Probabilistic Class Histogram Equalization for Robust Speech Recognition

Youngjoo Suh, Mikyong Ji, and Hoirin Kim, *Member, IEEE*

***Abstract*—In this letter, a probabilistic class histogram equalization method is proposed to compensate for an acoustic mismatch in noise robust speech recognition. The proposed method aims not only to compensate for the acoustic mismatch between training and test environments but also to reduce the limitations of the conventional histogram equalization. It utilizes multiple class-specific reference and test cumulative distribution functions, classifies noisy test features into their corresponding classes by means of soft classification with a Gaussian mixture model, and equalizes the features by using their corresponding class-specific distributions. Experiments on the Aurora 2 task confirm the superiority of the proposed approach in acoustic feature compensation.**

***Index Terms*—Feature compensation, histogram equalization, probabilistic class, robust speech recognition.**

I. Introduction

THE performance of automatic speech recognition (ASR) systems degrades severely when acoustic environments between training and test data differ from each other. The main cause of this acoustic mismatch is corruption by additive noise and channel distortion [1]. In robust speech recognition, the feature compensation approach has been widely employed due to such advantages as low computational complexity and effective performance improvement. Acoustic environments corrupted by additive noise and channel distortion cause a nonlinear transformation in the feature spaces of the cepstrum or log-spectrum [1]. For this reason, linear transformation-based feature compensation methods such as cepstral mean normalization [2] or cepstral mean and variance normalization [3] have substantial limitations, even though they yield significant performance improvement under noisy environments. Currently, piecewise linear approximation-based methods, such as interacting multiple model (IMM) [4] and stereo-based piecewise linear compensation for environments (SPLICE) [5], are the major approaches for coping with the nonlinear behavior of the acoustic mismatch.

As another efficient approach, a histogram equalization (HEQ) technique has been proposed. The basic idea of HEQ is to convert the probability density function (PDF) of the test features into that of the references [6]–[9]. Therefore, unlike other methods, HEQ can compensate for the acoustic mismatch by directly utilizing the nonlinear inverse transformation. Recent research has also shown that HEQ is quite effective in preventing the performance degradation of ASR under noisy environments [6], [7]. However, HEQ needs some fundamental assumptions for its full performance. The first assumption is that the acoustic mismatch should act as a monotonic transformation in the feature domain [7]. In other words, the ordering information of acoustic classes, defined in the acoustic modeling of speech recognition systems, along each feature axis should not be altered by the acoustic mismatch. The second is that distributions of acoustic classes, for both training and test data, should be identical or similar to each other. When any of these assumptions is not kept, the transformation by HEQ tends to impair the class separability of features by confusedly mapping to the regions of other classes. However, the transformation caused by the corruption from additive noise or channel distortion does not always guarantee the monotonic transformation. In addition, test speech utterances may be too short to make their acoustic class distributions identical or similar to those of training data. As a result, it is difficult to take full advantage of HEQ when the conventional HEQ is used to compensate for the acoustic mismatch in noisy environments.

In this letter, we propose a probabilistic class HEQ not only to compensate for the acoustic mismatch between the training and test data but also to remedy the limitations of the conventional HEQ. The proposed technique equalizes the test features by using class-specific reference and test cumulative distribution functions (CDFs) where the required class information is obtained from soft classification based on a Gaussian mixture model (GMM).

II. Conventional Histogram Equalization

For random reference variable x and test variable y, let $P_X(x)$ and $P_Y(y)$ denote their corresponding PDFs. A transform function $x = F(y)$ mapping $P_Y(y)$ into $P_X(x)$ is given in [6], [7] 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 random variable y.

One of the critical problems in HEQ is the reliable estimation of reference and test CDFs. In speech recognition applications, reference CDFs can be estimated quite reliably by computing cumulative histograms with a large amount of training data. However, when short utterances are used as test data, the length of each utterance may be insufficient for a reliable estimation. In these test environments, the test CDF estimation becomes much more important. When the number of estimation

Manuscript received June 19, 2006; revised August 29, 2006. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Frederic Bimbot.

The authors are with the School of Engineering, Information and Communications University, Daejeon 305-732, Korea (e-mail: yjsuh@icu.ac.kr; lindaji@icu.ac.kr, hrkim@icu.ac.kr).

Digital Object Identifier 10.1109/LSP.2006.884903

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 given as follows [6], [8].

