# Class-Based Histogram Equalization for Robust Speech Recognition

Youngjoo Suh and Hoirin Kim

*ABSTRACT—A new class-based histogram equalization method is proposed for robust speech recognition. The proposed method aims at not only compensating the acoustic mismatch between training and test environments, but also at reducing the discrepancy between the phonetic distributions of training and test speech data. The algorithm utilizes multiple class-specific reference and test cumulative distribution functions, classifies the noisy test features into their corresponding classes, and equalizes the features by using their corresponding class-specific reference and test distributions. Experiments on the Aurora 2 database proved the effectiveness of the proposed method by reducing relative errors by 18.74%, 17.52%, and 23.45% over the conventional histogram equalization method and by 59.43%, 66.00%, and 50.50% over mel-cepstral-based features for test sets A, B, and C, respectively.*

*Keywords—Acoustic feature compensation, class-based histogram equalization, robust speech recognition.*

## I. Introduction

The performance of speech recognition systems degrades severely when employed in acoustically mismatched environments, compared to when they are used in training environments. The main cause of this acoustic mismatch is corruption by additive noise and channel distortion. In this issue of robust speech recognition, the feature space-based approach has been widely employed due to such advantages as easy implementation, low computational complexity, and effective performance improvements. Acoustic environments corrupted by additive noise and channel distortion act as a nonlinear transform in the feature spaces of the cepstrum or log-spectrum [1]. Thus, classical linear feature space-based methods such as cepstral mean subtraction or cepstral mean and variance normalization have substantial limitations even though they yield significant performance improvements under noisy environments [2], [3].

As an alternative approach to coping with the drawbacks of linear transform-based methods, the histogram equalization (HEQ) technique has been employed for compensating the acoustic mismatch. While HEQ was originally used in image processing applications, recent researches have shown that it is also effective in preventing performance degradation in speech recognition systems under noisy environments [4]-[7]. The basic assumptions of HEQ in the compensation of the acoustic mismatch can be summarized as follows. First, the phonetic or acoustic class distributions in the feature vector should be identical for the training and test data. Then, the acoustic mismatch causes a nonlinear transformation of the test distributions. Finally, matching the training and test distributions provides the inverse transformation to reconstruct a clean speech representation. Therefore, the validity of HEQ can be weakened if the phonetic or acoustic class distributions of the training and test data are not identical. In this case, the inverse transformation tends to impair the class separability of features by confusedly mapping to the regions of other phonetic or acoustic classes. However, in many speech recognition applications, test speech utterances can be too short to make their phonetic or acoustic class distributions identical or similar to those of the training data. In this condition, the first assumption of HEQ becomes invalid. As a result, it is difficult to take full advantage of HEQ for compensating the acoustic mismatch in noisy environments. Here, we call the discrepancy between the two distributions a phonetic mismatch.

In this letter, we propose a new, class-based HEQ (CHEQ)

Manuscript received Jan. 09, 2006; revised Mar. 20, 2006.

Youngjoo Suh (phone: +82 42 866 6221, email: yjsuh@icu.ac.kr) and Hoirin Kim (email: hrkim@icu.ac.kr) are with the School of Engineering, Information and Communications University, Daejeon, Korea.

technique to reduce this limitation of the conventional HEQ. Instead of utilizing the global reference and test cumulative distribution functions (CDFs) as in the conventional HEQ, the proposed method employs multiple class-specific CDFs not only to compensate for the acoustic mismatch but also to reduce the phonetic mismatch between training and test data. The experimental results show that the proposed CHEQ is superior to the conventional HEQ for improving speech recognition accuracy in noisy environments.

## II. Conventional Histogram Equalization

The basic idea of HEQ is to convert the probability density function (PDF) of the original variable into its reference PDF. For a given random test variable $y$, whose PDF is given as $p_Y(y)$, the inverse transform function $x = F(y)$ mapping $p_Y(y)$ into reference PDF $p_X(x)$ can be given as in [5] as

$$x = F(y) = C_X^{-1}(C_Y(y)), \tag{1}$$

where $C_X^{-1}(x)$ is the inverse of reference CDF $C_X(x)$, and $C_Y(y)$ is the test CDF of the random variable $y$.

One of the critical issues in HEQ is the reliable estimation of reference and test CDFs. In speech recognition applications, the amount of training data is usually large enough that the reference CDFs can be reliably estimated by cumulative histograms. However, when short utterances are used as test data, the length of each utterance may be insufficient for a reliable estimation. Accordingly, test CDF estimation becomes much more important in these test environments. When the amount of estimation samples is small, the order-statistic-based CDF estimation is known to be more accurate than the cumulative histogram-based method, and its brief description is as follows [5].

