THE EFFECTIVENESS OF HISTOGRAM EQUALIZATION ON ENVIRONMENTAL MODEL ADAPTATION

Youngjoo Suh and Hoirin Kim

Information and Communications University
119 Munjiro, Yuseong-gu, Daejeon 305-714, Korea

ABSTRACT

In this paper, we introduce a new histogram equalization-based environmental model adaptation method for robust speech recognition in noise environments. The proposed method adapts initially-trained acoustic mean models of a speech recognizer into the environmentally matched models. The covariance models are adapted by using utterance-level local covariance matrices. We performed a series of experiments based on the Aurora2 framework to examine the effectiveness of the proposed environmental model adaptation technique. In both clean and multi-condition trainings, the proposed approach achieved substantial performance improvements over the baseline speech recognizers.

Index Terms—Histogram equalization, model adaptation, robust speech recognition.

1. INTRODUCTION

Speech recognizers usually show considerable performance deterioration when they are deployed in the acoustically mismatched environments compared to the training ones [1]. Thus, one of the major issues in automatic speech recognition is to provide the robustness against the mismatch between training and test environments. Of a couple of robust speech recognition approaches, an easiest way to provide the robustness against the acoustic mismatch is feature compensation [2]. Histogram equalization (HEQ) is known to be one of the most efficient feature compensation techniques due to its algorithmic simplicity [3]-[7]. Nevertheless, it shows considerable compensation effectiveness because of its nonlinear transformation characteristics which are fundamentally required in dealing with the logarithmic domain-based features such as cepstral coefficients [4]. However, since the noise corruption and feature extraction usually cause some forms of acoustic-phonetic information loss, it is difficult to fully recover the clean speech features from their corresponding noisy speech features by using the feature compensation approach. As a result, there would be some unavoidable discrepancies between the clean speech models of a speech recognizer and the compensated features. On the contrary, clean speech models can be completely adapted into acoustically matched models as far as the amount of adaptation data is provided enough in model adaptation. In this case, even though the information loss still exists in the noisy speech feature, it does not cause any discrepancies and can be disregarded in the process of acoustic-phonetic classification. For this reason, the model adaptation approach can be superior to the feature compensation one in coping with the mismatch between training and test environments.

In this paper, we propose a model adaptation method based on the histogram equalization technique to take advantage of its possible superiority to the feature compensation approach. In the proposed approach, the histogram equalization technique adapts the trained acoustic mean models of the speech recognizers into environmentally matched models. The covariance models are adapted by using utterance-level sample covariance estimates. Experiments on the Aurora2 framework confirmed the effectiveness of the proposed approach for model adaptation.

2. HISTOGRAM EQUALIZATION FOR FEATURE COMPENSATION

The utilization of histogram equalization techniques for feature compensation (HEQ-FC) begins with such an assumption that the acoustic mismatch between reference (or training) speech features and test speech ones results in the discrepancy between their corresponding probability density functions (PDFs). Then, HEQ-FC tries to compensate test features into reference features by transforming the test PDF into the reference one. Here, we assume that histogram equalization is conducted in a component-by-component basis for algorithmic simplicity. Then, a basic algorithm of HEQ-FC is as follows.

For given reference and test random variables x and y, respectively, a transformation function by HEQ-FC mapping the test PDF $P_Y(y)$ into the reference PDF $P_X(x)$ is given as

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

where C_X^{-1} is the inverse of the reference cumulative

distribution function (CDF) $C_X(x)$, $C_Y(y)$ is the test CDF, and $F(y)$ is the transformation function of HEQ-FC, which has monotonically nondecreasing characteristics. It can be noted in (1) that the effectiveness of HEQ-FC is closely related to the reliable estimation of both reference and test CDFs. In practice, the CDFs are approximated by their cumulative histograms. Therefore, the larger amount of sample data gives the better CDF estimation. Due to its relatively large amount of sample data in the training phase, the reference CDF can be well estimated by its cumulative histograms. However, current speech recognizers may employ short utterances as their input unit. In this case, the amount of sample data can be insufficient for the reliable estimation of the test CDF. Therefore, a reliable estimation of the test CDF can be an important issue for effective HEQ-FC in the short utterance-based test environments. When the amount of sample data is small, the order-statistic-based approach is known to provide more reliable CDF estimation with an improved resolution. A brief algorithm of the order-statistics-based CDF estimation is given as follows [3].

