Robust Histogram Equalization Using Compensated Probability Distribution

Sungtak Kim(ICU), Hoirin Kim(ICU)

<차 례>

<Abstract>

Robust Histogram Equalization Using Compensated Probability Distribution

Sungtak Kim, Hoirin Kim

A mismatch between the training and the test conditions often causes a drastic decrease in the performance of the speech recognition systems. In this paper, non-linear transformation techniques based on histogram equalization in the acoustic feature space are studied for reducing the mismatched condition. The purpose of histogram equalization(HEQ) is to convert the probability distribution of test speech into the probability distribution of training speech. While conventional histogram equalization methods consider only the probability distribution of a test speech, for noise-corrupted test speech, its probability distribution is also distorted. The transformation function obtained by this distorted probability distribution maybe bring about miss-transformation of feature vectors, and this causes the performance of histogram equalization to decrease. Therefore, this paper proposes a new method of calculating noise-removed probability distribution by using assumption that the CDF of noisy speech feature vectors consists of component of speech feature vectors and component of noise feature vectors, and this compensated probability distribution is used in HEQ process. In the AURORA-2 framework, the proposed method reduced the error rate by over 44% in clean training condition compared to the baseline system. For multi training condition, the proposed methods are also better than the baseline system.

* Keywords : Histogram equalization, Cumulative density function, Robust speech recognition.

1. Introduction

Noise robustness is one of major requirements for practical automatic speech recognition systems. Typically, a high recognition accuracy can be obtained if there is a good matching between the training and recognition conditions. In real world applications, however, the acoustically same training and recognition environments can not be guaranteed.

The easiest and straightforward way to improve noise robustness is to attempt to collect large amount of data from a wide rage of acoustic environments. Although this so-called multi-environment training works reasonably well up to a certain limit, it is obvious that the problem of noise robustness cannot be solved by simply collecting a huge amount of training data. It is impossible to collect a database which covers all possible usage environments. Furthermore, the use of large amounts of data often leads to HMMs with large variances, and hence, the models do not provide a high recognition accuracy in arbitrary environment. In addition to multi-environment training, more sophisticated techniques have also been developed for improving the noise robustness of speech recognition systems[1].

The methods proposed to make the speech recognition systems more robust against the noise are mainly focused on the minimization of the mismatch caused by noise. There are two broad categorized methods in order to overcome this problem. Some of them so called normalization try to reduce the mismatch by transforming the acoustic vectors. Another method based on adaptation tries to adapt the recognizer to noise conditions in order to match the noisy speech representation with noise models.

The noise causes a distortion of the feature space which usually presents a non-linear behavior. For instance, cepstral based representations suffer non-linear distortions when the speech signal is affected by additive noise. Linear normalization methods which are CMN[2] (Cepstral Mean Normalization) and CMVN[1] (Cepstral Mean and Variance Normalization) provide significant improvement for cepstral based representation. However, these linear normalization methods have critical limitations due to the non-linear distortion of noise. Methods oriented to compensation of the noise effects over the speech representation such as widely used MFCC should consider non-linear effects and should be able to estimate the non-linear transformation providing the best estimation of the clean speech given the noisy speech. Several histogram equalization-based approaches have been proposed in order to overcome this non-linear distortion[3]-[5]. The main characteristic of HEQ is that this can compensate higher moments in comparison with CMN and CMVN which compensate only two

moments that are mean and variance.

The aim of HEQ-based methods is to find a transformation which converts the probability distribution of a test speech into the probability distribution of training speech. However, conventional histogram equalization methods consider only the probability distribution of a test speech, but if a test speech is corrupted by noise, the probability distribution of test feature vectors is distorted. This distorted probability distribution causes the performance of histogram equalization to decrease. Hence, if the distortion caused by noise is compensated in histogram equalization, the performance of histogram equalization is better than conventional one. Therefore, we propose a new histogram equalization method that uses a compensated probability distribution of a test speech feature vectors by removing component of noise feature vectors from the CDF value of noisy speech feature vectors.

