the speaker specific characteristics. However, in TDSV system there is no modeling of speaker characteristics. It basically involves matching of two templates and TDSV system is mostly used for the co-operating scenario. This can find applications in biometric authentication system which is why the TDSV system has been chosen over TISV system. In addition to this, depending on the acoustic background and the presence of different kinds of degradation, poor speaker verification performance may be encountered. A practical case of degraded speech is taken in this work and speech enhancement is applied on the degraded speech so as to improve the speaker verification performance. In this work, the degraded speech condition is assumed to have a foreground/background model Deepak and Prasanna (2016). The reason is that for most cases of the speech taken for TDSV system, the speaker generally speaks close to the microphone. As a result, the speech which is recorded close to the microphone will be termed as foreground speech and the remaining other interfering signals captured by the sensor which are far from the microphone will be termed as background noise. Note that the background noise can be any kind of degradation which even includes background speech. It has been shown that the foreground speech enhancement performs well in Deepak and Prasanna (2016) for any kind of degradation present in the background as long as the interfering noises are far away from the microphone. In this work the foreground speech enhancement will be used for TDSV. The exploration of the enhancement strategy will be of two types. The first type is after doing the energy based end point detection, there may still be degradation present between

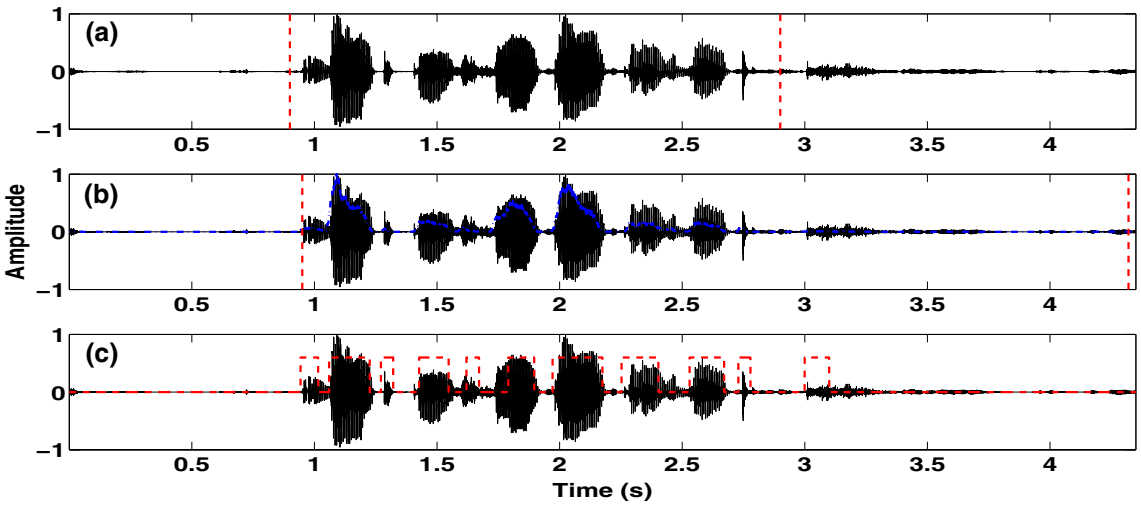

Fig. 1 Illustration of end point detection on degraded speech. **a** Degraded speech signal with the ground truth marked. **b** Energy based method. **c** VOP based method

the end points and hence foreground speech enhancement will be performed so as to obtain a good accuracy for the TDSV system. The second type consists of directly using foreground speech enhancement on the signal and the energy based end point detection algorithm is applied on the enhanced speech which may improve the accuracy of the TDSV system. In this context the performance of the TDSV under degraded conditions are explored using temporal and spectral enhancement methods Krishnamoorthy and Prasanna (2011).

Temporal enhancement involves modifying the degraded Linear Prediction (LP) residual by emphasizing the high signal to noise (SNR) regions resulting in ideally a clean residual. The modification is based on deriving a total weight function which is used to weight the residual. This total weight function is based on the combination of the weight functions obtained from voiced/unvoiced detection and high SNR region detection. This residual is then used to excite a time-varying vocal tract system which are the LP coefficients resulting in a temporally enhanced speech. Temporal enhancement has been shown to perform very well in Krishnamoorthy and Prasanna (2011). This kind of enhancement focuses on the high SNR regions without considering the type of noise present.

Another kind of enhancement which has been widely explored is the spectral subtraction Boll (1979). In this method the background noise is assumed to be stationary and uncorrelated. As a result the background noise can be modeled by calculating the average magnitude spectrum which is subtracted from the overall signal in the frame based approach to estimate the desired speech from the signal. The average magnitude spectrum is computed mostly during the speech pauses to obtain better noise modeling.

Detecting the end points after enhancement may be a better option since the goal of the enhancement methods is to obtain an enhanced speech which ideally is free from the effect of noise or in other words it has the characteristics similar to the clean speech. This work combines the enhancement strategies along with the end point detection algorithm so that improvements in the TDSV system can be obtained.

Even though the tools and techniques for TDSV system as well as enhancement methods proposed in this work are from existing work, however, combining the enhancement tools as well as the TDSV system gives a solution for the TDSV system which is robust for practical variabilities. These variabilities also involve degradation. Hence, combining speech enhancement with speaker verification helps to overcome some of the issues encountered for practical speaker verification like degradation. The rest of the work is organized as follows. Section 2 describes the enhancement strategies. The TDSV system along with the end point detection is described in Sect. 3. The experimental evaluation is given in Sect. 4. Finally the Conclusion is given in Sect. 5.

2 Speech enhancement

The degraded speech is passed through a certain level of enhancement. The enhancement strategies followed in this work involve the one which does not assume the noise characteristics which is the temporal enhancement and the other one which assumes the noise characteristics are stationary which is the spectral enhancement. Accordingly the details of the two types of enhancement are given below.

