Attribute-based histogram equalization (HEQ) and its adaptation for robust speech recognition

Xiong Xiao\(^1\), Eng Siong Chng\(^1,2\), Haizhou Li\(^1,2,3\)

\(^1\)Temasek Lab@NTU, Nanyang Technological University, Singapore
\(^2\)School of Computer Engineering, Nanyang Technological University, Singapore
\(^3\)Department of Human Language Technology, Institute for Infocomm Research, Singapore

xiaoxiong@ntu.edu.sg, aseschng@ntu.edu.sg, hli@i2r.a-star.edu.sg

Abstract

Histogram equalization (HEQ) is a simple and effective feature normalization technique for robust speech recognition. Recently, we proposed to adapt HEQ transform to each test utterance using a maximum likelihood (ML) criterion and observed improved performance. In this paper, we further study the effects of applying attribute-based HEQ and its ML adaptation. Instead of applying a global HEQ transform to the test utterance, we propose to apply different HEQ transforms to the 6 manners of speech, e.g. vowel and fricative. We also developed the ML adaptation algorithm of the attribute-based HEQ. Experimental results show that the attribute-based HEQ adaptation obtained 21.8% and 19.5% relative error rate reduction over the global HEQ baseline on the Aurora-2 and Aurora-4 benchmarking tasks, respectively.

Index Terms: histogram equalization, robust speech recognition, noise reduction, linear regression, maximum likelihood adaptation

1. Introduction

Speech recognition performance degrades significantly when there is mismatch between the training and testing environment, e.g. when we train the acoustic model on clean features and test it on noisy features. Feature normalization is a group of techniques designed to improve the robustness of speech recognition systems against noise and channel distortions. Generally speaking, feature normalization methods normalize the statistics of noisy features to those of clean features so that the noise effects can be partially removed. Examples include cepstral mean normalization (CMN) [1] that normalizes the feature means, mean and variance normalization (MVN) [2] that normalizes the means and variances, and histogram equalization (HEQ) [3] that normalizes the histogram (i.e. distribution) of the features. HEQ is the most general form of feature normalization and usually produces good results in noisy speech recognition.

However, HEQ faces two major limitations [4]. One limitation is that HEQ uses a monotonically increasing transformation function. As a result, the rank of features in a feature trajectory cannot be changed by HEQ. This is a limitation because the noise not only changes the values of features, but also the rank of features. Another limitation is that HEQ tries to normalize the histogram of a short utterance to the histogram of the whole training corpus. Such an approach is inappropriate, especially for short utterances, as the phone composition of an utterance may be very different from that of the training corpus. To overcome these limitations, class-based HEQ is proposed [4], in which the feature vectors of an utterance are first assigned into classes, and then the histogram of each class is normalized to their reference histograms. This approach allows a non-monotonical transformation function and is also less affected by mismatched phone composition of test utterance and training corpus. In another work [5], a two-class processing is proposed, one for silence and the other for speech frames. However, linear transform rather than HEQ is used in class-based processing. It is reported that class-based HEQ outperforms global HEQ significantly in both Aurora-2 [6] and Aurora-4 [7] benchmarking tasks, however, the performance in clean test case is seriously affected [4, 5].

Recently, we proposed a maximum likelihood (ML) adaptation algorithm [8] to fine tune global HEQ’s transformation to each test utterance based on the knowledge of a reference Gaussian mixture model (GMM) trained from clean features. A parametric implementation of HEQ (PHEQ [8, 9, 10]) is used and the parameters of PHEQ is optimized to make the processed features fit better to the clean GMM. It was shown that the ML adaptation of PHEQ produces significantly better results than unadapted PHEQ [8].

Motivated by the previous studies on class-based HEQ and ML adaptation of PHEQ, we propose an attribute-based PHEQ and adapt its parameters to each test utterance using the ML criterion. We use 6 manners of speech as classes, such as vowel, fricative, nasal, etc. For each test utterance, its feature vectors are first classified into these 6 manners, and then a separate PHEQ is applied to equalize the histogram of each class. In addition, we derive an algorithm to adapt the attribute-based PHEQ with a new regularization term different from that in [8].