This paper is organized as follows. In Section 2, we discuss the non-linear transformations which is based on histogram equalization. Section 3 is devoted to the proposed method which uses the compensated probability distribution which is noise-removed probability distribution. In Section 4, experimental results are presented. Finally, Section 5 presents conclusion.

2. Non-Linear Transformation

2.1 Histogram Equalization

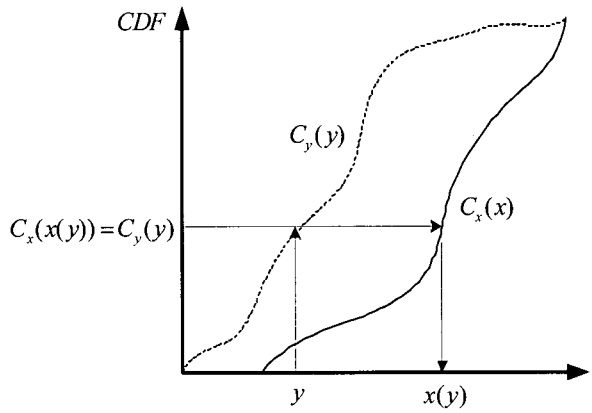
In order to effectively compensate the non-linear effects, histogram equalization (HEQ) techniques have been proposed[6]-[8]. The aim of these methods is to find a transformation which converts the probability distribution of a test speech into the probability distribution of training speech. It can be demonstrated that, if $x(y)$ transforms $p_y(y)$ into $p_x(x)$, then the cumulative histograms verify that

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

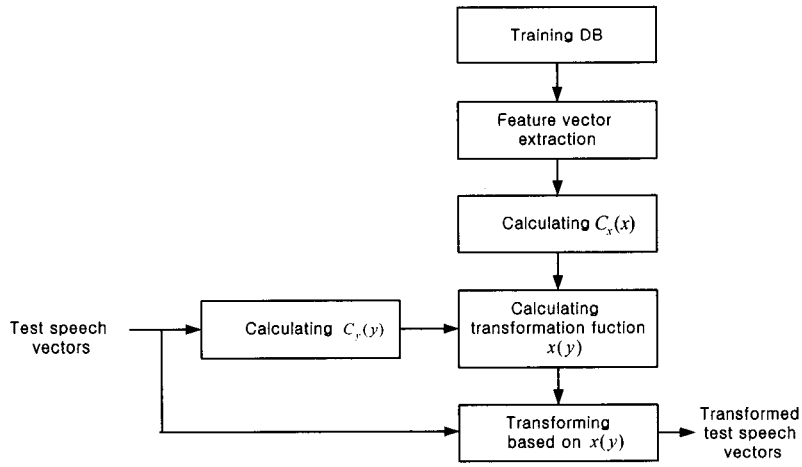
and therefore the transformation can be obtained from the target cumulative histogram of feature vectors of test speech and the reference cumulative histogram for feature vectors of the training speech as

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

The cumulative histogram-based transformation and block diagram are depicted in <Fig. 1> and <Fig. 2>.



<Figure 1> Principle of histogram equalization: Test data y are transformed such that the CDF $C_y(y)$ of test data matches the CDF $C_x(x)$ of training data



<Figure 2> Block diagram of histogram equalization

2.2 Order Statistic-Based CDF [9]

For the histogram equalization, the cumulative distribution function (CDF) is computed at every test utterance. More efficient algorithms can be formulated by exploiting the relation between order statistics and the values of CDF. This relation can be straightforwardly applied to construct a sample based estimate of the CDF.

Especially when the number of sample data is small as a test utterances, this relation can be useful.

