Fuzzy Reconstruction of Cluster-Based Missing Features Method for Robust Speech Recognition

Sadegh Masjoodi, Mansour Vali
Engineering Faculty of Shahed University of Tehran

***Abstract*—Despite one decade of the missing feature theory application in the domain of Robust Automatic Speech Recognition (ASR), this field is still an active area for researchers. In this report using fuzzy concepts, we will present a method for modifying the cluster-based reconstruction of unreliable components of the noisy speech spectrogram. In this simple but effective method using a fuzzy membership function the feature vector component reliability is fuzzified. In the next stage this new parameter is applied as a weighting parameter for summing new reconstructed components and their old noisy values. Experiments were done on the FarsDat database using two recognition models, a Neural Network (NN) and a Hidden Markova Model (HMM). The improvements in the recognition results using this new reconstruction method in low SNRs for the frame-based neural network was approximately 5% and for the phoneme-based HMM was between one and two percent.**

I. INTRODUCTION

Mismatch conditions between train and test environments of a speech recognition system degrades the accuracy of recognition significantly. Among different methods that have been presented for compensating these mismatches, missing features techniques have this advantage that are applicable in both domains of feature compensation and model modification. Another advantage of these methods is that these techniques are not dependent to the parameters of corrupting noise.

[1] has reviewed these missing feature methods in two groups of model modification and feature compensation. In that report, marginalization and class conditional imputation are in the group of model modification methods. Both methods have been shown to be effective in compensating noise effects but suffer from some drawbacks. Those methods are only applicable in the spectral domain and for this reason recognition must be done in the same spectral domain. But we know that speech recognition accuracy with cepstral parameters is significantly better than accuracy in spectral domain. The next drawback of model based missing feature approaches is related to their practical implementation. For implementing such algorithms recognition models should be modified so that they are applicable in situations that recognition models internally are accessible.

Second group of missing feature methods introduced in [1], are feature imputation techniques. These methods do not have disadvantages of model-based missing feature methods. Really the most important characteristic of feature imputation methods is that they are model independent. In these techniques feature vectors extracted from noisy speech signal before applying to recognition models are processed. In the first step of processing, time-frequency representation of noisy speech signal is formed. Then using a suitable criterion such as SNR criterion, different components of this representation (spectrogram) are classified to reliable and unreliable classes:

$$[S_r(t), S_u(t)] = Mask(S(t)) \qquad (1)$$

In the above relation $S(t)$ is the log spectral vector of 't'th frame. $S_r(t)$ and $S_u(t)$ are reliable and unreliable portions of that log spectral vector. Based on this classification every component of spectrogram is labeled as "one" and/or "zero". One means that the corresponding cell of spectrogram is reliable and zero means that cell is unreliable or missing. For example in the SNR criterion method SNR of all components of LFBE spectrogram are estimated then these values are compared with an optimum threshold. Those components of spectrogram that their corresponding SNRs are lower from that threshold are labeled as unreliable and are shown using "0". The remained components of the spectrogram are labeled as reliable and are shown using "1". In the second stage of feature imputation process missing or unreliable components of spectrogram are recovered using one of reconstruction methods. One of spectrogram reconstruction methods is referred to as covariance-based reconstruction. In this method that is completely described in [2] log spectral vectors of speech signal are applied. These feature vectors are considered as outputs of a Gaussian stationary random process. Statistical information of this process such as mean vector and covariance matrix are derived from clean training speech signals. For reconstructing of missing cells some adjacent reliable components are applied that their correlation with missing cells are greater than a threshold value. Estimations of unreliable components from such reliable components are calculated using a bounded version of MAP algorithm. Other reconstruction method that is more noticed and applied in reports [3,4,5] is cluster-based reconstruction method. In comparison to covariance based technique, derived results of cluster-based method are superior [2]. In this research we will present a method for modifying cluster-based reconstruction technique. In the presented modification, we will show how this technique

can improve recognition results in different levels of frame and phoneme recognition.

This Paper is organized as follows. In section 2 cluster-based reconstruction approaches are described. Fuzzy cluster-based reconstruction method is presented in section 3. Experimental results for recognition of fuzzy and nonfuzzy reconstructed noisy speech are explained in section 4 and finally section 5 will be on conclusion.

