I.J. Image, Graphics and Signal Processing, **2014, 6, 27-35**
Published Online May 2014 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijigsp.2014.06.04

Silence Removal and Endpoint Detection of Speech Signal for Text Independent Speaker Identification

Tushar Ranjan Sahoo
International Institute of Information Technology, Bhubaneswar, Odisha, India
E-mail: tushar@iiit-bh.ac.in

Sabyasachi Patra
International Institute of Information Technology, Bhubaneswar, Odisha, India
E-mail: sabyasachi@iiit-bh.ac.in

***Abstract*—**In this paper we propose a composite silence removal technique comprising of short time energy and statistical method. The performance of the proposed algorithm is compared with the Short Time Energy (STE) algorithm and the statistical method with varying Signal to Noise Ratio (SNR). In the presence of low SNR the performance of proposed algorithm is highly appreciable in compare to STE and statistical method. We have applied the proposed algorithm in the pre processing stage of speaker identification system. A comparison between the speaker identification rate including and excluding the silence removal technique shows around 20% increase in identification rate by the application of this proposed algorithm.

***Index Terms*—**End point detection, short time energy, Gaussian distribution, signal to noise ratio, speaker identification, mel frequency cepstral coefficient, Gaussian mixture model.

I. Introduction

The problem of separating the speech segments of an utterance from the background noise is called speech detection (or endpoint detection, or even voice activity detection). This is very important in many areas of speech signal processing: speech and speaker reorganization, speech compression and transmission. A proper locations of regions of speech (sometimes together with pause removal), not only reduces the amount of processing, but also increases the accuracy of speech processing system. Speech detection is an old and known problem; however, it has not been completely answered until today. The major difficulty of the speech detection task deals with the variability of the speech and background noise patterns. Pre-processing of the speech signal serves various purposes in any speech processing application. It includes Noise Removal, Endpoint Detection, Pre-emphasis, Framing, Windowing etc. Out of these the removal of silence/unvoiced portion is the fundamental step for applications like Speaker Identification. The information which is more important from the prospective of speaker identification is generally contained inside the voiced part of the speech signal. Therefore the process of isolating the redundant information especially in the unvoiced part in the preprocessing step bears a lot of importance; to reduce the dimension and hence the computational complexity in the subsequent stages, at the same time not hampering the speaker identification rate. Robust Speaker Identification also demands efficiency in feature extraction even in noisy environment, where the importance of silence removal is more than that of clean speech. Silence/unvoiced portion of speech is affected more by noise than voiced portion. There are three main events in speech i.e. silence (S), unvoiced (U) and voiced (V). It should be clear that the segmentation of the waveform into well-defined regions of silence, unvoiced and voiced is not exact; it is often difficult to distinguish a weak, unvoiced sound from silence, or weak voiced sound from unvoiced sounds or even silence. However, it is usually not critical to segment the signal to a precision much less than several milliseconds; hence, small errors in boundary locations usually have no consequence for most applications. Since for most of the practical cases the unvoiced part has low energy content and thus silence (background noise) and unvoiced part is classified together as silence/unvoiced and is distinguished from voiced part.

In this paper, present section briefly introduced the work along with the organization of this paper. Next section elaborates the relevant background theory of this paper. Then the speaker identification process is discussed with feature extraction and parameter estimation. Proposed algorithm for silence removal and end point detection is presented in the next section. Experimental evaluation and comparison of proposed method with existing methods shows its effectiveness. At last, the impact of composite silence removal technique on speaker identification method is summarized with a concluding remark.

II. Background

The methods used in speech detection can be classified

into two general categories, according to their basic principles:

Explicit methods, that detect the endpoints independently of the subsequent processes [1]; the main techniques are energy-based [2], pattern recognition-based [3], or use of Hidden Markov Models [4] or neural networks [5] to explicitly determine the endpoints.

Implicit methods, that use the application itself to detect the endpoints; this approach assumes that the endpoint detection is an implicit part of the speech processing system and is developed along with the application [6].

All these techniques may work well enough for high signal to noise (SNR) ratios, but most of them are not really adaptive and degrade their performances as the noise level increases. Energy based speech activity detector [7] and zero crossing rate [8] are two widely accepted silence removal techniques adopted in many speech related application. For specific applications other techniques based on pattern- recognition, Hidden Markov Model (HMM) and Neural Network [4, 5] are used. The Statistical behavior of the background noise is more useful to separate out the unvoiced portion of speech when speech signal starts with background noise. We have explained some of the important silence removal techniques in this section.

A. Speech Activity Detector by Short Time Energy (STE) of signal

The procedure of speech detection and the elimination of silence part for a recorded utterance using STE algorithm [9, 10, 11] consist of three steps:

(1) Pre-processing: Input speech signal is divided into frames, each of duration 15ms by using hamming window. Because the ZCR is highly susceptible to 50/60 Hz hum, very low frequency noise, DC offset, etc., the waveform is first high-pass filtered. The filter is used only for energy and ZCR computing and does not affect the input signal.