At a utterance, the time sequence of cepstral coefficients in a particular dimension is considered

$$Y=\{y_1, y_2, \ldots, y_T\} \tag{3}$$

Let us denote the order statistics of (3) by

$$y_{(1)} \leq y_{(2)} \leq \cdots \leq y_{(t)} \leq \cdots \leq y_{(T)} \tag{4}$$

Using the order statistics, an asymptotically unbiased point estimation of CDF can be defined [10] as

$$C_y(y)=\frac{r(y)-0.5}{T} \qquad \forall r(y)=1,2,\cdots T \tag{5}$$

using (5) and (2), an estimation of transformed value $x(y)$ can be obtained as

$$x(y)=C_x^{-1}[C_y(y)]=C_x^{-1}\left[\frac{r(y)-0.5}{T}\right] \tag{6}$$

where $r(y)$ denotes the rank of y that is obtained by counting the number of values less than or equal to y in T values of cepstral coefficients. In this paper, the CDFs of test data are calculated by using oder statistic-based CDF computation method.

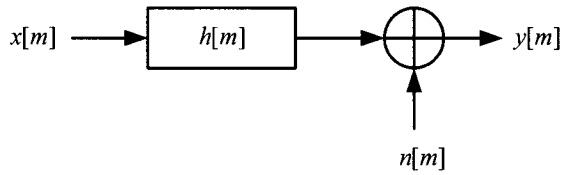
3. Compensated Probability Distribution

If a test speech is corrupted by noise, the probability distribution of a test speech feature vectors is also distorted. The performance of HEQ which uses this distorted probability distribution is decreased because this distorted distribution causes transformation function of HEQ to be incorrect. If this distorted probability distribution can be compensated, the HEQ with the compensated probability distribution has better performance than conventional HEQ. To compensate this distorted probability

distribution, we will expand a model of acoustical environment to CDF domain, and compensate the distorted CDF by using simple assumption within the framework of order statistic-based CDF computation.

<Figure 3> shows a commonly used model for the acoustical environment, which assumes the speech signal $x[m]$ is corrupted by additive noise $n[m]$ and channel noise $h[m]$

$$y[m] = x[m] * h[m] + n[m] \tag{7}$$



<Figure 3> A model of the acoustical environment

Equation (7) can be also presented using power spectral density (PSD)

$$|Y(w)| = |X(w)||H(w)|^2 + |N(w)| \tag{8}$$

where $|X(w)|$, $|H(w)|^2$, $|N(w)|$ and $|Y(w)|$ are PSD of $x[m]$, $h[m]$, $n[m]$ and $y[m]$, respectively. Taking natural logarithms on equation (8), we get

$$\begin{aligned} \log|Y(w)| &= \log[\,|X(w)||H(w)|^2 + |N(w)|\,] \\ &= \log|X(w)| + \log|H(w)|^2 + \log\left[1 + \frac{|N(w)|}{|X(w)||H(w)|^2}\right] \end{aligned} \tag{9}$$

For brevity, we use x_l, h_l, n_l and y_l instead of $\log|X(w)|$, $\log|H(w)|^2$, $\log|N(w)|$ and $\log|Y(w)|$, where subscript "l" denotes log-domain. After some algebraic manipulation, equation (9) is

$$y_l = x_l + h_l + \log\{1 + \exp(n_l - x_l - h_l)\} \tag{10}$$

In a similar way, we can present noisy environment in cepstral domain[11].

$$y_c = x_c + h_c + C\log\{1 + \exp(C^{-1}(n_c - x_c - h_c))\} \tag{11}$$

where subscript "c" representing cepstral domain, C denoting DCT matrix, and C^{-1} being inverse DCT matrix. From above observations, we find that corrupted speech signal (or feature vector) y, regardless of its domain, is can be expressed as

$$y = x + f(x, n, h) \tag{12}$$

We will apply the equation (12) to the CDF of feature vectors. We can assume that the CDF of noisy speech feature vectors consists of component of speech feature vectors and component of non-linear part of speech, additive noise, and channel distortion by using equation (12).