II. Cluster-Based Reconstruction of Unreliable Components

In the cluster-based reconstruction method log-spectral vectors are modeled as outputs of an independent, identically distributed (IID) random process. In this method log-spectral vectors of clean training data are clustered to a number of clusters and a Gaussian distribution function is specified to everyone. So if K is the total number of clusters, total distribution of all data will be a multivariate Gaussian mixture model:

$$P_{s(t)} = \sum_{K=1}^{N} C_K N(S(t); \mu_K, \Sigma_K) \qquad (2)$$

In the above relation C_K, μ_K and Σ_K are a priory probability, mean vector and covariance matrix of kth cluster, respectively. Their values can be trained from the training data corpus.

For compensating noisy components of any corrupted feature vector first of all corresponding cluster to that vector should be identified. In [2] some solutions for finding cluster memberships have been suggested such as interpolation along time, interpolation along frequency and marginal method. In the marginal cluster membership identification this maximum a posteriori probability must be estimated:

$$\hat{k}_{S(t)} = \arg\max_k \{P(k \mid S_r(t), S_u(t) \le Y_u(t))\} \qquad (3)$$

In the above relation $Y_u(t)$ is the observed value of unreliable part of the log spectral vector $S(t)$. Using Bays theory the above relation can be rewritten as:

$$\hat{k}_{S(t)} = \arg\max_k \{P(S_r(t), S_u(t) \le Y_u(t) \mid k) P(k)\} \qquad (4)$$

There is no a closed form for calculation of the above marginal probability but if covariance matrix assumed to be diagonal the following relation is calculable [2]:

$$P(S_r(t), S_u(t) \le Y_u(t) \mid k) = \int_{-\infty}^{Y_u(t)} P(S_r(t), S_u(t) \mid k) dS_m(t) \qquad (5)$$

Knowing cluster membership of the noisy vectors applying a bounded version of MAP algorithm missing components of corrupted vectors could be recovered.

III. Fuzzy Cluster Based Reconstruction of Unreliable Components

The idea of fuzzy masks for the first time was suggested in [6]. Following this idea we have presented a simple but effective technique for improving cluster-based reconstruction method. In the suggested method using a sigmoid membership function estimated signal to noise ratio of all spectrogram cells are fuzzified and are taken in to the range of [0,1]. Standard form of a sigmoid function is as follows:

$$\mu(SNR; a, c) = \frac{1}{1 + e^{-a(SNR-c)}} \qquad (6)$$

In our suggested method we will apply below relation for imputation of noisy components:

$$\tilde{Y}_{ls} = \mu_{SNR} Y_{ls} + (1 - \mu_{SNR}) \hat{Y}_{ls} \qquad (7)$$

In this relation μ_{SNR} is a correction factor calculated using relation (6). $\hat{Y}_{ls}(t)$ is the estimated vector from cluster-based reconstruction method and $Y_{ls}(t)$ is the input feature vector before any reconstruction. In this relation when signal to noise ratio is more than a predetermined upper threshold '*u*', $\mu_{SNR} \rightarrow 1$, and when SNR is less than a predetermined lower threshold '*l*', $\mu_{SNR} \rightarrow 0$. It means that if the reliability of spectrogram components are more than the upper threshold the reconstruction is not necessary and if reliability is under the lower threshold, imputation of corresponding cells should be done completely. For $l < \text{SNR} < u$ final reconstructed vector is determined by summing weighted versions of vectors Y_{ls} and $\hat{Y}_{ls}$. Optimal values for upper and lower thresholds '*l*' and '*u*' were determined using experiments as '-4' and '2', respectively. The parameters of sigmoid function, *'a'* and *'c'*, that correspond to *'l'* and *'u'* values are '1.4' and '-1', respectively.

