FRONT-END POST-PROCESSING USING HISTOGRAM EQUALIZATION COMBINED WITH ARMA FILTERING FOR NOISE ROBUST SPEECH RECOGNITION

Seyedeh Saloomeh Shariati, Seyed Mohammad Ahadi**, Karim Mohammadi**

* Department of Electrical Engineering, Iran University of Science and Technology, Narmak, 16846_13114, Tehran, Iran.
** Electrical Engineering Department, Amirkabir University of Technology, Hafez Avenue, 15914, Tehran, Iran.
phone: + 98 (21) 6454 3336, fax: + 98 (21) 6640 6469, email: ssaloomeh_shariati@ee.iust.ac.ir, sma@aut.ac.ir, mohammadi@iust.ac.ir

ABSTRACT

In this paper, we present a new feature set for robust speech recognition based on histogram equalization (HEQ) combined with auto regressive moving average (ARMA) filtering. Cepstral vectors extracted from the clean data, modified by Mean and Variance Normalization (MVN) have been used to generate a reference histogram for histogram equalization. The proposed post-processing module also consists of ARMA temporal filtering applied to normalized cepstral coefficients. HEQ compensates for nonlinear distortions caused by noise and ARMA filtering is used for smoothing the normalized feature vectors. The results on the AURORA2 task have shown noticeable improvements in the recognition of noisy speech. The proposed front-end achieved a relative error reduction of around 60% compared to the standard Mel-Cepstral front-end.

1. INTRODUCTION

The accuracy of speech recognition systems degrades when the acoustic mismatch between the training and test data is encountered in real conditions. This is generally caused by the additive or convolutional distortions introduced by background or channel conditions. Therefore, inaccurate training/test data representation is leading to inaccuracy in the performance of the speech recognition system.

Various efforts have been made with the purpose of developing robust speech recognition systems that maintain a high level of recognition in various noisy environments [1]. These compensation methods mainly try to remove the effects of mismatch caused by various conditions in the feature parameters and recognition models. Some of these methods propose noise robust models that are able to more accurately model the noisy speech. While, others try to modify the features extracted from noisy speech to better represent the clean speech, trying to minimize the mismatch between the training and recognition data [2].

Cepstral Mean Subtraction/Normalization (CMS/CMN) and Mean and Variance Normalization (MVN) are two widely used feature compensation methods for reducing the noise effect. They achieve that by trying to eliminate the irrelevant information that is contained in mean and variance of the signal. Cepstral parameter means usually contain the convolutional channel distortion effect that is not relevant and may be assumed invariant for all the frames. MVN normalizes mean and variance information of the speech signal [3]. Although these linear methods are effective for the compensation of channel distortion and some effects of additive noise, they are not able to treat the non-linear effect of additive noise especially in the low SNRs. The application of histogram equalization method has been found helpful for estimating non-linear transformations used for better noise compensation. HEQ uses the assumption that the shape of the entire distribution of the cepstral coefficients is invariant and should be normalized to the reference distribution [3, 4].

ARMA filtering has also been found as an efficient smoothing temporal filter trying to remove outlier (noise) frequencies from the speech modulation spectrum and is proved to enhance the recognition results of the mean-subtracted, variance-normalized features [5, 6].

The organization of this paper is as follows. Basic principles of histogram equalization and its role in compensating non-linear effect of the noise have been described in Section 2. In this section, we also present ARMA temporal filtering and its effect on enhancing the overall results of the speech recognition system. Experimental results are presented and analyzed in Section 3 and conclusions are given in the last section.

2. ALGORITHMS USED IN THE PROPOSED FRONT-END

In this section we discuss the methods that have been applied in our proposed front-end. First, Histogram equalization and its achievement in removing the noise effect on feature parameters distribution are discussed. Since filtering the feature parameters is proved to improve their efficiency, we investigated some temporal filters and found ARMA filter a relatively low cost and efficient filter to be applied to normalized feature vectors. ARMA filter and its use in eliminating the high frequency noise from the feature parameters spectrum are also described in this section.

