Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 628018, 8 pages
doi:10.1155/2010/628018

Research Article

Histogram Equalization to Model Adaptation for Robust Speech Recognition

Youngjoo Suh and Hoirin Kim

Korea Advanced Institute of Science and Technology, Yuseong-gu, Daejeon 305-701, South Korea

Correspondence should be addressed to Youngjoo Suh, yjsuh@kaist.ac.kr

Received 2 November 2009; Accepted 13 February 2010

Academic Editor: Sharon Gannot

We propose a new model adaptation method based on the histogram equalization technique for providing robustness in noisy environments. The trained acoustic mean models of a speech recognizer are adapted into environmentally matched conditions by using the histogram equalization algorithm on a single utterance basis. For more robust speech recognition in the heavily noisy conditions, trained acoustic covariance models are efficiently adapted by the signal-to-noise ratio-dependent linear interpolation between trained covariance models and utterance-level sample covariance models. Speech recognition experiments on both the digit-based Aurora2 task and the large vocabulary-based task showed that the proposed model adaptation approach provides significant performance improvements compared to the baseline speech recognizer trained on the clean speech data.

1. Introduction

Speech recognizers employed on the noisy environment usually show dramatic performance degradation [1]. This performance degradation has been the major obstacle in introducing the automatic speech recognition (ASR) technology to the real-world applications. For this reason, one of the hot issues in the current research areas of ASR is to provide robustness against performance degradation of speech recognizers in the noisy environments [1, 2]. Noisy environments encountered in ASR are usually different from training acoustic environments. Therefore, the performance degradation of speech recognizers in the noisy environments can be well accounted for the acoustic mismatch [3] between training and test environments. In this case, the acoustic mismatch is mainly due to the corruption of speech by additive noise and channel distortion in the test environments. A lot of robust speech recognition approaches have been proposed to reduce the acoustic mismatch in the past few decades and most of them can be categorized into feature compensation, model adaptation, and uncertainty-based approaches [2]. Of the three approaches, the easiest way to provide the robustness against the acoustic mismatch is feature compensation, where noisy test features are compensated or enhanced to remove noise effects and then decoded by speech recognizers trained on clean speech data [4]. Cepstral mean normalization [5] and cepstral mean variance normalization [6] are the popular techniques for feature compensation. However, it is generally known that model adaptation has the potential for greater robustness in noisy environments than feature compensation although feature compensation is simpler and more efficient to implement [2]. One reason for the possible superiority of model adaptation results from the fact that it can use very detailed knowledge of the underlying speech signal encoded in the acoustic models of the speech recognizer [2]. Because acoustic models in the speech recognizer are designed to represent their own acoustic-phonetic units, they can provide a much more detailed representation of speech. On the contrary, feature compensation methods usually make use of the much simpler model of speech such as a single Gaussian mixture model (GMM). For this reason, better performance can be expected by transforming these acoustic models to match current noise conditions. Another reason for the advantage of model adaptation may be due to the fundamental limitations of feature compensation. Because of the acoustic-phonetic information loss in both noise corruption and feature extraction, it is difficult to perfectly recover clean features from noisy features by using feature compensation algorithms. As a result, this information loss causes discrepancies between clean speech models and compensated features in the decoding process of ASR. On the other hand, clean speech models can be fully adapted into acoustically matched speech models as far as the amount of adaptation data is provided enough in model adaptation. Therefore, although the same information loss can occur in model adaptation, it does not cause undesirable discrepancies between acoustic models and speech features but it is just disregarded in the decoding process. Due to these advantages, numerous environmental model adaptation techniques have been proposed for robust speech recognition until recently. A well-known environmental model adaptation method is the parallel model combination technique [7], which combines both clean speech and noise models in the spectral domain to obtain noisy speech models. Another representative model adaptation technique is the vector Taylor series (VTS) approach [8], which linearly approximates noisy speech models from both clean speech and noise models by using the Taylor series expansion. Both methods are reported to be quite effective in providing the robustness against noise. The standard adaptation methods used for speaker adaptation, such as maximum *a posteriori* (MAP) [1, 2] and maximum likelihood linear regression (MLLR) [9], can be also used for environment model adaptation. Because MAP has the asymptotic property, it can offer performance similar to those of matched conditions. MLLR uses a set of linear transforms to map the initially trained models into the adapted models such that the likelihood of the adaptation data is maximized. This method is known to be quite robust and to achieve reasonable performance with about a minute of speech for minor mismatches.

Basically, the model adaptation approach needs to adapt the entire model parameters employed in the speech recognizer. Therefore, the amount of computation in model adaptation can be a serious problem even in the small vocabulary speech recognition task. Moreover, due to the temporal and spatial variations of acoustic environments, the environmental model adaptation needs to be performed in the input utterance or temporal segment level. Therefore, the model adaptation technique should have computational efficiency as well as noise robustness in its application to real-time speech recognition.

