Exploiting Spatial-Temporal Feature Distribution Characteristics for Robust Speech Recognition

Wei-Hau Chen, Shih-Hsiang Lin, Berlin Chen

Department of Computer Science & Information Engineering, National Taiwan Normal University, Taipei, Taiwan
696470631@ntnu.edu.tw, {shlin, berlin}@csie.ntnu.edu.tw

Abstract
Noise robustness is one of the primary challenges facing most automatic speech recognition (ASR) systems. Quite several speech feature histogram equalization (HEQ) methods have been developed to compensate for nonlinear noise distortions. However, most of the current HEQ methods are merely performed in a dimension-wise manner and without taking into consideration the contextual relationships between consecutive speech frames. In this paper, we present a novel HEQ approach that exploits spatial-temporal feature distribution characteristics for speech feature normalization. All experiments were carried out on the Aurora-2 database and task. The performance of the presented approach is tested and verified by comparison with the other HEQ methods. The experiment results show that for clean-condition training, our method yields a significant word error rate reduction over the baseline system, and also considerably outperforms the other HEQ methods compared in this paper.

Index Terms: noise robustness, speech recognition, histogram equalization, spatial-temporal distribution characteristics

1. Introduction

Varying environmental effects, such as ambient noises, noises caused by the recording equipments and transmission channels etc., often lead to severe mismatch between the acoustic conditions for the training and test speech data. Such mismatch no doubt will make the performance of an automatic speech recognition (ASR) system degrade dramatically. Substantial efforts have been made and also a number of techniques have been presented to cope with this issue and improve the ASR performance. Broadly speaking, these techniques fall into three main categories: (1) enhancement, (2) normalization, and (3) adaptation, while these approaches can be conducted either in the feature domain or in the model domain [1].

A wide variety of speech feature normalization methods have been developed over the decades. For example, the cepstral mean normalization (CMN) is a simple but effective way to remove the time-invariant distortions introduced by the transmission channel. A nature extension of CMN is the cepstral mean and variance normalization (CMVN) [2] that normalizes not only the features’ means but also their variances. Although these two methods do provide better ASR performance, they to some extent have their inherent limitation. They can only deal with linear distortions and cannot adequately compensate the non-linear environmental effects due to their linear property. On the other hand, in order to compensate the non-linear environmental effects, the histogram equalization (HEQ) methods have been proposed and extensively studied in the recent past [3-9], which have also been shown their superiority over the linear compensation approaches, such as CMN and CMVN. One of the remarkable features of the HEQ methods is that they not only attempt to match speech feature means or variances, but also completely match the feature distribution of the training and test data using transformation functions that are estimated based on the cumulative density functions (CDFs) of the training and test data. Even though HEQ has been shown its superiority for feature compensation, however, most of the current approaches still have room for improvement. For example, the table-lookup HEQ [3-5] typically needs a set of large tables kept in memory (the need of huge disk storage consumption) for performing the feature transformation, while the quantile based HEQ [6] instead needs on-line exhaustive search or optimization of the coefficients of the transformation function (the need of high computation cost) before the transformation is actually performed. Moreover, the aforementioned HEQ methods are merely performed in a dimension-wise manner, without taking into consideration the contextual relationships between consecutive speech frames [8].

In this paper, we present a novel feature normalization method, which exploits the spatial-temporal distribution characteristics of speech features for robust speech recognition (denoted as STHEQ). The proposed method is efficient in terms of time and memory complexity, in contrast to the conventional table-lookup or quantile based HEQ methods, and can leverage the contextual relationships of feature vector components, not only between different dimensions but also between consecutive speech frames, for speech feature normalization. The rest of this paper is organized as follows. Section 2 describes the basic concept of HEQ and three commonly used HEQ methods. Section 3 sheds light on our proposed feature normalization method. Then, the experiment settings and a series of ASR experiments conducted are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. Review of Histogram Equalization (HEQ)
Theoretically, HEQ has its roots in the assumptions that the transformed speech feature distributions of the test (or noisy) data should be identical to that of the training (or reference) data, and each feature vector dimension can be normalized independently of each other. Under the above two assumptions, the aim of HEQ is to find a transformation that can convert the distribution of each feature vector component of the input (or test) speech into a predefined target