2.1 Temporal enhancement

Temporal enhancement can be divided into two steps. One is the gross level processing which is similar to the VAD techniques. The other step is deriving the fine level processing which basically emphasizes the high signal to noise ratio regions. The gross level processing involves deriving the different speech production features. The details of this process can be found in Krishnamoorthy and Prasanna (2011). An example plot of the various speech production features and their role in the gross level processing can be seen in Fig. 2.

The fine level processing involves emphasizing the high signal to noise ratio regions and these regions mostly correspond to the location of the epochs. Hence a robust epoch extraction method needs to be utilized here and the epoch extraction method explored in this work is based on the zero frequency filtered signal (ZFFS). This method has been shown to be robust to different noise types Murthy and Yegnanarayana (2008). The method for obtaining the ZFFS can be found in detail in Murthy and Yegnanarayana (2008). Based on the gross weight function and the fine weight function derived from the gross and fine level processing, a total weight function needs to be derived and this total weight function can be used to weight the LP residual for the temporal enhancement. An example of obtaining the total weight function can be seen in Fig. 3.

The total weight function is multiplied by the LP residual signal shown in Fig. 3e to obtain the weighted LP residual shown in Fig. 3f. The temporally enhanced speech signal can be obtained by synthesizing as follows:

$$S_t(z) = \frac{R_w(z)}{1 + \sum_{k=1}^{p} a_k z^{-k}} \tag{1}$$

where $S_t(z)$ is the temporally enhanced speech and $R_w(z)$ is the weighted LP residual and a_k are the LP filter Coefficients. The plot of the speech which has been temporally enhanced is shown in Fig. 3g.

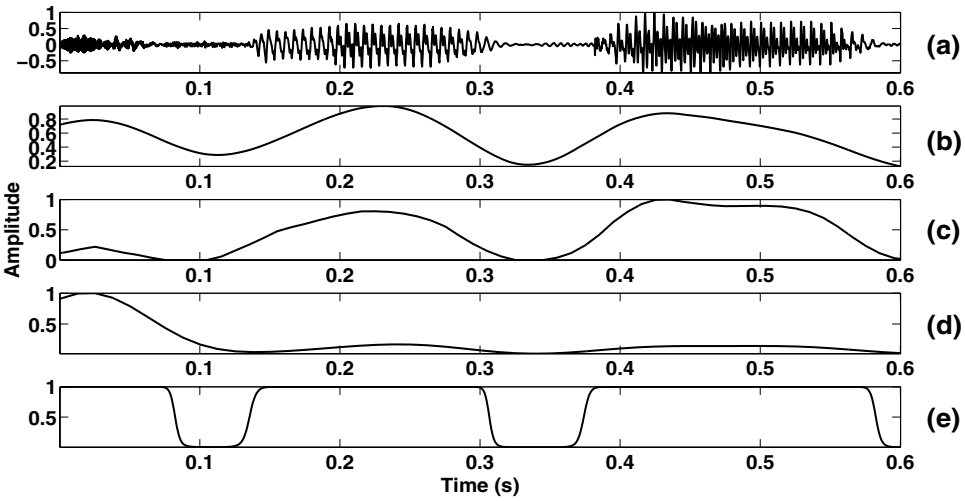

Fig. 2 Illustration of gross weight function derivation. **a** Speech degraded with background noise. **b** HE of LP residual. **c** Sum of ten largest peaks of DFT spectrum. **d** Modulation spectrum energy. **e** Gross weight function

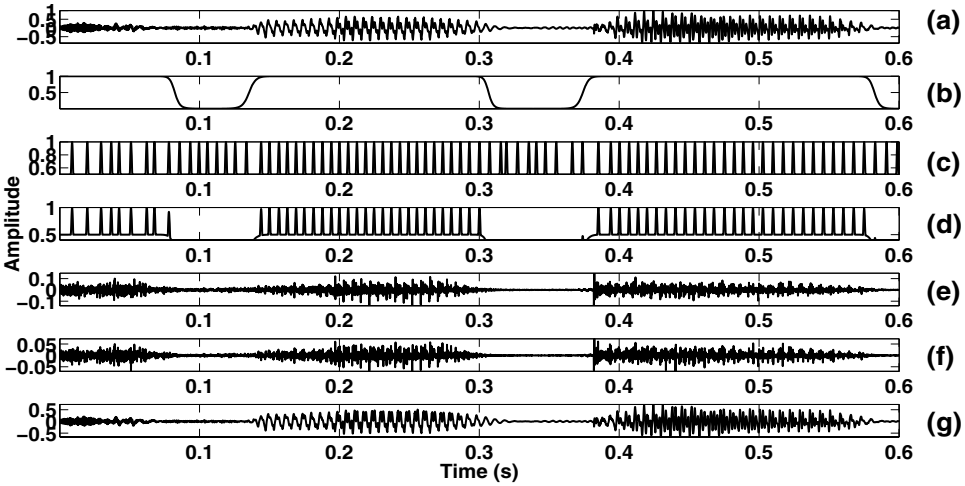

Fig. 3 Illustration of fine weight function derivation. **a** Speech degraded with background noise. **b** Gross weight function. **c** Epoch locations obtained from ZFF. **d** Overall weight function. **e** LP residual from degraded speech. **f** Weighted LP residual. **g** Temporally enhanced speech

2.2 Spectral enhancement

The conventional spectral processing methods have been explored here in which the short-term magnitude of the degradation and the degraded speech are estimated. A spectral gain function corresponding to the MMSE-LSA estimator Ephraim and Malah (1985), is applied to the magnitude spectra of the degraded speech, to obtain the enhanced speech spectra. The spectral gain function of this estimator is given by Ephraim and Malah (1985). The Fig. 4, illustrates the different methods of spectral and temporal enhancements and their respective spectrograms under degraded speech conditions.

$$H(k) = \frac{\zeta_k}{1+\zeta_k} exp(\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-x}}{x} dx) \quad (2)$$

where

$$v_k = \frac{\zeta_k}{1+\zeta_k}\gamma_k$$

ζ_k and γ_k are a priori SNR and a posteriori SNR,respectively.