(2) Speech boundaries estimation: The energy of each of the frames is calculated by the following equation:

$$E(m) = \sum_{n=1}^{N} x_m(n)^2 \tag{1}$$

There are N numbers of sample in a frame and M such frames are there in the input signal. Two energy thresholds, T1 and T2, are calculated over all M frames of the input signal, based on three typical values of signal energy which could be derived from the following equations:

$$Energy\_\max = \max(E(i)), i = 1,2,\cdots,M \tag{2}$$

$$Energy\_\min = \min(E(i)), i = 1,2,\cdots,M \tag{3}$$

$$T_1 = Energy\_\min(1 + 2\log_{10}\frac{Energy\_\max}{Energy\_\min}) \tag{4}$$

$$SL = \frac{\sum_i E(i)}{\sum_i 1}, \text{ where} \tag{5}$$

i is the index for all frames having $E > T_1$

$$T_2 = T_1 + 0.25\,(SL - T_1) \tag{6}$$

In the previous equations, *Energy_max* is the peak energy value of the input speech signal, *Energy_min* is the minimum energy value and SL is the average level of the signal above T1. The speech boundaries are approximately estimated based on the following energy criteria:

a. When energy exceeds T1 and subsequently exceeds T2, the crossing point with T1 level is declared a 'preliminary start point' (PS).

b. When energy falls below T2 and then falls below T1, the crossing point with T1 level is declared a 'preliminary endpoint' (PE).The region between the PS and the PE will be further called 'word'.

c. A short isolated 'word' is a 'noise spike'; it is marked as unnecessary and it will be rejected in the silence removal procedure.

(3) Silence elimination: All frames that are not detected as 'voice segments' (do not belong to 'word') by steps 2 are considered as 'silence' and removed from the given utterance.

B. Silence Removal using Zero Crossing Rate (ZCR)

Recordings of perfectly clean speech are very difficult. This means that often there is some level of background noise that interferes with the speech and leading a higher zero-crossings rate in the silence region as the signal changes sign from just one side of zero amplitude to the other and back again. Silence removal using zero crossing rates [8] has three steps:

(1) Pre-processing: Input speech signal is divided into 10ms frame by using hamming window.

(2) Calculation of zero crossing rate (Zm):

$$Z_m = \sum_{n=1}^{N-1} |\,\mathrm{sgn}[x_m(n+1)] - \mathrm{sgn}[x_m(n)]\,| \tag{7}$$

In equation 7, m represents the frame number, xm (n) stands for nth sample in frame m.

$$\mathrm{sgn}(x_m(n)) = \begin{cases} 1 & if\ x_m(n) \geq 0 \\ 0 & elsewhere \end{cases} \tag{8}$$

(3) Silence Elimination: All frames having zero crossing rates higher than a threshold are eliminated and they are categorized as silence or unvoiced part of speech.

C. Silence Removal and Endpoint Detection Algorithm by Statistical behavior of background noise.

One of the basic properties of any speech signal is, the first 100 to 200 ms or more (1600 to 3200 samples if the sampling rate is 16000 samples/sec) of a speech recording corresponds to silence or background noise, because the speaker takes some time to speak when recording starts. The silence or background noise is considered to be white noise and hence its distribution is normal distribution [10, 12]. Analytically,

$$p(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \quad (9)$$

The parameters μ and σ for the above distribution are calculated using first 3200 samples of the input speech as these samples are considered to be background noise. Probability of any data point x obeys the following three relations

$$p(|x-\mu| \le \sigma)\, is\, 0.68$$
$$p(|x-\mu| \le 2\sigma)\, is\, 0.95$$
$$p(|x-\mu| \le 3\sigma)\, is\, 0.997$$

So for any sample x if $\frac{|x-\mu|}{\sigma} \le 3$, then we are almost 100 percent sure that it belongs to the distribution of background noise and hence it can be eliminated from the speech part. Silence removal using this algorithm has four steps:

(1) Calculate the mean μ and standard deviation σ of the first 3200 samples of the given input signal. First 3200 samples represent background noise. So the background noise is characterized by this μ and σ.

$$\mu = \frac{1}{3200} \sum_{i=1}^{3200} x(i) \quad (10)$$

$$\sigma = \sqrt{\frac{1}{3200} \sum_{i=1}^{3200} (x(i)-\mu)^2} \quad (11)$$

(2) The one-dimensional Mahalanobis distance is estimated in the recorded speech signal for each sample and is classified as per the following conditions: $If\ \frac{|x-\mu|}{\sigma} > 3$, the sample is a voiced sample otherwise it is a silence/unvoiced sample.

(3) The voiced sample is marked as 1 and unvoiced sample as 0. The whole speech signal is divided into non-overlapping windows, each of duration 10ms. Now the complete speech is transferred to sequence of zeros and ones.

(4) The frame in which number of zeros exceeds number of ones treated as silence or unvoiced and it is eliminated from the speech part do not belong to 'word') by steps II are considered as 'silence' and removed from the given utterance.