2.1 Histogram Equalization

The main idea of histogram equalization method is to normalize the distribution of the speech signal to a fixed reference histogram. Hence, by using this method, the effect of

various acoustic environments on probability distribution of the feature vectors is removed and the mismatch between the training and test data is reduced.

The non-linearity introduced by additive noise leads to deformation of the distribution of the feature vector. It highly affects the low SNR part of the speech i.e. deviation from the clean speech distribution increases by reduction of the signal to noise ratio. For normalizing each dimension of the feature vectors that in this work were in cepstral domain, it is convenient to use cumulative density function (CDF) of each coefficient and normalize it to the reference CDF. The reference CDF is obtained by integrating the probability density function (PDF) corresponding to the relevant reference parameter. Therefore, the histogram normalization transformation can be established by applying the rule that for mapping the PDF of a random variable y ($P_y(y)$) to the reference PDF ($P_x(x)$) it is adequate to find a transformation ($x=F(y)$) that satisfies the equality of their cumulative density functions [7]. It can be declared as:

$$C_y(y) = C_x(x) = C_x(F(y)) \tag{1}$$

$$x = F(y) = C_x^{-1}(C_y(y)) \tag{2}$$

The transformation function F can be found as Eq. 2. Here, C_x^{-1} is the inverse of the reference CDF. This reference can be defined in various ways. In fact, it is the representation of the training data and can be extracted from a sufficient amount of training data, or alternatively, be estimated as a normal Gaussian distribution. In this work, the clean training features are used to better approximate the reference PDF that is not precisely identical to the standard Gaussian.

Our approach to normalize histograms of each cepstral coefficient has been carried out as follows: We first established the reference distribution corresponding to one cepstral coefficient with enough frames of speech signal. Before applying this reference, in order to use a unique reference vector for all of the coefficients, the reference feature vector and other cepstral coefficients were normalized by using MVN so that they would have zero means and unit variances. In this condition, the difference between the cepstral coefficients distributions gets negligible. Mean and variance are estimated for current utterance as:

$$\mu = \frac{1}{N}\sum_{n=1}^{N} V^{(n)} \tag{3}$$

and

$$\sigma^2[k] = \frac{1}{N}\sum_{n=1}^{N}\left(V^{(n)}[k] - \mu[k]\right)^2 \tag{4}$$

where $V^{(n)}$ is the feature vector for the n^{th} frame. N is the number of frames in the utterance to be recognized, μ is the mean and $\sigma^2[k]$ is an estimate of the variance of the k^{th} parameter of the feature vector.

Next step in the proposed HEQ algorithm is to equalize the histogram of individual coefficients with the reference histogram. As long as the coefficients are discrete values, a number of bins are to be used for producing the histogram of the reference vector and other coefficients. We considered 100 uniform intervals between the minimum and maximum values of each vector and found the HEQ transformation by minimizing the difference between the histogram of the given coefficient and that of the reference vector. This transformation was applied to each of the 12-cepstral coefficients and logE and the newly extracted features were then passed to the next module, i.e. ARMA filtering.

2.2 ARMA Filtering

ARMA (Auto Regressive Moving Average) filter is a low-pass filter used for smoothing any spikes (higher frequencies) in the time sequence. Therefore, for the noisy speech, ARMA filter most likely removes noise that is laid mainly in spikes appearing among the parameters.
The general formulation of ARMA filter is:

$$\tilde{C}_{td} = \frac{\sum_{i=1}^{M}\tilde{C}_{(t-i)d} + \sum_{j=1}^{M}\bar{C}_{(t+j)d} + \bar{C}_{td}}{2.M+1} \tag{5}$$

where M is the order of the ARMA filter and $\tilde{C}_{td}$ denotes ARMA filter output vector that is obtained in time t for d^{th} cepstral coefficient.

ARMA filter is proved to be useful when integrated with MVN technique for constructing the more robust feature vectors. As mentioned, HEQ is an extension of MVN process that normalizes more than the first two moments (mean and variance) of the probability distribution of feature vectors. Hence, one can anticipate that applying ARMA filter to HEQ-normalized features improves the efficiency of the front-end, especially for lower SNRs. This anticipation has been approved by our experimental results.