The enhanced magnitude and degraded speech phase spectra are then combined to produce an estimate of clean speech and the overlap-add method is normally used for the re-synthesis in time domain.

3 Text dependent speaker verification (TDSV) system

3.1 Energy based end point detection

Robust end point detection is a crucial task for achieving good performance in speaker verification systems since the error introduced in the end point detection gets propagated in the overall performance of the TDSV system. Detecting the begin and end points of the speech signal under the clean case is comparatively easier when compared to the degraded condition cases. In most of the real time

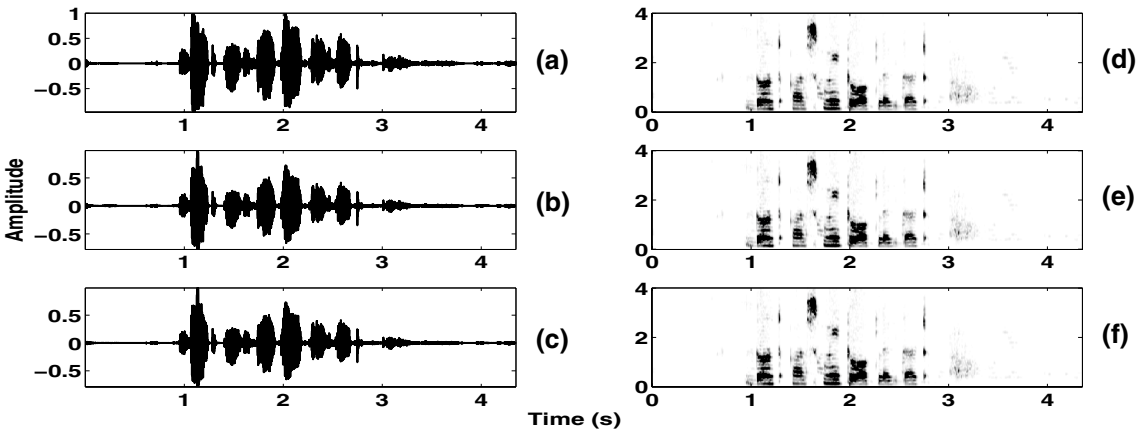

Fig. 4 Illustration of different stages of enhancement. **a** Speech degraded with with background noise. **b** Temporally enhanced speech. **c** Spectrally enhanced speech. **d** Spectrogram of **a**. **e** Spectrogram of **b**. **f** Spectrogram of **c**

applications the speech is mostly degraded since it may be recorded in different conditions. Hence there is a demand for robust begin and end point detection algorithms for the degraded conditions. The other advantages of end point detection is that there will be a reduction in the computational cost and response time of the overall system. This is because only the useful speech frames can be passed to the system to do the further processing for the TDSV system. The most popular end point detection system is based on using the energy as a feature.

The implementation of the energy based end point detection method is to calculate the energy of all the frames present in the speech signal and the average energy of the total frames is computed. Next the energy of each of the frames is compared with the threshold taken as 6 percent of average energy. The energy threshold is decided based on several speaker verification experiments (off-line). Different thresholds are considered and the one which is equal to 6 percent of average energy gave the best performance.

The comparison starts with the starting frame from the left side of the speech signal. The energy of the first three consecutive frames are compared to see if they are higher than the threshold which is equal to 6 percent of average energy. If this is true then the first frame is considered as the begin point. Similarly, starting with the last frame from the right side of the speech signal the energy of the three consecutive frames are compared with the 6 percent of average energy threshold. If this condition is true the last frame is considered as the end point. If the condition is false then the next frames are considered until the condition is satisfied and the begin and end points are detected. Finally, for the TDSV system, the speech signal considered between these start and end point is used for further processing.

3.2 Dynamic time warping based TDSV system

The details of the dynamic time warping based TDSV system is described in the next section where initially the feature extraction will be explained followed by the template matching.

3.2.1 Feature extraction

The uncontrolled database is collected in the real environment for the training and testing process. The speech signal is processed with a frame size of 20 ms and a frame shift of 10 ms. Each 20 ms of the signal is windowed with a Hamming window function and MFCCs are computed using the 22 spaced logarithmic filters. The first 13 coefficients are considered as a feature vectors. The Delta (Δ) and delta-delta ($\Delta\Delta$) features are computed using two preceding and two succeeding feature vectors from the current feature vector Davis and Mermelstein (1980). The final feature vector dimension is 39 which includes the first 13 coefficients, along with the first and second order derivatives of the 13 dimensional coefficients. These 39 dimensional feature vectors are normalized to have a zero mean and unit variance distribution using cepstral mean subtraction followed by cepstral mean variance Pradhan and Prasanna (2013). The extracted features are used for further speaker verification analysis purpose.

3.2.2 i-vector based system

The motivation behind this work is to implement and investigate the state-of-the-art speaker verification system by implicit modeling of speaker specific information in the *i*-vectors. For performance comparison, first a baseline *i*-vector based SV system is developed using energy based

end point detection Subhadeep Dey et al. (2014); Larcher et al. (2014).The *i*-vector system implementation, uses the total variability matrix based modeling as introduced in Dehak et al. (2011). The dimensionality of the GMM mean supervectors for a test speech is reduced by projecting it to a low rank subspace. The GMM mean supervector for a test speech supervectors are obtained by concatenating the mean vectors of the each user's model. The reduced dimension representation is called the *i*-vector. The implementation of the *i*-vector system described detailed in Subhadeep Dey et al. (2014); Larcher et al. (2014). The GMM mean supervector can be represented as,

$$M_s = m + Tw_s \tag{3}$$

where m is the user and channel independent supervector and M_s is the adapted GMM mean supervector and w_s is the linearly related to *i*-vector of the given test speech. The speaker verification done by comparing the *i*-vectors corresponding to test speech and claimed user's training utterance using the cosine kernel score between these two *i*-vectors. To further reduce the dimensionality of the *i*-vectors LDA and WCCN is applied, improves the performance of the TDSV system.

