IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering

ISO 3297:2007 Certified

Vol. 6, Issue 5, May 2017

Estimation of Noise Power Spectrum and Automatic Speaker Recognition System

Rajalakshmi. P[1], Anju. L[2]

2nd Year, M.E (Applied Electronics), Electronics and Communication Engineering, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India[1]

Assistant Professor, Electronics and Communication Engineering, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India[2]

Abstract: Unseen noise estimation is one of the challenging steps to make the speech enhancement algorithm work in adverse conditions. The prior knowledge known about the encountered noise is that it is different from the involved speech.The proposed work consists of two segments, Speech Enhancement and Automatic Speaker Recognition (ASR) system. The speech enhancement comprises of offline training and online enhancement processes. In offline training, the inputs clean speech data and noisy speech magnitude are collected and trained using Support Vector Machine (SVM). In online enhancement, the trained signals are compared and their noise spectrum is estimated using the Modified Spectral Subtraction (MSS) method which is also used for the removal of noises. Then the enhanced speech signal is obtained by transforming the estimated spectrum into time domain. The features are extracted from the obtained enhanced speech using Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). Finally the speaker recognition is done using k-NN and Gaussian Mixture Model (GMM) based Multi-SVM. The experimental results are compared and efficient system is obtained.

Keywords: Speech enhancement, SVM, MSS, k-NN, GMM and Multi-SVM.

I. INTRODUCTION

Speech Enhancement plays a vital role in improving the quality of speech and intelligibility of the noisy speech signal degraded in adverse environments. This paper is mainly aimed at the removal of noises along with speech enhancement and finally Automatic Speaker Recognition system is obtained. Speech Enhancement is aimed to improve the overall perceptual quality of degraded speech signal by using audio signal processing techniques. It is mainly used in many applications such as mobile phones, VoIP, teleconferencing systems, speech recognition, and hearing aids, etc., Thespeech enhancement is having the problem of separatingthe speech and noise, for which a commonly employed technique is estimation and removal of the noise spectrum from the input noisy speech spectrum. Another problem in the enhancement of speech is that mosttypeof noises is non-stationary. Its spectral properties are very difficult to predict which makes removal of noise, a challenge.

Speaker recognition is a process which enables the machines to understand,explain and verify the authenticity of a speaker with the help of a database.Speaker recognition is essentially a method of automatically identifying a speaker from a recorded or a live speech signal by analyzing the speech signal parameters.

The main goal of automatic speaker recognition systems is to extract the information, describe and recognize it in the speech conveying the speaker identity.The speaker recognitionsystems are commonly used in many areas like Access control, Law enforcement, Transaction authentication, speech data management and Personalization.

The only assumption taken hereis that the noise is different from the involved speech signal. A supervised learning algorithm called SVM is used for modeling the clean speech spectrum. An algorithm called MSS is used for the estimation of the noise power spectrum, removing the noises present and reconstructing the original clean speech signal.Then the feature extraction of the enhanced speech signals is done by usingthe two methods,MFCC and PLP and finally classification/identification of speakers by using k-NN and Multi-SVM algorithms.

The remainder of this paper is organized as follows: The methodologies used are described in Section II. The metrics for evaluation are given in Section III. Experimental results and discussion are done in Section IV. Finally, Section V represents the conclusion.

II.METHODOLOGY

The proposed system consists of two segments, Speech Enhancement and the ASR. Each segment comprises of two parts;speech enhancement consists of offline training and online enhancement parts. The ASR consists of Feature

DOI10.17148/IJARCCE.2017.6575

Extraction and Classifier parts. In offline training, the clean speech data and the noisy speech magnitude are collected and then trained using an algorithm named SVM. In online enhancement, these signals are first compared and then the noise is estimated and removed using MSS method. The spectrum is reconstructed in time domain.

