Robust Speaker Verification using Self Organizing Map

Pranab Das
Dept. Computer Science and Engineering
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India
pranab.das@dbuniversity.ac.in

Utpal Bhatacharjee
Dept. Computer Science and Engineering
Rajiv Gandhi University
Doimukh, Arunachal Pradesh, India
utpalbhattacharjee@rediffmail.com

Abstract— **This paper proposes a new approach of noise reduction based on the analysis of MFCC feature space using self-organizing map network. Here the U-matrix plot of the feature space is analyzed in presence of white noise at different signal to noise ratio. Based on the observation, boundary neurons separating clusters are identified in the feature space. For each such neuron in the boundary, its 2-D feature vector is extracted from the U-matrix and hit matrix. This collection of feature vectors based on the boundary neurons are eliminated from the original feature space. Thus the new feature space obtained is used to perform the tasks of visualization and speaker verification. Experiments were carried out by combining synthetic white noise with real world data sets.**

Keywords— Self Organizing maps, Mel frequency cepstral coefficient, Speaker verification

I. INTRODUCTION

Since the early years of speaker verification, researchers have faced the problem in the performance of the system when the test data were recorded in adverse condition. This problem is encountered each time mainly due to the presence of background noise and channel noise. In literature, many approaches have been developed which work at signal, feature or model level to improve the tolerance of the verification system with respect to noise.

At signal level various filtering techniques such as Wiener filtering [1], spectral subtraction [2] were used. Spectral subtraction enhances the speech signal by subtracting the estimated noise from the spectrum of the input noisy signal. The Wiener filter is a popular statistical approach based on the assumption that signal and the noise are stationary linear stochastic processes with known spectral characteristics. At feature level, feature mapping technique introduced by [3] all features are mapped to a common space so that they would work best with the same background model. Other feature compensation methods include cepstral mean subtraction [4], feature transformation techniques like feature warping [5], and short-time Gaussianization [6]. Techniques such as parallel model compensation (PMC) [7], vector Taylor series approximation are based on model compensation [8]. Parallel model combination (PMC) adjusts the clean speech acoustic models so that they reflect noisy speech with a noise distribution similar to the one measured on the test data. Maximum Likelihood Linear Regression (MLLR), and Maximum A Posterior (MAP) have been used in past for model level compensation.

In this paper we have focused on artificial neural network (ANNs) which is based on the model of biological neural network of brain. This approach has been widely used because of its capability to provide solutions with improved performance by the way of learning and pattern recognition. One such popular neural network architecture and learning algorithms used in the study is Kohonen's self organizing map (SOM) [9]. Developed as a visualization and analysis tool based on associative memory model, SOM is an unsupervised learning algorithm with topology preserved mapping from input to output space. The Self-Organizing Map (SOM) is a vector quantization method which places the prototype vectors on a regular low-dimensional grid in an ordered fashion. This makes the SOM a powerful tool for clustering, dimensionality reduction, and classification.

The structure of the paper is as follows section I introduces the speaker verification system in noisy environment; section II and section III describes Mel frequency cepstral coefficients and self organizing map respectively; section IV analyzes the U-matrix plot of the MFCC features. Section V proposes the new approach towards noise reduction; section VI gives description of the speaker recognition database. Section VII shows the results of the experiments and finally the paper is concluded in section VIII.

II. FEATURE EXTRACTION USING MFCC

For speaker recognition, the most commonly used acoustic features are mel-scale frequency cepstral coefficient (MFCC). MFCC is based on human perception of hearing which said to concentrate on certain frequency components. Therefore are considered best for speaker recognition. The feature extraction process is shown in figure 1.

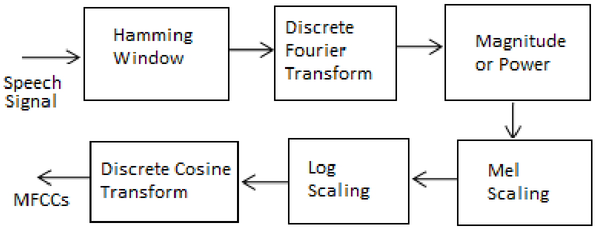


Fig. 1. MFCC feature extracttion steps

As the first step called pre-emphasis, the speech signal $s(n)$ is made to pass through a high-pass filter:

$$s_2(n) = s(n) - a * s(n-1) \qquad (1)$$

where $s_2(n)$ is the output signal and the value of a lies between 0.9 and 1.0. Each frame is then multiplied with a hamming window in order to keep the continuity of the first and the last points in the frame. The Hamming window is defined by:

$$w(n, \alpha) = (1-\alpha) - \alpha \cos\left(\frac{2\pi n}{N-1}\right) \qquad (2)$$
$$0 \leq n \leq N-1$$