91500794, Interspeech, 2008-319
distribution which corresponds to that of the training (or reference) speech. Accordingly, HEQ attempts not only to match the means and variances of the speech features, but also to completely match the speech feature distributions of training and test data. Put another way, HEQ normalizes all the moments of the probability distributions of the speech features. In practice, the equalization can be conducted either in a non-parametric way, such as the table-lookup histogram equalization (THEQ) [3-5], or in a parametric way, such as the quantile-based histogram equalization (QHEQ) [6].

2.1 Table-lookup Histogram Equalization (THEQ)

The cumulative histogram of each feature dimension \(d\) of all training data is computed and divided into a set of equally-probable bins, where the mean \(\gamma_d(l)\) of each bin \(l\) is taken as one of the representative outputs of the transformation function \(F(x_d)\) [3-5]. That is, each feature vector component \(x_d\) of the test utterance is replaced by the mean of a specific bin in the cumulative histogram of the training speech data that corresponds to the same bin position of \(x_d\) in the histogram of the test data. However, normalization of the test data alone results in only moderate gain of performance improvement. It is usually necessary to normalize the training data in the same way to avoid mismatch and to achieve good performance [9]. Moreover, because a set of cumulative histograms of all speech feature dimensions of the training data have to be kept in memory for the table-lookup of restored feature values, such an approach usually needs huge disk storage consumption and the table-lookup is also time-consuming.

2.2 Quantile-based Histogram Equalization (QHEQ)

In [6], a parametric type of histogram normalization, referred to as the quantile based histogram (QHEQ) approach, has been proposed. QHEQ attempts to calibrate the CDF of each feature vector component of the test data to that of the training data in a quantile-corrective manner instead of full-matching of the cumulative histogram as that done by the table-lookup approach described above. A transformation function \(H(x_d)\) is applied to each feature vector component \(x_d\) to the CDF of the equalized feature match that observed in training:

\[
H(x_d) = Q_{d, k} \left( a_d \left( \frac{x_d}{Q_{d, k}} \right) ^{\gamma_d} + (1 - a_d) \frac{x_d}{Q_{d, k}} \right),
\]

where \(K\) is the total number of quantiles; \(Q_{d, k}\) is the \(k\)-th quantile of a specific feature vector dimension \(d\) calculated from the utterance; and \(a_d\) and \(\gamma_d\) are transformation coefficients. For each feature vector dimension \(d\), \(a_d\) and \(\gamma_d\) are optimized using the following equation:

\[
\{a_d, \gamma_d\} = \arg\min_{\{a_d, \gamma_d\}, k=1} \left\{\sum_{k=1}^{K} \left[H(Q_{d,k}) - Q_{d,k}^{\min}\right]^2\right\},
\]

where \(Q_{d,k}^{\min}\) is the \(k\)-th quantile of the same feature vector dimension calculated from the training speech. QHEQ allows the estimation of the transformation function performance merely on the basis of a single test utterance (or eventually, a short utterance), without using additional adaptation data [5]. However, in order to find the optimal transformation coefficients for each feature vector dimension, an exhaustive grid search is required, which in fact is time-consuming.