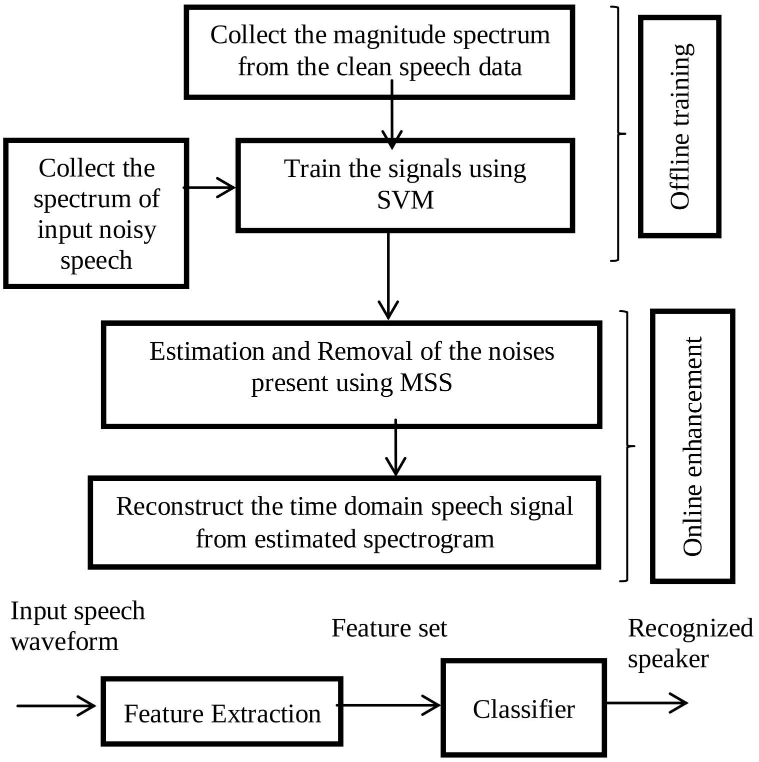
The methods for speech enhancement of the signals include the removal of background noise and echo suppression. When the background noise is suppressed, it is crucial not to harm the speech signal.Background noise suppression has many applications.

Thespeech signal is enhanced mainly for coding and recognition purposes. The codes have been optimized for speech and they usually make the background noise sound weird. Moreover, enhanced speech can be compressed in fewer bits than non-enhanced. Most of the recognition systems whose operation relies on the features extracted from speech will be disturbed by extra noise sounds.

The speech enhancement methods are mainly aimed at the termination of the background noise which is naturally based on the estimate of it. Several speech enhancement methods were developed over the past several years. Spectral subtraction is one of the methods which subtractthe noise spectrum to produce a spectrum of the clean speech.

Feature Extraction deals with an initial set of measured data and builds derived values. Feature extraction is related to dimensionality reduction. When the input data to an algorithm is too large to be processed then it can be transformed into a reduced set of features. Classification is another important part of speaker recognition system where the datasets are classified into different classes. During this stage, the decisions are taken based on the similarity measures from training sets. In classification, the data sets are separated into train and test sets for easy access and validation.

a. Block for Speech Enhancement


Collect the magnitude spectrum from the clean speech data
Collect the spectrum of input noisy speech
Train the signals using SVM
Offline training
Estimation and Removal of the noises present using MSS
Reconstruct the time domain speech signal from estimated spectrogram
Online enhancement
Input speech waveform
Feature set
Recognized speaker
Feature Extraction
Classifier

b. Block for ASR system

Figure 1 Block Diagram for the Proposed Method

A. Support Vector Machine(SVM)

SVM are supervised machine learning algorithms that analyze data for classification and regression analysis. The goal of SVM is to decide which class is to be selected for a new data point. SVM algorithm builds a model which assigns new data points into one or other category. SVM can only be segregated into two classes.

SVM can be used for both classification and regression challenges. However, it is mostly used in classification problems. In this algorithm, each data item is plotted as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, the classification is done by finding the hyper-plane that differentiates the two classes very well. The decision function is fully specified by a subset of training samples, i.e., the support vectors. Support vectors are the data points that lie closest to the decision surface. They have direct bearing on the optimum location of the decision surface.