Let us define a sequence S consisting of N frames of a certain test feature component as

$$S = \{y_1, y_2, \ldots, y_n, \ldots, y_N\} \tag{2}$$

where y_n is the test feature component at the nth frame.

The order statistics of (2) can be represented as

$$y_{T(1)} \leq y_{T(2)} \leq \cdots \leq y_{T(r)} \leq \cdots \leq y_{T(N)} \tag{3}$$

where $T(r)$ denotes the original frame index of feature component $y_{T(r)}$ in which its rank is r when the elements of sequence S are sorted in ascending order.

The order-statistic-based direct estimate of the test CDF is given as

$$\hat{C}_Y(y_n) = \frac{R(y_n) - 0.5}{N} \tag{4}$$

where $R(y_n)$ denotes the rank of y_n ranging from 1 to N.

An estimate of the reference feature component by the conventional HEQ given test feature component y_n is then obtained as

$$\hat{x}_n = C_X^{-1}[\hat{C}_Y(y_n)] = C_X^{-1}\left[\frac{R(y_n) - 0.5}{N}\right]. \tag{5}$$

III. Class Histogram Equalization

The proposed approach to reducing both acoustic mismatches and the limitations of the conventional HEQ consists of utilizing multiple class-specific CDFs at both reference and test sides. From the viewpoint of utilizing class information, the class HEQ (CHEQ) has two approaches: hard-CHEQ using vector quantization [10] and probabilistic or soft-CHEQ based on a GMM as follows.

A. Hard-CHEQ

In this class-based approach, reliably assigning class information to each feature component is a prerequisite condition for ensuring the effectiveness of CHEQ. In most HEQ methods, the equalization is performed on a feature component basis for more reliable estimation of the distributions with limited amounts of sample data [9]. However, acoustic class modeling or feature parameterization in current ASR is usually processed on a vector basis. For this reason, the acoustic class information utilized in each feature component is extracted on a feature vector basis as follows.

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

$$V_n = \begin{bmatrix} y_n^{(1)} & y_n^{(2)} & \ldots & y_n^{(k)} & \ldots & y_n^{(K)} \end{bmatrix}^T \tag{6}$$

where $y_n^{(k)}$ is the kth test feature component, and T stands for transpose.

Then, acoustic class index $\hat{i}$ assigned to noisy test feature vector V_n by hard classification can be obtained as

$$\hat{i} = \arg\min_i d(\hat{V}_n, z_i), \quad 1 \leq i \leq I_H \tag{7}$$

where $d(\cdot,\cdot)$ denotes the Mahalanobis distance measure, z_i stands for the centroid vector of the ith hard-class by the k-means algorithm [1], I_H is the number of hard-classes, and $\hat{V}_n$ is the histogram equalized version of V_n by the conventional HEQ in order to reduce the adverse noise effects in classification [10] and is given as

$$\begin{aligned} \hat{V}_n &= \begin{bmatrix} \hat{x}_n^{(1)} & \hat{x}_n^{(2)} & \ldots \hat{x}_n^{(K)} \end{bmatrix}^T \\ &= \left[C_X^{-1}\left(\hat{C}_Y\left(y_n^{(1)}\right)\right) C_X^{-1}\left(\hat{C}_Y\left(y_n^{(2)}\right)\right) \right. \\ &\qquad \left. \ldots C_X^{-1}\left(\hat{C}_Y\left(y_n^{(K)}\right)\right)\right]^T. \end{aligned} \tag{8}$$

An estimate of the reference feature component by the hard-CHEQ given test feature component y_n is then defined as

$$\hat{x}_{H,n} = C_{H,X(\hat{i})}^{-1}[\hat{C}_{H,Y(\hat{i})}(y_n)] = C_{H,X(\hat{i})}^{-1}\left(\frac{R_{\hat{i}}(y_n) - 0.5}{N_{\hat{i}}}\right) \tag{9}$$

where $\hat{C}_{H,Y(\hat{i})}(y_n)$ and $R_{\hat{i}}(y_n)$ denote the hard-class-based test CDF estimate and the rank of y_n at the $\hat{i}$th hard-class, respectively. In addition, $C_{H,X(\hat{i})}^{-1}(\cdot)$ and $N_{\hat{i}}$ represent the inverse of hard-class-based reference CDF, $C_{H,X(\hat{i})}(\cdot)$, obtained by the cumulative histogram computed from the training data, and the number of frames at the $\hat{i}$th hard-class, respectively.