2.3 Polynomial-fit Histogram Equalization (PHEQ)

Recently, we have presented an efficient method exploring the use of data fitting to approximate the inverse of the CDFs of the speech feature vector components for HEQ, named polynomial-fit histogram equalization (PHEQ) [8]. PHEQ makes use of data fitting (or so-called least squared error regression) to estimate the inverse functions of the CDFs of the training speech. For a specific speech feature vector dimension \(d\) of the clean training utterances, given the pair of the CDF value \(c_{d,j}\) of the vector component \(x_{d,j}\) and \(x_{d,j}\) itself at the frame \(j\), the linear polynomial function \(Q(c_{d,j})\) with output \(x_{d,j}\) can be expressed as:

\[
Q(c_{d,j}) = \sum_{m=0}^{M} a_{d,m} (c_{d,j})^m,
\]

where the coefficients \(a_{d,m}\) of each dimension \(d\) can be estimated by minimizing the sum of squared errors expressed in the following equation:

\[
E_d = \sum_{j=1}^{N} (x_{d,j} - \sum_{m=0}^{M} a_{d,m} (c_{d,j})^m)^2,
\]

where \(N\) is the total number of training speech feature vectors. During the training phase, the polynomial functions of all dimensions are obtained by minimizing the sum of squared errors expressed in Eq. (4). During the recognition phase, for each feature vector dimension, the feature vector components of the test utterance are simply sorted in ascending order of their values to obtain the approximate CDF values, which can be then taken as the inputs to the inverse function to obtain the corresponding restored component values.

3. Proposed HEQ Method (STHEQ)

It is true that the independence assumption of speech feature components made by HEQ would be invalid when the feature vector components of different dimensions are not completely de-correlated. On the other hand, since speech signals are slowly time-varying, the contextual relationships of consecutive speech frames might provide additional information clues for feature normalization. With these observations in mind, in this paper we devise an alternative HEQ method that can exploit both spatial and temporal feature distribution characteristics simultaneously for speech feature normalization (denoted as STHEQ). Given the feature vector sequence \(x_{1}, \ldots, x_{L}\) of an utterance, where \(x_{j} = [x_{d,j}, \ldots, x_{D,j}]\), the corresponding CDF values \(c_{d,j}\) of the feature vector components \(x_{d,j}\) of each dimension \(d\) at each frame \(j\) are first computed, resulting in an CDF vector sequence \(c_{1}, \ldots, c_{L}\) of the utterance being constructed, where \(c_{j} = [c_{1,j}, \ldots, c_{D,j}]\). The resultant CDF vector \(c_{j}\) is further concatenated with its \(\kappa\) preceding and \(\kappa\) succeeding vectors to form a spliced CDF vector \(S_{j}\) of \((2\kappa+1)D\) dimensions. Then, the restored feature vector \(\hat{x}_{j}\) of each \(x_{j}\) can be obtained by applying a linear transformation on \(S_{j}\):

\[
\hat{x}_{j} = A_{j} S_{j},
\]
where $A$ is a $(2K \times D)$ by $D$ transformation matrix, which can be estimated by minimizing the sum of squared errors expressed in the following equation:

$$E^2 = \frac{1}{N} \sum_{l=1}^{N} \|X_l - \hat{X}_l\|^2,$$

where $N$ is the total number of training clean speech feature vectors. Once the transformation matrix $A$ is obtained, it can be used to transform the feature vectors in the test data. In implementation, we use the order-statistics based approach instead of the cumulative-histogram based approach to obtain the approximate CDF values [5, 10]. For the feature vector component sequence $x_{d,1}, \ldots, x_{d,ij}, \ldots, x_{d,L}$ of a specific dimension $d$ of a speech utterance, the corresponding CDF value of each feature component $x_{d,ij}$ can be computed approximately through the following two steps [5]:

- **Step 1:** The sequence $x_{d,1}, \ldots, x_{d,ij}, \ldots, x_{d,L}$ is first sorted in ascending order according to the values of the feature vector components.
- **Step 2:** The order-statistics based approximation of the CDF value of a feature vector component $x_{d,ij}$ is then given as:

$$e_{d,ij} = \frac{\text{Pos}(x_{d,ij}) - 0.5}{L}$$

where $\text{Pos}(x_{d,ij})$ is a function that returns the rank of $x_{d,ij}$ in ascending order of the values of the feature vector components of the sequence $x_{d,1}, \ldots, x_{d,ij}, \ldots, x_{d,L}$. Therefore, for each utterance, Eq. (7) can be used to approximate the CDF values of the feature vector components of all dimensions.

