International Journal of Engineering Research in Electronic and Communication Engineering (IJERECE) Vol 2, Issue 11, November 2015

Speaker Identification And Verification Of Noisy-Echoed Speech Using Gaussian Mixture Models

Veena K V
Dept. of Applied Electronics and Instrumentation Engineering
Rajagiri School of Engineering and Technology
Kochi, India
k.v.veena29@gmail.com

Abstract: **— The two major applications of speaker recognition applications are speaker verification and speaker identification. But in most of the cases the signal is corrupted with background interferences such as noise and echo. This paper proposes the method of speaker recognition and identification after the noise separation and echo cancellation. Support vector machine(svm) classification based signal separation is adopted here and general kalman filter is used for echo cancellation. Adapted gaussian mixture models along with universal background model(ubm) is used for speaker verification and identification tasks.**

Keywords **- GMM-UBM; CASA; SVM; General Kalman Filter(GKF);**

I.INDRODUCTION

Speech is one of the most private form of person's communication and is widely used in many speech processing tasks like speaker identification and verification. In speaker identification and verification, we use the recorded speech to classify the speaker among a set of individuals. These tasks find use in surveillance and authentication applications respectively.

But the current speaker recognition systems may not preserve the privacy of speaker. This allows an eavesdropper to access the voice samples and try to impersonate the speaker by generating fake voice samples using enrolled speech samples. This paper proposes a privacy preserving speaker verification and identification system using Gaussian Mixture Models.

In real world applications, speech is corrupted with background interferences such as noise, echo etc.. Due to the presence of noise, one cannot predict whether the signal contains valid information or not through the direct observation. Voice activity detection (VAD)is used to detect the presence or absence of human speech in a speech sample.

In order to improve the performance of speaker recognition system, the signal should be pre-processed to remove the noise. Monaural speech technique is used for the speech separation. Various approaches for monaural speech separation have been proposed. Speech enhancement approaches [5], [6] utilize the statistical properties of the signal to enhance speech that has been degraded by additive non speech noise and the model based approaches [7],[8] use trained models to capture the characteristics of individual signals for separation. On the other hand, computational auditory scene analysis (CASA) aims to separate a sound mixture. For sound separation, the ideal binary mask (IBM) has been recently proposed as a main goal of CASA.

After the noise and speech separation [2],[3],the preprocessed signal is used for speaker recognition task . The system is built around the likelihood ratio test for verification, using GMMs for likelihood functions, a universal background model (UBM) for alternative speaker representation, and a form of Bayesian adaptation to derive speaker models from the UBM. In privacy preserving systems input speech is given in the encrypted format. The system does not observe the signal in the original form.

II. SPEECH SEPARATION

For speech and noise separation computational auditory scene analysis (CASA) method is used. Ideal Binary Mask(IBM) is treated as the main goal of CASA. IBM can be constructed using the premixed clean speech and noise signal. IBM is a binary time-frequency (T-F) matrix where 1 represents the signal-to-noise ratio within the T-F unit is greater than a local SNR criterion (LC) and 0 represents the SNR is less than the local SNR criterion.

s(t; f) and n(t; f) are the signal and noise power in decibels.

$$IBM(t,f) = \begin{cases} 1; \text{ if } s(t;\ f) - n(t;\ f) > LC \\ 0; \text{ otherwise} \end{cases} \quad (1)$$

These approaches subsist of an Support Vector Machine (SVM) training, widely used in classification problems. SVM is a maximum margin classifier which separates the training data into separate classes.

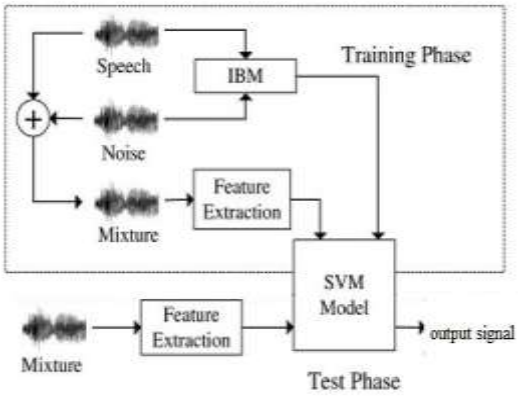

Figure 1: Block diagram for the speech separation

III. ALGORITHM DESCRIPTION

Speech separation is done through two stages, training phase and a test phase [3]. In the training phase, the speech and the noise are used to create the IBM, which provides the desired output for training. The features in each T-F unit are extracted from the mixture and then used to train an SVM model in each frequency channel. In the test phase, we first use the trained SVM to initially classify T-F units,

The output of this stage is shown in figure(2).