Assuming that the signal within the frame is periodic and continuous, FFT is applied to obtain the magnitude frequency response of each frame. Next the log energy of each triangular bandpass filter is obtained by multiplying the magnitude frequency response by a set of 20 triangular bandpass filters. The triangular filters are used as they can smoothen the harmonics .There are 10 filters spaced linearly below 1000 Hz, and the remaining filters spread logarithmically above 1000 Hz. Next spectral envelope and spectral details are separated from the spectrum by taking logarithm which converts the multiplication of the magnitude into addition. Use of log makes the feature estimates less sensitive to input variations caused by speaker mouth movement. Finally the Mel Frequency cepstral coefficients (MFCC) are obtained by converting the log-compressed filter outputs to time domain using the Discrete Cosine Transform (DCT). DCT is calculated as

$$C(u) = a(u) \sum_{n=0}^{N-1} f(n) \cos\left[\frac{\pi(2n+1)u}{2N}\right] \qquad (3)$$

where u and n are 0,1,2,......N-1 and

$$a(u) = \sqrt{\frac{1}{N}} \text{ for } u=0$$
$$a(u) = \sqrt{\frac{2}{N}} \text{ for } u \neq 0$$

III. Self Organizing Map

The SOM consists of a two-dimensional lattice that contains a number of neurons. These neurons are usually arranged in a rectangular or hexagonal way. Typically there are three major phases in SOM training process. During the first phase of training, which is also called as the competition phase, all the neurons compete amongst themselves to be activated, that is each neuron computes a discriminant function and the neuron with largest discriminant function is declared as the winner. In the second phase called Cooperation, the winning neuron determines the spatial location of topological neighbourhood of excited neurons, thereby strengthening not only the winning neuron but also its neighbouring neurons. And finally in the third phase synaptic adaptation enables the excited neurons to increase their individual values of the discriminant function such that the mapping from input space to output space preserves the topology.

Before the training process can start, connection weights $\{w_1, w_2, \ldots\ldots, w_n\}$ are initialized with small random values. Where w_i is the weight vector associated with neuron, n is the number of neurons and r_i is the location of neuron i in the lattice.

Algorithm:

Repeat

1. At each time t, present an input x(t), and select the winner or the best matching unit (BMU),

$$v(t) = \arg\min_{i \in \mathcal{J}} \|x(t) - w_i(t)\| \qquad (4)$$

2. Update the weights of the best matching unit (BMU) and its neighbours,

$$\Delta w_i(t) = \alpha(t)\eta(v, i, t)[x(t) - w_v(t)] \qquad (5)$$

Until the map converges

Where the neighbourhood function is represented by

$$\eta(v, i, t) = \exp\left[-\frac{\|r_v - r_i\|^2}{2\sigma(t)^2}\right] \qquad (6)$$

with σ representing the range of neighbourhood , $\mathcal{J}$ is the set of neuron index and α(t) called the learning rate.
The result of step 2 in the algorithm is represented in figure 2.

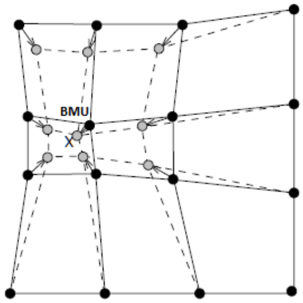
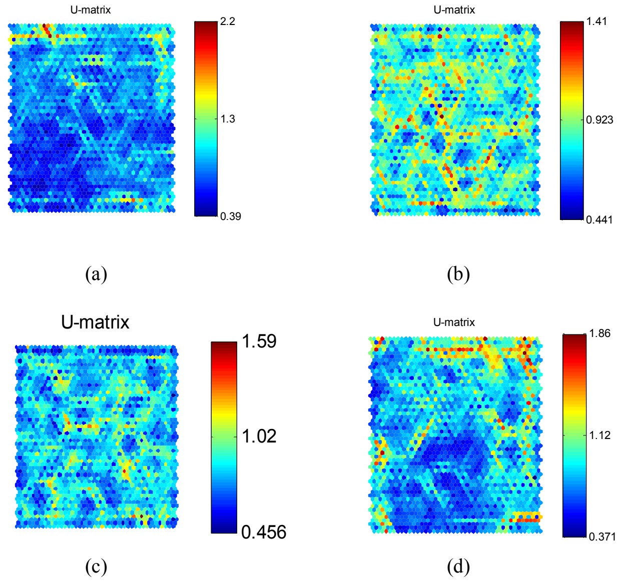


Fig. 2. The best matching unit (BMU) and its neighbors are updated marked by solid and dashed lines respectively.