In this paper, we propose a new efficient model adaptation approach based on the histogram equalization (HEQ) technique [10]. HEQ is basically a nonlinear transformation-based approach. In this sense, it can fundamentally cope with the nonlinearity of noisy features in the case of logarithmic space-based features such as cepstral features and is reported to provide considerable performance improvements in both speech recognition and speaker recognition in noisy environments [11–17]. In addition, HEQ is computationally efficient because most of its algorithm consists of sort and search routines with relatively narrow depth and scope. Since its first application to speech recognition [18], HEQ has been mainly used in feature compensation. However, due to the potential superiority of the model adaptation approach, it is expected that the use of HEQ in model adaptation can provide more robustness in noisy environments. In the proposed approach, HEQ transforms the trained mean models of a speech recognizer into environmentally matched models. The transformation function of HEQ is obtained by using reference and test cumulative distribution functions (CDFs) of the training data and test utterance, respectively. A signal-to-noise ratio (SNR)-dependent linear interpolation-based method is used to efficiently adapt the covariance models of a speech recognizer to achieve further performance improvements in heavily noise conditions.

2. HEQ for Feature Compensation

2.1. Basic Algorithm. The application of HEQ to feature compensation begins with such an assumption [18] that the acoustic mismatch between clean reference (or training) features and noisy test features results in the statistical difference between their corresponding probability density functions (PDFs). Then, the idea of HEQ for feature compensation (HEQ-FC) is to conduct a transformation that converts the PDF of the original or test features into that of reference or training features to reduce the effects of noise corruption. In practice, reference and test PDFs are replaced by their corresponding histograms and the test histogram is equalized by using the transformation given by the HEQ algorithm [10]. Here, we assume that HEQ-FC is applied to each feature on a component-by-component basis for algorithmic simplicity. This assumption can be well accepted in the orthogonal transformation-based features such as cepstral features due to their low correlation. In this case, the algorithm of HEQ-FC is described as follows [10]. For given reference and test random features x and y, respectively, a transformation function of HEQ-FC mapping test PDF $P_Y(y)$ into reference PDF $P_X(x)$ is obtained by equating their corresponding CDFs defined as

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

$$x = F(y) = C_X^{-1}[C_Y(y)], \tag{2}$$

where C_X^{-1} is the inverse of the reference CDF $C_X(x)$, $C_Y(y)$ is the test CDF, and $F(y)$ is the transformation function of HEQ-FC and has single valued, monotonically nondecreasing characteristics.

2.2. Order Statistic-Based CDF Estimation. In (2), it is noted that the effectiveness of HEQ-FC is directly related to the reliable estimation of both reference and test CDFs. A better CDF estimation can be achieved by using a larger amount of sample data. Due to its relatively large amount of sample data in the training phase, the reference CDF can be well estimated by the classical cumulative histogram approach. However, current speech recognizers frequently employ a short utterance or word as their input unit. In this case, the amount of sample data can be insufficient for the reliable estimation of the test CDF. In this test environment, the reliable estimation of the test CDF is an important issue for the effective HEQ-FC. When the amount of sample data is small, the order statistic-based CDF estimation method can be more reliable than the classical histogram-based approach due to its enhanced probabilistic resolution. A brief algorithm of the order statistics-based CDF estimation is given as follows [13].

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

$$S = \{y_1, y_2, \ldots, y_n, \ldots, y_N\}, \qquad (3)$$

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

The order statistics of the sequence S in (3) is given by sorting its elements in ascending order as

$$y_{T(1)} \leq y_{T(2)} \leq \cdots \leq y_{T(r)} \leq \cdots \leq y_{T(N)}, \qquad (4)$$

where $T(r)$ denotes the original frame index of feature component $y_{T(r)}$ in which its rank is given by r. Then, the order statistic-based test CDF estimate of test feature component y_n is given by

$$\hat{C}_Y(y_n) = \frac{R(y_n)}{N}, \qquad (5)$$

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 (4). Given test feature y_n, an estimate of the reference feature by HEQ-FC using the order statistic-based test CDF estimation is obtained by assigning (5) to (2) as

$$\hat{x}_n = C_X^{-1}\left[\frac{R(y_n)}{N}\right]. \qquad (6)$$

Because the reference CDF is approximated by its cumulative histogram, the inverse reference CDF transformation in (6) is performed with a linear interpolation by considering the relative position of test CDF estimate within the reference histogram bin to reduce the quantization error.