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

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

where y_n is a certain test feature component at the nth frame. The order-statistics of the sequence S in (2) is given by rearranging its elements in ascending order as

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

where $T(r)$ denotes the original frame index of the feature component $y_{T(r)}$ in which its rank is r when the elements of the sequence S are sorted in ascending order. Then, the order-statistic-based test CDF estimate of the test feature component y_n is given as

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

where $R(y_n)$ denotes the rank of y_n among the feature components composing the sequence S according to the order-statistics defined in (3). From (1) and (4), an estimate of the reference feature component by HEQ-FC given the test feature component y_n is obtained as

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

3. HISTOGRAM EQUALIZATION FOR MODEL ADAPTATION

To employ the histogram equalization technique in the model space, we interpret the acoustic mismatch between test and reference environments as a transformation function $y = G(x)$, which is the inverse of the transformation function employed in HEQ-FC. In model adaptation, the trained acoustic models of a speech recognizer are transformed into the adapted models to match test environments acoustically. Therefore, with the transformation function $y = G(x)$, the histogram equalization technique for model adaptation (HEQ-MA) can transform the trained acoustic models into the environmentally matched models. If the acoustic models under training and test environments are denoted as Φ_X and Φ_Y, respectively, the transformation by HEQ-MA is given by mapping the reference PDF $P_x(\Phi_X)$ into the test PDF $P_Y(\Phi_Y)$ as

$$\Phi_Y = G(\Phi_X) = F^{-1}(\Phi_X) = C_Y^{-1}(C_X(\Phi_X)). \tag{6}$$

We assume that the mean vector and covariance matrix of a trained acoustic model in a speech recognizer are based on the Gaussian distribution and is given by μ and Σ, respectively. If we further assume that HEQ-MA is applied to each mean vector on a component-by-component basis as in HEQ-FC, the adaptation rule for a trained acoustic mean component by HEQ-MA is given by using (6) and a linear interpolation between two test feature components in the sequence S which are nearest to the trained mean component in terms of the CDF value as

$$\begin{aligned}\hat{\mu}(k) &= C_{Y(k)}^{-1}(\hat{C}_{X(k)}(\mu(k))) \\ &= \begin{cases} \alpha(k) y_{T(m)}(k) + (1-\alpha(k)) y_{T(m+1)}(k), & 1 \le m < N \\ \alpha(k) y_{T(N)}(k) + (1-\alpha(k))(y_{T(N)}(k) + \rho(k)), & m = N, \end{cases}\end{aligned} \tag{7}$$

where $\hat{\mu}(k)$ and $\mu(k)$ denote the kth components of the adapted and trained mean vectors, respectively, $C_{Y(k)}^{-1}$ is the inverse of the test CDF for the kth test feature component, $\hat{C}_{X(k)}(\mu(k))$ is the reference CDF estimate of the kth mean component $\mu(k)$, m is the rank index satisfying the relationship

$$\hat{C}_{Y(k)}(y_{T(m-1)}(k)) < \hat{C}_{X(k)}(\mu(k)) \le \hat{C}_{Y(k)}(y_{T(m)}(k)), \tag{8}$$

$\rho(k)$ stands for a positive value at the kth mean component for the boundary condition, and $\alpha(k)$ is the linear interpolation factor of the kth mean component that is based on the order-statistics-based test CDF of the sequence S defined in (4) and is given by

$$\begin{aligned}\alpha(k) &= \frac{\hat{C}_{Y(k)}(y_{T(m)}(k)) - \hat{C}_{X(k)}(\mu(k))}{\hat{C}_{Y(k)}(y_{T(m)}(k)) - \hat{C}_{Y(k)}(y_{T(m-1)}(k))} \\ &= m - N\hat{C}_{X(k)}(\mu(k)),\end{aligned} \tag{9}$$

where the test CDF estimate of the undefined feature component $y_{T(0)}$ is assumed to be zero to satisfy its boundary condition.

As the noise corruption increases, the dynamic range of certain features such as cepstral coefficients tends to shrink due to the spectral whitening effect. Because the dynamic range is directly related to the covariance, we expect that the covariance shrinkage occurs in noisy features. For this reason, it is generally known that the improvements gained by using mean and covariance adaptation over mean adaptation only becomes significant in noisy environments, although adapting the means provides the major effect on performance in the cepstral feature-based speech recognition [8]. The proposed HEQ-MA technique focuses its adaptation target on the mean models. Therefore, to cope with the covariance shrinkage effect in the noisy environments, we introduce an efficient adaptation rule for covariance matrices, which is given by a linear interpolation of the trained covariance matrix and the sequence-level sample covariance matrix as

$$\hat{\Sigma}(k,l) = \beta(\gamma)\Sigma^{s}(k,l) + (1-\beta(\gamma))\Sigma(k,l), \quad (10)$$

where $\beta(\gamma)$ is a signal-to-noise ratio (SNR)-dependent smoothing factor to deal with the higher covariance shrinkage effect at the heavier noise conditions and is given by $\beta(\gamma) = a\gamma + b$, where γ is the averaged SNR value of the sequence S and a and b are empirically chosen slope and bias constants, respectively, and Σ^{s} denotes the global sample covariance matrix of the sequence S obtained by

$$\Sigma^{s}(k,l) = \frac{1}{N}\sum_{n=1}^{N}(y_n(k)-\nu(k))(y_n(l)-\nu(l)), \quad (11)$$

where $\nu(l)$ represents the sequence-level sample mean of the lth feature component.