IV. Analysis of U-matrix

An experiment is carried out by taking 50 clean speech samples. Each of the samples is contaminated with additive white Gaussian noise at SNR 0dB, 10dB and 20dB resulting in a total of 200 speech samples. The experiments were carried out separately for each noise level. A 39-dimensional feature vector, made up of 13 mel-frequency cepstral coefficient (MFCC) and their first and second order derivatives were obtained. For visualisation and analysis SOM toolbox is used. The U-matrix plot of SOM is used for evaluation. A sample U-matrix plot of a speech signal is given in figure 3.



(a)

(b)

(c)

(d)

Fig. 3. U-matrix plot of a sample speech signal (a) Clean Speech (b) White noise added at 0dB with clean speech (c) White noise added at 10dB with clean speech (d) White noise added at 20dB with clean speech

Observation: The SOM is colored by the values of u-matrix elements. The color scale is displayed near the SOM denotes the distances between neighbouring neurons. As can be observed from color scales, dark color such as blue or dark blue signifies that codebook vectors are close to each other in the input space. Whereas areas such as red or yellow signifies codebook vectors are far in the input space. As it was observed from the U-matrix plot of Fig 3 (b), (c) and (d) and the remaining 199 samples, that as the SNR decreases from 20dB to 0dB the distance between the neurons keeps on increasing so the number of clusters. Also more and more cluster separators in the form of yellow and red regions tend to increase with decrease in SNR. In case of clean speech, number of cluster is very less as compared with the same with white noise at various SNR. So it can be clearly observed that cluster boundary separators in the map are because of the presence of white noise at lower SNR resulted in increase in the number of clusters.

V. Proposed Methodology

In the U-matrix yellow and red regions which correspond to larger distance between the neurons resulted because of adding white noise at 0dB is eliminated. Here the neurons having higher values in the distance matrix are identified and the corresponding feature vectors of those neurons in the original feature data space are eliminated. SOM clustering algorithm is applied to this new feature space. Based on the Davies Bouldin index a reduced feature set is obtained by taking centroids of each of the cluster.

Algorithm

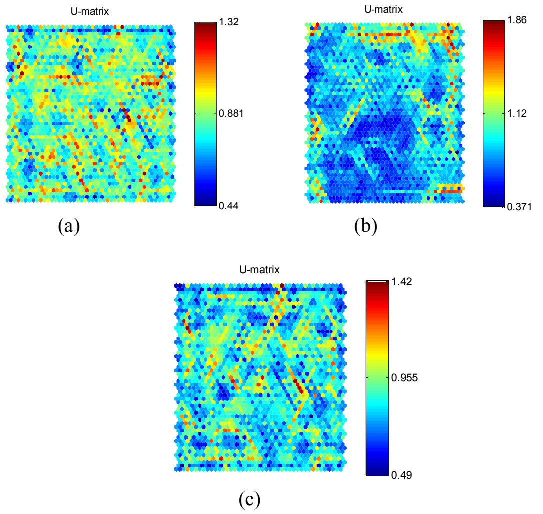
1. Apply SOM algorithm on the MFCC feature vector to obtain the topology preserved map. In the map so obtained, sort the neurons in decreasing order of their values based on the distance matrix.
2. Based on the map size select 5% of the neurons from the sorted list for evaluation. The hit matrix is also calculated for the selected neurons.
3. Next select those neurons for which the number of hits is greater than or equal to average hits of the entire map.
4. Identify the feature vectors corresponding to the neurons selected in step 3.
5. Delete the feature vectors obtained in step 4 from the original feature space.
6. Apply SOM clustering algorithm on the feature vector obtained in step 5. Here Davies Bouldin index (DBI) metric is used to evaluate the number of clusters in this new feature space.
7. Take the centroid of each cluster to form the new reduced feature vector.