3. HEQ for Model Adaptation

3.1. Basic Algorithm. To employ the HEQ technique in the model space, we interpret the acoustic mismatch between the noisy test environment and the clean reference environment as a transformation function $y = G(x)$, which is the inverse function of the transformation used in HEQ-FC. In model adaptation, the trained acoustic models of a speech recognizer are adapted to be acoustically matched into the test environment. Therefore, the HEQ technique for model adaptation (HEQ-MA) transforms the trained acoustic models into the test environment-matched models such that their transformation function follows $y = G(x)$. If the acoustic models under training and test environments are denoted as Φ_X and Φ_Y, respectively, the transformation function of HEQ-MA is obtained by mapping the reference PDF $P_x(\Phi_X)$ into the test PDF $P_Y(\Phi_Y)$ as

$$\Phi_Y = G(\Phi_X) = F^{-1}(\Phi_X) = C_Y^{-1}(C_X(\Phi_X)). \qquad (7)$$

3.2. Mean Model Adaptation. Our model adaptation approach is aimed to provide the speech recognizer with robustness against acoustic noise. Therefore, the actual adaptation is applied on the acoustic models of the speech recognizer. In most cases, the acoustic model adaption is focused on the mean vectors and covariance matrices of the acoustic models in the speech recognizer due to their dominant effectiveness compared with other model parameters [1, 2]. Hence, we confine the adaptation scope to both mean and covariance models in this paper.

Let μ denote the mean vector of a trained acoustic model in a speech recognizer obtained from the clean speech data. It is then assumed that HEQ-MA is applied to each mean vector of all trained acoustic models in the speech recognizer on a component-by-component basis as in HEQ-FC. Under these assumptions, the adaptation rule for HEQ-MA is given by using (7) and a linear interpolation between two test feature components in the sequence S which are the nearest to the trained mean component in terms of their CDF values such as

$$\begin{aligned}\hat{\mu}(k) &= C^{-1}_{Y(k)}\left(\hat{C}_{X(k)}(\mu(k))\right)\\ &= \begin{cases}\alpha(k)y_{T(m)}(k) + (1-\alpha(k))y_{T(m+1)}(k), \\ \qquad\qquad 1 \leq m < N, \\ \alpha(k)y_{T(N)}(k) + (1-\alpha(k))\left(y_{T(N)}(k) + \rho(k)\right), \\ \qquad\qquad m = N,\end{cases}\end{aligned} \qquad (8)$$

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

$$\hat{C}_{Y(k)}(y_{T(m-1)}(k)) < \hat{C}_{X(k)}(\mu(k)) \leq \hat{C}_{Y(k)}(y_{T(m)}(k)), \qquad (9)$$

$\rho(k)$ is a linear extrapolation factor for the boundary condition at the kth mean component and is set to the interval between the two last order statistic values in our case, and $\alpha(k)$ is the linear interpolation factor of the kth mean component and is determined as the relative position of $\hat{C}_{X(k)}(\mu(k))$ between the two boundary values in (9) such as

$$\alpha(k) = \frac{\hat{C}_{Y(k)}(y_{T(m)}(k)) - \hat{C}_{X(k)}(\mu(k))}{\hat{C}_{Y(k)}(y_{T(m)}(k)) - \hat{C}_{Y(k)}(y_{T(m-1)}(k))}, \qquad (10)$$

where the test CDF estimate of the undefined feature component $y_{T(0)}$ is assumed to be zero to satisfy its boundary condition. By using the order statistics-based test CDF estimate defined in (5), the interpolation factor in (10) can be further simplified as

$$\alpha(k) = m - N\hat{C}_{X(k)}(\mu(k)). \qquad (11)$$

When the acoustic models are estimated by using the logarithmically scaled features such as cepstral coefficients, the transformation function driven by the acoustic mismatch defined in (8) is known to be the form of a nonlinear function. In this nonlinear case, the mean of transformed features is not generally the same as the transformed mean value. However, it can be assumed that the transformed features belonging to each acoustic model are distributed in its relatively small acoustic space due to the detailed definition of acoustic models. In this case, the transformation within each acoustic model space can be approximated linearly even though the overall transformation through the entire acoustic model space has nonlinear characteristics. Under this assumption, the HEQ algorithm is applied for mean model adaptation as in (8).