Let us define a sequence consisting of $N$ frames of a particular feature component as

$$V_k = \{y_k(1), y_k(2), \Lambda\ , y_k(n), \Lambda\ , y_k(N)\}, \tag{2}$$

where $y_k(n)$ is the $k$-th feature component at the $n$-th frame.

The order statistics of (2) can be represented as

$$y_k([1]) \le \Lambda\ \le y_k([r_k]) \le \Lambda\ \le y_k([N]), \tag{3}$$

where $[r_k]$ denotes the original frame index of feature component $y([r_k])$ in which its rank is $r_k$ when the elements of sequence $V_k$ are sorted in ascending order.

The order-statistic-based estimate of test CDFs is given as

$$\hat{C}_{Y(k)}(y_k([r_k])) = \frac{r_k - 0.5}{N},\ 1 \le r_k \le N,\ 1 \le k \le K, \tag{4}$$

where $K$ denotes the dimension of the feature components.

An estimate of $x_k(n)$ by transforming $y_k(n)$ with the HEQ is obtained as

$$\begin{aligned}\hat{x}_k(n) &= C_{X(k)}^{-1}(\hat{C}_{Y(k)}(y_k(n))) \\ &= C_{X(k)}^{-1}\left(\frac{r_k(y_k(n)) - 0.5}{N}\right).\end{aligned} \tag{5}$$

## III. Class-Based Histogram Equalization

The proposed approach to reducing both acoustic and phonetic mismatches consists of utilizing multiple class-specific CDFs at both the reference and test sides. By dividing the global distributions into sets of multiple class distributions, classifying feature components into their classes and then transforming them using their corresponding class CDFs, a phonetic mismatch can be effectively reduced because of the increased similarity between the reference and test distributions of the same class. In this approach, however, reliably assigning class information to each feature component is a prerequisite condition for providing the validity of CHEQ. In most HEQ methods, the equalization is applied to each feature component. In this respect, phonetic classification can be performed on a feature component basis. However, utilizing a feature vector instead of only a specific feature component is more useful in phonetic classification and is adopted in the proposed method. Nevertheless, it may still be a critical problem to accurately classify feature vectors into their corresponding phonetic classes in noisy environments. In this sense, a histogram equalized feature vector is used in the classification instead of the original noisy feature vector. The detailed idea of the proposed CHEQ is described as follows and is depicted in Fig. 1.

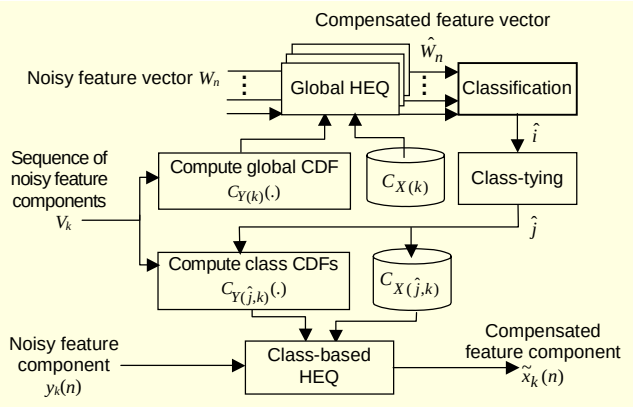


Fig. 1. Block diagram of the class-based HEQ method.

Let us define feature vector $W_n$ consisting of $K$-dimensional components at time frame $n$ as

$$W_n = \{y_1(n), y_2(n), \Lambda\ , y_K(n)\}^T, \tag{6}$$

where $T$ stands for transpose.

Then, the phonetic class index $\hat{i}$ assigned to the noisy feature vector $W_n$ is obtained as

$$\hat{i} = \arg\min_i \; d(\hat{W}_n, z_i), \quad 1 \le i \le I, \tag{7}$$

where $d(\cdot,\cdot)$ denotes the Mahalanobis distance measure, $z_i$ stands for the centroid of the $i$-th class computed by the $k$-means algorithm, $I$ is the number of classes, and $\hat{W}_n$ is the histogram equalized version of $W_n$ by the conventional HEQ given as

$$\begin{aligned}\hat{W}_n &= \{\hat{x}_1(n), \Lambda\ , \hat{x}_K(n)\}^T \\ &= \left\{C^{-1}_{X(1)}(\hat{C}_{Y(1)}(y_1(n))), \Lambda\ , C^{-1}_{X(K)}(\hat{C}_{Y(K)}(y_K(n)))\right\}^T.\end{aligned} \tag{8}$$