3.2.3 Template matching

The existing TDSV system uses MFCC features and DTW based template matching for speaker verification of a claimed identity. The main objective of the DTW was to exploit the linear time normalization alignment by implicitly assuming that the speaking rate variation is proportional to the time, speed and duration of the speech utterance. It is also independent of the speech pattern and mostly involves a sequence of short-time acoustic representations being spoken by the claimed identity. It is believed that the timing difference between train utterance and test speech utterances are minimized by using the warping in the time axis of one so that the best alignment is obtained with the test utterance. In other words, the test utterances are sometimes stretched and compressed so as to find the proper alignment that results in best possible match between the train utterance and test utterance on frame by frame basis using the MFCC features. The distance between the MFCC features of the train and test utterances can be calculated by using the minimum residual (Euclidean) distance Mahanta et al. (2016); Pandit and Kittler (1998).

In time series analysis, the TDSV systems are usually based on template based matching methods in which the time axis of the test speech utterance and reference models of the registered speakers are initially aligned using the DTW algorithm. Next the distances of the aligned frames between the training and testing utterances are computed. Finally the distances computed between the aligned frames are accumulated from the beginning of the utterance to the end of the utterance Yegnanarayana et al. (2005); Prasanna et al. (2003); Furui (1981); Rabiner and Juang (1993b). TDSV systems performs well when the linguistic content of the utterance is known and also when the characteristics of the speaker specific information is contained in the speech signal. From practical applications point of view, the DTW system should be designed in such a way that it should be language dependent, sensor dependent and session independent.

The test feature vectors are compared with reference template of the claimed model by the DTW algorithm. The DTW algorithm Mahanta et al. (2016) calculates the accumulated distance score between reference model train template X and test template Y of different lengths by considering a warping path as, The equation representing the DTW method is as follows,

$$DTW_{\phi}(X, Y) = \sum_{p=1}^{N} DTW(\phi_X(p), \phi_Y(p))m(p)/M_{\phi} \tag{4}$$

where the $DTW_{\phi}(X, Y)$ is total accumulated cost distance of p, $DTW(\phi_X(p), \phi_Y(p))$ is the shortest time spectral distortion, m(p) is a non-negative weighting co-efficient of the warping path and M_{ϕ} is the warping path normalizing factor. The claimed speaker speech data is matched to the claimed reference model using DTW algorithm,to give the shortest distance score. The cohort speakers are selected randomly for each user from the enrolled users excluding the corresponding user and kept fixed for decision logic.The decision is based on the obtained DTW score with respect to a set of four cohort speakers by comparing the distance score against the claimed model. This claimed user score compared to the scores obtained from the set of four cohort speakers of the claimed user and the decision of the claimed user is to accept/reject a claim.

The accuracy of a speaker verification system depends on the statistical match between the speaker model and test features Kinnunen and Li (2010). The DTW algorithm is applied on the features derived from the speech signal collected using the voice-server. The performance of the system after applying the proposed end point detection algorithm on this speech signal is evaluated. It can be observed in Fig. 5a, b that the DTW warping path for an utterance before and after using the proposed end point detection algorithm is different. The main observations are that without using the end point detection the warping path deviates from the regression line and using the end point detection actually helps in bringing the DTW warping path closer to the regression line.

From the practical usability point of view, the users are allowed to give test speech freely with no constraints

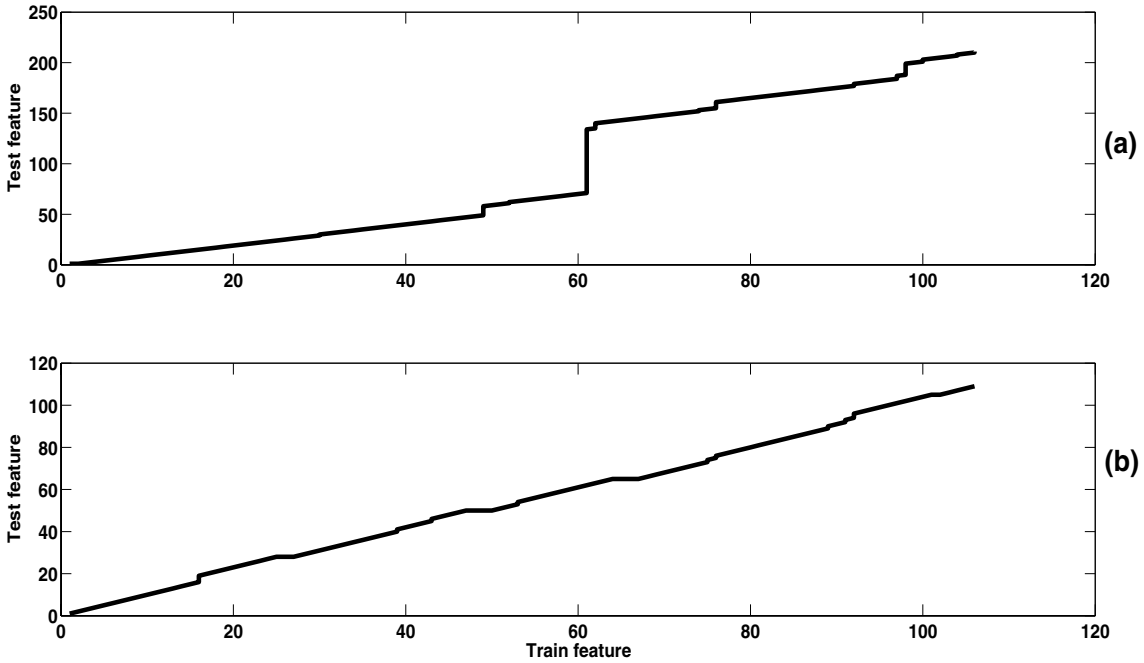

Fig. 5 Evidences for degraded speech utterance for *Don't ask me to walk like that* **a** speech utterance showing the non-overlapping background noise and the warping path which is deviated away from the regression line before applying the enhancement and end point detection, **b** showing the warping path nearer to the regression line after applying the enhancement followed by end point detection and follows the DTW constraints