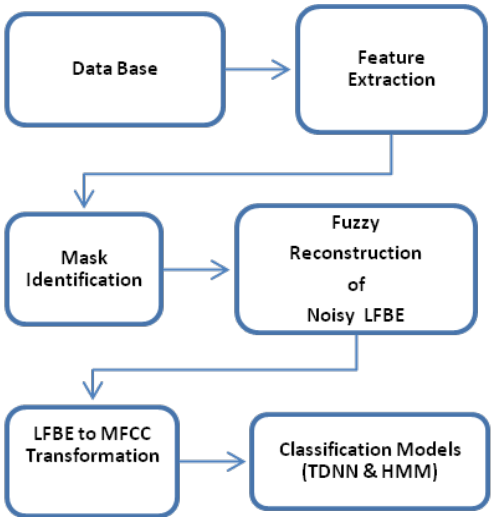

Fig. 1 Block diagram for implementing the noisy speech recognition system

IV. Experimental results

Block diagram of implementation process has been shown in figure 1. In this research for evaluating presented method we have used FarsDat database [7]. Framing of speech signal were done using hamming windows. We applied 20 channel triangular filter banks for extracting LFBE features. For implementing the Mask Identification stage of recognition we used oracle masks. Details of this method have been described in [2]. In the recognition stage we used MFCC features. By using DCT transform we extracted 13 MFCC parameters from reconstructed LFBE features. By adding their delta and acceleration coefficients, MFCC feature vectors include 39 parameters.

The applied neural network was a Time Delay Neural Network (TDNN). Schematic of this neural network is demonstrated in figure 2.

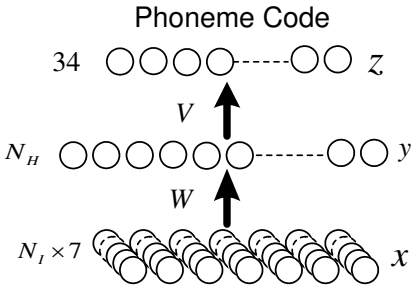

Fig. 2. A simple Schematic for applied Neural Network

Really, it is a frame-based recognizer. It has a hidden layer with 100 neurons. The number of input neurons is 7x39 so that could receive the parameters of current frame and the parameters of 3 frames before and 3 frames after it. The network's Output layer had 34 neurons which everyone corresponded to one phoneme.

The employed Hidden Markov Model (HMM) here included 34 left to right 3 state models. We used HTK toolkit for modeling of HMMs. A simple Schematic diagram for a typical 3 state left-right HMM is shown in figure 3.

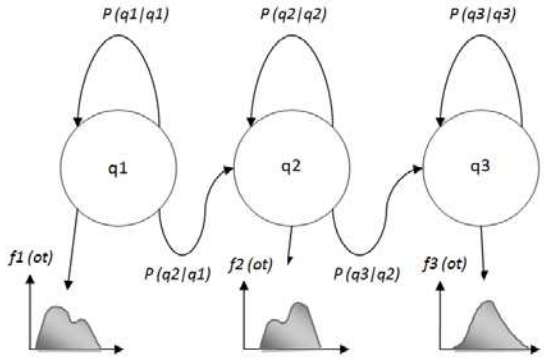

Fig. 3. Schematic diagram of a typical 3 state left-right HMM

Having two recognition model and two reconstruction methods we would present our experiments in 4 sections. In all these experiments the numbers of clusters were 1, 2,8,16 and 32. Finally we will compare the best results of fuzzy and non-fuzzy methods on each recognition model separately.

A. Nonfuzzy Reconstruction Method Evaluation

In the first step for implementing of presented method we evaluate the common cluster-based reconstruction method using introduced Neural Network model and the Hidden Markov Model. First of all recognition of unreconstructed noisy speech was done using both models. Recognition accuracy of the unreconstructed noisy speech is shown by "noisy" label in the all following figures. In the next stage recognition of reconstructed speech data which were reconstructed using that common nonfuzzy cluster-based method for k clusters and k = 1,2,8,16,32 was done. Evaluations are shown in figures 4 and 5 for both frame level model and phoneme level model.