B. Soft-CHEQ

It is more reasonable to assume that a feature vector belongs to a number of acoustic classes than only a single dominant class as in the hard-CHEQ. In this way, a more generalized form of CHEQ is derived by using a soft-class concept, where the relation between the given feature vector and each acoustic class is determined probabilistically by using a GMM-based posterior probability. Given conventional HEQ-based transformed feature vector $\hat{V}_n$, the posterior probability of soft-class ω_i is defined by

$$P(\omega_i \mid \hat{V}_n) = \frac{\alpha_i \mathcal{N}(\hat{V}_n; \mu_i, \Sigma_i)}{\sum_{m=1}^{I_S} \alpha_m \mathcal{N}(\hat{V}_n; \mu_m, \Sigma_m)} \tag{10}$$

where I_S denotes the number of soft-classes, α_i represents the mixture component weight for the ith soft-class, and $\mathcal{N}(\hat{V}_n; \mu_i, \Sigma_i)$ is the K-dimensional normal distribution with mean vector μ_i and covariance matrix Σ_i at the ith soft-class.

By using the idea of CHEQ in (9) and taking into account the probabilistic relations to all soft-classes, an estimate of the

reference feature component by soft-CHEQ given test feature component y_n is defined by

$$\hat{x}_{S,n} = \sum_{i=1}^{I_S} P(\omega_i \mid \hat{V}_n) C_{S,X(i)}^{-1} \left[\hat{C}_{S,Y(i)}(y_n) \right] \tag{11}$$

where $\hat{C}_{S,Y(i)}(y_n)$ denotes the test CDF estimate at the ith soft-class computed as

$$\hat{C}_{S,Y(i)}(y_n) = \frac{\left(\sum_{r=1}^{R(y_n)-1} P(\omega_i \mid \hat{V}_{T(r)}) \right) + 0.5 P(\omega_i \mid \hat{V}_n)}{\sum_{r=1}^{N} P(\omega_i \mid \hat{V}_r)}. \tag{12}$$

The reference CDF estimate at the ith soft-class is similarly obtained by accumulating the posterior probabilities of the corresponding reference histogram bins.

C. Class-Tying

According to the CHEQ scheme, the nonmonotonic transformation and distribution mismatch of acoustic classes can be effectively reduced by using a larger number of acoustic classes, provided reliable acoustic classification is possible. However, the classification accuracy is inevitably degraded in noisy environments. In this condition, increasing the number of classes further deteriorates the classification accuracy due to increased class candidates. For these reasons, the effectiveness of CHEQ increases to a certain number of classes, and then it tends to decrease. Therefore, classification accuracy plays a critical role in CHEQ. When the number of classes cannot be increased arbitrarily, one possible way to improve classification accuracy is to model each class in more detail through the union of a number of small classes rather than by a large coarse class. In this sense, the class-tying technique is employed such that the $\hat{j}$th tied-class for the given $\hat{i}$th untied-class in the hard-CHEQ is obtained as

$$\hat{j} = \arg\min_j d(z_{\hat{i}}, Z_j), \quad 1 \le j \le J_H \tag{13}$$

where Z_j represents the centroid vector of the jth tied-hard-class computed from those of all untied-hard-classes defined in (7), and J_H is the number of tied-hard-classes.

By using (9) and (13), an estimate of the reference feature component by the tied-hard-CHEQ given test feature component y_n is defined as

$$\begin{aligned} \hat{x}_{HT,n} &= C_{HT,X(\hat{j})}^{-1} [\hat{C}_{HT,Y(\hat{j})}(y_n)] \\ &= C_{HT,X(\hat{j})}^{-1} \left(\frac{R_{\hat{j}}(y_n) - 0.5}{N_{\hat{j}}} \right) \end{aligned} \tag{14}$$

where $\hat{C}_{HT,Y(\hat{j})}(y_n)$ and $R_{\hat{j}}(y_n)$ denote the test CDF estimate and the rank of feature component y_n at the $\hat{j}$th tied-hard-class, respectively. $C_{HT,X(\hat{j})}^{-1}(\cdot)$ and $N_{\hat{j}}$ represent the inverse of tied-hard-class reference CDF, $C_{HT,X(\hat{j})}(\cdot)$, obtained by the cumulative histogram computed from all training data of the feature components, and the number of frames at the $\hat{j}$th tied-hard-class, respectively.