It is worth mentioning that the use of the linear transformation together with the contextual information of consecutive speech frames in fact is not new for the speech processing community. For example, the principal component analysis (PCA), the linear discriminant analysis (LDA) [11, 12], the heteroscedastic linear discriminant analysis (HLDA) [13], etc., have been designed to use the linear transformation and the frame contextual information for discriminative feature extraction. However, most of them in essence are feature-to-feature transformation approaches, while the STHEQ method instead is a distribution-to-feature transformation approach.

4. Experiment Results

4.1 Experiment Setup

The speech recognition experiments were conducted under various noise conditions using the Aurora-2 database and task [14]. The Aurora-2 database is a subset of the TI-DIGITS, which comprises a set of variable-length continuous English digit strings spoken into a close-talking microphone. The Aurora-2 task consists of the recognition of the continuous digit string utterances interfered with by various noise sources at different signal-to-noise ratios (SNRs). Eight different types of real-world additive noises are artificially added into Test Set A (subway, babble, car, and exhibition) and Test Set B (restaurant, street, airport, and station) under a wide range of SNRs ($-5$ dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, and Clean).

In contrast, Test Set C contains one noise from Test Set A (subway) and one from Test Set B (street) under the same range of SNRs and also includes extra channel (convolutional) distortion. For the baseline system, both the training and recognition tests are performed using the HTK recognition toolkit [15], following the original setup defined for the ETSI AURORA evaluations [14]. More specifically, each digit was modeled as a left-to-right continuous density HMM with 16 states and three diagonal Gaussian mixtures per state. Two additional silence models were defined. One had three states with six Gaussian mixtures per state for modeling the silence at the beginning and at the end of each utterance. The other one had one state with 6 Gaussian mixtures for modeling the interword short pause.

In the baseline front-end speech analysis, a 39-dimensional feature vector was extracted at each time frame, including 12 Mel-frequency cepstral coefficients (MFCCs), the logarithm of the energy and the corresponding delta and acceleration coefficients. The frame length is 25 ms and the corresponding frame shift is 10 ms [14]. The HEQ methods investigated in this paper are conducted on feature vectors comprising the 12 MFCCs and the energy, and their corresponding delta and acceleration coefficients are derived from the restored feature vectors afterward. All the experiment results reported below are based on clean condition training, i.e., the acoustic models were trained only with the clean training utterances. The average word error rate (WER) result obtained by the MFCC-based baseline system is on average 41.04%, which is an average of the WER results of the test utterances respectively contaminated with eight types of noises under different SNR levels (0db to 20dB) for the three test sets (Sets A, B and C).

4.2 Experiments on STHEQ

We first evaluate the performance of STHEQ with respect to different numbers (the values of $\kappa$ described in Section 3) of the preceding and the succeeding frames being used to form the spliced CDF vectors. The corresponding results are shown in Table 1, which are averaged for the three sets (Sets A, B and C) and the SNR levels between 0 dB and 20 dB. As can be seen, the WER of STHEQ is significantly improved when the value of $\kappa$ is greater than or equal to 1, which reveals that the additional use of the distribution characteristics from the neighboring speech frames is beneficial. However, the improvements seem to saturate for most cases when the value is set to two, which is equivalent to the condition that the CDF vectors of five consecutive speech frames are considered together for feature normalization. If we compare the best result of STHEQ with that of the MFCC-based baseline system, it can be found that STHEQ can provide a relative WER reduction of about 50% over the baseline system.