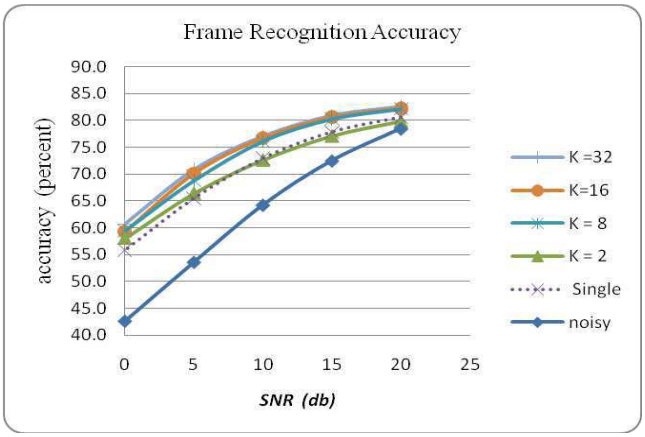

Fig. 4. Improvement of frame-based NN recognition accuracies by increasing the number of clusters in the nonfuzzy cluster-based reconstruction method.

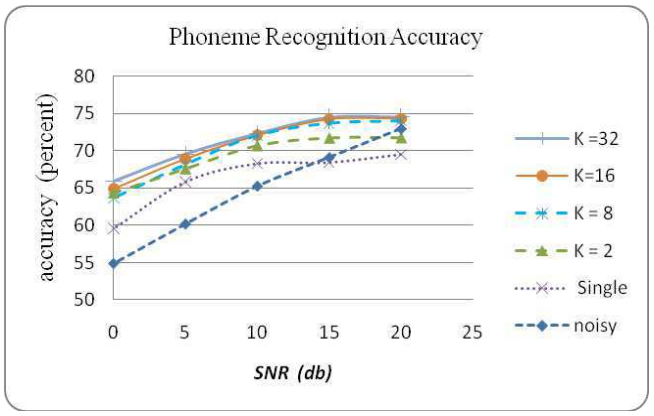

Fig. 5. Improvement of phoneme-based HMM recognition accuracies by increasing the number of clusters in the nonfuzzy cluster-based reconstruction method.

We can see in two above figures that increasing the number of clusters results in the improvement of recognition accuracies. It is clear in figures 4 and 5 that the best recognition accuracies at both levels of frame and phoneme could be derived applying 32 clusters. In the following of this section we will evaluate the recognition results of the presented fuzzy cluster-based reconstruction method.

B. Fuzzy Reconstruction Method Evaluation

Recognition results of presented fuzzy cluster-based reconstructed speech applying frame level NN is shown in figure 6. Similar results by applying phoneme level HMM are shown in figure 7. As can be seen in both figures the effect of increasing the number of clusters is similar to the nonfuzzy cluster-based method. Similar to those methods the best results were derived using 32 clusters.

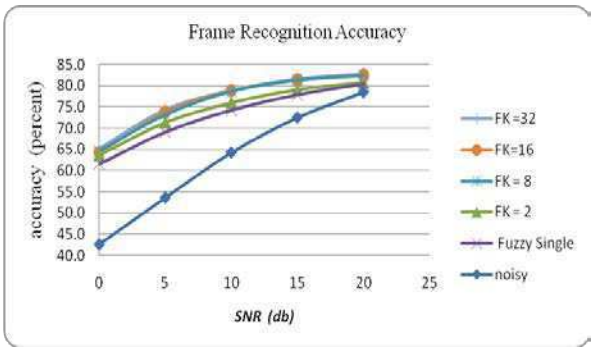

Fig. 6. Improvement of frame-based NN recognition accuracies by increasing the number of clusters in the fuzzy cluster-based reconstruction method.

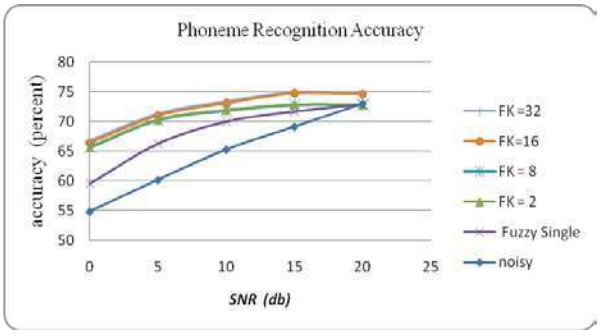

Fig. 7. Improvement of phoneme-based HMM recognition accuracies by increasing the number of clusters in the fuzzy cluster-based reconstruction method.