The goal of the SVM is to train a model that assigns new unseen objects into a particular category. It achieves this by creating a linear partition of the feature space into two categories. Based on the features in the new unseen objects (e.g. documents/emails), it places an object "above" or "below" the separation plane, leading to a categorization. However, much of the benefit of SVMs comes from the fact that they are not restricted to being linear classifiers.

B. Modified Spectral Subtraction(MSS)

Spectral subtraction (SS) is based on the principle that an estimate of the clean signal spectrum is obtained by subtracting an estimate of the noise spectrum from the noisy speech spectrum. The spectral subtraction method is a simple and effective method of noise estimation and removal. In this method, signal spectrum and noise spectrum are estimated in parts of the recording and subtracted from each other, so that average signal-to-noise ratio (SNR) is improved. It is assumed that the signal is deformed by a wide-band, stationary, and additive noise. The noise estimate and the phase aresame during the analysis and the restoration.

$$x(m) = y(m) - d(m) \qquad (1)$$

where y(m) – noisy speech , x(m) – speech signal and d(m) – noise

The noisy speech is first divided intooverlapping frames. Then the Hamming window is applied on each frame and a set of Fourier coefficients using short-time fast Fourier transform has been generated. Noise spectrum is obtained during periods when there is no speech in the input signal. Voice Activity Detector (VAD) identifies the no speech segment which produces a control signal that permits the updating of store with spectrum when the speech is absent. This spectrum is smoothed and then used to update a spectral noise estimate, which consists of a portion of the previous and current noise segment. Thus this spectrum transforms to the changes in the actual noise spectrum. After noise estimation and removal, the root of the output provides the corresponding Fourier Amplitude. The time-domain signals are reconstructed by an inverse Fourier transform. Thus the speech segments are overlapped toprovide the reconstructed time domain output signal.

C. Mel Frequency Cepstral Coefficients (MFCC)

Mel Frequency Cepstral Coefficients (MFCCs) are widely used in automatic speech and speaker recognition systems. MFCC's are cepstral coefficients computed on a wrapped frequency scale based on known human auditory perception. It is a nonparametric frequency domain approach which is based on human auditory perception system.

The first step in MFCC feature extraction is to boost the amount of energy in the high frequencies. Then windowing of the speech is done. Most commonly used window is hamming window. The next step is to extract spectral information for the windowed signal. This is done by using FFT or DFT. The next step is the filter-bank processing. Finally DCT is applied to produce highly uncorrelated features.

The MFCC's are so popular because it is efficient to compute, it incorporates a perceptual Mel frequency scale and it also separates the source and the filter.

D. Perceptual Linear Prediction (PLP)

PLP is modeled based on the psychophysics of hearing. It discards the irrelevant information of the speech and improves the recognition rate. PLP is similar to LPC except its spectral characteristics which have been transformed to match the characteristics of human auditory system.

PLP features are robust when there is an acoustic mismatch between training and test data sets. PLP consists of the following steps, first the power spectrum is computed from the windowed speech signal. Then the frequency warping, smoothing and sampling are integrated into a single filter-bank named Bark filter-bank and it is carried out. The resulting spectrum is then processed by linear prediction (LP). Applying LP to the auditorily warped line spectrum means that it is able to compute the predictor coefficients of a signal that has this warped spectrum as a power spectrum. Finally, the cepstral coefficients are obtained from the predictor coefficients by a recursion followed by an inverse Fourier transform.

E. k-NNAlgorithm

The k-Nearest Neighbor algorithm (k-NN) is a non parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. In k-NN classification, the output is a class membership.

K-nearest neighbors uses the local neighborhood to obtain a prediction. The parameters of the algorithm are the number k of neighbors and the procedure for combining the predictions of the k examples. k-Nearest Neighbor is an example of instance-based learning, in which the training data set is stored, so that a classification for a new unclassified record may be found simply by comparing it to the most similar records in the training set.