4. EXPERIMENTAL RESULTS

The performance of the proposed model adaptation approach is evaluated on the Aurora2 speech database. We employed the ETSI Aurora-2 experimental framework in experiments as follows [9]. We trained the acoustic models of two baseline speech recognizers with both of the clean and multi-condition training sets, respectively. In evaluations, we used the three Aurora2 test sets, where test set A is added by four kinds of noise (subway, babble, car, and exhibition), test set B is corrupted by another four types of noise (restaurant, street, airport, and train station), and test set C is contaminated by two of the eight kinds of noise (subway and street) and channel distortion together. Each of the three test sets is further composed of 6 noisy sub-sets with SNR levels of 20, 15, 10, 5, 0, and -5 dB. In feature extraction, speech signals are firstly blocked into a sequence of frames, each 25ms in length with a 10ms interval. Next, speech frames are pre-emphasized by a first-order FIR filter with a factor of 0.97, and a Hamming window is applied to each frame. From a sequence of 23 mel-scaled log filter-bank energies, the 39-dimensional mel frequency cepstral coefficient (MFCC)-based feature vectors, each consisting of 12 MFCCs, log energy, and their delta and acceleration features, are extracted. The baseline speech recognizer employs 13 whole-word hidden Markov models (HMMs), which consist of 11 digit models with 16 states, a silence model with three states, and a short-pause model with a single state. Each state in digit models consists of 3 Gaussians while those in silence and short-pause models have 6 Gaussians. Diagonal covariance matrices are used in the HMMs. In the performance evaluation, we examined the effectiveness of the HEQ-MA compared to the baseline speech recognizers trained on the clean as well as multi-conditioned speech data, respectively, and HEQ-FC. HEQ-FC is applied to all of the 39-dimensional MFCCs independently for both training and test data. In model adaptation, the HEQ-based mean adaptation and the proposed variance adaptation techniques are applied to the 39-dimensional mean vectors and diagonal covariance matrices, respectively, of all trained HMMs in the baseline speech recognizers. The number of histogram bins in reference CDFs was empirically chosen as 64. The SNR-dependent smoothing parameters a and b in the covariance adaptation are empirically set to -0.03 and 0.9, respectively. Each utterance, the averaged SNR value γ was estimated as the ratio of the averaged frame energy to the averaged noise energy of the initial silence region. The histogram equalization is conducted on an utterance-by-utterance basis in both feature compensation and model adaptation.

Figure 1 shows the recognition results for the three Aurora2 test sets at various SNR conditions with the clean-condition training of the baseline speech recognizer. In the figure, we observe that HEQ-MA gives significant improvements compared to the baseline speech recognizer and also yields meaningful performance gains over HEQ-FC. The improvements obtained by HEQ-MA are more notable at the lower SNR range of 0-10dB. The figure indicates that HEQ-MA can be a very effective technique when the acoustic models to be adapted are trained on the clean speech data.

Figure 2 represents the recognition results for the three Aurora2 test sets at various SNR conditions when the baseline speech recognizer is trained on the multi-condition speech data. In this figure, we observe that HEQ-MA is not as effective as HEQ-FC although it still outperforms the baseline speech recognizer.

Tables I and II show the word error rates obtained by the three techniques for the Aurora2 test sets when the baseline speech recognizers are trained in clean and multi-conditions, respectively. In clean-condition training, HEQ-MA provides an error reduction of 62.83% over the baseline speech recognizer. This error reduction can be regarded significant compared to the reduction of 51.48% gained by HEQ-FC. In multi-condition training, HEQ-MA provides an error reduction of 10.73% over the baseline speech recognizer and