3.3. Covariance Model Adaptation. As noise corruption increases, the dynamic range of certain features such as cepstral features tends to shrink due to the spectral whitening effect. Because the dynamic range is directly related to the covariance matrices of the acoustic models, it is expected that the covariance shrinkage [14] can occur in noisy features. For this reason, it is generally known that the improvements gained using mean and variance adaptation over mean adaptation only are especially large in noisy environments, although adapting the means provides the greater effect on performance [19]. The proposed HEQ-MA technique focuses its adaptation target on the mean models. Therefore, to cope with the covariance shrinkage effect in noisy environments, it is required to introduce an efficient adaptation rule for the covariance models. Because the covariance shrinkage tends to increase at the severer noise corruption, the covariance adaptation rule needs to take into account the SNR condition of the input utterance. For this purpose, an efficient covariance adaptation rule is given by a linear interpolation between the trained covariance and the sequence-level sample covariance by SNR dependently as

$$\hat{\Sigma}(k,l) = (1-\beta(\gamma))\Sigma(k,l) + \beta(\gamma)\frac{\Sigma^{S}(k,l)}{\Sigma^{G}(k,l)}\Sigma(k,l), \qquad (12)$$

where $\hat{\Sigma}(k,l)$ and $\Sigma(k,l)$ are the adapted and trained covariance coefficients, respectively, and $\beta(\gamma)$ is an SNR-dependent smoothing factor to deal with the higher covariance shrinkage effect at the lower SNR conditions and is approximated by a linearly decreasing function $\beta(\gamma) = a\gamma + b$, ranging between 0 and 1, where γ is the averaged SNR value of the sequence S, and a and b are empirically chosen slope and bias constants, respectively. The parameters $\Sigma^{S}(k,l)$ and $\Sigma^{G}(k,l)$ are the sequence-level sample covariance coefficient obtained from the test sequence S and the global sample covariance coefficient computed from the whole training features, respectively. Equation (12) indicates that the proposed covariance model adaptation rule tries not only to make the trained covariance models less changed at the higher SNR condition but also to make them shrunk by the ratio of the sequence-level sample covariance to the global sample covariance at the lower SNR condition.

4. Experimental Results

4.1. Experimental Setup. In the experiments, we used two speech databases, the Aurora2 speech database [20] converted from the TI-DIGITS database and the Korean phonetically optimized words (KPOW) database [21] consisting of 37,993 utterances of 3,848 Korean words, to examine the effectiveness of the proposed approach in the small as well as the large vocabulary speech recognition tasks. The trained acoustic models of two baseline speech recognizers were separately obtained from the clean speech training sets of the two databases. For performance evaluation, we used the three test sets of the Aurora2 noisy speech database, where test set A was added by four kinds of noise (subway, babble, car, and exhibition), test set B was corrupted by another four types of noise (restaurant, street, airport, and train station), and test set C was contaminated by two kinds of noise (subway and street) and channel distortion (MIRS) together [20]. Additionally, we used two test sets of the KPOW noisy speech database, which were generated by artificially adding the same kinds of the Aurora noise used in the Aurora2 test sets A and B to the KPOW clean speech test set composed of 7,609 utterances. Each of the three Aurora2 test sets and the two KPOW test sets is further composed of 6 noisy subsets with the SNR levels of 20, 15, 1, 5, 0, and −5 dB.

We employed the ETSI Aurora2 experimental framework [20] in our experiments as follows. In feature extraction, speech signals are firstly blocked into a sequence of frames, each 25 ms in length with a 10 ms interval. Next, speech frames are preemphasized 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, 12-dimensional mel-frequency cepstral coefficients (MFCCs) are extracted. The final 39-dimensional feature vector for each frame consists of 12 MFCCs, log energy, and their delta and acceleration coefficients. The baseline speech recognizer for the Aurora2 task employs 13 whole-word-based 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. The states for digit models are composed of 3 Gaussian mixture components while those for silence and short-pause models have 6 Gaussian mixture components, respectively. The baseline recognizer for the KPOW task has 6,776 tied-state triphone-based HMMs, where each HMM has 3 states and each state is modeled with 8 Gaussian mixture components. Diagonal covariance matrices are used in all of the HMMs.

In the performance evaluation, the performances of the baseline speech recognizer trained on the clean speech data, CMN, CMVN, and HEQ-FC, and HEQ-MA were examined. Additionally, the performance of the standard model adaptation technique based on the MLLR method was also evaluated by using the HTK toolkit [22] and compared with those of the above mentioned techniques. The number of regression classes in MLLR was set to 8 for the Aurora2 task and 16 for the KPOW task. In the MLLR-based model adaptation, we adopted the unsupervised adaptation method where the acoustic mean models were incrementally adapted with each test utterance. In feature compensation, HEQ-FC is applied to all of the 39-dimensional MFCCs independently for both training and test data after estimating the reference CDFs from all training data. In the HEQ-based model adaptation, the HEQ and 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 on a component-by-component basis. The number of histogram bins in the reference CDFs was empirically chosen as 64. Due to the adoption of the linear interpolation in the inverse CDF transformation of both HEQ-FC in (6) and HEQ-MA in (8), a further increase in the number of histogram bins did not show any meaningful performance improvements. The SNR-dependent smoothing parameters a and b in the adaptation of covariance matrices are set to -0.028 and 0.9, respectively, to make the smoothing factor $\beta(\gamma)$ become 0 and 1 at the SNRs of 30 dB and -5 dB, respectively. 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 in each test utterance. To cope with the time-varying nature of environmental noise, the histogram equalization was conducted on a single utterance basis in both feature compensation and model adaptation.