III. Speaker identification process

The process of speaker identification is divided into two main phases e.g. the enrollment phase and the identification phase. During the enrollment phase also known as speaker training, speech samples are collected from the speakers, and they are used to train their models. The collection of enrolled models is also called a speaker database. In the second phase, identification phase, a test sample from an unknown speaker is compared against the speaker database. Both the phases involve a common first step i.e. feature extraction, where the speaker dependent features are extracted from the speech sample. The main purpose of this step is to reduce the amount of test data while retaining speaker discriminative information. Then in the enrollment phase, these features are modeled and stored in the speaker database. In the identification step, the extracted features are compared against the models stored in the speaker database. Based on results obtained from these comparisons the final decision about speaker identity is made.

A. Feature Extraction: Mel Frequency Cepstral Coefficient

The acoustic speech signal embeds various kinds of information about speaker e.g. "high-level" properties such as dialect, context, speaking style, emotional state of speaker etc and also some "low-level" properties such as pitch (fundamental frequency of the vocal cord vibrations), intensity, formant frequencies and their bandwidths, spectral correlations, short-time spectrum and others. The amount of data, generated during the speech production, is quite large while the essential characteristics of the generated speech changes quite slowly therefore, requires relatively less data to represent the characteristics of speech and the person who has spoken it. According to these matters feature extraction is a process of reducing data while retaining the speaker discriminative information of the speakers.

Mel-frequency cepstrum coefficients (MFCC) are well known features used to describe speech signal [13, 14, 15]. They are based on the known evidence that the information carried by low-frequency components of the speech signal is phonetically more important for human perception than carried by high-frequency components. Technique of computing MFCC is based on the short-term analysis, and thus from each frame a MFCC vector is computed. MFCC extraction is similar to the cepstrum calculation except that one special step is inserted, namely the frequency axis is warped according to the mel-scale.

B. Speaker modeling and Feature matching: Gaussian Mixture Model (GMM)

Classification of speakers is a decision making process for validating the speaker given a speech signal based on the previously stored or learned information. This step is usually divided into two parts, namely modeling and matching. The modeling is a process of enrolling speaker to the identification system by constructing a model of his/her voice, based on the features extracted from his/her speech sample. The matching is a process of computing a matching score, which is a measure of the similarity of the features extracted from the unknown speech sample and speaker model. Vector Quantization and Gaussian Mixture Model are the two widely used classifiers for speaker identification [16,17,18]. We have given a brief description on GMM on the following paragraph.

Gaussian mixture modeling (GMM) belongs to the stochastic modeling and is based on the modeling of

statistical variations of the features. Therefore, it provides a statistical representation of how speaker produces sounds. For the identification each speaker is represented by his/her GMM, which is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. The number of components must be determined, either by some clustering algorithm or by automatic speech segmenter. An initial model can be obtained by estimating the parameters from the clustered feature vectors whereas proportions of vectors in each cluster can serve as mixture weights. Means and covariance are estimated from the vectors in each cluster. After the estimation, the feature vectors can be reclustered using component densities (likelihoods) from the estimated mixture model and then model parameters are recalculated. This process is iterated until model parameters converge. This algorithm is called Expectation Maximization (EM) [19]. In the identification phase, speaker with maximum likelihood is selected as the author of a speech sample.

C. Decision Process

The next step after computing of matching scores for every speaker model enrolled in the system is the process of assigning the exact classification mark for the input speech. This process depends on the selected matching and modeling algorithms. In template matching, decision is based on the computed distances, whereas in stochastic matching it is based on the computed probabilities. More details about decision process can be found in [20].

IV. Proposed Algorithm

STE uses the fact that the energy content in the voiced sample is higher than silence/unvoiced sample. However, it is not specific about how much greater it needs to be for proper classification and varies from case to case. In the transitional region some frames may have higher energy due to some noise spike and STE algorithm classified it as a voice where as it should have been categorized as a silence part. ZCR based silence removal techniques used a threshold for the classification of the voiced and unvoiced parts, which itself is subjected to variation across various speakers. However it cannot be varied from speaker to speaker, resulting in improper classification. In silence removal by statistical method, it is assumed that first 3200 samples of the input speech is background noise, where as this condition is not always satisfied. Moreover there can be some cases where some portion of the voiced part would have same distribution as that of the background noise present in the beginning. So, that portion of the voiced part gets wrongly classified as background noise.

In this thesis work we have used both the concept of energy based algorithm and statistical classification of voiced and unvoiced signals. The input speech signal is classified into voiced and noise parts using STE algorithm. Due to improper selection of energy levels, classification at the transition region may not be proper. So in our method six frames are taken, three from each side in the transition region as transitional frames. One side of the transitional frame contains voiced signals and other side contains the unvoiced ones. Bhattacharya distances of transitional frames are calculated from both noisy distribution and voiced signal distribution and the frames belong to the distribution which has less distance. Although the proposed hybrid method leads to higher computational complexities than the traditional ones, still the increase in this processing time is negligible as compared to the overall time required for sp speaker identification where as it leads to tremendous improvement in speaker identification rate.