Our proposed front-end can be summarized as in Fig.1. According to this figure, MFCC features are first normalized by MVN. HEQ is then applied to the new features and finally the features are filtered by ARMA temporal filtering.

Various M values in the transfer function of the ARMA filter have been examined and the best order of the ARMA filter used after HEQ module has been found to be M=5. We also modified the original formulation and weighted the current value of the vector more than those preceding or succeeding the existing frame. The proposed weighting process is defined in Eq. 6:

$$\tilde{C}_{td} = \frac{W\sum_{i=1}^{M}\tilde{C}_{(t-i)d} + W\sum_{j=1}^{M}\bar{C}_{(t+j)d} + \bar{C}_{td}}{2.W.M+1} \tag{6}$$

It was shown that using the modified ARMA filter would lead to a better performance with respect to the regular ARMA filter defined in Eq. 5.

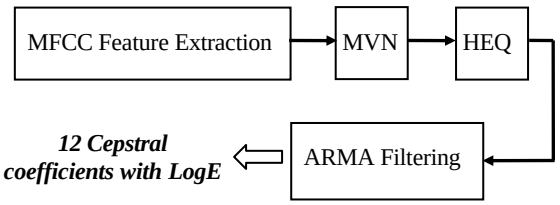

Figure 1– Block diagram of the proposed front-end for robust feature extraction.

3. EXPERIMENTAL RESULTS

3.1 Experimental Setup

The proposed front-end was evaluated on Aurora 2 Task [8]. This is a speaker-independent connected-digit recognition task that contains speech contaminated with 8 various types of additive noise in 6 different SNRs ranging from -5 dB to 20 dB and clean speech. There are three test sets in the database: Set A (with subway, babble, car, and exhibition noises added to speech), set B (with restaurant, street, airport and train station noises) and set C. In set C the channel effect is also added to the additive noises (subway and street). MFCC features including 12 cepstral coefficient and log energy have been obtained from Aurora 2 front-end. The dynamic parameters have been appended to these parameters later. The Hidden Markov model Toolkit (HTK) has been used for speech recognition evaluations [9]. Recognition experiments have been carried out using a set of continuous density left-right Hidden Markov Models (HMMs) trained with clean speech data. Digit models had 16 states with mixtures of 3 Gaussians per state.

3.2 Results with Different Front-End Configurations

As depicted in the block diagram in Fig.1, first, the log energy and 12 cepstral coefficients were modified with the proposed method. Normalized cepstral coefficients were then passed to the back-end module where 1st order and 2nd order time derivatives of them were also used for training and recognition processes. To compare the performance of the mentioned methods, different front-end experiments have been carried out: Baseline, MVN, MVA (MVN+ARMA), HEQ and HEQ+ARMA. We apply the utterance to be recognized to estimate mean and variance for MVN and histograms for HEQ. Normalization of feature components has been performed for both the training and test data in all experiments.
Histogram equalization was only applied to the 12 cepstral coefficients (C_0-C_{12}) appended with log energy. The first and the second derivatives were obtained from HEQ-normalized features during the recognition process.
We investigated different bin numbers for either evaluating the histograms of the reference and test vectors or mapping each coefficient's histogram to the reference. Number of bins was selected as 100 according to the best empirical results. We also concluded from these results that selecting reference histogram and method with which HEQ is implemented highly affected the performance of the algorithm.

ARMA filter was also applied to mean-variance-normalized and histogram-compensated features. Fig. 2 shows the recognition rates obtained by various front-end configurations for different noise levels.

Table 1 – Comparison of different feature extraction method for Aurora 2 Task.

