Environmental Model Adaptation Based on Histogram Equalization

Youngjoo Suh, Member, IEEE, and Hoirin Kim, Member, IEEE

Abstract—In this letter, a new environmental model adaptation method is proposed for robust speech recognition under noisy environments. The proposed method adapts initial acoustic models of a speech recognizer into environmentally matched models by utilizing the histogram equalization technique. Experiments performed on the Aurora noisy environment showed that the proposed technique provides substantial improvement over the baseline speech recognizer trained on the clean speech data.

Index Terms—Environmental model adaptation, histogram equalization, robust speech recognition.

I. INTRODUCTION

Currently, most speech recognizers trained on clean speech data usually show serious performance degradation when they are tested in acoustically mismatched noisy environments [1], [2]. The main cause of the acoustic mismatch between training and test environments is known as the corruption by additive noise and channel distortion. Therefore, one of the hot issues in automatic speech recognition is to provide speech recognizers with robustness against the acoustic mismatch between the two environments [3]. Current robust speech recognition techniques 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 which compensates noisy test features into compensated test features, which are then decoded by speech recognizers trained on clean speech data [3]. The histogram equalization (HEQ) technique is known to be one of the most efficient feature compensation approaches because most of its algorithm consists of sorting and search routines with narrow depths [4]–[8]. Additionally, it provides considerable effectiveness in feature compensation because of its nonlinear transformation characteristics which can fundamentally cope with the nonlinearity of noisy speech features in the logarithmic feature spaces such as cepstral coefficients [4], [8]. However, due to some acoustic-phonetic information loss in noisy speech features, there are unavoidable discrepancies between clean speech models and compensated features. On the contrary, the clean speech models can be fully adapted into acoustically matched speech models as far as the amount of adaptation data is enough in model adaptation [1]. Therefore, although the same information loss can occur in model adaptation, it does not cause undesirable discrepancies between acoustic models and speech features in the decoding process. For this reason, the model adaptation approach can be more effective than the feature compensation one in reducing the mismatch between acoustic environments. Two well-known model adaptation methods are parallel model combination and vector Taylor series [1], [2]. These methods use both clean speech and noise models to approximate noisy speech models and are reported to be very effective in reducing the mismatch. However, in the computational aspect, they are not very efficient due to the transformations between log-spectral and cepstral domains.

In this letter, we propose a new efficient model adaptation approach based on the histogram equalization technique to take advantage of its effectiveness in providing noise robustness as well as computational efficiency in the model space. In the proposed approach shown in Fig. 1, the histogram equalization technique transforms initial acoustic models of a speech recognizer into environmentally adapted acoustic models. The transformation function of the histogram equalization technique is obtained by using the reference and test cumulative distribution functions (CDFs) of the training data and utterance input, respectively. Then, the speech recognizer based on the adapted acoustic models decodes the utterance input into the recognized output.

II. HISTOGRAM EQUALIZATION FOR FEATURE COMPENSATION

A basic assumption in utilizing HEQ for feature compensation (HEQ-FC) is that the acoustic mismatch between clean reference features and noisy test ones is directly related to the discrepancy between their corresponding probability density functions (PDFs). Then, the idea of HEQ-FC is to convert the test PDF into the reference one to recover clean reference features.
from noisy test features. Here, HEQ-FC is applied to each feature component independently for algorithmic simplicity under the assumption of statistical independence which can be accepted in the orthogonal features such as cepstral coefficients. In this case, the algorithm of HEQ-FC is given as follows.

For given reference and test random variables $x$ and $y$, respectively, a transformation function $x = F(y)$ by HEQ-FC mapping the test PDF $P_X(y)$ into the reference PDF $P_X(x)$ is obtained by equating their corresponding CDFs defined in [4] as

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

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 an inverse transformation function of the acoustic mismatch and has monotonically nondecreasing characteristics. From (2), it can be inferred that the effectiveness of HEQ-FC is directly related to the reliable estimation of both reference and test CDFs. In practice, each CDF in (2) is approximated by its cumulative histogram. Therefore, a better CDF estimation is achieved by a larger amount of sample data. In this sense, the reference CDF can be obtained quite reliably because a large amount of sample data can be available at the training phase. On the contrary, when short utterances are used as the test data in speech recognition, the amount of sample data may be insufficient for the reliable estimation of the test CDF. Therefore, a reliable estimation of the test CDF becomes an important issue for effective HEQ-FC in this test environment. 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. A brief algorithm of the order-statistics-based is given as follows [4], [6].