A. Statistical classification of transitional frames in an energy based voiced and unvoiced classifier

This is a hybrid algorithm which first classifies the input signal using short time energy into smaller parts; each part is either voiced or unvoiced. The transitional frames between voiced and unvoiced part are then classified properly according to their statistical behavior. We have already discussed in section 3.2.3 that background noise has Gaussian distribution and it is characterized by its mean and variance. The speech signal has also normal distribution, so it is also represented in terms of mean and variance. We have to find out whether the signals in the transitional frames have a distribution like that of the neighboring speech signal or neighboring background noise. The algorithm has two parts, e.g. preliminary classification using short term energy (STE) and exact classification of transitional frames using statistical behavior. Silence removal using this algorithm has six steps:

(1) Classify the input signal into smaller segments by using STE algorithm. Each segment is either voiced or unvoiced.
(2) Calculate the mean μ and variance σ of each segment.
(3) Select the transitional frames by taking 3 frames on either side of the transition region between voiced and unvoiced segment.
(4) Calculate the mean μt and variance σt of all transitional frames.
(5) The segment present on the left side of transitional frames is named as left segment with its mean μl and variance σl and the segment present on the right side of transitional frames is named as right segment and its mean μr and variance σr.
(6) Calculate the Bhattacharya distance between left segment and transitional frame and it is denoted by dl (distance left). Analytically.

$$dl = \frac{1}{2}\ln\frac{\sigma_l + \sigma_t}{2(\sigma_l\sigma_t)^{1/2}} + \frac{1}{8}(\mu_l - \mu_t)^2\left(\frac{\sigma_l + \sigma_t}{2}\right)^{-1} \quad (12)$$

(7) Calculate the Bhattacharya distance between right segment and transitional frame and it is denoted by dr (distance right). Analytically.

$$dr = \frac{1}{2}\ln\frac{\sigma_r + \sigma_t}{2(\sigma_r\sigma_t)^{1/2}} + \frac{1}{8}(\mu_r - \mu_t)^2\left(\frac{\sigma_r + \sigma_t}{2}\right)^{-1} \quad (13)$$

(8) Classify the transitional frame on the basis of following rule.
If dl < dr : transitional frame belongs to the left segment. So if the left segment is unvoiced segment it is classify as unvoiced frame else it is voiced.
Else : transitional frame belongs to the right segment. So if the right segment is unvoiced segment it is classify asunvoiced frame else it is voiced.
(9) Eliminate all unvoiced frames from the input speech signal.

B. Speech Database

Selection of an appropriate database plays an important role in the evaluation of any speaker identification system. In this paper we have used TIMIT, NTIMIT and two other noisy databases for closed set speaker identification. These databases are chosen for various reasons. Firstly, the TIMIT database is widely used and publicly accessible, facilitating our need to compare our results with those of others. Noisy databases are used to simulate the results in real world noisy environment. TIMIT (Texas Instrument Massachusetts Institute of Technology) [21] database allows identification to be done under nearly ideal conditions. Hence the errors in speaker recognition resulting from the use of TIMIT database must be from the overlapping of speaker distributions. The TIMIT database consists of 630 speakers, out of which 70% are male and 30% are female from 10 different dialect regions in America. Each speaker has approximately 30 seconds of speech spread over 10 utterances. The speech is recorded using a high quality microphone in a sound proof booth at a sampling frequency of 16 KHz, with no session intervals between recordings. The speech is designed to have a rich phonetic content, which consists of 2 dialect sentences (SA), 450 phonetically compact sentences (SX) and 1890 phonetically diverse sentences (SI). The dialect sentences developed by SRI are spoken by all speakers and were designed to show the variability introduced by the different dialects of the speakers. The potentially compact sentences are designed by MIT and their purpose was to provide a good coverage of phoneme pairs. Each speaker read five of these sentences and each sentence is read by seven speakers. The speakers spoke three phonetically diverse sentences those were directly acquired from existing text sources namely Brow Corpus and the Play Wrights Dialog. NTIMIT database is nothing but the speech in the TIMIT database passing it through local or long distance telephone loop. Through the use of an "artificial mouth", each sentence is directly coupled to a carbon button telephone. The speech is then relayed to a local or long distance central office where it is looped back and recorded. The NTIMIT database can be considered to be TIMIT speech suffering from degradation due to carbon button transducers and actual telephone line conditions. In this work two noisy databases are prepared to measure the robustness of speaker identification system in noisy environment. One of the databases is prepared by the addition of white noise with the clean speech of TIMIT database and the other by the application of channel noise on the same. White noise is dynamically generated using MATLAB. Channel noise is collected by using modem spy through the telephone line. The experiments were conducted on these two generated databases where the SNR is set as 10 dB, 20 dB and 30 dB.