DOI10.17148/IJARCCE.2017.6575

The purpose of the k Nearest Neighbor algorithm is to use a database in which the data points are separated into several separate classes to predict the classification of a new sample point.It gives the maximum likelihood estimation of the class posterior probabilities.

F. Gaussian Mixture Model (GMM)

GMM is used for generating a fixed set of features for each set. GMM is adapted from a Universal Background Model (UBM) and is currently the most popular approach for speaker verification. A GMM with several mixture components is used to model the speech signal characteristics for each speaker. The model consists of trained datasets. GMMs can be trained by maximum likelihood using an efficient algorithm. The likelihood is given by mean and covariance matrix.

G. Multi-SVM

The Multi-class SVM is a discriminative method, modeling the boundaries directly between different classes in some feature space rather than by the difficult intermediate step of estimating class densities. For the task of speaker recognition, four multi-class SVMsmethods were designed.They are the All-at-once, One-against-all, One-against-one, and the Directed Acyclic Graph SVM (DAGSVM).

The One-against-One method is performed the best, achieving a high accuracy for multi-speaker speech.It performs well for the real-time applications. In this method, one SVM is constructed for each pair of classes. SVMs are trained to distinguish the samples from one class to another class. Here each training point belongs to one of N different classes. The goal is to construct a function which when given a new data point or set, will correctly predict the class to which it belongs.

III. METRICS FOR EVALUATION

Three metrics were computed to evaluate the performance of the algorithms.

A. PESQ Score

PESQ stands for Perceptual Evaluation of Speech Quality. PESQ is the popular ITU-T standard for the measurement of the quality of voice in the communication networks. It measures the subjective quality of the speech. It is calculated by the comparing the output speech with the input reference speech. The value of this score ranges from -0.5 to 4.5. It analyses the speech signal in sample by sample manner. It provides numerical measure of the quality of human speech. PESQ can be widely used in many applications since it is fast and repeatable.

B. Peak SNR (PSNR)

It refers to Peak Signal-to-noise ratio. It is the ratio between the maximum possible powers of the speech signal to the power of corrupting noise. Generally it is expressed in terms of the logarithmic decibel scale. It is the most easily defined parameter via the mean squared error (MSE). Lower the error, higher will be the PSNR. The PSNR value obtaining 40 dB or more refers to good quality of the signal.

$$\mathrm{PSNR} = 20 \log 10\,(\mathrm{MAX}_j) - 10 \log 10\,(\mathrm{MSE}) \qquad (2)$$

WhereMSE – mean squared error and

MAX_j– max possible value of the signal

$$\mathrm{MSE} = \frac{1}{M}\sum_{j=0}^{M}(xj - yj)^2 \qquad (3)$$

Where xjand yj are the original and noisy signals and M – no of signal samples

C. Identification Rate

The average identification rate for the data sets are computed as,

$$\%\ \text{Identification Rate} = M_C / M_T\% \qquad (4)$$

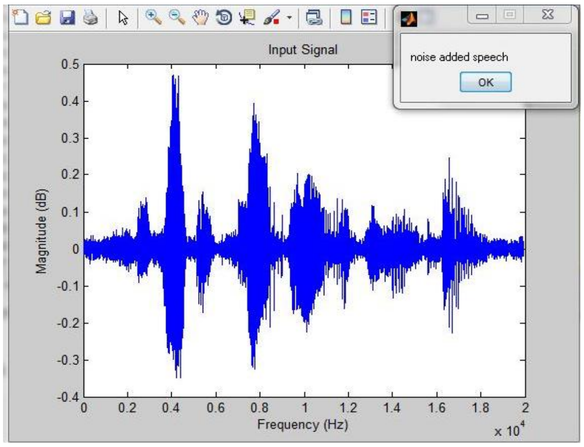
Where M_C is the number of correctly identified sets andM_T the total number of sets

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were carried out on the MATLAB. Two databases were used – noizeus and super sededdatabases. The sample sets were experimented under both supervised and unsupervised conditions for speech enhancement process. For the ASR system, the sample sets were given as trained and test sets.