In the tied-soft-CHEQ, the j_Sth tied-class for the given ith untied-class is obtained by using a symmetric version of the Kullback–Leibler distance measure between the two Gaussian PDFs [11] as

$$\begin{aligned} j_S = \arg\min_j \{ & tr\left(\Sigma_i \Sigma_j^{-1}\right) + tr\left(\Sigma_j \Sigma_i^{-1}\right) - 2K \\ & + tr\left[\left(\Sigma_i^{-1} + \Sigma_j^{-1}\right)(u_i - U_j)(u_i - U_j)^T \right] \} \\ & 1 \le j \le J_S \end{aligned} \tag{15}$$

where U_j and Σ_j represent the mean vector and covariance matrix of the jth tied-soft-class, respectively. J_S is the number of tied-soft-classes, and $tr(\cdot)$ stands for the trace of a matrix.

By using (11) and (15), an estimate of the reference feature component by the tied-soft-CHEQ given test feature component y_n is defined as

$$\hat{x}_{ST,n} = \sum_{j=1}^{J_S} P(\omega_j \mid \hat{V}_n) C_{ST,X(j)}^{-1} \left[\hat{C}_{ST,Y(j)}(y_n) \right] \tag{16}$$

where $\hat{C}_{ST,Y(j)}(y_n)$ denotes a test CDF estimate of feature component y_n at the jth tied-soft-class, and $C_{ST,X(j)}^{-1}(\cdot)$ represents the inverse of tied-soft-class reference CDF, $C_{ST,X(j)}(\cdot)$, obtained by the cumulative histogram computed from the training data of feature components at the jth tied-soft-class. With a formulation similar to (12), each estimate of the test CDF is obtained by using the posterior probability of tied-soft-class ω_j given $\hat{V}_n$, which is defined as

$$P(\omega_j \mid \hat{V}_n) = \frac{\sum_{i \in j} \alpha_i \mathcal{N}(\hat{V}_n; \mu_i, \Sigma_i)}{\sum_{m=1}^{J_S} \sum_{i \in m} \alpha_i \mathcal{N}(\hat{V}_n; \mu_i, \Sigma_i)}. \tag{17}$$

IV. Experimental Results

In the performance evaluation, the Aurora 2 database converted from the TI-DIGITS database is used. Only clean condition training is used in the experiments. Test sets A and B, each containing four kinds of additive noise, and test set C, contaminated by two kinds of additive noise and different channel distortion (MIRS), are chosen for the test. The 39-dimensional MFCC-based feature vectors, each consisting of 12 MFCCs, log energy, and their first and second derivatives, are used in the recognition experiments. These are extracted with a frame length of 25 ms and an interval of 10 ms. Hidden Markov model (HMM)-based speech recognizers are used where each digit-based HMM consists of 16 states and each state has three mixture components. Diagonal covariance matrices are used in the HMM and GMM. The number of histogram bins in the reference CDFs was empirically chosen as 64 for both the conventional HEQ and CHEQ. The equalization was conducted on all

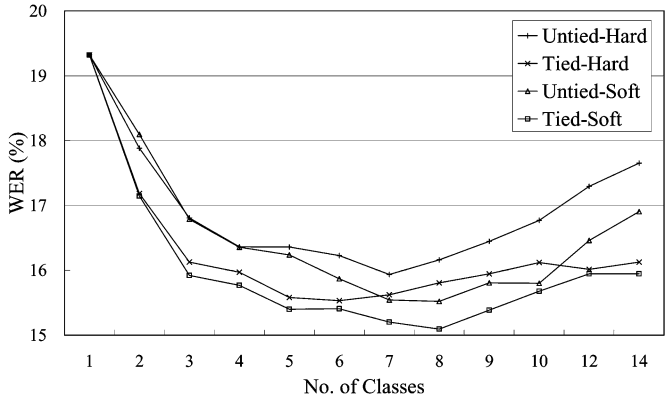

Fig. 1. Recognition results of hard or soft-CHEQ with or without the class-tying technique for various numbers of classes on the Aurora 2 task (averaged between 0 and 20 dB SNRs for test sets A, B, and C).

of the 39-dimensional MFCCs for both the training and test data on an utterance-by-utterance basis after estimating the reference CDFs from the training data.