For a good comparison of the classic nonfuzzy cluster-based reconstruction with the new presented method we have brought figures 8 and 9 for both recognition models. In the figure 8 that is for frame level recognition we can compare the best result of the nonfuzzy cluster-based reconstruction for k=32 and the fuzzy cluster-based reconstruction. In the same figure we can see that the covariance-based results are under both fuzzy and nonfuzzy cluster-based reconstruction methods. Similarly in the figure 9 we can compare the best derived results of fuzzy and nonfuzzy cluster-based reconstruction methods. We can see from figures 8 and 9 that the improvements of recognition accuracies of the suggested fuzzy method with respect to that classic method for the frame-level NN in the low signal to noise ratios are about 5%. These improvements for the phoneme based HMM results are between one and two percents for low SNRs.

CONCLUSION

In this paper we presented a method for improving the fuzzy method for the cluster-based reconstructing instead of fully imputation of missing features are more effective in improving recognition accuracies. Finally we observed that implementing the suggested fuzzy method would lead to considerable improvements of recognition accuracies in both phoneme-based HMM and frame-based NN.

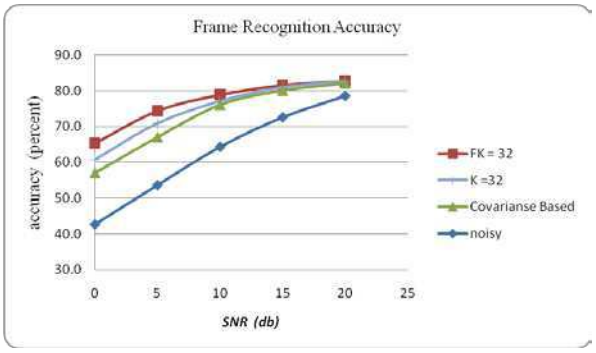

Fig. 8. Frame-based NN Recognition accuracy comparison of Fuzzy Cluster-based, nonfuzzy cluster-based reconstruction using 32 clusters with covariance-based reconstruction method.

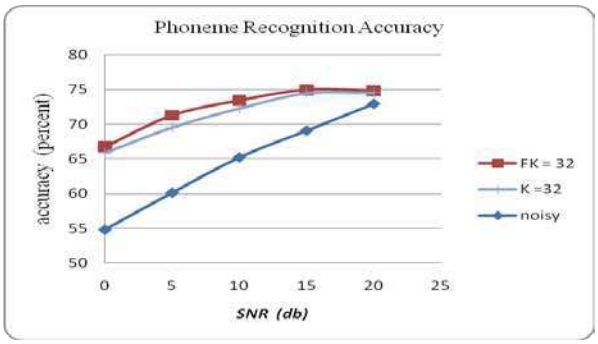

Fig. 9. Phoneme-based HMM Recognition accuracy comparison of Fuzzy Cluster-based with nonfuzzy cluster-based reconstruction using 32 clusters.

REFERENCES

[1] B. Raj, M. L. Seltzer, and R. M. Stern, "Reconstruction of missing features for robust speech recognition," Speech Communication, 43(4): 275-296, 2004

[2] B. Raj, "Reconstruction of incomplete spectrograms for robust speech recognition," Ph.D. dissertation, ECE Dept., Carnegie Mellon Univ., Pittsburgh, PA, Apr. 2000.

[3] W. Kim, R.M. Stern, "Band-independent mask estimation for missing-feature reconstruction in the presence of unknown background noise" in Proc. ICASSP, 2006.

[4] A.H. Hadjahmadi, M.Homayounpour, M.Ahadi "A Neural Network based local SNR estimation for estimating spectral masks", International Symposium on Telecommunications 2008.

[5] A. Mohammadi, F. Almasganj, A. Taherkhani, F. Naderkhani "Using Phoneme Segmentation in Conjunction with Missing Feature Approaches for Noise Robust Speech Recognition", IEEE International Symposium on Signal Processing and Information Technology 2007.

[6] J. Barker, L. Josifovski, M. Cooke, and P. Green, "Soft decisions in missing data techniques for robust automatic speech recognition," in Proc. ICSLP 2000, Beijing, China, Sept. 2000, pp. 373–376.

[7] FARSDAT, Persian speech database, Available from: http://www.elda.org/catalogue/en/speech/S0112.html.