DOI10.17148/IJARCCE.2017.6575

A. Offline Training Results



noise added speech

Figure 2 Output of offline training process

The above figure 2 shows the spectrum of an input speech of the offline training process. It shows the training of the input speech which is given as input to the system of speech enhancement.

B. Online Enhancement Results

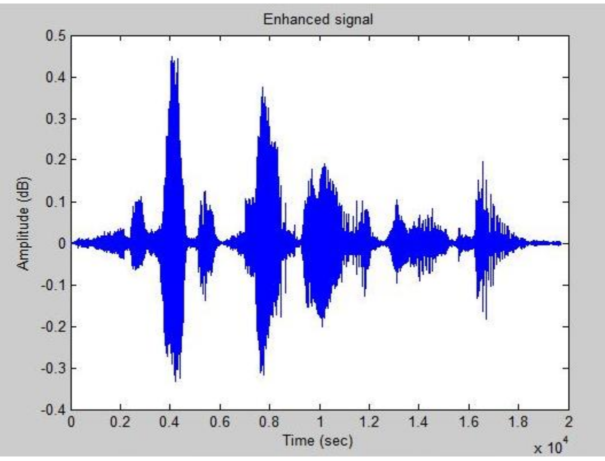
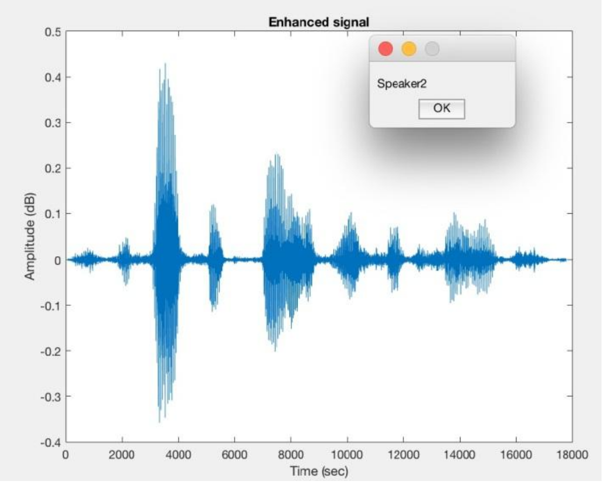


Figure 3 Enhanced output

The above figure 3 shows the enhanced output of the process. It is the reconstructed time domain signal after noise estimation and noise removal.

C. k-NN Classification Results



Speaker2

Figure 4 k-NN Speaker Identification

DOI10.17148/IJARCCE.2017.6575

The above figure 4 shows the spectrum of the identified speaker using k-NN algorithm. The train sets of data are correctly identified. The test sets are mostly identified.

D. Multi-SVM Classification Results

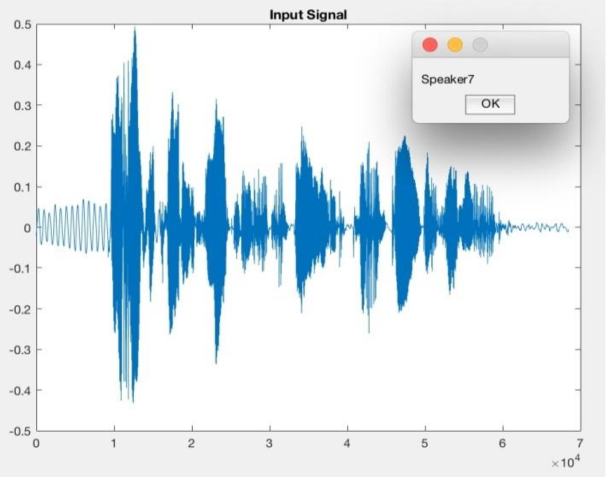


Figure 5 Multi -SVM Speaker Identification

The above figure 5 shows the spectrum of the identified speakerusing GMM based Multi-SVM. The train and test sets of data are correctly identified.