Fig. 1 shows the recognition results by the CHEQ method with respect to various numbers of classes ranging from 1 (i.e., the conventional HEQ case) to 14 when hard or soft classification is used alone or in combination with the class-tying technique. The results represent average word error rates (WERs) for the noisy speech data between 0 and 20 dB signal-to-noise ratios (SNRs) of the three test sets as suggested by the Aurora Group. In the experiments of tied-class cases, the corresponding untied-classes are empirically chosen to be between 20 and 120 untied-classes. In this figure, we observe that CHEQ provides significant improvements over the conventional HEQ when the number of classes exceeds two. In addition, the figure illustrates that the proper number of classes yielding the best performance is six to eight. For more than this number of classes, the recognition accuracy tends to deteriorate due to degrading classification accuracy in noisy environments.

Table I shows the recognition results in terms of WER for the clean and noisy data of test sets A, B, and C obtained by MFCC, HEQ, and CHEQ, respectively, where hard or soft classification is used with or without class-tying. The table shows that the CHEQ techniques produce outstanding improvements over MFCC with error reductions ranging from 60.13% to 62.27% and substantial improvement over the conventional HEQ with error reductions between 17.51% and 21.92% on noisy speech data, although they yield some degradation on clean speech data. Moreover, soft-CHEQ provides meaningful performance improvement over hard-CHEQ, and the class-tying technique yields additional improvement for both hard and soft-CHEQ approaches on noisy speech data. In robust speech recognition, more emphasis is usually put on the results of the noisy speech data. In this sense, we can conclude from the results that the proposed CHEQ provides consistent effectiveness in compensating for the acoustic mismatch in noisy environments for speech recognition applications.

TABLE I
COMPARISONS OF WORD ERROR RATES (%) OF HEQ-BASED FEATURE COMPENSATION TECHNIQUES ON THE AURORA 2 TASK (NOISY RESULTS ARE AVERAGED BETWEEN 0 AND 20 dB SNRS)

Test Sets		MFCC	HEQ	Hard-CHEQ		Soft-CHEQ	
				Untied	Tied	Untied	Tied
A	Clean	1.06	1.01	1.73	1.34	1.53	1.44
	Noisy	38.88	19.34	15.96	15.44	15.53	15.05
B	Clean	1.06	1.01	1.73	1.34	1.53	1.44
	Noisy	44.43	18.24	15.44	15.14	15.05	14.71
C	Clean	1.01	1.10	1.74	1.33	1.39	1.43
	Noisy	33.32	21.46	16.90	16.51	16.45	15.96
Avg.	Clean	1.05	1.03	1.73	1.34	1.50	1.44
	Noisy	39.99	19.32	15.94	15.53	15.52	15.10

V. CONCLUSION

In this letter, we propose a new feature compensation approach called a probabilistic CHEQ method that can alleviate the fundamental limitations of the conventional HEQ in reducing the acoustic mismatch for robust speech recognition. Class information is obtained by using GMM-based soft classification. The class-tying technique is additionally employed to provide higher classification accuracy. Compared to the conventional HEQ and hard-CHEQ, the probabilistic CHEQ yielded improved speech recognition accuracy in noisy environments.

REFERENCES

[1] X. Huang, A. Acero, and H. -W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*. Englewood Cliffs, NJ: Prentice-Hall, 2001.
[2] A. E. Rosenberg, C. -H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proc. ICSLP*, 1994, pp. 1835–1838.
[3] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Commun.*, vol. 25, pp. 133–147, 1998.
[4] N. S. Kim, Y. J. Kim, and H. W. Kim, "Feature compensation based on soft decision," *IEEE Signal Process. Lett.*, vol. 11, no. 3, pp. 378–381, Mar. 2004.
[5] J. Droppo, L. Deng, and A. Acero, "Evaluation of the SPLICE algorithm on the Aurora2 database," in *Proc. Eurospeech*, 2001, pp. 217–220.
[6] 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 Process. Lett.*, vol. 11, no. 5, pp. 517–520, May 2004.
[7] 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 Audio Process.*, vol. 13, no. 3, pp. 355–366, May 2005.
[8] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, 2001, pp. 213–218.
[9] S. Chen and R. Gopinath, "Gaussianization," in *Proc. Neural Information Processing Systems*, 2000, pp. 423–429.
[10] Y. Suh and H. Kim, "Class-based histogram equalization for robust speech recognition," *ETRI J.*, vol. 28, no. 4, pp. 502–505, 2006.
[11] P. J. Moreno and P. P. Ho, A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels, Hewlett Packard Research Lab, Cambridge, MA, 2004, Tech. Rep. HPL-2004-7.