C. Speaker Reorganization System Parameters

All experiments use 24 seconds of speech to train the system. During TIMIT/NIMIT experiments 2 SA sentences, 3 SI sentences and 3 SX sentences are concatenated to produce one 24 seconds utterance containing 8 sentences for each speaker. The remaining two SX sentences are used as two independent tests segments. Two sets of experiment have been done to evaluate the performance of proposed speaker identification system. In the first experiment all 630 speakers (438 males and 192 females) of TIMIT/NIMIT database are used for training as well as testing. In the second set of experiment 200 speakers (112 males and 88 females) are selected alphabetically from the TIMIT database. During training each speaker is trained by clean speech of TIMIT database where as the testing is done individually on TIMIT database contaminated with white noise and channel noise respectively. Having acquired the testing and training utterances, it is now the role of the feature extractor to extract the acoustic features from the speech.

D. Feature Extraction and Parameter Estimation

In this paper, we investigate the use of the MFCC feature set for speaker identification. In chapter 2 we have given a brief description on the extraction of MFCC features from the speech waveform. The MFCC feature extractor converts an utterance into a sequence of MFCC feature vectors. It involves three steps, namely pre emphasis, frame blocking and windowing sections. But during the extraction of testing features from noisy data, an additional silence removal step is inserted for the removal of the background noise. In windowing, the input speech signal cuts into overlapping windows of equal length. Throughout the experiment a Hamming window of 16 ms length with the overlapping of 8 ms is fixed. The spectrum is calculated by using an FFT algorithm and the number of points used in the FFT algorithm is taken as the power of 2 greater than or equal to the frame size. The resulting power spectrum is windowed by a set of 26 triangular filters (mel filters) which are equally spaced apart by 1500 mels and each one having width of 3000 mels. An estimation of the power of each window was done for calculation of the MFCC coefficients. Typical values for the cepstral order used are 12, 16, 24 etc.

E. Speaker Modeling

Each speaker is modeled using one Gaussian Mixture Model (GMM) with 32 mixture components. Each mixture component is characterized by its weight, mean vector and (diagonal) covariance matrix. The GMMs are trained using the EM algorithm with an approximate model ($\lambda 0$) derived by a K-means algorithm. 30 iterations of the EM algorithm were used. During identification phase these models are used to identify the speaker from the given test utterance.

F. Speaker Identification framework

The block diagram of the speaker identification framework is presented in Figure. 1. In this figure, F means Feature vectors; NF means New Feature vectors after PCA transform; W means transform matrix V; and M means trained Model. Dotted line represents the ENROLLMENT phase and solid line represents the IDENTIFICATION phase. The first block is the feature extraction block which is common for both the ENROLLMENT phase and IDENTIFICATION phase. Feature extraction block is meant for the extraction of 24 dimensional MFCC feature vectors from the input speech signal. Feature vectors obtained after are used to train speaker model and the trained model is stored in the database associated with the ID of the new speaker too. When to identify the speaker of an input voice sample, the processes are as follows: First, general feature extraction is made similar to that of the ENROLL process. According to the definition of speaker identification, each speaker enrolled in the enrollment database should be compared with the test utterance to determine the identified speaker. So, after all scores are obtained, the identified speaker is determined by the best scores obtained.

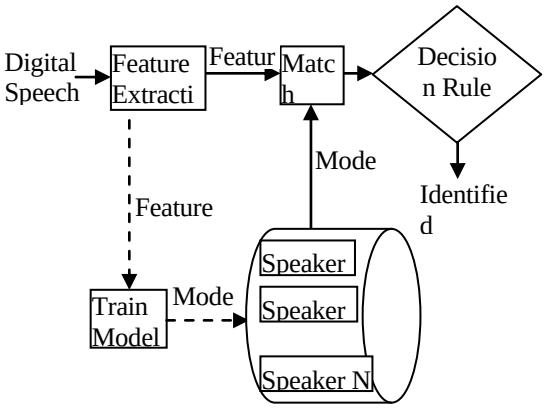

Fig 1: Speaker Identification framework

G. Performance of the Speaker Recognition System

In a speaker identification system, we are ultimately concerned with its ability to identify speakers; the performance of the system is measured using the speaker identification rate. The speaker identification rate can be described as

$$\%\,SIR = \frac{no\,of\ \ correctly\ identified\,segments}{total\,no.of\ \ segments} \times 100\%$$

Where, SIR =Speaker Identification Rate

V. Experiment Simulation and Result Analysis

To verify the performance of the proposed algorithm, we have compared the performance of our algorithm with STE algorithm and silence removal by the statistical behavior of background noise. All the three algorithms are applied on the same sentence taken from TIMIT and NTIMIT database. Figures 2 to 13 show the output speech signals after the preprocessing step using three silence removal algorithms applied on a same phrase ("Don't ask me to carry an oily rag like that" present in the file TIMIT\ test\dr6\fmgd0\sa2.wav), uttered by the same speaker, in two databases: TIMIT (clean speech) and NTIMIT (standard telephone speech), and TIMIT sentence added with white noise in 20db SNR and 10 dB SNR; the white zones indicate speech regions and the grey zones indicate the rejected 'silence' regions.