TABLE I Comparison of PSNR Values for Super seded Database

Noises	**PSNR(dB) of Supervised samples**	**PSNR(dB) of Unsupervised samples**
Noise 1	83.396	80.594
Noise 2	86.641	81.375
Noise 3	88.560	88.390
Noise 4	89.308	74.816
Noise 5	81.529	74.639
Noise 6	81.813	76.739
Noise 7	88.802	83.137
Noise 8	80.450	79.492

The above Table I shows the comparison of the obtained PSNR values of the supervised and unsupervised samples of the Super seded database. It is inferred that the performance of the latter is quite close to that of the first. Hence enhancement is efficiently carried out.

TABLE II Comparison of Identification Rates of k-NN and Multi-SVM algorithms

Speakers	**Identification Rate for Test Sets using k-NN**	**Identification Rate for Test Sets using GMM**
Speaker 1	55%	70%
Speaker 2	70%	85%
Speaker 3	50%	75%
Speaker 4	65%	80%
Speaker 5	75%	85%
Speaker 6	55%	75%
Speaker 7	45%	70%
Speaker 8	70%	85%
Speaker 9	65%	80%
Speaker 10	55%	75%

The above Table II shows the comparison of the obtained Identification Rates of the test sets of both the methods. It is inferred that the performance of the latter is efficient than that of the first. Hence the speaker identification is efficiently carried out.

DOI10.17148/IJARCCE.2017.6575

V.CONCLUSION

Modified spectral subtraction method was proposed for the estimation and removal of noise for speech enhancement. Initially the input signals are trained using SVM. Then estimation and removal of noises are done using MSS. When compared to previous methods, this method copes up with both stationary and non-stationary noises. Both supervised and unsupervised learning methods were investigated. Experimental evaluation on PESQ score and PSNR on the two databases i.e., Noizeus and Super seded are demonstrated. The results of unsupervised samples are similar to the supervised one. This method is highly efficient for learning real world datasets. The noises are reduced without affecting the signal power and the SNR is improved.

Then theAutomatic Speaker Recognition system is carried out.First the features are extracted from the sample data sets using MFCC and PLP. Then based on the features it is classified. The classification or identification of the speakers is done using k-NN and Multi-SVM algorithms. The results were obtained and efficiency is calculated.The efficiency of the latter is higher than that of the earlier method. Thus .ASR system is constructed and verified.