$$C_y(y) = \widehat{C}_x(y) + \widehat{C}_f(y) \tag{13}$$

There have been many methods to get noise information in cepstrum domain, but, instead of using these sophisticate methods, we use a simple assumption that a few frames in front of speech frames consist of channel distortion and additive noise, and feature vectors from these frames are also feature vectors of channel distortion and additive noise. If noise information is obtained by using above assumption, the obtained $\widehat{C}_f(y)$ is not related with speech but related with additive noise and channel distortion. So $\widehat{C}_f(y)$ can be estimated by using the distribution of the ordered noise feature vectors and the above simple assumption as the following equation.

$$\begin{aligned} \widetilde{C}_{Compen.}(y) = \widetilde{C}_x(y) &= C_y(y) - \widetilde{C}_f(y) \\ &= \frac{r(y) - 0.5}{T} - \frac{N_{below}(y)}{T} \end{aligned} \tag{14}$$

where $N_{below}(y)$ is the number of noise features less than y. Within the framework of order statistic-based CDF computation, $N_{below}(y)$ can be considered as the noise factor that influences the CDF of clean speech feature vectors. Therefore, in equation (8), $\widetilde{C}_{Compen}(y)$ is noise-removed CDF of noisy speech feature vectors, y, because

the noise factor is expelled from the CDF of noisy speech feature vectors. If the $\widetilde{C}_{Compen.}(y)$ is used in stead of $C_y(y)$, the performance of HEQ will be increase compared with conventional HEQ. Finally, transformation function for HEQ can be obtained as

$$x_{robust}(y) = C_x^{-1}\left[\frac{r(y)-0.5}{T} - \frac{N_{below}(y)}{T}\right] \tag{15}$$

4. Experimental Result

In this paper, four feature vector normalization methods have been compared in recognition experiments under noise conditions using AURORA-2 database and task[12]. The task consists of the recognition of connected digits in English. The speech is artificially corrupted at several SNRs with ten different conditions. The recognition results at each SNR are averaged over all the kinds of noise.

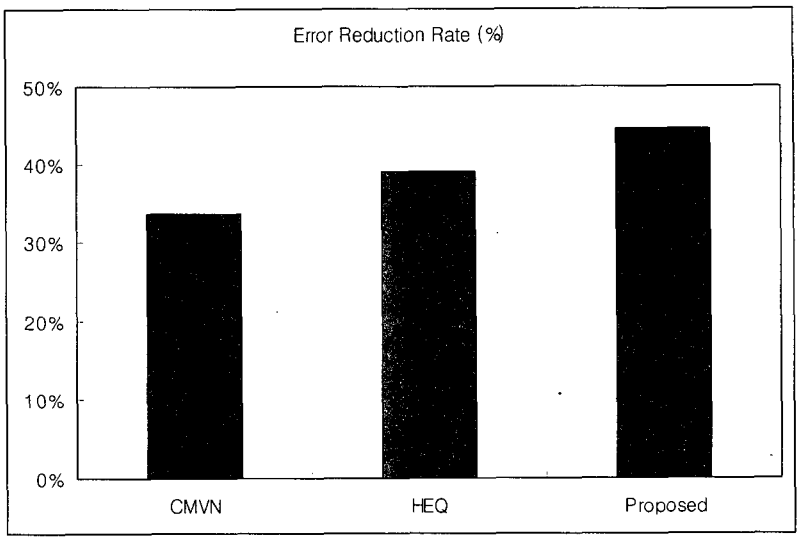
Speech representation is based on MFCC parameterization. The speech signal, sampled at 8kHz, is segmented into frames and each frame is represented as a feature vector containing 39-components feature vector. This feature vector includes 12 MFCC plus the energy and the corresponding delta and acceleration coefficients. Features are extracted at a frame-rate of 10ms with 25ms frame size. Continuous density left-to-right HMMs are acoustic models. Digits are modeled with 16 emitting states and a three Gaussian mixtures per states. There are also two pause models. The first one consists of the three states with six Gaussian mixtures per state, and models beginning and end pauses. The second one models inter digit pause and has only one state.