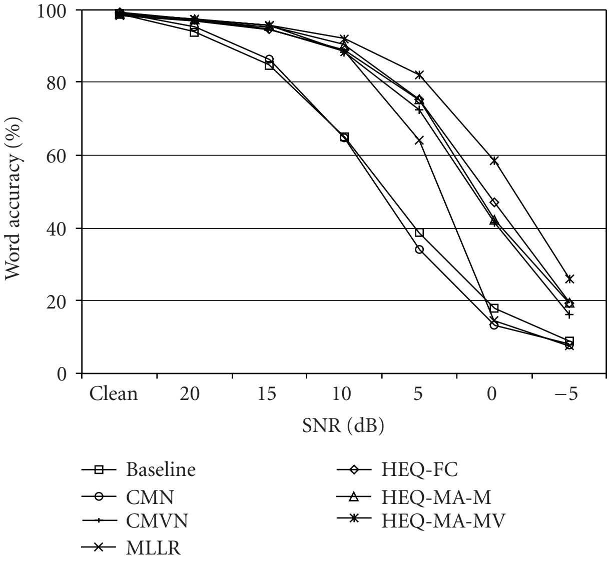

Figure 1: Recognition results on the Aurora2 data at various SNR conditions by the baseline speech recognizer, CMN, CMVN, MLLR, HEQ-FC, HEQ-MA-M (mean-only adaptation), and HEQ-MA-MV (mean and variance adaptation) techniques.

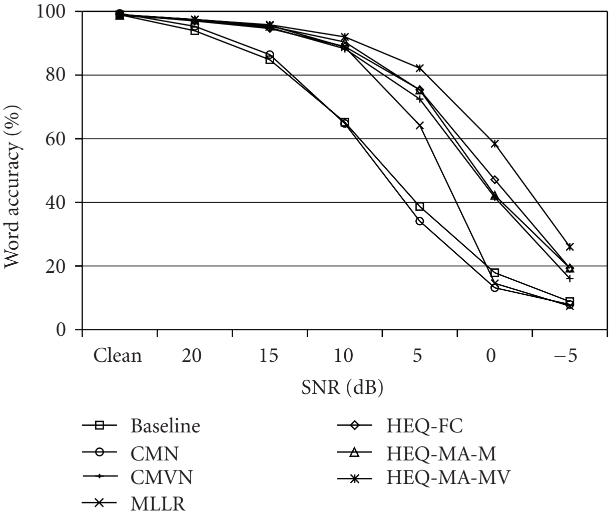

Figure 2: Recognition results on the Korean POW data at various SNR conditions by the baseline speech recognizer, CMN, CMVN, MLLR, HEQ-FC, HEQ-MA-M (mean-only adaptation), and HEQ-MA-MV (mean and variance adaptation) techniques.

4.2. Test with SNR Conditions. Figure 1 illustrates the recognitions results on the Aurora2 test sets at various SNR conditions in terms of the averaged word accuracy for all of the three test sets. The figure indicates that the CMN technique produces almost the same performance as the baseline speech recognizer trained on the clean speech data. The recognition performance is slightly improved in the higher SNR conditions above 10 dB but degraded noticeably in the lower SNR conditions. On the contrary, it is observed that the CMVN technique produces meaningful performance improvements, especially at the higher SNRs. It is also indicated that both HEQ-FC and the two HEQ-MA approaches, HEQ-MA with mean adaptation only (HEQ-MA-M) and HEQ-MA with mean and variance adaptation (HEQ-MA-MV), provide significant performance improvements over all SNR conditions. The MLLR technique yields the comparable performance compared to HEQ-FC and HEQ-MA at the SNR conditions above 10 dB but it degrades sharply at the lower SNRs. When the MLLR adaptation technique is conducted on the single utterance basis, it usually produces poor performance. For this reason, the MLLR adaptation technique in our experiment was performed incrementally through the test utterances. HEQ-MA-M is better than HEQ-FC at higher SNRs than 5 dB but it also becomes inferior at the SNR condition of 0 dB. On the contrary, HEQ-MA-MV yields substantial improvements over HEQ-FC for all SNR conditions. Figure 2 shows the recognition results on the KPOW test sets at various SNR conditions. The CMN technique in this task produces notable performance gains and the CMVN technique consistently provides substantial performance improvements over the baseline speech recognizer. The relative gains obtained by the CMVN technique over the CMN technique suggest that the variance normalization be a quite effective means for reducing the acoustic mismatches. The results provided by the MLLR technique look very similar to those from the Aurora2 task, which confirms that the MLLR with mean-only adaptation technique is effective only in the slightly corrupted noise environments. Compared to HEQ-FC, HEQ-MA with both mean-only adaptation and mean and variance adaptation shows higher performance.

