ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 3, April 2017

Feature Extraction and Speaker Identification in Automatic Speaker Recognition System

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

PG Student [APE], Dept. of ECE, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India [1]

Assistant Professor, Dept. of ECE, Sri Venkateswara College of Engineering, Chennai, Tamilnadu, India [2]

ABSTRACT: Although the speaker recognition has been extensively studied, the efficiency for the test data sets has remained a challenge. Thus, this paper is proposed in order to improve the efficiency. The proposed method for Automatic Speaker Recognition (ASR) system consists of two parts, Feature Extraction and Classifier/Identification. The feature extraction is done using Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). The classifier part is done using k-Nearest Neighbor (k-NN) algorithm and Gaussian Mixture Model (GMM). The experimental results show the improved efficiency. The performances are analysed and the efficient ASR system is obtained.

KEYWORDS: ASR, MFCC, PLP, k-NN and GMM

I. INTRODUCTION

Speaker Recognition is the task of establishing identity of an individual based on his/her voice. Speaker Recognition is a process that enables machines to interpret, understand and verifies the authenticity of a speaker with the help of a database. Speaker Recognition or Identification is essentially a method of automatically identifying a speaker from a recorded or a live speech signal by analysing the speech signal parameters.

The goal of Automatic Speaker Recognition systems is to extract, characterize and recognize the information in the speech signal conveying speaker identity. The areas where speaker recognition technologies are used are Access Control, Law Enforcement, Transaction Authentication, Speech Management and Personalization. It has a significant potential as a convenient biometric method for telephony applications.

For noisy datasets degraded in adverse conditions, the noise is estimated and removed and speech enhancement is done to improve the speech quality and intelligibility. Modified Spectral Subtraction (MSS) is used for the removal of noise and enhancement of the speech.

II. METHODOLOGY

The proposed system consists of two parts, feature extraction and classifier. The feature extraction is done by using Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction (PLP). The classifier is done using k-NN algorithm and GMM.

The central methods for enhancing speech are the removal of background noise. This is done using the Modified Spectral Subtraction (MSS) method where noise is estimated, removed and finally speech is enhanced. For coding and recognition purposes, speech enhancement is done. Moreover, enhanced speech can be compressed in fewer bits than non-enhanced.

Feature Extraction starts from an initial set of measured data and builds derived values (features) intended to be informative and non-redundant, facilitating the subsequent learning and generalization steps, and in some cases leading to better human interpretations. Feature extraction is related to dimensionality reduction.When the input data to an algorithm is too large to be processed and it is suspected to be redundant, then it can be transformed into a reduced set of features. Classification is another important part of speaker recognition system since the datasets are classified into

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 3, April 2017

different classes. During this stage, the decisions are made using the similarity measures from training sets. In almost all classification methods, the data is separated into train and test sets.

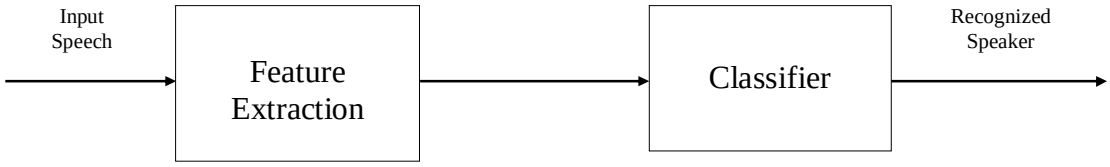


Figure 1 Block Diagram for the Proposed Method

A. *Modified Spectral Subtraction (MSS)*

Spectral subtraction (SS) is based on the principle that one can obtain an estimate of the clean signal spectrum 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 reduction.

$$X(n) = Y(n) - D(n) \qquad (1)$$

where

Y(n) – noisy speech , X(n) – speech signal and D(n) – noise

The noisy speech is segmented into overlapping frames. Then Hamming window is applied on each segment and a set of Fourier coefficients using short-time fast Fourier transform is generated. Noise spectrum is estimated during periods when no speech is present in the input signal. This condition is recognized by Voice Activity Detector (VAD) to produce a control signal which permits the updating of store with spectrum when speech is absent from the current segment. This spectrum is smoothed by making each frequency samples of the average of adjacent frequency samples. This smoothed spectrum then will be used to update a spectral estimate of noise, which consists of a proportion of the previous noise and a portion of the smoothed short-term spectrum of current segment. Thus the noise spectrum gradually adapts to changes in the actual spectrum noise. After noise estimation and subtraction , the a root of the output terms is taken to provide corresponding Fourier amplitude components and the time-domain signal segments reconstructed by an inverse Fourier transform unit from these along with phase components φ directly from the FFT unit. The windowed speech segments are overlapped to provide the reconstructed output signal at an output.

B. *Mel Frequency Cepstral Coefficients (MFCC)*

Mel Frequency Cepstral Coefficients (MFCCs) are a feature widely used in automatic speech and speaker recognition. 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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 3, April 2017

C. *Perceptual Linear Prediction (PLP)*

PLP models the human speech based on the concept of psychophysics of hearing. PLP discards irrelevant information of the speech and thus improves recognition rate. PLP is identical to LPC except that its spectral characteristics have been transformed to match characteristics of human auditory system.