on duration, recording condition, channel, lexical content and quality of the speech. Compared to other variabilities, lexical variability can be manageable in the uncontrolled environmental conditions. The performance of the TDSV systems are influenced by many of these other possible variabilities. Using the DTW, matching of short duration phrases in the training and testing utterances with high accuracy makes it an attractive option for commercial speech based multilevel person authentication systems. The accuracy of a speaker verification system depends on the matching between the speaker model and test features Yegnanarayana et al. (2005); Furui (1981); Rabiner and Juang (1993b). In the present work, a TDSV system is developed using 39 dimensional MFCC features and template matching technique of the DTW algorithm Rabiner et al. (1978); Sakoe and Chiba (1978).

4 Experimental results and analysis

The experimental setup used for the TDSV systems are described in the following section.

Two databases, the RSR2015 Larcher et al. (2014), and the IITG-MV database Haris et al. (2011); Pradhan and Prasanna (2011), which represents speakers from two different countries- Singapore and India, are used for the task of TDSV system. The RSR2015 database contains audio recordings, tablets and recorded on mobile phones, in a closed office environment, with age ranging from 17 to 42. The RSR2015 database is one of the most publicly available database, especially designed for the TDSV system based studies under different duration and lexical contraints, which contains 300 speakers (143 female and 157 male), each of the speakers have 9 sessions of each phrase, out of 9 sessions, three sessions are used for training, three sessions are used for calculation of speaker model and remaining three session for testing. Part I of the RSR2015 database is considered for the evaluation.

The RSR2015 database is collected in a closed room considered as clean database. In order to evaluate the performance of the TDSV system under degraded conditions, the test speech files of the RSR2015 database are corrupted with Bable noise from the NOISEX-92 database Varga and Steeneken (1993). The energy level of the noise is scaled such that the overall SNR of the noise added to test speech signal is maintained at 20, 15, 10, 5 and 0 dB, respectively. The performance of the RSR2015 database under clean and different noise degraded test speech conditions are evaluated using the i-vector and DTW algorithm.

The IITG-MV database is collected on the IVR system and the user callflow through (integrated services digital network-primary rate interface) ISDN-PRI line which can handle telephone channel calls through computer telephone interface (CTI) card. The database collected from

the pre-defined mobile set calls of the students over the telephone channel. The IVR is facilitated on a voice-server which runs the Asterisk software, used for private branch exchange (PBX). PBX can be used to allow the server handling incoming and outgoing calls to and from public switched telephone network (PSTN) or voice over internet protocol (VoIP) services. This callflow is made for execution of the database collected from the students of our department at our institute level. The users have to give their attendance by making a call from few pre-defined mobile handsets. In the initial stage an IVR system guides a user in enrollment and an enrolled user for testing. On the regular basis, the registered students are allowed to call to the toll-free number and mark their attendance in a practical setting. The overview of the database collection process over the telephone network which is developed for speech biometric based attendance system described in Subhadeep Dey et al. (2014).

The RSR2015 database mainly focuses on TDSV task where the users pronounce same pass-phrases for verifying the claimed mode. To evaluate the performance of the TDSV task, three sentences have been selected from the RSR2015 databse among the 30 sentences slected from the TIMIT database, used for the evaluation. The average duration per sentance varies from 2.73 to 3.65 s. These sentances are selected to evaluate the impact of the different lexical contant with a short duration by considering overall the speakers and sessions. They are

"Only lawyers love millionaires"-(TD-1),
*"I know I did not meet her early enough"-(TD-2)*and
"The Birthday party has cupcake and ice-cream"-(TD-3).

In the RSR2015 databse, for each speaker, 9 utterances for every sentence, out of 9 utterances 3 are used for training and remaining 6 for testing.

The IITG-MV practical databse collected as a part of the development of a student attendance system, in the EEE department of IITG, Guwahati, India. The 325 students (276 male and 49 female), represent an adult age group between 22 and 30 years. In the training, 3 sentences (sessions) each for 3 speech sentences, were recorded for each speaker in clean environment and controlled conditions during the beginning of the semester. In the testing phase, the speakers are allowed to move freely within, in and out of an open hall where the pre-defined mobile handset were kept. There are multiple users made simultaneous calls to the voice-server to mark their attendance. The databse collected in the real and practical environment. The average duration of the each sentance of the IITG-MV databse vary between 3 and 5 s overall the users and sessions. The three sentences are

"Don't ask me to walk like that"-(TD-1),
"Lovely picture can only be drawn"(TD-2), and
"Get into the hole of tunnels"-(TD-3).

For the IITG-MV database, the 3 sentences recorded in the enrollemnt phase are used for training. The number of sentences for a particular sentences, claimed by a particular user for testing, vary randomly. 30 sentences are available for every sentence on an average for every user, for testing.

The test speech data is collected in an uncontrolled environment at every other time. Due to the various factors like background noise, background speech, clipped data, sensor mismatches, reverberation, real environmental noise, blurred noise and some unwanted signals, degradation may be present in the test speech signal Haris et al. (2011); Pradhan and Prasanna (2011). A preliminary study is performed by using clean speech for training as well as for testing without the presence of background noise, background speech and any other real time environmental noise. Next the testings are made on a regular basis by the users which give the voice biometric based attendance system. As a result there are no constraints in the environment. This resulted in degradation of the speech signal due to various environmental conditions and poor performance in automatic speech processing tasks like speaker verification and speech recognition is obtained. These failures occur in TDSV system mainly because of the end-point detection problem, background noise, background speech and speech from the other competing speakers as well as impulses at the start and end of the speech which leads to degrading the overall performance of the system in terms of perceptual quality and intelligibility. The experimental evaluation is also taken under such degraded conditions to evaluate the performance of the system.