In order to apply the four normalization methods, each transformation is applied to each component of cepstral vector. CMN is based on the estimation of the mean for each component, and CMVN is based on the estimation of the mean and the variance for each component. In the case of HEQ-based methods, the considered reference probability function is a Gaussian probability distribution with zero mean and unit variance, and we use first 20ms of test utterance in order to get noise information. For these normalization methods, the estimations of the mean, the variance, and the cumulative density function are obtained using all the frames in a utterance. In the

case of all normalization methods, normalization is applied for both training and recognition.

<Table 1> Word accuracies (%) in clean training condition

Method / dB	Baseline	CMN	CMVN	HEQ	Proposed
Clean	98.94	99.05	98.99	99.06	99.03
20	94.45	97.64	97.07	97.32	97.45
15	85.35	88.14	94.47	94.91	95.54
10	65.21	67.18	87.53	89.39	90.40
5	38.91	36.19	71.41	75.67	78.62
0	17.10	15.21	41.45	46.97	53.20
-5	8.04	8.36	15.85	18.94	24.26

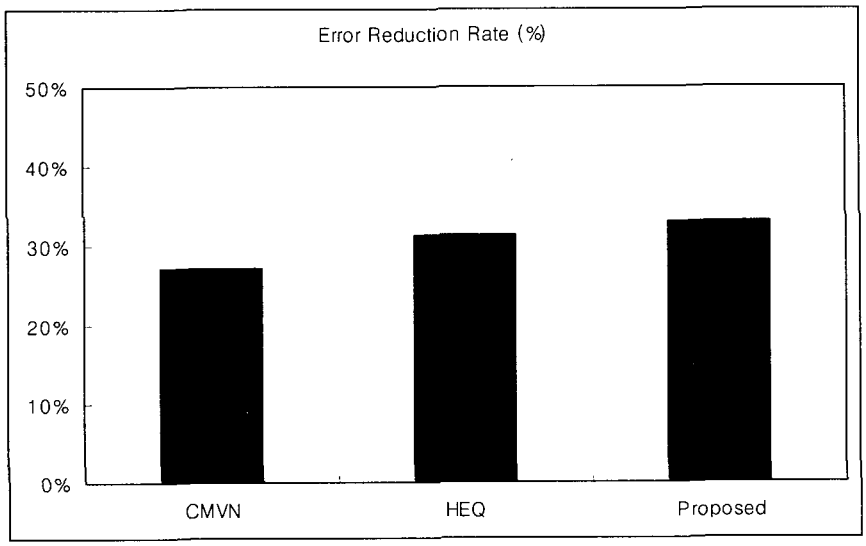


<Figure 4> ERRs in clean training condition

In clean training condition, the ERR of proposed method is 44.6% with respect to baseline system which is without any normalization techniques. In multi training condition, the proposed method also has the higher recognition performance than the conventional HEQ method even though we used a simple noise estimation method. In multi training condition, since the multi training condition already guarantees the matched condition between training and test, the proposed methods which are to reduce the mismatch don't have much better performance improvement than that of clean training condition.

<Table 2> Word accuracies (%) in multi training condition

Method / dB	Baseline	CMN	CMVN	HEQ	Proposed
Clean	98.90	99.04	98.75	98.74	98.77
20	97.64	96.61	98.15	98.15	98.07
15	96.30	96.18	96.99	96.86	97.05
10	92.89	92.67	94.65	94.56	94.69
5	78.93	82.10	87.08	88.10	88.32
0	43.21	51.65	66.70	68.67	69.47
-5	18.08	20.46	31.01	35.47	37.01



<Figure 5> ERRs in multi training condition

<Table 3> shows the performance of a conventional HEQ and the proposed HEQ in the case of using feature vectors which are not compensated with mean and variance in clean training condition.

From <table 3>, we can also know that the proposed method has better performance than conventional method in the case of using non-compensated feature vectors.