PLP features are reported to be more robust when there is an acoustic mismatch between training and test data. PLP consists of the following steps, first the power spectrum is computed from the windowed speech signal. Then the three steps frequency warping, smoothing and sampling are integrated into a single filter-bank called Bark filter-bank is done. The resulting auditorily warped line spectrum is further processed by linear prediction (LP). Precisely speaking, applying LP to the auditorily warped line spectrum means that we compute the predictor coefficients of a (hypothetical) signal that has this warped spectrum as a power spectrum. Finally, cepstral coefficients are obtained from the predictor coefficients by a recursion that is equivalent to the logarithm of the model spectrum followed by an inverse Fourier transform.

D. *k-NN ALGORITHM*

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. The k-NN algorithm is among the simplest of all machine learning algorithms. In k-NN classification, the output is a class membership.

K-nearest neighbors uses the local neighborhood to obtain a prediction. The K memorized examples more similar to the one that is being classified are retrieved. 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.

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 is memory-based, no explicit training or model is required, hence called as "lazylearning". In its basic form one of the most simple machine learningmethods used commonly. It gives the maximum likelihood estimation of the class posterior probabilities. It can be used as a baseline method as it is having many extensions.

III. METRICS FOR EVALUATION

Two metrics were computed to evaluate the performance of the speech enhancement algorithms.

A. *PESQ Score*

PESQ stands for Perceptual Evaluation of Speech Quality. PESQ is the new ITU-T standard for measuring the voice quality of communications networks. It measures the subjective speech quality. It is calculated by comparing the enhanced speech with the clean reference speech. The value ranges from -0.5 to 4.5. It analyses the speech signal sample by sample. It provides numerical measure of the quality of human speech. PESQ can be used in a wide range of measurement 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 power of a signal and the power of corrupting noise. Usually it is expressed in terms of the logarithmic decibel scale. It is most easily defined via the mean squared error (MSE). Lower the error, higher will be the PSNR. The PSNR value 40 dB or more refers to good quality of the signal.

$$PSNR = 20 \log 10 (MAX_i) – 10 \log 10 (MSE) \qquad (2)$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 3, April 2017

Where

MSE – mean squared error

MAX_i – max possible value of the signal

$$\text{MSE} = \frac{1}{N}\sum_{i=0}^{N}(xi - yi)2 \qquad (3)$$

Where

x_i and y_i are the original and noisy signals

N – no of signal samples

The metric used for the analysis of k-NN algorithm is as follows,

C. *Identification Rate*

The average identification rate for the sets are computed as,

$$\%\ \text{Identification Rate} = N_{\text{correct}} / N_{\text{total}}\ \% \qquad (4)$$

Where

N_{correct} is the number of correctly identified sets

N_{total} 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 seded. The samples were experimented under both as trained and test datasets.

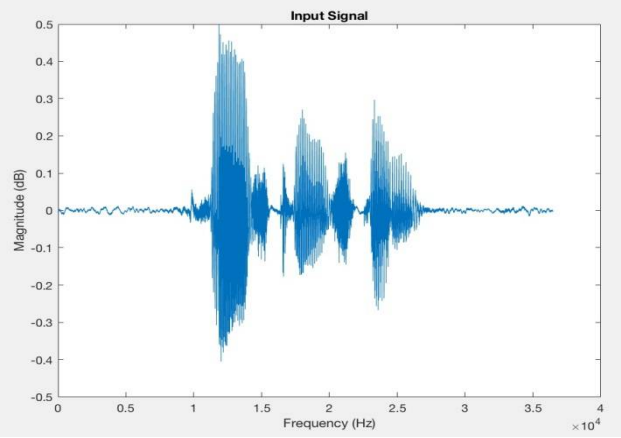


Figure 2 Spectrum of the Input

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

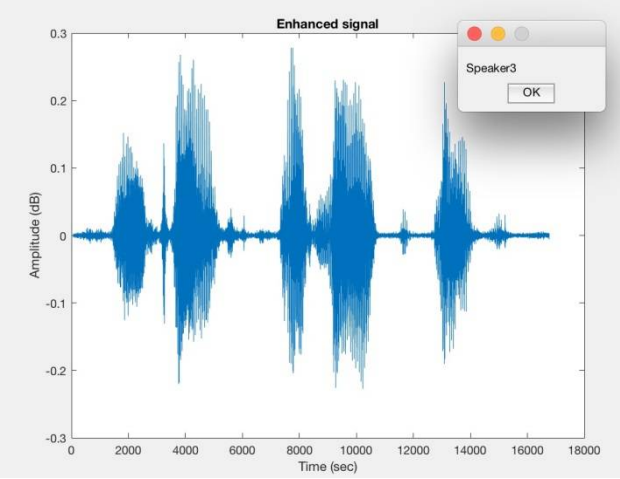
(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 3, April 2017

The above figure 2 shows the spectrum of an input speech.