VI. Database Description

To carry out the experiments, a speaker verification database was developed and all the testing and evaluation of the speaker recognition system has been done using that database. The database consists of 50 speakers, 27 male and 23 female. Each speaker is recorded for 3 minutes duration in typical office environment using table mounted microphone. The training set consists of speech data of length 120 seconds per speaker. The test set consists of speech data of length 15 seconds and 30 seconds. Each speaker model is tested against 11 speakers of which one is the actual speaker and rest 10 are the imposters. The test segments are contaminated with additive Gaussian noise of SNR 0dB, 10dB and 20dB. The experiments were carried out separately for each noise level.

VII. Experimental Results

All the experiments reported in this paper are carried out using the database described in section VI. In the first set of experiment, the proposed method is applied to the speech signal contaminated with white Gaussian noise. The U-matrix plot of a sample speech signal corrupted with white noise at 0dB is shown in figure 4(a) and the result after applying the proposed method is shown in 4(c). In the second set of experiment a speaker verification system is designed to evaluate the performance of the proposed method. In the first step silence intervals from the input speech are removed based on an envelope threshold. Then Gaussian mixture model with 1024 Gaussian components were used for both the UBM and speaker model. The UBM was created by training the speaker model with speaker's data with Expectation Maximization (EM) algorithm. The speaker models were created by adapting only the mean parameters of the UBM using maximum a posteriori (MAP) approach with the speaker specific data. The detection error trade-off (DET) curve is plotted using log likelihood ratio between the claimed model and the UBM. The equal error rate (EER) obtained from the DET curve has been used as a measure, for the performance of the speaker verification system. The results of the experiments are shown in Table 1.



(a) (b)

(c)

Fig. 4. U-matrix plot of MFCC feature of a sample speech (a) Noisy Speech at 0dB (b) Clean Speech (c) After applying proposed technique.

TABLE I. EQUAL ERROR RATE FOR SPEAKER VERIFICATION SYSTEM AT DIFFERENT SNR LEVEL

	Clean Speech	**SNR 0dB**	**SNR 10dB**	**SNR 20dB**
Baseline System	09.67	31.41	24.28	21..31
Proposed Method	Not Applicable	15.92	16.63	17.57

VIII. CONCLUSION

This paper investigated the problem of speaker verification in noisy conditions assuming absence of information about the noise. The works reported in this paper are carried out using a database contaminated by simulated white Gaussian noise at three different SNR. From the result shown in Table 1 it is clearly observed that the system recognition rate decreases significantly with lower SNR. Though similar pattern is observed when the proposed method is applied, a significant improvement in recognition rate is observed in case of SNR at 0dB around 15%. It is also observed that the recognition rate is better at 0dB than at 10dB and is somewhat similar at 20dB. The reason for the recognition rate to be higher at 10db than at 0dB may be because of the cluster separators are less when the noise level in the speech signal is less. Also from U matrix plot in figure 3 it is observed that the numbers of clusters are reduced and the red and yellow regions which resulted because of white noise is also reduced to a greater extent after applying the proposed method. Hence it can be concluded from the results that the system performance can be improved to a greater extent by the proposed method in presence of white noise at lower SNR.

REFERENCES

[1] Vaseghi, S.V., Milner, and B.P.: Noise compensation methods for hidden Markov model speech recognition in adverse environments. IEEE Trans. Speech Audio Process. 5 (1), 11–21 (1997).

[2] Boll, S.F.: Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoustics, Speech Signal Process. 27 (2), 113–120 (1979).

[3] Reynolds, D.a.: Channel robust speaker verification via feature mapping. In: Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on, vol. 2, pp. II–53–6 (2003)

[4] S. Furui, "Cepstral analysis technique for automatic speaker verification," IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 29, pp. 254–272, 1981

[5] Pelecanos, J., Sridharan, S.: Feature Warping for Robust Speaker Verification. In: A Speaker Odyssey - The Speaker Recognition Workshop, pp. 213–218 (2001)

[6] Chen, S., Gopinath, R.A.: Gaussianization. In: Neural Information Processing Systems (NIPS) (2000)

[7] M. J. F. Gales and S. Young, "HMM recognition in noise using parallel model combination," in Proc. Eurospeech'93, Berlin, Germany, 1993, pp. 837–840.

[8] Moreno, P.J., Raj, B., Stern, R.M.: A vector Taylor series approach for environment independent speech recognition. In: ICSLP, Philadelphia, PA, pp. 733–736 (1996).

[9] Kohonen T (1982) Self-organised formation of topologically correct feature map. Biological Cybernetics, 43: pp. 56–69.