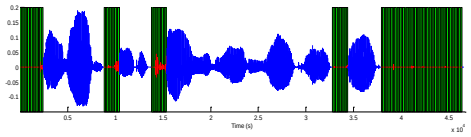

Fig 2: silence removal of TIMIT sentence by STE algorithm

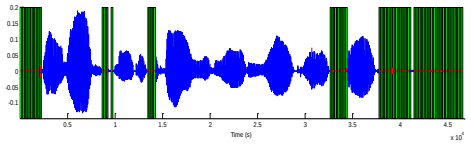

Fig 3: silence removal of TIMIT sentence by Statistical behavior of background noise.

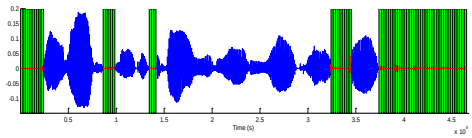

Fig 4: silence removal of TIMIT sentence by proposed algorithm

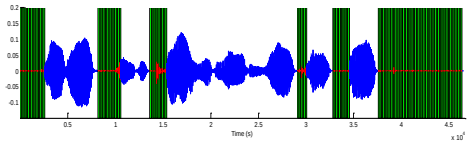

Fig 5: silence removal of NTIMIT sentence by STE algorithm

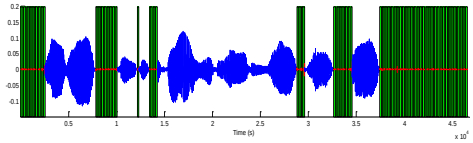

Fig 6: silence removal of NTIMIT sentence by Statistical behavior of background noise

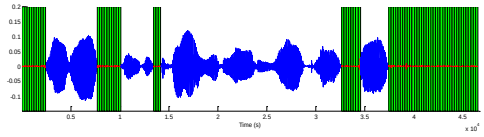

Fig 7: silence removal of NTIMIT sentence by proposed algorithm

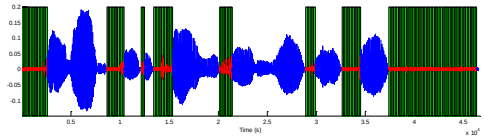

Fig 8: silence removal of TIMIT sentence with 20dB SNR by STE algorithm

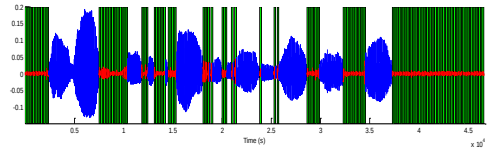

Fig 9: silence removal of TIMIT sentence with 20dB SNR by Statistical behavior of background noise

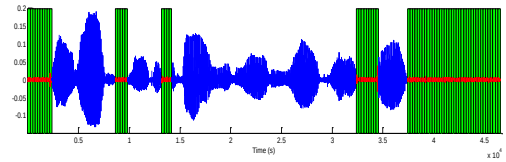

Fig 10: silence removal of TIMIT sentence with 20dB SNR by proposed algorithm

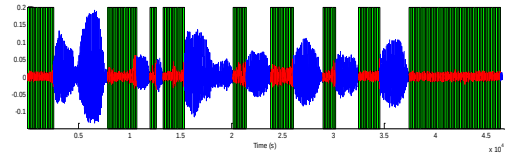

Fig 11: silence removal of TIMIT sentence with 10dB SNR by STE algorithm

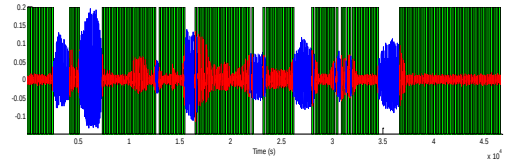

Fig 12: silence removal of TIMIT sentence with 10dB SNR by Statistical behavior of background noise.

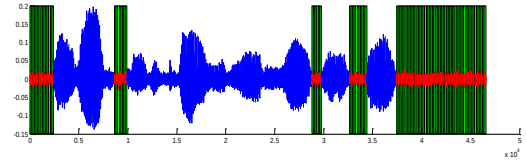

Fig 13: silence removal of TIMIT sentence with 10dB SNR by hybrid algorithm

From the above result we can draw the conclusion that with the increase in signal to noise ratio the effectiveness of proposed algorithm increases. In the clean sentence taken from the TIMIT database the performance of proposed algorithm is nearly same as STE algorithm and statistical method. But with higher SNR the difference is more prominent; when SNR is 10dB both STE algorithm and the statistical method eliminate most of the speech parts as silence where the proposed algorithm eliminates only the unvoiced part. Silence removal in noisy signal is an essential part of any speaker identification system. Because silence/unvoiced part of speech signal is more affected by noise than the voiced parts and the features extracted from these region is prominently due to noise. There is no speech or speaker dependent information in those features, which leads to misclassification. The proposed algorithm can be adopted for any signal to noise ratio. It is also independent of the recognition algorithm and it is proved to be very efficient for a robust speaker identification system.