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

$$S = \{y_1, y_2, \ldots, y_h, \ldots, y_N\}$$

where $y_h$ is a test feature component at the $n$th frame.

The order-statistics of the sequence $S$ in (3) is represented by rearranging its elements in ascending order as

$$y_{T(1)} \leq y_{T(2)} \leq \cdots \leq y_{T(i)} \leq \cdots \leq y_{T(N)}$$

where $T(r)$ represents the original frame index of the feature component whose rank is denoted by $r$. Then, the order-statistic-based test CDF estimate of the test feature component $y_h$ is given as

$$\hat{C}_Y(y_h) = \frac{R(y_h)}{N}$$

where $R(y_h)$ denotes the rank of $y_h$ among the feature components composing the sequence $S$ according to the order-statistics defined in (4). From (2) and (5), an estimate of the reference feature by HEQ-FC given the test feature $y_h$ is obtained as

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

III. HISTOGRAM EQUALIZATION FOR MODEL ADAPTATION

Unlike HEQ-FC, the histogram equalization for model adaptation (HEQ-MA) technique models the acoustic mismatch between the acoustic model under the noisy test environment $\Phi_Y$ and that under the clean reference one $\Phi_X$ as the transformation function $\Phi_Y = G(\Phi_X)$ in the model space. The transformation function of HEQ-MA is obtained by mapping the reference PDF $P_X(\Phi_X)$ into the test PDF $P_Y(\Phi_Y)$ by the following rule, which is the inverse of the transformation function of HEQ-FC, as

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

Let $\mu$ and $\Sigma$ denote the mean vector and covariance matrix of an initial acoustic model in a speech recognizer trained on clean speech data, respectively. We then assume that HEQ-MA is applied to each mean vector of all acoustic models in the speech recognizer on a component-by-component basis. 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 nearest to the initial mean component in terms of the CDF value as (8), shown at the bottom of the page, where $\hat{\mu}(k)$ and $\mu(k)$ denote the $k$th components of the adapted and initial mean vectors, respectively, $C_Y^{-1}$ is the inverse of the test CDF for the $k$th test feature component, $\hat{C}_X(k)(\mu(k))$ is the reference CDF estimate of the $k$th mean component $\mu(k), m$ is the smallest rank index satisfying the relationship $\hat{C}_X(k)(\mu(k)) \leq \hat{C}_Y(k)(y_{T(m)}(k))$, $\rho(k)$ is a linear extrapolation factor for the boundary condition at the $k$th mean component and is set to the interval between the last two order statistic values, and $\alpha(k)$ is the linear interpolation factor of the $k$th mean component that is based on the order-statistics-based test CDF of the sequence $S$ defined in (5) and is given by

$$\alpha(k) = \frac{\hat{C}_Y(k)(y_{T(m)}(k)) - \hat{C}_X(k)(\mu(k))}{\hat{C}_Y(k)(y_{T(m)}(k)) - \hat{C}_Y(k)(y_{T(m-1)}(k))} = m - N\hat{C}_X(k)(\mu(k))$$

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

In environmental model adaptation, 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 [9]. Therefore, unlike HEQ-FC, HEQ-MA needs to adapt not only mean vectors but also covariance matrices. For this purpose, an efficient covariance adaptation rule, which depends on the degree of noise corruption, is given by a linear interpolation of the initial covariance matrix and the sequence-level global sample covariance matrix as

\[
\hat{\Sigma}(k, l) = \beta(\gamma)\Sigma^a(k, l) + (1 - \beta(\gamma))\Sigma^g(k, l)
\]  

where \(\beta(\gamma)\) is a signal-to-noise ratio (SNR)-dependent smoothing factor given by

\[
\beta(\gamma) = a\gamma + b,
\]

where \(\gamma\) is the averaged SNR value of the sequence \(S\) and \(a\) and \(b\) are empirically chosen slope and bias constants, respectively, and \(\Sigma^g\) denotes the global sample covariance matrix of the sequence \(S\) from the maximum likelihood estimation criterion as

\[
\Sigma^g(k, l) = \frac{1}{N} \sum_{n=1}^{N} (y_n(k) - \nu(k))(y_n(l) - \nu(l))
\]

where \(\nu(f)\) represents the global sample mean of the \(f\)th feature components in sequence \(S\).