4.3. Variance Shrinkage Effect. Figure 3 illustrates the change of MFCC-based feature covariance with regard to various SNR conditions. The results are represented as the global diagonal covariance values for the three Aurora2 test sets. In

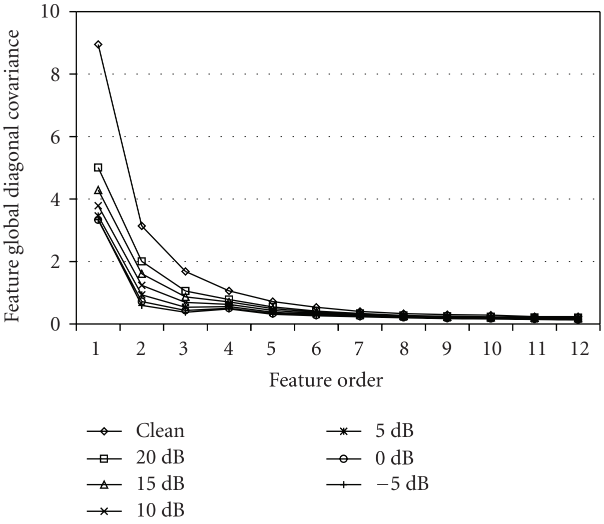

Figure 3: Changes in feature variance with various SNR conditions.

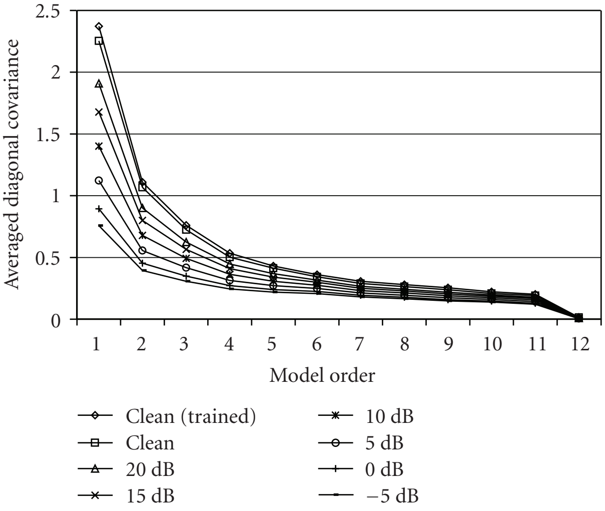

Figure 4: Changes in average values of covariance models after HEQ-MA.

this figure, it is observed that the variance decreases sharply as the SNR condition is lowered. The variance value even at the slight noise corruption of 20 dB SNR seems to be reduced by half compared with the value at the clean condition. As a result, this figure strongly suggests that variance models should be adjusted according to their corresponding noisy conditions to reduce the variance mismatch.

Figure 4 shows the change of the diagonal covariance at various SNR conditions when the proposed variance adaptation technique was used to adapt the variance models. The results are obtained as the average values of all diagonal covariance models of the speech recognizer used in the Aurora2 task. In Figures 3 and 4, we observe some differences in scale between the test feature variance and the averaged variance model. These scale differences are resulted from the fact that the feature variance in Figure 3 is computed globally from the entire test features while the averaged variance model is represented as the average value of the individual variance models, each of which covers its small acoustic model space. Therefore, if the feature variance is computed in its acoustic unit level, it should be closer in scale to the average variance model. After considering these differences, we observe in Figure 4 that the averaged variance model decreases quite linearly with the SNR conditions of test utterances to compensate for the mismatch between the input feature variance and the variance models.

4.4. Test with Various Test Sets. Tables 1 and 2 show the recognition results for the Aurora2 and KPOW tasks, respectively. The results are represented in terms of the word error rates which are averaged between 0 and 20 dB SNRs as proposed by the Aurora Group [20]. In the Aurora2 task, the performance of the HEQ-MA with mean-only adaptation technique is similar to that of HEQ-FC. From Figure 1, it is supposed that the poor performance gain in HEQ-MA compared with HEQ-FC is mainly due to its performance degradation at the lower SNR conditions. On the contrary, the HEQ-MA with mean and variance adaptation technique provides relative error reductions of 62.83%, 63.36%, 31.52%, 46.71%, and 23.40% over the baseline recognizer, CMN, CMVN, MLLR, and HEQ-FC, respectively. Compared to the multicondition training scheme, it produces slightly worse results in terms of the overall performance. However, it is seen that most of the gains in the multicondition training scheme are obtained from test set A, where noise conditions are the same as those employed in the training set. In the cases of test sets B and C, where noise types are not exposed to the training phase, the proposed approach is even better than the multicondition training scheme.