REFERENCES

1. Meng Sun, Xiongwei Zhang, Hugo Van hamme, and Thomas Fang Zheng, "Unseen Noise Estimation Using Separable Deep Auto Encoder for Speech Enhancement," IEEE/ACM Transactions on Audio , Speech and Language Processing, Vol 24, no 1, pp 93-104, January 2016.
2. Furong Yan, Aidong Men, Bo Yang, and Zhuqing Jiang, "An Improved Ranking-Based Feature Enhancement Approach for Robust Speaker Recognition," IEEE Access, Vol 4, pp 5258-5267, 2016.
3. Jun Du, YanhuiTu, Li-Rong Dai, and Chin-Hui Lee, "A Regression Approach to Single-Channel SpeechSeparation Via High-Resolution Deep NeuralNetworks," IEEE Transactions On Audio, Speech And Language Processing, 2016.
4. Niko Moritz, JörnAnemüller, and BirgerKollmeier,"An Auditory Inspired Amplitude Modulation Filter Bank for Robust Feature Extraction in Automatic Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language processing, 2015.
5. Sarika S. Admuthe and ShubhadaGhugardare, "Survey Paper on Automatic Speaker Recognition Systems," International Journal of Engineering And Computer Science, Vol 4, Issue 3, pp 10895-10898, March 2015.
6. Y.Xu, J.Du, L-R. Dai, and C-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio , Speech and Language Processing, Vol 23,no 1, pp 7-19, January 2015.
7. Sujay D. Mainkar," Performance comparison of EMD based noise classification for different SNR using GMM and k-NN classifiers," International Journal of Emerging Technology and Advanced Engineering,2015.
8. N.Mohammadiha, P.smaragdis, and A.Leijon, "Supervised and unsupervised speech enhancement using non negative matrix factorisation," IEEE/ACM Transactions on Audio , Speech and Language Processing, Vol 21,no 10, pp 2140-2151, October 2014.
9. Karam M., Khazaal H.F., Aglan H. and Cole C., "Noise Removal in Speech Processing Using Spectral Subtraction," Journal of Signal and Information Processing, Vol 5, pp. 32-41,2014.
10. MdSahidullahand GoutamSaha, "A Novel Windowing Technique for EfficientComputation of MFCC for Speaker Recognition," IEEE Signal Processing Letters, Vol. 20, No. 2, February 2013.
11. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising auto encoder," in Proc. INTERSPEECH, 2013, pp.436–440.
12. Z. Chen and D. P. Ellis, "Speech enhancement by sparse, low-rank, and dictionary spectrogram decomposition," in Proc. IEEE Workshop Application Signal Process. Audio Acoust., 2013, pp. 1–4.
13. Alexey Ozerov, Mathieu Lagrange and Emmanuel Vincent,"GMM-based classification from noisy features," The Journal of Systems and Software, Elsevier, 2013.
14. Ekaterina Verteletskaya, and Boris Simak, "Noise Reduction Based on Modified Spectral Subtraction Method," IAENG International Journal of Computer Science, Vol 38, pp. 231-239,2011.
15. J. Bai and M. Brookes, "Adaptive hidden Markov models for noise modelling," in Proc. 19th Eur. Signal Process. Conf. (EUSIPCO'11),Aug. 2011, pp. 494–499.
16. K. Paliwal, K. Wjcicki, and B. Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," Speech Commun., vol. 52, no. 5, pp. 450–475, 2010.
17. Vijendra Raj Apsingekar and Phillip L. De Leon, "Support Vector Machine based Speaker Identification Systems usingGMM Parameters," IEEE, 2009.
18. Enrico Marchetto, Federico Avanzini, and Federico Flego, "An Automatic Speaker Recognition System for Intelligence Applications," 17th European Signal Processing Conference (EUSIPCO 2009), pp 1612-1616, August 2009.
19. D.Y.Zhao, W.B. Kleijn, A.Ypma, and B.de Vries, "Online noise estimation using stochastic-gain HMM for speech enhancement," IEEE/ACM Transactions on Audio , Speech and Language Processing, Vol 16,no 4, pp 835-846, May 2008.
20. Yang Lu and Philipos C. Loizou, "A geometric approach to spectral subtraction," in speech communication,2008.
21. S. Srinivasan, J. Samuelsson, and W. B. Kleijn, "Codebook driven short-term predictor parameter estimation for speech enhancement," IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 1, pp. 163–176,January 2006.
22. P.Smaragdis, "Non-negative matrix factor deconvolution: Extraction of multiple sound sources from monophonic inputs,"5thInt.Conf. Ind. Compon. Anal., September 2004,pp 494-499.
23. AnujaChougule and V. V. Patil , " Survey of Noise Estimation Algorithms for Speech Enhancement Using Spectral Subtraction," in International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2, Issue: 12, pp. 157-168,2004.
24. K. Lebart, and J. M. Boucher, "A New method based on spectral subtraction for speech enhancement," Acustica, Vol. 87, pp. 359-366,2001.
25. Y. Ephraim, "A signal subspace approach for speech enhancement," IEEE Transactions on speech and audio processing, pp. 251-266,1995.