Initially the clean speech is used for training and clean test speech is taken recorded in controlled environment to evaluate the performance of the TDSV system. In the text dependent module of the speaker verification system, the three selected predefined text dependent prompts used during training and testing are used. The scores are generated for each speaker test trial using DTW algorithm. The score level analysis is done on the three different basis (stringent, moderate and less moderate). The verification of the claimed identity is performed during the testing phase using the sequence of the extracted feature vectors from the test speech and the scores are obtained using the DTW approach against the 4 cohort speakers for each of the enrolled speakers. The controlled database results are used as the baseline performance using the cohort based method.

Next the database collected in the practical and uncontrolled environment which is mostly degraded speech. This is used to evaluate the performance of the TDSV system. Using various methods for the speech recorded in uncontrolled environment, different sets of experiments are conducted on the voice biometric based attendance for the

TDSV system in order to find the differences between the speaker verification performances for the clean speech database condition and the degraded speech condition. A test speech is taken under such degraded conditions to evaluate the performance.

4.1 Evaluation of TDSV system

The database collected from the IITG voice-server, is practical and under the influence of real time environmental conditions. This database contains the degraded speech. The text-dependent module of the speaker verification system is based on the three predefined prompts used during the training and testing speech utterances. Using DTW algorithm, the optimal warping path that provides the best matching under two speech segments is obtained. By carefully observing the score level analysis, it reveals that among the accumulated distances computed, the genuine claimed identity gives minimum distance compared to the cohortset scores.

The scores are generated for each test utterance using DTW algorithm. The algorithm calculates the distance of the test utterance with each of the three training models and four fixed cohort set utterances for a particular speaker. The decision of a claimed test utterance is either accepted or rejected based on the cohort set distance scores. For each of the enrolled speakers, there exist a fixed pre-defined cohort set. The DTW distance score is compared to the scores generated from the claimed model and four cohort set for arriving the scores at the decision level. The test utterance is tested against the claimed model (which consists of three generated claimed scores and twelve cohort set scores obtained from the four cohort set of the claimed speakers). Hence, for a particular test utterance, 15 scores are generated which are sorted in the ascending order of their distance scores. The claimed identity is accepted if one out of the three distance scores (1/3), (2/3) and (3/3) generated for the claimed speakers, occupies a place in the top three position in the array of the 15 ascending distance scores. The TDSV system is evaluated and the performances are reported in EER and minimum DCF.

4.2 Results using energy based method on clean speech

The performance of the RSR2015 database under clean condition is evaluated on the RSR2015 Part I out of the three different cases for the challenging task.

At the first level, optimal value of energy threshold for the baseline TDSV system is selected on the RSR2015 database and the performance of the TDSV system under clean speech condition is evaluated using the energy based end point detection. The database is collected under controlled conditions and considered as the clean speech. The experiments on the clean speech are performed and the performance of the database is measured in equal error rate (*EER*) and minimum *DCF*. Table 1, shows the performance of the *i-vector* system on the RSR2015 database using development data set of part I, in terms of *EER* and minimum *DCF*. Table 2, shows the *i-vector* system performance on the RSR2015 database using evaluation data set of part I, in terms of *EER* and minimum *DCF* on male and female dataset.

Table 1 The performance of the *i-vector* system on the RSR2015 database using development set of part I shows the evaluation in terms of equal error rate (*EER*) and minimum *DCF* for different test trials

Data	Male		Female	
Method	*i-vector*		*i-vector*	
Metrics	EER	DCF	EER	DCF
Development	3.41	0.35	6.94	0.47

Table 2 The performance of the *i-vector* system on the RSR2015 database using evaluation set of part I shown in terms of equal error rate (*EER*) and minimum *DCF* for different test trials

Data	Male		Female	
Method	*i-vector*		*i-vector*	
Metrics	EER	DCF	EER	DCF
Evaluation	3.4	0.32	3.88	0.36

As can be seen from the Table 1 all the genuine utterances are giving good performance in the development data set as compared to the imposter claim. Note that the performance for the clean condition is expected to be better than the degraded speech conditions which will be evaluated later and this validates the TDSV system performance. By comparing the male and female data sets of the RSR2015 databse, male speakers are giving better performance as compared to the female speakers in both development data set and evaluation data set.

4.3 Results using energy based method on degraded speech

For all the experimental conditions considered using DTW algorithm SV systems provide better performance in clean condition and as the level of noise increases, the TDSV system performance reduces for each system decreases in RSR2015 database as shown in the 3rd and 4th columns of the (Table 3). For instance the RSR2015 database, the DTW algorithm based SV system EER decreases from 3.73% to 12.54% as the noise varies from clean speech to degraded speech in terms of SNR ranging from 20-0 dB in steps of 5 dB. The performance improvement in the proposed system can be observed for the IITG-MV database using energy based end point detection method for the TDSV system

Table 3 Performance of TDSV systems in terms of *EER* and minimum *DCF* using RSR2015 database using different methods

Method	TD-1	Energy based		Temporal enhancement		Spectral enhancement		Combination	
Database	SNR	EER	DCF	EER	DCF	EER	DCF	EER	DCF
RSR2015	Clean	3.73	0.06	4.37	0.07	3.32	0.04	2.37	0.06
	B20	5.19	0.08	4.92	0.08	4.49	0.09	4.24	0.08
	B15	5.64	0.09	4.92	0.08	5.38	0.10	4.84	0.09
	B10	7.53	0.14	6.54	0.11	6.23	0.13	5.97	0.14
	B5	9.52	0.15	7.89	0.13	6.72	0.22	6.28	0.14
	B0	12.54	0.32	12.57	0.33	12.43	0.33	11.58	0.31