	Recognition Accuracy			
	Set A	Set B	Set C	Ave.
Baseline	61.12	55.57	66.68	61.12
MVN	70.43	71.13	66.75	69.44
MVN+ARMA *M*=2	75.79	76.01	72.48	74.76
MVN+HEQ	80.28	81.62	81.58	81.16
MVN+HEQ+ARMA *M*=5	80.80	82.43	81.67	81.63
MVN+HEQ+ARMA *M*=5, *W*=.8	80.81	82.46	81.74	81.67

According to the procedure followed in the Aurora evaluation framework, we have calculated the average recognition accuracy for each test set from the results corresponding to the SNRs of 0dB to 20dB excluding the clean and -5dB recognition results. The average results obtained from mentioned methods are shown in Table 1.

As depicted in this table, ARMA filter has further improved the performance of MV-Normalized features, which is the result of smoothing the spectral spikes that mostly represent the additive noise. However, in increasing the order of ARMA filter, *M*, a trade-off should be reached between missing spectral peaks corresponding to short-term cepstral information and eliminating the effect of additive noise. Thus a small *M* retains more speech information together with more corrupting noise. It suggests using a relatively small positive *M*. According to our experiments, the best results have been obtained by *M*=5.

In order to obtain better performance, especially for higher SNRs, weighing modification is applied to the ARMA filter. *W*=.8 was found to be the best weighting coefficient. The results of histogram equalization (HEQ) outperforms the jointly use of MVN+ARMA as shown in Table 1. Moreover, adding ARMA filter to the features compensated with HEQ caused a significant improvement with respect to ARMA+MVN feature compensation method. This is due to this fact that HEQ is the inherent extension to MVN and can compensate for the non-linear effect of additive noise that cannot be dealt with by linear normalization methods such as MVN. Although not significantly, ARMA filtering can further improve the performance of histogram equalization as shown in Table 1.

4. CONCLUSION AND FUTURE WORKS

Based on our experimental results, it is found that HEQ performs better compared to other normalization methods such as CMS and MVN for the reason that it provides compensation of rather than first two moments (mean and variance) of the distributions. However, the computational load is higher and may be compensated by either improving the HEQ algorithm or using more powerful processors.