The rest of the paper is organized as follows. In Section 2, we first review the global PHEQ, then introduce the attribute-based PHEQ, followed by its adaptation. In Section 3, we present experimental studies and discussions. Finally, we conclude in section 4.

2. Attribute-based PHEQ Adaptation

2.1. Review of Global PHEQ

Let \(\mathbf{x}_t\) be the feature vector of dimension \(D\) at frame \(t\) and \(x_t^{(d)}\) be its \(d^{th}\) element. The HEQ processes each feature element independently as follows [3]:

\[
y_t^{(d)} = C_{w,d}^{-1}(C_{e,d}(x_t^{(d)})), \quad d = 1, \ldots, D
\] (1)
where $C_{\text{ref},d}^{-1}(\cdot)$ is the inverse reference cumulative distribution function (CDF) and $C_{\text{ref},d}(\cdot)$ is the CDF of $x_i^{(d)}$, both for element $d$. The objective of the processing in (1) is to equalize the CDF of $y_i^{(d)}$ to the reference CDF.

To implement (1), $C_{x,d}(\cdot)$ can be estimated from the rank of $x_i^{(d)}$ among all available frames [11]:

$$
C_{x,d}(x_i^{(d)}) = (R(x_i^{(d)}) - 0.5)/T
$$

where $T$ is the number of available frames and $R(x_i^{(d)}) \in [1, T]$ is the rank of $x_i^{(d)}$. For $C_{\text{ref},d}(\cdot)$, we adopt a parametric approximation as follows [8]:

$$
y_i^{(d)} \approx \text{a}^{(d)} \cdot z_i^{(d)}
$$

where $\text{a}^{(d)} = [a_0^{(d)}, a_1^{(d)}, ..., a_M^{(d)}]$ is a row vector of parameters for element $d$, $z_i^{(d)} = [1, \text{sigm}(C_{x,d}(x_i^{(d)})), ..., \text{sigm}(C_{x,d}(x_i^{(d)}))]^T$ is a vector of order statistics, and $\text{sigm}(\cdot)$ represents matrix or vector transpose. $\text{sigm}(x) = [1 + \exp(-\gamma(x - \theta_m))]^{-1}$ is the $m$th sigmoid function centered at $\theta_m$. Although the processed feature is a linear combination of order statistics, the overall PHEQ processing is nonlinear as the order statistics is a nonlinear transformation of the original features. Hence, PHEQ is a flexible feature transform and may be able to reduce nonlinear distortions. For more detailed description of PHEQ, please refer to [8].

### 2.2. Attribute-based PHEQ

Motivated by class-based HEQ [4], we extend the global PHEQ to attribute-based PHEQ. In [4], the clusters are obtained by clustering clean feature vectors by k-means or EM. However, in our preliminary study, we found that such definition of clusters lead to inconsistent results as the clusters depend on the random initialization. Furthermore, the clusters found by k-means or EM has no clear meaning and sometimes even mix silence frames with weak speech frames. In this paper, we use speech attribute classes [12], specifically 6 manners as classes for attribute-based PHEQ. The 6 manners include vowel, fricative, nasal, silence, stop, and approximant. We hypothesize that the manner classes have different feature distributions and hence should be processed by a separate PHEQ transforms. To get the class label of clean training vectors, we first force align the phone labels to the training utterances, and then map the phones to the manners. Hence, no clustering is necessary. For each manner class, we train a PHEQ parameter vector and also a diagonal covariance Gaussian. The Gaussians are used to classify frames into classes in testing. Note that although we use manners as classes, the attribute-based PHEQ technique presented in this section is also applicable to other definition of feature classes.