IV. EXPERIMENTAL RESULTS

In the performance evaluation, we used two speech databases, the Aurora2 speech database [10] converted from the TI-DIGITS database and the KPOW (Korean phonetically optimized words) database [11] consisting of 37,993 utterances of 3848 Korean words, to examine the effectiveness of the proposed approach in the small as well as large vocabulary speech recognition tasks. The initial acoustic models of two baseline speech recognizers were separately trained on the clean speech training sets of the two databases. In the evaluations, we used the three test sets of the Aurora2 noisy speech database, where test set A was added by four kinds of noise, test set B was corrupted by another four types of noise, and test set C was contaminated by two kinds of the noise and channel distortion together [10]. Additionally, we used two test sets of the KPOW noisy speech database, which were generated by artificially adding the same eight kinds of the Aurora noise to the KPOW clean speech test set composed of 7609 utterances. We employed the ETSI Aurora2 experimental framework following [10]. 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 pre-emphasized by 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 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 three Gaussian mixture components while those for silence and short-pause models have six Gaussian mixture components, respectively. The baseline recognizer for the KPOW task has 6776 tied-state triphone-based HMMs, each has three states and eight Gaussian mixture components per state. Diagonal covariance matrices are used in all of the HMMs. In the performance evaluation, the performance of the baseline speech recognizer trained on the clean speech data, HEQ-FC, and HEQ-MA were examined. In feature compensation, HEQ-FC is applied to all of the 39-D MFCCs independently for both training and test data after estimating reference CDFs from the training data. In model adaptation, the HEQ and proposed variance adaptation techniques are applied to the 39-D mean vectors and diagonal covariance matrices, respectively, of all initial HMMs in the baseline speech recognizers on a component-by-component basis. The number of histogram bins in reference CDFs was empirically chosen as 64. The SNR-dependent smoothing parameters \(a\) and \(b\) in the adaptation of covariance matrices are empirically set to \(-0.03\) and \(0.9\), respectively. The averaged SNR value \(\gamma\) was estimated as the ratio of the averaged frame energy to the averaged noise energy of the initial silence region in each utterance. The histogram equalization is conducted on an utterance-by-utterance basis in feature compensation and model adaptation.

Fig. 2 shows the recognition 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 HEQ-MA approaches, HEQ-MA with mean adaptation only (HEQ-MA-M) and HEQ-MA with mean and variance adaptation (HEQ-MA-M&V), provide significant improvements over the baseline speech recognizer. HEQ-MA-M is better than HEQ-FC at high SNRs but it becomes inferior at the SNR condition of 0 dB. On the contrary, HEQ-MA-M&V yields substantial improvements over HEQ-FC even at the lower SNRs. Therefore, we believe that the inferior result of HEQ-MA-M at the lower SNR is mainly due to the mismatch in the covariance models and thus the variance adaptation plays a very crucial role at the lower SNRs. Fig. 3 illustrates the recognition results on the KPOW test sets at various SNR conditions. The results indicate that the proposed techniques are also effective in the large vocabulary task although their performance improvements are not as remarkable as those at the Aurora2 task. In Figs. 2 and 3, we observe that the...
Fig. 3. Recognition results on the Korean POW data with regard to various SNR conditions by the baseline speech recognizer, HEQ-FC, HEQ-MA-M (mean adaptation only), and HEQ-MA-M&V (mean and variance adaptation) techniques.

V. CONCLUSION

In this letter, we propose a new environmental model adaptation method for robust speech recognition. The proposed approach utilizes the histogram equalization technique which transforms initial mean vectors of the speech recognizer into the environmentally matched acoustic models on a single utterance basis. The covariance matrices are adapted SNR-dependently by using sample covariance matrices estimated in the input utterance level. According to the experimental results, the proposed model adaptation approach provides substantial effectiveness in reducing the acoustic mismatch between initial acoustic models of speech recognizers and test environments.

TABLE II

<table>
<thead>
<tr>
<th>Test Sets</th>
<th>Baseline</th>
<th>HEQ-FC</th>
<th>HEQ-MA</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean only</td>
<td>Mean &amp; Var.</td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>64.64</td>
<td>40.22</td>
<td>41.01</td>
</tr>
<tr>
<td>B</td>
<td>56.45</td>
<td>35.20</td>
<td>33.90</td>
</tr>
<tr>
<td>Average</td>
<td>60.54</td>
<td>37.71</td>
<td>37.46</td>
</tr>
</tbody>
</table>

VI. REFERENCES