A. Performance of the Speaker Identification System in Noisy databases

To further investigate the performance of the proposed method, the Speaker Identification system is applied on an artificially made noisy database prepared by using TIMIT database, white Gaussian noise and channel noise. The experiment is conducted with 200 speakers (112 males and 88 females) selected alphabetically from the TIMIT database. The system is tested using 32 mixture components per speaker using 24 MFCC coefficients. The model for each speaker is trained by clean speech of approximately 24 seconds containing 8 sentences (formed by the concatenation of 2 SA sentences, 3 SI sentences and 3 SX sentences). The remaining two SX sentences contaminated with white noise and channel noise are used as two independent tests segments. Testing has been carried out on both the databases in two sets, one with silence removal technique and another without it. The system is tested across three different SNR. The results obtained from the experiment are listed in Table-1 and Table-2.

Table 1: SPEAKER IDENTIFICATION RATES FOR SPEECH DEGRADED BY ADDITIVE WHITE GAUSSIAN NOISE

SNR	30 dB SNR	20 dB SNR	10 dB SNR
without silence removal technique	96.5%	71%	33.5%
with silence removal technique	97.5%	78.5%	42%

Table 2: Speaker identification rates for speech degraded by channel noise

SNR	30 dB SNR	20 dB SNR	10 dB SNR
without silence removal technique	99%	9.5%	40.5%
with silence removal technique	99.5%	81%	47.5%

From the above tables, it is apparent that, with the decrease in SNR the performance of silence removal technique is quite prominent in terms of speaker identification rate. In presence of white noise of 20 dB SNR, the identification rate is 7% better than without silence removal technique, where as with the silence removal technique the improvement raise to 15% at 10 dB SNR. Similarly in the presence of channel noise of 20 dB SNR, the identification rate is 8% better than without silence removal technique, where as with the silence removal technique the improvement raise to 18% at 10 dB SNR. From the above discussions it is apparent that silence removal techniques improve the speaker identification rate.

VI. CONCLUSION

This section summarizes the key issues and results covered in this paper, and a few suggestions are made for possible directions for future research in this area. We have

evaluated the performance of proposed algorithm on the standard TIMIT/NTIMIT database and to emulate the real world noise, white noise and channel noise are injected into the clean speech of TIMIT database. These databases are chosen because of the large amount of continuous speech they contain under a wide variety of conditions. The TIMIT databases are specially chosen due to their wide use and availability, serving as a means to compare our results with those of others.

Silence removal plays a key role in feature extraction in noisy environment for robust speaker identification. The information which is more important from the prospective of speaker identification is generally contained inside the voice part of the speech signal. Therefore the process of isolating the redundant information especially in the unvoiced part in preprocessing step bears a lot of importance. A hybrid algorithm is proposed to remove the silence part from the speech signal. Proposed algorithm is a hybrid form of energy based method and the statistical method and it overcome the limitations of both the existing method. Although the proposed hybrid method leads to a higher computational complexity than the traditional ones, still the increase in this processing time is negligible as compared to the overall time required for speaker identification, where as it leads to a tremendous improvement in speaker identification rate. In low SNR conditions both the energy based method and the statistical method fails to properly classify the silence part and the voiced part. The existing methods eliminate most of the speech parts as silence, where as the proposed algorithm eliminates the silence part only. Speaker Identification rate is improved by 15 % to 20 % due to the application of proposed silence removal technique in a conventional GMM based classifier using MFCC feature vectors.

Our main purpose was to develop and test a robust speech detection algorithm for a speaker recognition system having the following distinctive features:

(1) Possibility of use in any speaker recognition system and therefore independent of the recognition algorithm;
(2) Adaptability on the background noise level;
(3) Accuracy "as high as possible" of the detected speech boundaries; this means to include all significant acoustic events within the detected speech segments (and then to eliminate the non-speech regions of an utterance);
(4) Possibility of detecting and rejecting some typical background noises: isolated short external noises (including human artifacts), or very low frequency noise (for example 50/60 Hz hum);
(5) Reasonable complexity and consequently operation in real-time on a previously acquired utterance.
(6) Considering these fundamental requirements, we designed and implemented a novel explicit energy-based speech detection algorithm. Tested on the available TIMIT and NTIMIT databases and also using a particular speaker recognition program, it proved to be an accurate, adaptive and fast speech detection algorithm. As it was mentioned, the algorithm was designed for an off-line processing; nevertheless, it could be easily changed for an on-line version, with only a few minor modifications.