In attribute-based PHEQ, the feature vectors of a test utterance is first classified into manner classes, and then each class is processed individually. As suggest in [4], a soft-decision clustering performs better than a hard-decision clustering, so the final processed feature is defined as a weighted combination of class dependent PHEQ-processed features:

$$
y_i^{(d)} = \sum_{i=1}^{t} p(i|\mathbf{x}_i) y_i^{(d)}
$$

$$
y_i^{(d)} = \sum_{i=1}^{t} p(i|\mathbf{x}_i) \text{a}^{(d)}_i \cdot z_i^{(d)}
$$

where $I = 6$ is the number of classes, $p(i|\mathbf{x}_i)$ is the posterior probability of the current frame belonging to class $i$ and is used as the weight of class dependent $y_i^{(d)}$. $\text{a}^{(d)}_i$ is the weight vector of class $i$ and feature element $d$. $z_i^{(d)}$ is the order statistics of feature element $d$ and class $i$ at frame $t$. The posterior are computed as follows:

$$
p(i|\mathbf{x}_i) = \frac{p(\mathbf{x}_i|i)p(i)}{\sum_{i=1}^{T} p(\mathbf{x}_i|i)p(i)}
$$

The manner class dependent distribution $p(\mathbf{x}_i|i)$ is represented by a diagonal covariance Gaussian. Both the class prior $p(i)$ and class dependent distribution are estimated from clean feature vectors.

The order statistics depends on the CDF of test features, which in turn depends on the class ID $i$. Following [4], the CDF of test features in class $i$ are computed as:

$$
C_{x,d,i}(x_i^{(d)}) = \sum_{t \in \Phi_i^{(d)}} p(i|\mathbf{x}_i) (R(x_i^{(d)}) - 0.5)/T
$$

where $\Phi_i^{(d)} = \{t | x_i^{(d)} < x_i^{(d)}\}$ is the set of frames with feature value in dimension $d$ that is less than $x_i^{(d)}$. It can be seen that the class dependent test CDF is only depending on the posteriors of classes and the ranking information of features. Once the test CDF is obtained, order statistics $z_i^{(d)}$ can be computed using sigmoid functions as usual.

The weighted sum of class dependent processed features in (4) can also be represented by a simple linear regression as follows:

$$
y_i^{(d)} = \alpha^{(d)} \cdot \bar{z}_i^{(d)}
$$

where $\alpha = [\alpha_1^{(d)}, ..., \alpha_6^{(d)}]$ is the concatenated weight vector and $\bar{z}_i^{(d)} = [p(1|\mathbf{x}_i)|z_i^{(d)}|_1^T, ..., p(6|\mathbf{x}_i)|z_i^{(d)}|_6^T]_T$ is the concatenated order statistics weighted by class posteriors. This representation is useful as it indicates that both the attribute-based PHEQ and global PHEQ are a linear combination of order statistics, and the only difference is the computation of order statistics. This means that the algorithm for estimating parameters of global PHEQ can also be used to estimate the parameters of attribute-based PHEQ by just plugging the new order statistics vectors.

### 2.3. Minimum Mean Squared Error Estimation

To estimate the attribute-based PHEQ’s parameters from clean features, we can minimize the following mean squared error (MSE) [9]:

$$
\text{MSE} = \sum_{t=1}^{T} \sum_{d=1}^{D} (y_i^{(d)} - x_i^{(d)})^2
$$

where $x_i^{(d)}$ is the clean training feature at frame $t$ feature element $d$ and $y_i^{(d)}$ is its reconstruction by attribute-based PHEQ. By minimizing the MSE in (8), we can learn the parameters of all class dependent PHEQ parameters in $\alpha^{(d)}$ in one time. In this study, for training, the class posteriors is either 0 or 1. Hence, estimating the parameters of all classes jointly is exactly the same as estimating the class dependent parameters separately. However, for non 0/1 posteriors, the two estimation methods have minor difference.
The MSE can be minimized by using least square method for each feature element independently. The solution is
\[
\mathbf{a}_{\text{MMSE}}^{(d)} = (\mathbf{p}_{\text{MMSE}}^{(d)})^T (\mathbf{A}_{\text{MMSE}}^{(d)})^{-1}, \quad d = 1, \ldots, D
\]
where
\[
\mathbf{A}_{\text{MMSE}}^{(d)} = \sum_{t=1}^{T} \mathbf{z}_t^{(d)} (\mathbf{z}_t^{(d)})^T
\]
\[
\mathbf{p}_{\text{MMSE}}^{(d)} = \sum_{t=1}^{T} \mathbf{z}_t^{(d)} \mathbf{x}_t^{(d)}
\]
For global PHEQ, the solution is the same as (9)-(11), except that \( \mathbf{z} \) is replaced with \( \mathbf{z} \), i.e. the global PHEQ’s order statistics vectors.