In the large vocabulary-based KPOW task, the HEQ-MA with mean and variance adaptation technique reduces recognition errors by 43.04%, 38.77%, 15.05%, 41.39%, and 8.56% over the baseline recognizer, CMN, CMVN, MLLR, and HEQ-FC, respectively. From these results, we see that HEQ-MA produces substantially better performance than the other approaches. Compared to the results in the Aurora2 task, the reduced performance gains in the KPOW task imply that the proposed adaptation technique is more suitable for the small vocabulary task due to the fewer possibilities of unobserved acoustic models in the HEQ-based adaptation process. The performance gap between both mean-only adaptation and mean and variance adaptation confirms both the importance of variance adaptation in the SNR conditions lower than 10 dB and the effectiveness of our proposed variance adaptation approach.

5. Summary Analysis

5.1. Recognition Performance. In Figures 1 and 2, the slight performance gain in the high SNR conditions by CMN indicates that CMN can improve recognition performance in the moderate noisy conditions. In addition, its performance degradation in the low SNR conditions also implies that CMN is not very effective in the heavily noisy conditions. This result can be interpreted to mean that the performance degradation caused by the variance mismatch can be larger

Table 1: Word error rates (%) on the Aurora2 task (Results are averaged between 0 and 20 dB SNRs).

Test Sets	Baseline	CMN	CMVN	Multicondition	MLLR	HEQ-FC	HEQ-MA (mean-only)	HEQ-MA (mean & var.)
A	38.87	42.06	21.43	12.71	29.96	19.41	21.13	15.19
B	44.43	40.79	20.43	14.49	24.88	18.32	17.63	14.00
C	33.32	37.13	24.80	16.88	29.77	21.55	21.95	15.93
Average	39.98	40.56	21.70	14.26	27.89	19.40	19.90	14.86

Table 2: Word error rates (%) on the Korean POW task (Results are averaged between 0 and 20 dB SNRs).

Test Sets	Baseline	CMN	CMVN	MLLR	HEQ-FC	HEQ-MA (mean-only)	HEQ-MA (mean & var.)
A	64.64	58.19	43.57	43.46	40.22	41.01	36.33
B	56.45	54.43	37.62	38.88	35.20	33.90	32.64
Average	60.54	56.31	40.59	58.83	37.71	37.46	34.48

than the performance gain obtained by the mean adaptation in the heavily noisy conditions, which results in the overall degradation of recognition performance. Similar results are obtained in the MLLR-based mean-only model adaptation experiments. We believe that the superior performance of MLLR at the higher SNRs is also largely resulted from the mean model adaptation. Similar to the case of CMN, the variance mismatch can be regarded as the main cause of performance degradation in both MLLR and HEQ-MA with mean-only adaptation approaches under heavily noisy conditions. Therefore, it can be said that the variance adaptation plays the more crucial role at the lower SNR conditions. The importance of variance compensation is well confirmed by CMVN as well as HEQ-FC, both of which noticeably improve the recognition performance compared to CMN. Because the HEQ-MA with mean and variance adaptation technique tries to reduce the variance mismatch by the proposed variance adaptation technique, it can provide further performance gain compared to the HEQ-MA with mean-only adaptation approach as observed in Figures 1 and 2. From the results, it can be said that the proposed techniques are also effective in the large vocabulary task although the performance gains obtained by HEQ-MA over HEQ-FC are not as remarkable as those at the Aurora2 task shown in Figure 1. We think that the reduced performance improvements in the KPOW task are mainly resulted from the reason that because the adaptation is performed on a single utterance basis, the amount of adaptation data in each test utterance becomes not enough to fully adapt the much larger number of acoustic models in this large vocabulary task. In Figures 1 and 2, it is also observe that the performance gains obtained by HEQ-MA-MV over HEQ-FC are more notable at the lower SNR conditions. These results support our previous suggestion that model adaptation is more effective than feature compensation in serious noise conditions where it becomes more difficult to compensate noisy speech features into clean speech features due to the increased loss of acoustic-phonetic information.

5.2. Computational Complexity. The computational loads in HEQ-MA are directly related to the number of acoustic mean models whereas those in HEQ-FC are dependent upon the utterance length, that is, the number of frames on the given utterance. The usual speech recognition tasks require the whole phonetic units in acoustic modeling. In this case, the number of acoustic mean models tends to be much larger than the number of frames. Therefore, it can be said that HEQ-MA usually requires much larger amounts of computational load than HEQ-FC. However, the computational loads in HEQ-MA can be comparable to those in HEQ-FC in the domain-constrained speech recognition task such as digit recognition task which employs a small number of acoustic models. Although HEQ-MA has much larger computational complexity than feature compensation techniques, it can be still regarded as an efficient model adaptation technique compared to other more complex model adaptation techniques such as MLLR due to its predominantly simple algorithmic complexity.