Table 4 Performance of the TDSV systems in terms of the *EER* and minimum *DCF* using IITG-MV database using different methods

Method		Energy based		Temporal enhancement		Spectral enhancement		Combination	
Database	Prompt	EER	DCF	EER	DCF	EER	DCF	EER	DCF
IITG-MV	TD-1	5.15	0.082	4.47	0.079	4.23	0.068	4.08	0.065
	TD-2	4.76	0.0674	4.51	0.0665	4.19	0.0624	4.05	0.069
	TD-3	4.99	0.074	4.72	0.064	4.28	0.063	3.92	0.058

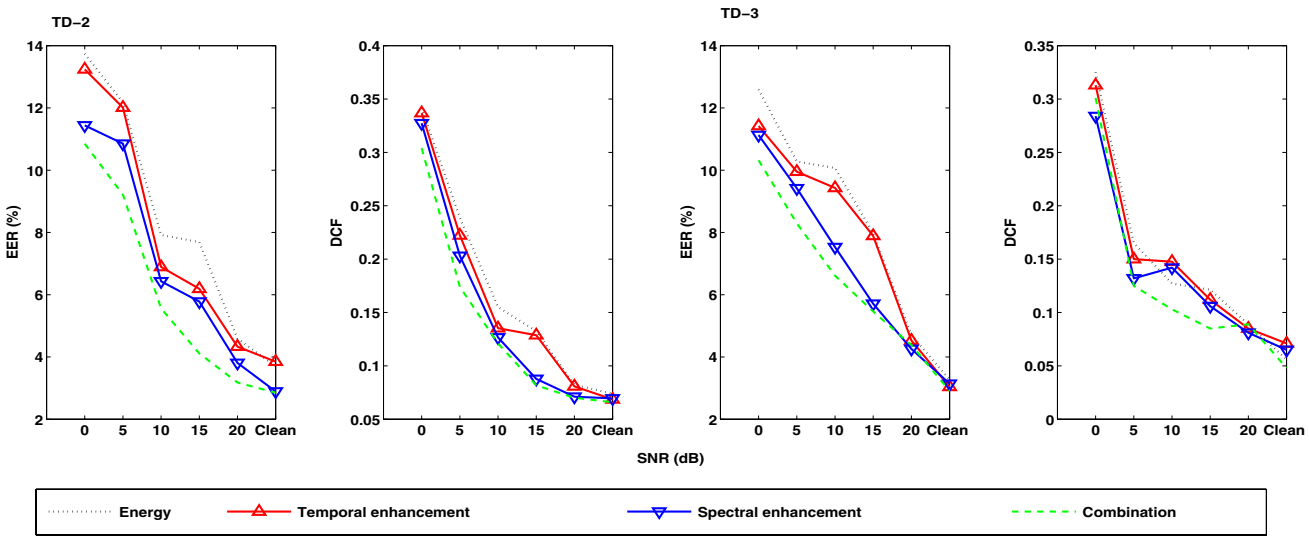

Fig. 6 Summary of TD-2 and TD-3 test trials, DTW based TDSV systems performance in terms of the *EER* and minimum *DCF* for different experimental setup on RSR2015 database

shown in Table 4. The Tables 3 and 4 are for the TD-1 case. The performance for the TD-2 and TD-3 case is shown in Fig. 6. The trend in the results are similar to the TD-1 case.

4.4 Results using energy based end point detection followed by speech enhancement

In this section the performance of the TDSV system when end point detection is first performed followed by enhancement is analyzed named as energy based end point detection method. The performance can be found in the 5th to 8th columns of the Table 3. Initially the temporal enhancement is performed on the files after end point detection and the performance of the TDSV system after temporal enhancement is found to be better compared to the case when only energy based end point detection is performed. The reason being that the temporal enhancement is able to suppress the noise components in between the end points. The enhanced file may have characteristics of the clean speech and hence better performances are achieved. Similarly, the spectral enhancement which is basically spectral subtraction is performed after end point detection and this process gives a

better result than the temporal case. The results for this case are given in the 5th to 8th columns of the Table 4.

4.5 Results using speech enhancement followed by energy based end point detection

This section presents the results of the same enhancement strategies but they are applied before the end point detection for the TDSV system. The results are given in Tables 3 and 4. It can be seen that enhancement followed by end point detection shows better performance than the case earlier. The trend in the results of Tables 3 and 4 showed that the combined temporal and spectral enhancement performed better compared to the temporal and spectral enhancement cases. The combined results are given in the last two column of the Tables 3 and 4. The EER and DCF showed the improved performance when compared to the baseline energy based end point detection method. The contribution of the enhancement abilities of the temporal and the spectral case are combined thus giving a better noise suppression. The best performance is obtained with the combination of temporal and spectral enhancement followed by end point detection, respectively. The reason for the better performances of the enhancement followed by the end point detection is again attributed to the noise suppression. First of all the TDSV system depends on the end point detection and a good accuracy end point detection system will give good results for the TDSV system.

In the earlier case where the end point detection is first performed prior to the enhancement, the end points obtained may not be accurate due to the presence of noise. However in the case when the end point detection is performed after enhancement, the enhancement modules suppress the noise and hence better end point detection can be done on the enhanced signal, since now the signal will be having the quality nearer to the clean speech case. As a result the end point detection using energy will be better when performed on the enhanced case compared to directly performing on the degraded case. Also the enhancement helps in noise suppression throughout the signal. The overall improvements of the TDSV system can be attributed to the better end points detected as well as the enhanced signal.