2.4. Maximum Likelihood Adaptation

The concept of ML adaptation of attribute-based PHEQ is to adjust \( \mathbf{a}_{\text{MMSE}}^{(d)} \) to increase the likelihood of the processed features on a reference GMM that represents the clean feature vector’s distribution. In [8], we showed that the ML adaptation of global PHEQ outperforms MMSE PHEQ significantly. In this section, we derive the ML adaptation algorithm for attribute-based PHEQ.

As we have formulated the attribute-based PHEQ as a linear combination of order statistics vectors, the ML adaptation of attribute-based PHEQ is exactly the same as the adaptation of global PHEQ in [8]. However, in attribute-based PHEQ, the problem of overfitting is more serious as the order statistics vector’s dimensionality is much higher than that of global PHEQ. For example, for \( I = 6 \) and \( M = 10 \), the order statistics vectors will have a dimensionality of \( I \times (M + 1) = 66 \). This means that there are 39 × 66 free parameters in the ML adaptation of attribute-based PHEQ if \( D = 39 \). For short utterances, it is impossible to estimated the parameters reliably. Hence, in this paper, we use a more strict regularization term than that in [8], i.e. we use the MMSE PHEQ processed features as the prior of the ML PHEQ processed features, and only allow moderate changes in the features. Mathematically, we propose to maximize the following constrained ML objective function:
\[
\mathbf{a}_{\text{ML}}^{(d)} = \arg \max_{\mathbf{a}^{(d)}} \left\{ \frac{1}{C} \sum_{c=1}^{C} \log p(\mathbf{z}_c | \Lambda) \right\}
\]
\[
- \alpha \sum_{t=1}^{T} \sum_{d=1}^{D} \left( \frac{y_t^{(d)} - a_t^{(d)}}{\sigma_t^{(d)}} \right)^2
\]
into (12), the objective function can be rewritten as:
\[
Q = \sum_{t=1}^{T} \sum_{c=1}^{C} \log P(c)E_c \exp \left\{ -\alpha \sum_{d=1}^{D} \frac{(\mathbf{a}_{\text{ML}}^{(d)} - \mu_{c}^{(d)})^2}{2 \sigma_{c}^{(d)}^2} \right\}
\]
where \( P(c) \), \( E_c \), \( \mu_{c}^{(d)} \) and \( \sigma_{c}^{(d)}^2 \) are the prior weight, normalizing constant, mean, and variance of the \( d \)-th element of the \( c \)-th Gaussian mixture.

Note that the problem in (13) is basically a linear regression problem with L2 regularization, so a closed form solution is available. Due to the space limitation of this paper, we will not derive the solution, which is:
\[
\mathbf{a}_{\text{ML}}^{(d)} = (\mathbf{p}_{\text{ML}}^{(d)} + 2\alpha \mathbf{A}_{\text{ML}}^{(d)} (\mathbf{A}_{\text{ML}}^{(d)})^T) \mathbf{A}_{\text{ML}}^{(d)} + 2\alpha \mathbf{A}_{\text{ML}}^{(d)} \right)^{-1}
\]
\[
\mathbf{A}_{\text{ML}}^{(d)} = \sum_{t=1}^{T} \sum_{c=1}^{C} \gamma_c(t) \mathbf{z}_t^{(d)} (\mathbf{z}_t^{(d)})^T / \sigma_{c}^{(d)}^2
\]
\[
\mathbf{p}_{\text{ML}}^{(d)} = \sum_{t=1}^{T} \sum_{c=1}^{C} \gamma_c(t) \mu_{c}^{(d)} / \sigma_{c}^{(d)}^2 = \mathbf{z}_t^{(d)}
\]
where \( \gamma_c(t) \) is the posterior probability of Gaussian mixture \( c \) at frame \( t \) and is computed from the initial features \( \mathbf{y}_t \). When computing \( \mathbf{a}_{\text{ML}}^{(d)} \) using (10), the order statistics should be derived from the test features instead of clean training features. From the solution in (14), it is clear that when \( \alpha = \infty \), \( \mathbf{a}_{\text{ML}}^{(d)} = \mathbf{a}_{\text{MMSE}}^{(d)} \), i.e. there is no adaptation of attribute-based PHEQ parameters. By tuning \( \alpha \), we can control the degree of adaptation.