5.3. Implementation Issues. The feature compensation and model adaptation techniques employed in this experiment are basically conducted in the utterance-by-utterance basis to estimate the required statistics such as mean and variance. Therefore, all these approaches produce some amount of time delay in the real applications. However, a segmental estimation approach utilizing a sliding window can be used to achieve the real-time processing of feature compensation and model adaptation without any significant performance degradation. In this approach, it is reported that the appropriate size of a sliding window producing comparable results compared to the utterance-by-utterance based approach is about 600 ms for these feature compensation techniques [13].

In the HEQ-MA with mean and variance adaptation approach, we used an SNR-dependent covariance model adaptation technique. In this approach, the more accurate frame-level SNR estimation is required for the better covariance model adaptation. In our experiments, we employed a simple SNR estimation method, where the noise power estimated from the initial silence region is used through the entire utterance without any update procedures. Therefore, it can be said that the estimated noise power has some degree of estimation error, which causes the resulting covariance models to be adapted less accurately. More reliable noise power estimation algorithm employed in the voice activity detection techniques can be used for better SNR estimation.

It is worthwhile conducting a further research activity utilizing this kind of more reliable SNR estimation technique.

6. Conclusion

We proposed a new environmental model adaptation method for robust speech recognition. The proposed approach utilizes both the histogram equalization technique for matching the acoustic mean models and an SNR-dependent linear interpolation-based method for adapting the covariance models into test environments. According to the experimental results, the proposed model adaptation approach provides substantial effectiveness in reducing the mismatch between trained acoustic models and test environments. The experimental results also indicate that the mean model adaptation plays the major role in improving the performance of the speech recognizer in noisy environments. Additionally, the variance model adaption is especially important for improving the recognition performance in the heavily noisy conditions. Due to its computational efficiency as well as noise robustness, the proposed technique can be another model adaptation approach to robust speech recognition under noisy environments. Further study about more sophisticated variance adaptation techniques is needed for enhancing the performance of the proposed approach more.

References

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

[2] M. Gales and S. Young, "The application of hidden Markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, no. 3, pp. 195–304, 2007.

[3] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 3, pp. 190–202, 1996.

[4] 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.

[5] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Cepstral channel normalization techniques for HMM-based speaker verification," in *Proceedings of the 2nd International Conference on Spoken Language Processing*, pp. 1835–1838, 1992.

[6] O. Viikki and K. Laurila, "Cepstral domain segmental feature vector normalization for noise robust speech recognition," *Speech Communication*, vol. 25, no. 1–3, pp. 133–147, 1998.

[7] M. J. F. Gales and S. J. Young, "Cepstral parameter compensation for HMM recognition in noise," *Speech Communication*, vol. 12, no. 3, pp. 231–239, 1993.

[8] P. J. Moreno, B. Raj, and R. M. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '96)*, vol. 2, pp. 733–736, Atlanta, Ga, USA, May 1996.

[9] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171–185, 1995.

[10] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, Prentice-Hall, Upper Saddle River, NJ, USA, 2002.

[11] F. Hilger and H. Ney, "Quantile based histogram equalization for noise robust large vocabulary speech recognition," in *Proceedings of the 7th European Conference on Speech Communication and Technology*, pp. 1135–1138, 2006.

[12] J. Pelecanos and S. Sridharan, "Feature warping for robust speaker verification," in *Proceedings of the Speaker Odyssey*, pp. 213–218, 2001.

[13] 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.

[14] 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 Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 355–366, 2005.

[15] Y. Suh and H. Kim, "Class-based histogram equalization for robust speech recognition," *ETRI Journal*, vol. 28, no. 4, pp. 502–505, 2006.

[16] 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 67870, 9 pages, 2007.

[17] 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.

[18] S. Dharanipragada and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proceedings of the 6th International Conference on Spoken Language Processing*, pp. 556–559, 2000.

[19] M. J. F. Gales and P. C. Woodland, "Mean and variance adaptation within the MLLR framework," *Computer Speech and Language*, vol. 10, no. 4, pp. 249–264, 1996.

[20] D. Pearce and H.-G. Hirsch, "The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proceedings of the 6th International Conference on Spoken Language Processing*, pp. 29–32, 2000.

[21] Y. Lim and Y. Lee, "Implementation of the POW (phonetically optimized words) algorithm for speech database," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '95)*, vol. 1, pp. 89–92, Detroit, Mich, USA, May 1995.

[22] S. Young, et al., *The HTK Book for Version 3.2.1*, Cambridge University Engineering Department, Cambridge, UK, 2002.