In the proposed CHEQ scheme, the phonetic mismatch can be effectively reduced by increasing the number of phonetic classes sufficiently. Additionally, the accuracy of phonetic classification increases as more phonetic classes are used. However, the larger the number of phonetic classes, the smaller the amount of classified data for each class, which results in poor test CDF estimation. Hence, the number of phonetic classes cannot be chosen arbitrarily. Modeling each phonetic class by using the union of multiple small classes can be more accurate than using a single large class in the phonetic classification. In this sense, the class-tying technique is employed in the CDF estimation such that tied-class index $\hat{j}$ is obtained by

$$\hat{j} = \arg\min_j d(z_{\hat{i}}, Z_j), \; 1 \le j \le J, \tag{9}$$

where $Z_j$ represents the centroid of the $j$-th tied-class computed from those of all untied-classes defined in (7), and $J$ is the number of tied-classes.

Then, the proposed CHEQ formulation is defined as

$$\begin{aligned}\tilde{x}_k(n) &= C^{-1}_{X(\hat{j},k)}(C_{Y(\hat{j},k)}(y_k(n))) \\ &= C^{-1}_{X(\hat{j},k)}\left(\frac{r_{\hat{j}k}(y_k(n)) - 0.5}{N_{\hat{j}}}\right),\end{aligned} \tag{10}$$

where $C_{Y(\hat{j},k)}(y)$ and $r_{\hat{j}k}(y)$ denote the test CDF and the rank at the $\hat{j}$-th tied-class and $k$-th feature component, respectively. $N_{\hat{j}}$ is the number of frames classified as the $\hat{j}$-th tied-class, and $C^{-1}_{X(\hat{j},k)}(x)$ represents the inverse of reference CDF $C_{X(\hat{j},k)}(x)$ obtained by the cumulative histogram computed from all training data of the $k$-th feature components classified as the $\hat{j}$-th tied-class.

## IV. Experimental Results

In the performance evaluation, the TI-DIGITS-based Aurora 2 database is used. Only clean speech data are used in the training of all experiments. Test sets A and B, each containing four kinds of additive noises, and test set C, contaminated by two kinds of additive noises and different channel distortion, are chosen for evaluation. The feature extraction procedure is conducted based on the ETSI Aurora formula as follows. Speech signals are blocked into a sequence of frames, each 25 ms in length with a 10 ms interval. Speech frames are pre-emphasized using a factor of 0.97, and a Hamming window is then applied. From a set of 23 mel-scaled filter-bank log energies, a 39-dimensional mel-frequency cepstral coefficient (MFCC)-based feature vector consisting of twelve MFCCs, the log energy, and their first and second derivatives is extracted. Prior to the derivative computation, 22-order cepstral liftering is applied to the static MFCCs [8], [9]. Each digit-based hidden Markov model consists of sixteen states, and each state has three mixtures. The number of histogram bins in the reference CDFs was chosen as 64 in both the conventional HEQ and CHEQ because a further increase did not show any meaningful performance improvements. The equalization was conducted on all of the 39 components of the MFCC feature vector for the training and test data with utterance-by-utterance estimation of the test CDFs.

Figure 2 shows the recognition results with respect to various numbers of classes when untied and tied-class CHEQs are used. The results are represented in terms of the averaged word error rate (WER) for the three test sets.

For each number of tied-classes, there is a corresponding number of untied classes, which is empirically chosen from the experiments. In this figure, we observe that both CHEQ methods provide significant improvements over the conventional

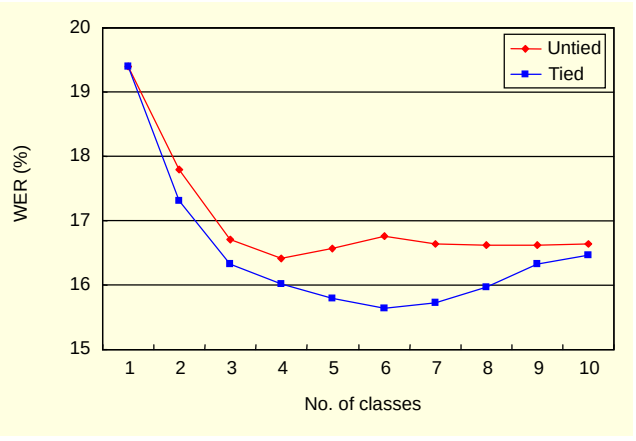


Fig. 2. WERs of untied/tied-class CHEQ methods with respect to various numbers of classes on Aurora 2 Task (averaged between 0 and 20 dB SNRs for test sets A, B, and C).