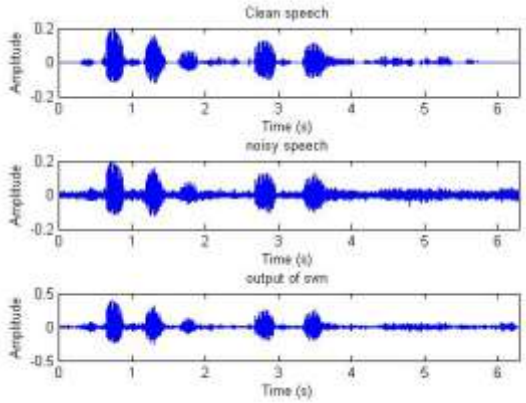

Figure 2:Output of sp10.wav using RASTA-PLP

IV. ECHO CANCELLATION

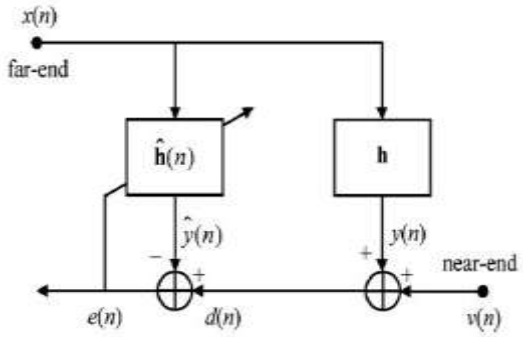

Figure 3. General configuration for echo cancellation

In the context of echo cancellation (Fig.1)[2], the microphone or desired signal at the discrete-time index *n* is

$$d(n) = x^T h(n) + v(n) \quad (2)$$

where

$$x(n) = [x(n)\ x(n-1) \ldots x(n-L+1)]^T \quad (3)$$

is a vector containing the L most recent time samples of the input (loudspeaker) signal *x(n)* , superscript T denotes transpose of a vector or a matrix,

$$h = [h_0\ h_1 \ldots\ h_{L-1}]^T \quad (4)$$

is the impulse response (of length L) of the system (from the loudspeaker to the microphone) that we need to identify, and $\upsilon(n)$ is a zero-mean stationary white Gaussian noise signal. The variance of this additive noise is σ_{υ}^2. The signal is called the echo in the context of echo cancellation that we want to cancel with an adaptive filter.

Then, our objective is to estimate or identify **h** with an adaptive filter:

$$\hat{\boldsymbol{h}}(\boldsymbol{n}) = [\boldsymbol{h_0}(\boldsymbol{n})\ \boldsymbol{h_1}(\boldsymbol{n}) \ldots \boldsymbol{h_{L-1}}(\boldsymbol{n})]^{\boldsymbol{T}} \quad \textbf{(5)}$$

in such a way that for a reasonable value of n , we have for the (normalized) misalignment:

$$\frac{\|\hat{\boldsymbol{h}}(\boldsymbol{n})-\boldsymbol{h}\|_2^2}{\|\boldsymbol{h}\|_2^2} \leq \boldsymbol{\iota} \quad \textbf{(6)}$$

where ι is a predetermined small positive number.

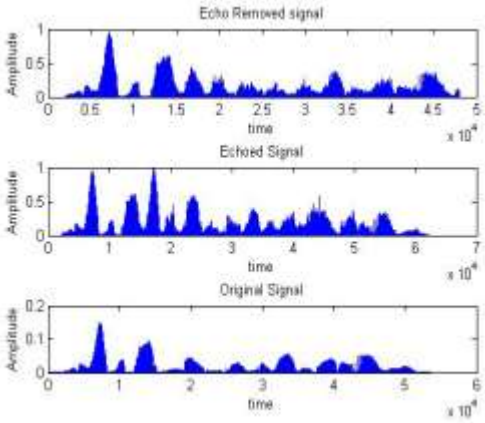

Figure 4: Echo removed signal

V. SPEAKER IDENTIFICATION AND VERIFICATION

Signal parameterization is the primary stage of speaker recognition system. Mel Frequency Campestral Coefficients are the most commonly used and seem to be more feasible for speaker identification and verification tasks.. Gaussian Mixture Model (GMM) is a commonly used generative model for density estimation in speech and language processing [1]. The probability of the model generating an example is given by a mixture of Gaussian distributions.