3. Experiments

The proposed attribute-based PHEQ and its adaptation is evaluated on two common benchmarking tasks, i.e. the Aurora-2 connected digits [6] and the Aurora-4 large vocabulary speech recognition tasks [7].

3.1. Tasks Description

The Aurora-2 [6] is a noisy connected English digits recognition task. In this study, we use the clean condition training scheme, i.e. the acoustic model is trained from 8440 clean utterances. The settings of the acoustic model follow the “simple back-end” defined in [6]. There are 10 test cases in Aurora-2 task, each represents a particular additive noise and channel distortion combination. Each test case is further divided into 7 signal to noise ratio (SNR) levels, i.e. clean and 20dB to -5dB with 5dB step. The noises are recorded in real environments, such as subway, restaurant, and then added to clean speech signal to generate all noisy test utterances.

The Aurora-4 [7] task is generated by adding noises to the Wall Street Journal (WSJ) corpus [13] in a similar way as the Aurora-2. Similar to Aurora-2, we also use the clean condition training scheme. A triphone based acoustic model is trained from 7138 clean utterances. About 3000 tied triphone states are generated by using decision tree based state clustering. There are 8 Gaussian mixtures in each tied states. The standard WSJ
the global PHEQ, except for clean test case. This could be due to that the utterances in Aurora-2 is only 1.8s long on average. Hence, the histogram of test features in each manner class is not reliably estimated. As a result, the performance in clean test case is degraded. However, the degradation of our attribute-based PHEQ is much smaller than that of class-based HEQ in [4]. For noisy test cases, the gain from multiclass processing is larger than the degradation caused by unreliable histogram estimation, so we see improved results.

When ML adaptation is applied, attribute-based PHEQ also outperforms global PHEQ, except for clean and 20dB case. One reason for the poor performance at high SNR levels may be due to that the number of free parameters is too many and the Aurora-2 utterances are too short. However, for low SNR levels, such as ≤ 5dB, the gain of attribute-based ML adaptation is quite big. On average, the attribute-based PHEQ with ML adaptation obtains a WER of 14.63%, which represents a 21.8% relative WER reduction over the global MMSE PHEQ baseline.

### 3.4. Results on Aurora-4

The performance of global and attribute-based PHEQ on Aurora-4 is shown in Table 1. By comparing row “Global PHEQ (MMSE)” and “Class PHEQ (MMSE)”, we can see that the WER of all test cases, including the clean test 1, are significantly reduced. As the sentences in Aurora-4 have an average length of 7s, it’s likely that each manner class have enough frames to estimate the class histogram reliably. Hence the attribute-based PHEQ outperforms global PHEQ even for clean test case. On average, the WER is reduced from 34.0% to 29.0%, representing a 14.8% relative WER reduction.

By comparing “Class PHEQ (+ML)” and “Class PHEQ (MMSE)”, we observe that the ML adaptation produces lower WER than MMSE in all but the test case 8. On average, the ML adaptation reduces the WER by 1.6% absolute over the MMSE-based processing. This shows that the ML adaptation is also effective in the attribute-based PHEQ framework. The best results of attribute-based PHEQ is 27.4%, which represents a 19.5% relative WER reduction over the global MMSE PHEQ baseline. To our best knowledge, this is the best reported results of HEQ on the Aurora-4 task using clean condition training.

### 4. Conclusions

In this paper, we proposed an attribute-based PHEQ processing to reduce noise and channel effects in speech recognition. We also derived the ML adaptation of attribute-based PHEQ parameters to each test utterance. This work is an extension of our previous work on PHEQ and ML adaptation of PHEQ. Experimental results show that the attribute-based PHEQ consistently outperforms global PHEQ, except for clean test case for short utterances in Aurora-2. In the future, a possible research direction is the robust estimation of class histograms for short utterances.