HEQ only when the number of classes exceeds two. Also note that the use of the class-tying technique produces further improvements with an error rate reduction (ERR) of 4.65% over the untied-CHEQ.

Table 1 shows the recognition results for test sets A, B, and C obtained by the MFCC, conventional HEQ, and CHEQ, respectively. The tied-class parameters, *I* and *J*, are empirically chosen as 60 and 6, respectively, from the results shown in Fig. 2. For sets A and B, the CHEQ shows outstanding improvements over the MFCC with average ERRs of 59.43% and 66.00%, respectively, and substantial improvements over the conventional HEQ with 18.74% and 17.52%, respectively. The results for set C also indicate that the CHEQ achieves significant improvements over the MFCC and conventional HEQ with average ERRs of 50.50% and 23.45%, respectively. From these results, we note that the proposed method shows consistent effectiveness for compensating the acoustic mismatch caused by both additive noises only, and additive noises and channel distortion together.

Table 1. WERs and ERRs for Aurora 2 Task (averaged between 0 and 20 dB SNRs).

| Noises | | WER (%) | | | ERR (%) over | |
|---|---|---|---|---|---|---|
| | | MFCC | HEQ | CHEQ | MFCC | HEQ |
| A | Subway | 30.14 | 18.68 | 15.44 | 48.79 | 17.38 |
| | Babble | 49.76 | 18.60 | 16.69 | 66.47 | 10.28 |
| | Car | 40.13 | 18.51 | 13.45 | 66.48 | 27.31 |
| | Exhibition | 35.47 | 21.85 | 17.51 | 50.63 | 19.84 |
| | **Average** | **38.88** | **19.41** | **15.77** | **59.43** | **18.74** |
| B | Restaurant | 48.49 | 18.02 | 16.58 | 65.80 | 7.98 |
| | Street | 38.48 | 18.21 | 14.68 | 61.86 | 19.41 |
| | Airport | 46.67 | 17.58 | 14.47 | 69.00 | 17.72 |
| | Station | 44.08 | 19.45 | 14.70 | 66.65 | 24.41 |
| | **Average** | **44.43** | **18.32** | **15.11** | **66.00** | **17.52** |
| C | Subway | 32.77 | 22.42 | 17.41 | 46.88 | 22.34 |
| | Street | 33.87 | 20.68 | 15.58 | 54.00 | 24.64 |
| | **Average** | **33.32** | **21.55** | **16.50** | **50.50** | **23.45** |

## V. Conclusion

For compensating the acoustic mismatch, the conventional HEQ has a fundamental limitation when the phonetic class distributions of the training and test data differ from each other. The proposed class-based HEQ not only compensates for the acoustic mismatch but also reduces the phonetic mismatch by utilizing multiple class-specific reference and test CDFs.

Experimental results showed the effectiveness of the class-based HEQ over the conventional HEQ for compensating an acoustic mismatch.

## References


[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice-Hall, 2001.

[2] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Cepstral Channel Normalization Techniques for HMM-Based Speaker Verification," *Proc. ICSLP*, 1994, pp. 1835-1838.

[3] O. Viikki and K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," *Speech Communication*, vol. 25, 1998, pp. 133-147.

[4] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice-Hall, New Jersey, 2002.

[5] J. C. Segura, C. Benítez, Á. de la Torre, A. J. Rubio, and J. Ramírez, "Cepstral Domain Segmental Nonlinear Feature Transformations for Robust Speech Recognition," *IEEE Signal Processing Letters*, vol. 11, May 2004, pp.517-520.

[6] Á. de la Torre, A. M. Peinado, J. C. Segura, J. L. Pérez-Córdoba, M. C. Benítez, and A. J. Rubio, "Histogram Equalization of Speech Representation for Robust Speech Recognition," *IEEE Trans. Speech and Audio Processing*, vol. 13, May 2005, pp. 355-366.

[7] S. Molau, F. Hilger, D. Keysers, and H. Ney, "Enhanced Histogram Normalization in the Acoustic Feature Space," *Proc. ICSLP, 2002*, pp. 1421-1424.

[8] S. Young et al., *The HTK Book, ver. 3.2.1*, Cambridge University Engineering Department, 2002.

[9] H.-Y. Jung, "Filtering of Filter-Bank Energies for Robust Speech Recognition," *ETRI J.*, vol. 26, no. 3, June 2004, pp. 273-276.