A GMM λ comprises of M multivariate Gaussians each with a mean and covariance matrix. If the mean vector and covariance matrix of the j[th] Gaussian are respectively μ_j and Σ_j , for an observation x, we have

$$\boldsymbol{P}(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_j \boldsymbol{w_j}\ \boldsymbol{N}(\boldsymbol{\mu_j}, \Sigma_j) \quad \textbf{(7)}$$

where w_j are the mixture coefficients that sum to one. The above mentioned parameters can be computed using the Expectation Maximization (EM) algorithm[4].Given a collection of training vectors, maximum likelihood model parameters are estimated using the iterative expectation–maximization (EM) algorithm . The EM algorithm iteratively refines the GMM parameters to monotonically increase the likelihood of the estimated model for the observed feature vectors and also the UBM parameters are trained. Generally, five iterations are sufficient for parameter convergence. the feature vectors of X are assumed independent, so the log likelihood of a model for a sequence of feature vectors, $X = x_1, \ldots, x_T$, is computed as

$$\boldsymbol{p}(\boldsymbol{x}|\boldsymbol{\lambda}) = \sum_{t=1}^{T} \log \boldsymbol{p}(\boldsymbol{x_t}|\boldsymbol{\lambda}) \quad \textbf{(8)}$$

The advantages of using a GMM as the likelihood function are that it is computationally inexpensive, is based on a well-understood statistical model, and, for text-independent tasks, is insensitive to the temporal aspects of the speech, modeling only the underlying distribution of acoustic observations from a speaker. The latter is also a disadvantage in that higher levels of information about the speaker conveyed in the temporal speech signal are not used.

VI. EXPERIMENT

TIMIT database is used for the speaker verification and identification tasks and the algorithm is developed using MATLAB software. Experiment is conducted for 20 speakers. For each speaker 10 speech samples are taken, out of these 10 samples 7 samples are taken for training and remaining 3 samples are taken for text independent speaker identification tasks.UBM is developed with 32 Gaussian Mixture Components from the training samples and perform MAP adaptation with the training data for individual speakers to obtain the speaker models.

VII. CONCLUSION

GMM is one of the most powerful tool and technique used for speaker identification and recognition applications.

But in the presence of background noises and echo, GMM is almost completely unable to predict the right speaker for a test speech signal even in the speaker verification process. So the signal is subjected to different preprocessing stages.SVM classifier is used to separate signal and noise. General Kalman Filter is used for echo cancellation. The features of signal obtained after these preprocessing treatments are used by the GMM to find the gmm score. The obtained accuracy is 75 % .

Input	Speaker Verification (Accuracy in %)	Speaker Identification (Accuracy in %)
Clean Speech	100	97.6
Noisy-Echoed Speech	96.8	75

Table1: Speaker identification and verification results obtained for 20 speakers

REFERENCES

[1] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B.Dunn "Speaker Verification Using Adapted Gaussian Mixture Models"M.I.T. Lincoln Laboratory, 244 Wood St., Lexington, Massachusetts 02420

[2] Constantin Paleologu, Jacob Benesty, and Silviu Ciochina "Study of the General Kalman Filter for Echo Cancellation in "IEEE Trans. on Audio, Speech, Lang. Process. " vol. 21, no.8,August 2013.

[3] Kun Han, and DeLiang Wang "Towards Generalizing Classification Based Speech Separation, " in "IEEE Trans. on Audio, Speech, Lang. Process. " vol. 21, no. 1,January 2013.

[4] Frank Dellaert "The Expectation Maximization Algorithm ,"Georgia Institute of Technology Technical Report number GIT-GVU-02-20 February 2002

[5] Y. Ephraim and D. Malah, "Speech enhancement using a minimum- mean square error short-time spectral amplitude estimator,"IEEE Trans. Acoust., Speech, Signal Process., vol. ASSP-32, no. 6, pp. 1109–1121, Dec. 1984.

[6] R. C. Hendriks, R. Heusdens, and J. Jensen, " MMSE based noise PSD tracking with low complexity, in Proc. IEEE ICASSP, 2010, pp. 4266–4269.

[7] A. Ozerov, P. Philippe, F. Bimbot, and R. Gribonval, "Adaptation of bayesian models for single-channel source separation and its application to voice/music separation in popular songs," IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 5, pp. 1564–1578, Jul. 2007.

[8] A. Ozerov, E. Vincent, and F. Bimbot, "A general flexible framework for the handling of prior information in audio source separation,",IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 4, pp. 1118–1133, May 2012.