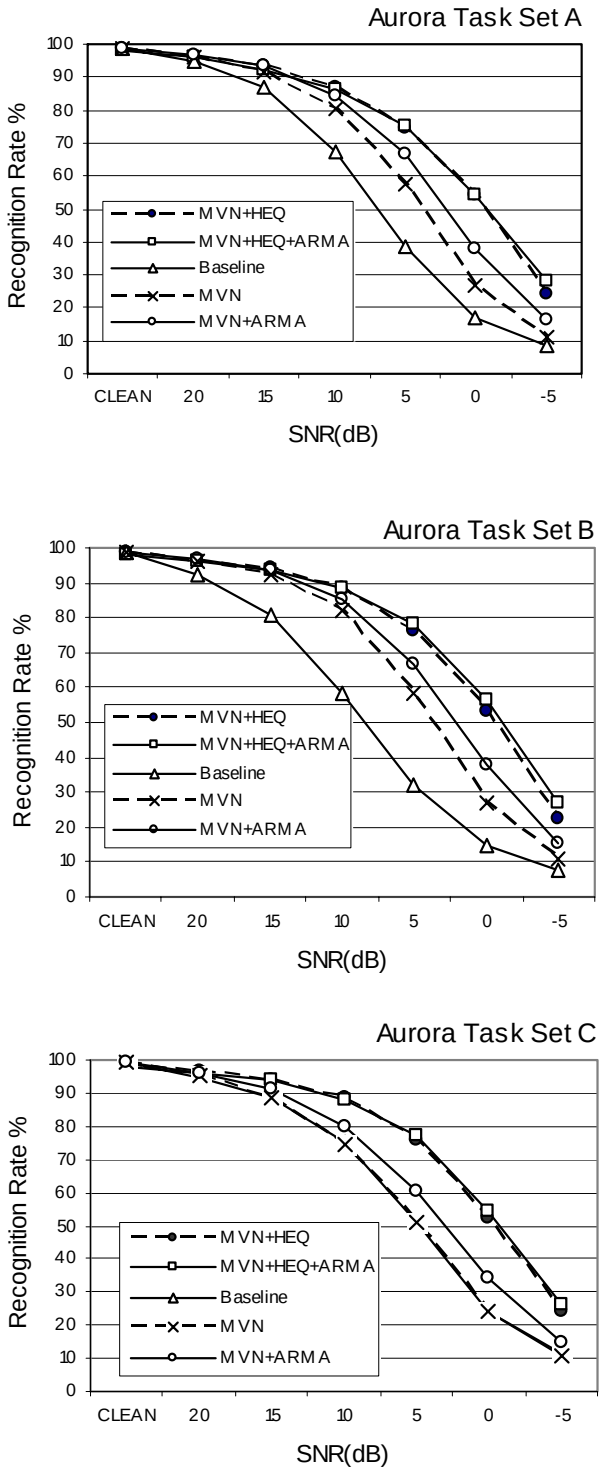

Figure 2 – Aurora 2 recognition results for sets A, B and C. Recognition accuracy is displayed as a function of the SNR under various noise compensation methods for robust feature extraction.

In this work, histogram equalization combined with ARMA filtering is shown to improve the performance of the front-end in comparison to using MVN or MVN+ARMA.

The achievement of our proposed method is because of independent effects of the mixed algorithms. It means they perform different noise compensation for feature parameters. MVN is mainly responsible for removing the DC component of the feature vectors spectrum and reducing the effect of convolutional noise. HEQ can compensate the adverse effect of additive noise on feature parameters distribution. ARMA filter is efficient to reduce the contribution of high frequency noise in the feature parameters spectrum. Hence, it can be observed that their combination have led to improvement of the recognition accuracy in the presence of various environmental noise.

Since energy distribution is different from other cepstral coefficients, further works on energy normalization may improve the overall results of the HEQ compensation. Another field that requires more investigation is to find more appropriate reference histograms able to better represent most of the feature vector properties. We can also enhance the ARMA filter to be more efficient to remove the HEQ normalized features corresponding to noise.

ACKNOWLEDGMENT

This work was in part supported by Iran Telecommunication Research Center (ITRC).

REFERENCES

[1] R. M. Stern, B. Raj, and P.J. Moreno, “Compensation for environmental degradation in automatic speech recognition,” in *Proc. ESCA_NATO Tutorial Res. Workshop Robust Speech Recognition Unknown Communication Channels*, Apr. 1997, pp. 33-42.

[2] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Perez, M. C. Benitez, and A. Rubio, “Histogram equalization of the speech representation for robust speech recognition,” *IEEE Trans. On Acoustic, Speech and Signal Processing*, vol. 13, no. 3, pp.355–366, 2005.

[3] A. de la Torre, J. C. Segura, M. C. Benitez, A. M. Peinado, and A. J. Rubio, “Non-linear transformations of the feature space for robust speech recognition,” in *Proc. ICASSP 2002*, Orlando, Florida, May 2002, pp. 401–404.

[4] Y. Obuchi, R. M. Stern, “Normalization of time-derivative parameters using histogram equalization,” in *Proc. EUROSPEECH 2003*, Geneva, Switzerland, Sep 2003, pp.665-668.

[5] C-P. Chen, J. Bilmes & D.P.W. Ellis, “Speech feature smoothing for robust ASR,” in *Proc. ICASSP 2005*, Philadelphia, March 2005, pp. I-525-528.

[6] C-P. Chen and J. Bilmes, “MVA processing of speech features,” *University of Washington Electrical Engineering Technical Report*, UWEETR-2003-0024.

[7] J. C. Segura, M. C. Benitez, A. de la Torre, and A. J. Rubio, “Cepstral domain segmental nonlinear feature transformations for robust speech recognition,” *IEEE Signal Processing Letters*, vol. 11, no. 5, pp. 517–520, 2004.

[8] H-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. ISCA ITRW ASR-2000*, Paris, France, Sept. 2000, pp.181–188.

[9] The hidden Markov model toolkit available from http://htk.eng.cam.ac.uk.