5 Conclusion and future work

In this work the performance of the TDSV system on the degraded speech conditions is analyzed. The temporal and spectral enhancement strategies are performed on the signal either before or after the energy based end point detection. The enhancement strategies tend to improve the performance of the TDSV system where the case when enhancement is performed after end point detection showed the best performance.

The future work can employ other forms of enhancement for improving the TDSV system. These enhancement strategies can be in terms of the source, vocal tract system and suprasegmental characteristics of speech. Additionally better versions of the end point detection algorithms can be used in place of the energy based end point detection to improve the accuracy of the overall TDSV system.

Acknowledgements The authors would like to thank Mr. Rajib Sharma, in the EEE department of IITG, Guwahati, for his help in making this work come to fruition.

References

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *27*(2), 113–120.

Chakrabarty, D., Prasanna, S. R., Mahadeva, Das, & Kumar, Rohan. (2013). Development and evaluation of online text-independent speaker verification system for remote person authentication. *International Journal of Speech Technology*, *16*(1), 75–88.

Das, C. K., Sanaullah, M., Sarower, H. M. G., & Hassan, M. M. (2009). Development of a cell phone based remote control system: An effective switching system for controlling home and office appliances. *International Journal of Electrical and Computer Sciences IJECS*, *9*(10), 37–43.

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

Deepak, K. T., & Prasanna, S. R. M. (2016). Foreground speech segmentation and enhancement using glottal closure instants and mel cepstral coefficients. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(7), 1204–1218.

Dehak, N., Kenny, P. J., Dehak, R., Dumouchel, P., & Ouellet, P. (2011). Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(4), 788–798.

Ephraim, Y., & Malah, D. (1985). Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *33*(2), 443–445.

Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *29*, 254–272.

Haris, B., Pradhan, G., Misra, A., Shukla, S., Sinha, R., Prasanna, S., (2011). Multi-variability speech database for robust speaker recognition. In *Communications (NCC), 2011 National conference on IEEE, pp. 1–5.*

Hébert, M., (2008). Text-dependent speaker recognition. In *Springer handbook of speech processing, pp. 743–762.*

Kinnunen, T., & Li, H. (2010). An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication*, *52*(1), 12–40.

Krishnamoorthy, P., & Prasanna, S. R. M. (2011). Enhancement of noisy speech by temporal and spectral processing. *Speech Communication*, *53*(2), 154–174.

Larcher, A., Lee, K. A., Ma, B., & Li, H. (2014). Text-dependent speaker verification: Classifiers, databases and rsr2015. *Speech Communication*, *60*, 56–77.

Mahanta, D., Paul, A., Ramesh K Bhukya, Rohan K Das, Sinha, R, Prasanna, S.R.M., (2016). Warping path and gross spectrum information for speaker verification under degraded condition. In *Communication (NCC), 2016 Twenty Second National Conference on IEEE, pp. 1–6.*

Marinov, S., (2003). Text dependent and text independent speaker verification system: Technology and application. Overview article.

Murthy, K. S. R., & Yegnanarayana, B. (2008). Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech, Language Processing*, *16*(8), 16021613.

Onukwugha, C., & Asagba, P. (2013). Remote control of home appliances using mobile phone: A polymorphous based system. *African Journal of Computing and ICT*, *6*(5), 81–90.

Pandit, M., Kittler, J., (1998). Feature selection for a dtw-based speaker verification system. In *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on IEEE, Vol. 2., pp. 769–772.*

Piyare, R., Tazil, M., (2011). Bluetooth based home automation system using cell phone. In *Consumer Electronics (ISCE), 2011 IEEE 15th International Symposium on IEEE, pp. 192–195.*

Pradhan, G., & Prasanna, S. M. (2011). Speaker verification under degraded condition: A perceptual study. *International Journal of Speech Technology*, *14*(4), 405.

Pradhan, G., & Prasanna, S. M. (2013). Speaker verification by vowel and nonvowel like segmentation. *IEEE Transactions on Audio, Speech, and Language Processing*, *21*(4), 854–867.

Prasanna, S. M., Zachariah, J. M., Yegnanarayana, B., (2003). Begin-end detection using vowel onset points. In *Workshop on Spoken Language Processing.*

Prasanna, S. R. M., Zachariah, J. M., Yegnanarayana, B. (2003). Begin-end detection using vowel onset points. In *Workshop on Spoken Language Processing*, (TIFR, Mumbai, India).

Rabiner, L., & Juang, B.-H. (1993a). *Fundamentals of speech recognition*. New Jersey: Pearson Education.

Rabiner, L. R., & Juang, B. H. (1993b). *Fundamentals of speech recognition*. Upper Saddle River: Prentice-Hall.

Rabiner, L. R., Rosenberg, A. E., & Levinson, S. E. (1978). Considerations in dynamic time warping algorithms for discrete word recognition. *The Journal of the Acoustical Society of America*, *63*(S1), S79–S79.

Sakoe, H., & Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *26*(1), 43–49.

Savoji, M. H. (1989). A robust algorithm for accurate endpointing of speech. *Speech Communication*, *8*, 45–60.

Shahriyar, R., Hoque, E., Sohan, S., Naim, I., Akbar, M. M., & Khan, M. K. (2008). Remote controlling of home appliances using mobile telephony. *International Journal of Smart Home*, *2*(3), 37–54.

Subhadeep Dey, Sujit Barman, Ramesh K Bhukya, Rohan K Das, Haris, BC, Prasanna, S.R.M., Sinha, R, (2014). Speech biometric based attendance system. In *Communications (NCC), 2014 Twentieth National Conference on IEEE, pp. 1–6.*

Tsao, C., Gray, R. M., (1984). An endpoint detection for lpc speech using residual look-ahead for vector quantization applications. In *IEEE International conference on acoustic, speech, signal processing.*

Varga, A., & Steeneken, H. J. (1993). Assessment for automatic speech recognition: Ii. noisex-92: A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Communication*, *12*(3), 247–251.

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

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.