In the next set of experiments, we compare STHEQ with two commonly-used HEQ methods (i.e., THEQ and QHEQ) and our previously proposed HEQ method (i.e., PHEQ). The corresponding results of these methods are shown in Table 2, where the results obtained by the ETSI standard system [13], CMVN, and the combination of HLDA and CMVN (HLDA-CMVN) [8, 16], respectively, are also listed for reference. As compared with the best result of STHEQ ($\kappa=2$) shown in Table 1, it can be observed that STHEQ is considerably more effective than THEQ, QHEQ and PHEQ, yielding a relative WER reduction of 20%, 21% and 13%, respectively. If we look into the detailed WER results of STHEQ and the other HEQ methods at different SNR levels, as graphically
presented in Figure 2, it is evident that the improvement made by STHEQ over the other HEQ methods is steadily increased as the noise condition becomes worse. These results indeed confirm our expectation that the spatial and temporal relationships between speech feature vector components of different dimensions and consecutive frames might provide helpful information clues for speech feature normalization. In the meantime, we are actively working on the ways to further improve the performance of STHEQ, including trying different methods for CDF estimation [10], investigating the possibility of using the other objective functions for deriving the distribution-to-feature transformation, etc. We are also investigating the application of STHEQ to more complicated ASR tasks.

5. Conclusions

In this paper, we have proposed a new HEQ based method that exploits spatial-temporal feature distribution characteristics for speech feature normalization. A distribution-to-feature linear transformation was designed for such a purpose. The performance of the presented method have been extensively tested and verified by comparison with the other HEQ methods. The experiment results have demonstrated that for clean-condition training, our method can yield significant word error rate reduction over the baseline system, and also considerably outperforms the other HEQ methods compared in this paper.

6. References


Table 1. The average WER results (%) of STHEQ with respect to different numbers (the values of $\kappa$ described in Section 3) of preceding and succeeding frames used to form the spliced CDF vectors.

<table>
<thead>
<tr>
<th>$\kappa$</th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>20.6</td>
<td>19.41</td>
<td>20.93</td>
<td>20.31</td>
</tr>
<tr>
<td>1</td>
<td>18.33</td>
<td>16.98</td>
<td>19.34</td>
<td>18.22</td>
</tr>
<tr>
<td>2</td>
<td>18.02</td>
<td>16.98</td>
<td>18.97</td>
<td>17.99</td>
</tr>
<tr>
<td>3</td>
<td>19.04</td>
<td>17.95</td>
<td>19.43</td>
<td>18.80</td>
</tr>
<tr>
<td>4</td>
<td>18.05</td>
<td>17.34</td>
<td>20.13</td>
<td>18.51</td>
</tr>
<tr>
<td>5</td>
<td>18.14</td>
<td>17.35</td>
<td>20.50</td>
<td>18.66</td>
</tr>
</tbody>
</table>

Table 2. The average WER results (%) of three HEQ methods, as well as the ETSI system, CMVN, and the combination of HLDA and CMVN.

<table>
<thead>
<tr>
<th>Method</th>
<th>Set A</th>
<th>Set B</th>
<th>Set C</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>41.06</td>
<td>41.52</td>
<td>40.03</td>
<td>41.04</td>
</tr>
<tr>
<td>THEQ</td>
<td>22.76</td>
<td>21.16</td>
<td>23.47</td>
<td>22.47</td>
</tr>
<tr>
<td>QHEQ</td>
<td>23.53</td>
<td>21.90</td>
<td>22.36</td>
<td>22.64</td>
</tr>
<tr>
<td>PHEQ</td>
<td>20.98</td>
<td>20.17</td>
<td>21.43</td>
<td>20.75</td>
</tr>
<tr>
<td>ETSI</td>
<td>38.69</td>
<td>44.25</td>
<td>28.76</td>
<td>38.93</td>
</tr>
<tr>
<td>CMVN</td>
<td>27.73</td>
<td>24.60</td>
<td>27.17</td>
<td>26.37</td>
</tr>
</tbody>
</table>

Figure 2: The detailed WER results (%) of STHEQ and the other HEQ methods at different SNR levels.