---

**Table 1: Recognition WER (%) on Aurora-4 task using different variants of PHEQ.** Avg. represents the average results over all test cases. R.I. is the relative WER reduction over global PHEQ (MMSE). Numbers in bold represent the best results for that test case.

<table>
<thead>
<tr>
<th>Test Case</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>Avg.</th>
<th>R.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Global PHEQ (MMSE)</td>
<td>11.9</td>
<td>19.2</td>
<td>33.3</td>
<td>35.2</td>
<td>34.2</td>
<td>33.1</td>
<td>37.1</td>
<td>17.1</td>
<td>29.5</td>
<td>41.6</td>
<td>46.3</td>
<td>49.4</td>
<td>43.1</td>
<td>45.7</td>
<td>34.0</td>
<td>-</td>
</tr>
<tr>
<td>Global PHEQ (+ML)</td>
<td>11.7</td>
<td>17.9</td>
<td>30.8</td>
<td>32.7</td>
<td>32.5</td>
<td>31.1</td>
<td>34.9</td>
<td>17.0</td>
<td>25.2</td>
<td>37.3</td>
<td>42.3</td>
<td>44.5</td>
<td>38.5</td>
<td>41.5</td>
<td>31.3</td>
<td>8.1</td>
</tr>
<tr>
<td>Class PHEQ (MMSE)</td>
<td>10.9</td>
<td>15.0</td>
<td>28.3</td>
<td>33.1</td>
<td>31.2</td>
<td>30.4</td>
<td>31.8</td>
<td>14.3</td>
<td>21.7</td>
<td>33.0</td>
<td>41.0</td>
<td>41.0</td>
<td>35.8</td>
<td>38.3</td>
<td>29.0</td>
<td>14.8</td>
</tr>
<tr>
<td>Class PHEQ (+ML)</td>
<td><strong>10.8</strong></td>
<td><strong>14.2</strong></td>
<td><strong>27.1</strong></td>
<td><strong>30.9</strong></td>
<td><strong>30.1</strong></td>
<td><strong>26.8</strong></td>
<td><strong>31.7</strong></td>
<td><strong>14.6</strong></td>
<td><strong>20.4</strong></td>
<td><strong>31.1</strong></td>
<td><strong>38.8</strong></td>
<td><strong>38.4</strong></td>
<td><strong>31.7</strong></td>
<td><strong>36.7</strong></td>
<td><strong>27.4</strong></td>
<td><strong>19.5</strong></td>
</tr>
</tbody>
</table>

**Table 2: Performance on Aurora-2 task.** Results are averaged across all 10 test cases. “+ML Adapt” refers to the adaptation of global or attribute-based PHEQ to each test utterance. Numbers in bold represent the best results for that SNR level. R.I. is the relative WER reduction over the basic PHEQ (global).

<table>
<thead>
<tr>
<th>SNR</th>
<th>Global PHEQ</th>
<th>Class PHEQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MMSE</td>
<td>+ML Adapt</td>
</tr>
<tr>
<td>clean</td>
<td>0.97</td>
<td>0.97</td>
</tr>
<tr>
<td>20dB</td>
<td>2.63</td>
<td>2.11</td>
</tr>
<tr>
<td>15dB</td>
<td>4.80</td>
<td>3.99</td>
</tr>
<tr>
<td>10dB</td>
<td>10.15</td>
<td>8.50</td>
</tr>
<tr>
<td>5dB</td>
<td>23.90</td>
<td>20.93</td>
</tr>
<tr>
<td>0dB</td>
<td>52.10</td>
<td>48.20</td>
</tr>
<tr>
<td>-5dB</td>
<td>80.37</td>
<td>78.82</td>
</tr>
<tr>
<td>0-20dB</td>
<td>18.71</td>
<td>16.74</td>
</tr>
<tr>
<td>RI (%)</td>
<td>-</td>
<td>10.3</td>
</tr>
</tbody>
</table>
5. References