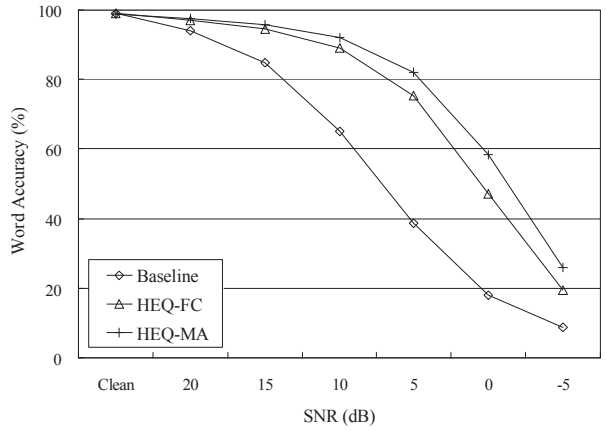

Fig. 1. Recognition results with various SNR conditions by the baseline speech recognizer, HEQ-FC, and HEQ-MA on the Aurora2 task with clean-condition training.

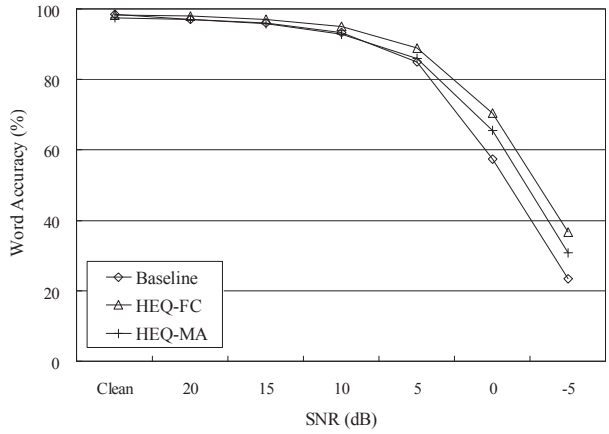

Fig. 2. Recognition results with various SNR conditions by the baseline speech recognizer, HEQ-FC, and HEQ-MA on the Aurora2 task with multi-condition training.

it does not reach the reduction of 29.33% gained by HEQ-FC. Consequently, in the clean-condition training, the results support our previous suggestion that the model adaptation approach can give fundamentally better results than the feature compensation method. However, the results obtained from the multi-condition training experiments do not match our suggestion well. One reason for the inferior results may be the insufficient amount of adaptation data due to the utterance-by-utterance adaptation basis in the multi-condition training, where the acoustic models could have very diverse environmental conditions.

5. CONCLUSION

We propose a histogram equalization-based environmental model adaptation technique. It adapts the acoustic mean models of speech recognizers into the environmentally matched models by using the histogram equalization algorithm. Covariance models are adapted by using an SNR-dependent linear interpolation with the utterance-level sample covariance matrix. In the Aurora2 experimental task, the proposed approach showed significant improvements with high computational efficiency. Further study about the less improvement in the multi-condition training is needed.

Table I. Word error rates by baseline speech recognizer, HEQ-FC, and HEQ-MA on the Aurora2 task with clean-condition training.

Test Sets	Baseline	HEQ-FC	HEQ-MA
A	38.87	19.41	15.19
B	44.43	18.32	14.00
C	33.32	21.55	15.93
Average	39.98	19.40	14.86

Table II. Word error rates by baseline speech recognizer, HEQ-FC, and HEQ-MA on the Aurora2 task with multi-condition training.

Test Sets	Baseline	HEQ-FC	HEQ-MA
A	12.72	10.07	11.89
B	14.47	9.55	13.63
C	16.87	11.13	12.56
Average	14.25	10.07	12.72

6. REFERENCES

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development*, Upper Saddle River, NJ: Prentice Hall, 2001.

[2] N.S. Kim, Y.J. Kim, and H.W. Kim, "Feature compensation based on soft decision," *IEEE Signal Processing Letters*, vol. 11, no. 3, pp. 378-381, 2004.

[3] 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, no. 5, pp. 517-520, 2004.

[4] 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 Processing*, vol. 13, no. 3, pp. 355-366, 2005.

[5] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proc. Speaker Odyssey*, 2001, pp. 213-218.

[6] Y. Suh, S. Kim, and H. Kim, "Compensating acoustic mismatch using class-based histogram equalization for robust speech recognition," *EURASIP Journal on Advances in Signal Processing*. Vol. 2007, Article ID 67070, 9 pages, doi: 10.1155/2007/67870.

[7] Y. Suh, M. Ji, and H. Kim, "Probabilistic class histogram equalization for robust speech recognition," *IEEE Signal Processing Letters*, vol. 14, no. 4, pp. 287-290, 2007.

[8] M.J.F. Gales and P.C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, pp. 249-264, 1996.

[9] H.-G. Hirsch and D. Pearce, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. Int. Conf. Spoken Language Processing*, Oct. 2000, pp. 29-32.