REFERENCES

[1] L. Lamel, L. Rabiner, A.E. Rosenberg, J.G. Wilpon, "improved endpoint detector for isolated word recognition", IEEE Transactions on Acoustics, Speech and Signal Processing, Volume:29, Issue: 4, Aug, 1981.
[2] Sen Zhang, Graduate Sch., Chinese Acad. of Sci., Beijing, "an energy-based adaptive voice detection approach", 8th International Conference on Signal Processing , Volume: 1, 2006.
[3] M. Liscombe, A. Asif, "A new method for instantaneous signal period identification by repetitive pattern matching", Multitopic Conference, INMIC 2009. IEEE 13th International, Publication Year: 2009 , Page(s): 1-5.
[4] Deisher, E. Michael, A. S. Spanias, "HMM-based speech enhancement using harmonic modeling", IEEE International Conference on Acoustics, Speech, and Signal Processing, 1997. ICASSP-97, 1997 , Volume: 2, Page(s): 117 -1178.
[5] A. Hussain, S.A. Samad, Liew Ban Fah, "Endpoint detection of speech signal using neural network", TENCON 2000. Proceedings, Volume: 1, Page(s): 271-274
[6] J. Ramirez, J.C. Segura, J.M. Gorriz, L. Garcia, "Improved Voice Activity Detection Using Contextual Multiple Hypothesis Testing for Robust Speech Recognition", IEEE Transactions on Audio, Speech, and Language Processing, Volume: 15 , Publication Year: 2007 , Page(s): 2177- 2189.
[7] Dong Enqing, Liu Guizhong, Zhou Yatong, Cai Yu, "Voice activity detection based on short-time energy and noise spectrum adaptation", 6th International Conference on Signal Processing, 2002, Volume: 1, Publication Year: 2002, Page(s): 464-467.
[8] D. G. Childers, M. Hand, J. M. Larar, "Silent and Voiced/Unvoied/Mixed Excitation(Four-Way), Classification of Speech", IEEE Transaction on ASSP, Vol-37, No-11, pp. 1771-74, Nov 1989.
[9] Dragos Burileanu1, Lucian Pascalin1, Corneliu Burileanu1 and Mihai Puchiu, "An Adaptive and Fast Speech Detection Algorithm", Proceedings of the Third International Workshop on Text, Speech and Dialogue, 2000, Vol. 1902, pp. 177-182.
[10] G. Saha, Sandipan Chakroborty, Suman Senapat , "A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications", Proceedings of the NCC 2005, Jan. 2005.
[11] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition", in Proc. ICASSP2002, vol. 4, 2002, pp. 3808–3811.
[12] R.B. Blazek, Wei-Tyng Hong, "Robust Hierarchical Linear Model Comparison for End-of-Utterance Detection under Noisy Environments", International Symposium on Biometrics and Security Technologies (ISBAST), 2012.
[13] M.G. Sumithra, A.K. Devika, "A study on feature extraction techniques for text independent speaker identification", International Conference on Computer Communication and Informatics (ICCCI), 2012, Page(s): 1-5.
[14] Shahzadi Farah, Azra Shamim, "Speaker recognition system using mel-frequency cepstrum coefficients, linear prediction coding and vector quantization", International Conference on Computer,Control & Communication (IC4), 2013, Page(s): 1-5.

[15] H. Ezzaidi, Jean Rouat, "Pitch and MFCC dependent GMM models for speaker identification systems", Canadian Conference on Electrical and Computer Engineering, 2004., Volume: 1, Page(s): 43-46.
[16] D.A. Reynolds, R.C. Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models", IEEE Transactions on Speech and Audio Processing, Volume: 3, Issue: 1 Publication Year: 1995 , Page(s): 72-83.
[17] Chee-Ming Ting, S.H. Salleh, Tian-Swee Tan A.K. Ariff, "Text independent Speaker Identification using Gaussian mixture model", International Conference on Intelligent and Advanced Systems, ICIAS 2007 , Page(s): 194-198.
[18] Abdul Manan Ahmad, Loh Mun Yee, "Vector quantization decision function for Gaussian Mixture Model based speaker identification", International Symposium on Intelligent Signal Processing and Communications Systems, 2008, ISPACS 2008., Page(s): 1-4.
[19] T.F. Covoes, E.R. Hruschka, "Unsupervised learning of Gaussian Mixture Models: Evolutionary Create and Eliminate for Expectation Maximization algorithm", IEEE Congress on Evolutionary Computation (CEC), 2013, Page(s): 3206 – 3213.
[20] D. A. Reynolds, "An overview of automatic speaker recognition technology", ICASSP, pp. 4072-4075, 2002.
[21] J.S. Garofolo, et. al., .DARPA TIMIT: Acoustic-Phonetic Continuous Speech Corpus, New Jersey: NIST Publications, 1993.

Tushar Ranjan Sahoo, male, works in IIIT Bhubaneswar, India, since 2011 August. He received his Bachelor degree in Computer Science in the year 2004 from PIET Rourkela and Masters in same discipline in 2010 from IIT Kharagpur. His research interest includes: Fault-tolerant Analysis for Embedded Controls in Sefty-critical Applications, Formal Verification for Digital VLSI, CUDA for General Purpose GPU Programming, Speaker Identification & Verification.

Sabyasachi Patra, male, was born in 9th March, 1983. He received his B Tech degree in Instrumentation & Electronics Engineering in 2005 and M.Sc (Engg) in Super Computer Education and Research Center in 2008 from IISc Bangalore. He has joined IIIT Bhubaneswar as Assistant Professor since 22.08.2011. Prior to this he was an assistant professor in School of Computer Engineering, KIIT University. His research areas of interest are Information Security, Image Processing and Speech Processing.