<Table 3> Word accuracies (%) in the case of using non-compensated feature vectors in clean training condition

Method / dB	HEQ	Proposed
Clean	98.76	98.77
20	97.18	97.10
15	94.65	95.17
10	88.95	90.06
5	75.67	78.41
0	48.23	54.18
-5	19.75	24.21

5. Conclusion

In this paper, we have studied several linear and nonlinear transformation in cepstral domain. Compared with the linear transformations, the performances of the non-linear transformation methods is more efficient under mismatched conditions. Even though simple noise estimation method is used, the proposed method is effective on noise environment.

For clean training condition, the proposed method performs better than the conventional HEQ because the compensated CDF removes the effects of noise from the CDF of noisy speech feature vectors. In multi training condition, the proposed methods have also the best recognition performance. From the results, we can conduct that the proposed non-linear transformation techniques give noise robust transformation methods.

References

[1] O. Viikki, K. Laurila, "Cepsral domain segmental feature vector normalization for noise robust speech recognition", *Speech Communication*, Vol. 25, pp. 133-147, 1998.

[2] C. R. Jankowski, Jr. Hoang-Doan, and R. P. Lippmann, "A comparison of signal processing front ends for automatic word recognition", IEEE Trans. *Speech and Audio Processing*, Vol. 3, No. 4, pp. 286-293, 1995.

[3] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition", *in Poc. ICSLP*, pp. 556-559, 2000.

[4] F. Korkmazsky, D. Fohr, I. Illina, "Using Linear Interpolation to Improve Histogram

Equalization for Speech Recognition", in *Proc. ICSLP*, pp. 2082-2092, 2004.

[5] Y. Obuchi, "Improved Histogram-Based Feature Compensation for Robust Speech Recognition and Unsupervised Speaker Adaptation", *in Proc. ICSLP*, pp. 2065-2068, 2004.

[6] Y. Obushi, and R. M. Stern, "Normalization of Time-Derivative Parameters Using Histogram Equalization", *in Proc. Eurospeech*, pp. 665-668, 2003.

[7] J. C. Segura, M. C. Bentez, "Feature Extraction Combining Spectral Noise Reduction and Cepstral Histogram Equalization for Robust Speech Recognition", *in Proc. ICSLP*, pp. 225-228, 2002.

[8] Á. de la Torre, J. C. Segura *et al.*, "Non-Linear Transformations of the Feature Space for Robust Speech Recognition", *in Proc. ICASSP*, pp. 401-404, 2002.

[9] J. C. Segura, C. Bentez, Á. de la Torre, J. Rubio, and J. Ramrez, "Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition", *IEEE Signal Processing Letters*, Vol. 11, No. 5, pp. 571-520, May, 2004.

[10] R. Suoranta, K.-P. Estola *et al.*, "PDF estimation using order statistic filter bank", *in Proc. ICASSP*, pp. 625-628, 1994.

[11] D. Y. Kim, C. K. Un and N. S. Kim, "Speech Recognition in noisy environments using first-order vector Taylor series", *Speech Communication*, Vol. 24, pp. 39-49, 1998.

[12] H. Hirsch and D. Pearce, "The AURORA experimental framework for the performance evaluation of speech recognition systems under noise conditions", *In Proc. ARS*, pp. 181-188, 2000.

접수일자 : 2005년 8월 10일
게재결정 : 2005년 9월 23일

▶ 김성탁(Sungtak Kim)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교 공학부 음성인식기술연구실
전화: 042) 866-6221
Fax: 042) 866-6245
E-mail: stkim@icu.ac.kr

▶ 김회린(Hoirin Kim)
주소: 305-732 대전광역시 유성구 문지로 119번지 한국정보통신대학교
소속: 한국정보통신대학교 공학부 음성인식기술연구실
전화: 042) 866-6139
Fax: 042) 866-6245
E-mail: hrkim@icu.ac.kr