A. *k-NN Classification Results*



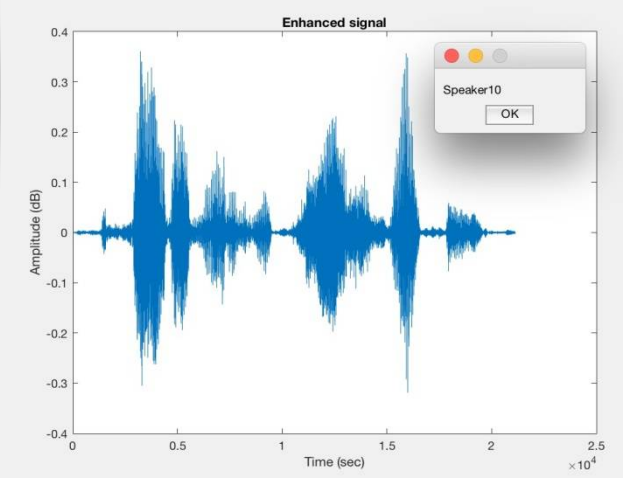


Figure 3 Speaker IdentificationFigure 4 Speaker Identification

The above figure 3 shows the spectrum of the identified speaker. The train sets of data are correctly identified. The test set is identified based on the features of the train sets.

The above figure 4 shows the spectrum of the identified speaker. The train sets of data are correctly identified. The test sets are mostly identified.

TABLE I

Comparison of Identification Rates of train and test sets

Speakers	**Identification rate for train sets**	**Identification rate for test sets**
Speaker 1	90%	54%
Speaker 2	95%	70%
Speaker 3	85%	50%
Speaker 4	90%	65%
Speaker 5	95%	75%
Speaker 6	85%	55%
Speaker 7	90%	45%
Speaker 8	95%	70%
Speaker 9	90%	65%
Speaker 10	80%	55%

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Special Issue 3, April 2017

The above Table I shows the comparison of the obtained identification rates of the train and test datasets. It is inferred that the performance of the latter is not as efficient as the trained one. Hence more improved algorithm should be used for efficient outcome.

V. CONCLUSION AND FUTURE WORK

Modified spectral subtraction method was proposed removal of noise for speech enhancement. Experimental evaluation on PESQ score and PSNR on the two databases i.e., Noizeus and Super seded are demonstrated. This method is highly efficient for learning real world datasets. The noises are reduced without affecting the signal power and the SNR is improved. This enhanced speech are also used as input for ASR systems as train and test datasets. The features are extracted using MFCC and PLP. The classification is done by k-NN algorithm and the results are analyzed for both train and test sets.

In the future work, Automatic Speaker Recognition system will be carried out with GMM algorithm. The experimental results are analysed and efficient method is obtained. Future applications of automatic speaker recognition will contribute substantially to the quality of life among deaf children and adults, as well as public and private sectors of the business community who will benefit from this technology.

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. M.Kalamani, Dr.S.Valarmathy, S.Anitha, "Automatic Speech Recognition using ELM and KNN Classifiers," International Journal of Innovative Research in Computer and Communication Engineering, Vol. 3, Issue 4, April 2015.
3. 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.
4. Munish Bhatia, Navpreet Singh, and Amitpal Singh, "Speaker Accent Recognition by MFCC Using KNearestNeighbour Algorithm: A Different Approach," International Journal of Advanced Research in Computer and Communication Engineering, Vol. 4, Issue 1, January 2015.
5. 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.
6. AnujaBombatkar, GayatriBhoyar, KhushbuMorjani, ShalakaGautam, and Vikas Gupta, "Emotion recognition using Speech Processing Using k-nearest neighbor algorithm," International Journal of Engineering Research and Applications (IJERA),pp 68-71, April 2014.
7. 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.
8. Muhammad Rizwan and David V. Anderson, "Using k-Nearest Neighbor and Speaker Ranking for Phoneme Prediction," 13th International Conference on Machine Learning and Applications, IEEE, 2014.
9. X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising auto encoder," in Proc. INTERSPEECH, 2013, pp.436–440.
10. 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.
11. Ekaterina Verteletskaya, and BorisSimak, "Noise Reduction Based on Modified Spectral Subtraction Method," IAENG International Journal of Computer Science, Vol 38, pp. 231-239,2011.
12. 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.
13. 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.
14. 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.
15. Yang Lu and Philipos C. Loizou, "A geometric approach to spectral subtraction," in speech communication,2008.
16. 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.
17. 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.
18. 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.
19. K. Lebart, and J. M. Boucher, "A New method based on spectral subtraction for speech enhancement," Acustica, Vol. 87, pp. 359-366,2001.
20. Y. Ephraim, "A signal subspace approach for speech enhancement," IEEE Trans. on speech and audio processing, pp. 251-266,1995.