IMPROVED HISTOGRAM EQUALIZATION (HEQ) FOR ROBUST SPEECH RECOGNITION

Shih-Hsiang Lin¹, Shih-Hung Liu², Yao-Ming Yeh², Berlin Chen²

¹Graduate Institute of Information and Computer Education,  
²Department of Computer Science & Information Engineering,  
National Taiwan Normal University, Taipei, Taiwan  
69308027@cc.ntnu.edu.tw, {g93470185, ymyeh, berlin}@csie.ntnu.edu.tw

ABSTRACT

With the rapid development of Intelligent Transportation Systems (ITS), how to provide users with a natural and efficient human-machine interface is now becoming a crucial issue for driver safety. It is no doubt that speech will be one of the best mediators of human-machine interaction; however, the performance of automatic speech recognition (ASR) always radically degrades when the input speech is corrupted by varying noises. In this paper, we consider the use of histogram equalization (HEQ) for robust ASR. A novel data fitting scheme was presented to efficiently approximate the inverse of the cumulative density function of training speech for HEQ, which has the merits of lower storage and time consumption compared to the conventional table-lookup or quantile based HEQ approaches. Moreover, a more elaborate attempt of using multiple inverse functions for different noise conditions was investigated as well. All experiments were carried out on the Aurora-2 standard database and task. Very encouraging results were obtained. The proposed robustness technique has also been properly integrated into our prototype system for in-vehicle traffic information retrieval using spoken queries.

1. INTRODUCTION

Intelligent Transportation Systems (ITS), which aim to bring together and integrate state-of-art technologies from different fields, such as the electronic, computer, information, telecommunication, automatic control and vehicle-sensing etc., have attracted a lot of attention over the past decades. Accompanying with the rapid spread of various ITS applications, Advanced Traveler Information Systems (ATIS) are now an active focus of much research in the development of ITS, since they can provide drivers with the pre-trip and en-route information, including trip planning, road condition, route guidance, weather forecast and attractions directory etc., using either text, image or graph representations. However, there still remain several key issues for its popularization. One of them is the design of a better human-machine interface (HMI). For example, most of the currently existing ATIS systems are purely text-driven such that users can only interact with them through text commands or queries, which indeed is impractical and dangerous when users are concentrating on driving. As a result, seeking for more appropriate solutions has become a central concern for the development of both ITS and ATIS.

On the other hand, as we know, speech is the primary and the most convenient means of communication between people. Due to the successful development of much smaller electronic devices and the popularity of wireless communication and networking, it is widely believed that speech will play a more active role and will serve as the major HMI for the interaction between people and different kinds of smart devices in the near future [1]. Therefore, it is expected that automatic speech recognition (ASR) will have significant impact in the interaction between drivers and ATIS, for that it can enable drivers to naturally and efficiently interact with the systems without diverting their attention from the road scene [2].

Nevertheless, how to effectively alleviate the performance deterioration resulting from varying in-vehicle noises is actually the primary challenge facing most ASR systems. Substantial efforts have been made and presented to cope with this issue to improve the ASR performance in the last two decades. In general, they fall into three main categories [3]:

- Speech enhancement, which removes the noise from the speech signal;
- Robust speech feature extraction, which searches for noise resistant and robust features;
- Acoustic model adaptation, which transforms acoustic models from training (clean) space to the test (noisy) space.

Techniques of each of the above three categories have their own superiority and limitations. In the practical implementation, acoustic model adaptation often yields the best performance, because it directly adjusts the acoustic model parameters (like the mean vectors or covariance matrices of Gaussian mixture models) to accommodate the uncertainty caused by noisy environments. However, most of such methods generally require extra adaptation data (either with or without reference transcripts) and longer computation time in comparison with the other two approaches. Moreover, most of the speech enhancement techniques target at enhancing the signal-to-noise ratio (SNR) but not necessarily at improving the speech recognition accuracy. On the other hand, the robust speech feature methods can be further divided into two subcategories, i.e. model-based compensation and feature space normalization. Model-based compensation assumes the mismatch between clean and noisy conditions can be modeled by a stochastic process. The associated compensation models can be obtained in the training phase, and then used to restore the feature vectors in the test phase. Feature normalization is believed to be a simpler and more effective way to compensate the mismatch caused by noises, and it has also been demonstrated its capability in preventing performance degradation of speech recognition systems under various environments.

With these observations in mind, in this paper we propose an effective but efficient feature normalization scheme for robust speech recognition. A novel data fitting scheme is presented to
efficiently approximate the inverse of the Cumulative Density Function (CDF) of training speech for HEQ, which has the merits of lower storage and time consumption compared to the conventional table-lookup or quantile based HEQ approaches. Moreover, a more elaborate attempt of using multiple inverse functions for different noise conditions is investigated as well. The rest of this paper is organized as follows. Section 2 reviews the related work. Section 3 elucidates our proposed robust speech feature extraction approaches. The experimental settings and evaluation results are detailed in Section 4, while Section 5 describes our prototype system. Finally, conclusions are drawn in Section 6.

2. RELATED WORK

Histogram equalization (HEQ) is one of the most effective robustness approaches that have been extensively studied for robust feature extraction of speech recognition in the recent past [4-7]. HEQ also has been shown its superiority over the conventional linear compensation approaches, such as cepstral mean normalization (CMN), and cepstral mean and variance normalization (CMVN) [6]. Theoretically, HEQ has its roots in the assumption that the transformed speech feature distributions of the test (or noisy) data should be identical to that of the training (or reference) data, for which the speech features can be estimated either from the Mel-frequency filter bank outputs [5, 7] or from the cepstral coefficients [6]. Under this assumption, the aim of HEQ is to find a transformation that can convert the distribution of each feature vector component of the input sentence (test speech) into a predefined target distribution which corresponds to that of the training speech. Accordingly, HEQ attempts not only to match speech feature means or variances, but also to completely match the feature distributions of the training and test data. Phrased another way, HEQ will normalize each moment of feature vectors.

The HEQ-based feature normalization can be conducted either in a non-parametric way (e.g., using a cumulative histogram to approximate the CDF of each utterance) [4, 5] or in a parametric way (e.g., using a piecewise transformation function where the parameters was estimated online in a quantile-corrective manner) [7]. 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 [4] 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 [7] 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.

3. IMPROVED HEQ APPROACHES

3.1. Polynomial-Fit Histogram Equalization (PHEQ)

In contrast to the above table-lookup or quantile based approaches, we proposed a Polynomial-Fit Histogram Equalization (PHEQ) approach which explored the use of the data fitting scheme to efficiently approximate the inverse of the CDF of training speech for HEQ [8]. Data fitting is a mathematical optimization method which, when given a series of data points \((u_i, v_i)\) with \(i = 1, 2, ..., N\), attempts to find a function \(G_{\theta}(v)\) whose output \(\hat{v}\) closely approximates to \(v_i\). That is, it minimizes the sum of the squares error (or the squares of the ordinate differences) between the points \((u_i, v_i)\) and their corresponding points \((\hat{u}_i, \hat{v}_i)\) in the data.

Figure 1. A schematic description of cluster-based polynomial-fit histogram (CPHEQ).

The function \(G_{\theta}(v)\) to be estimated can be either linear or nonlinear in its coefficients. For each speech feature vector dimension of the training data, given the pair of the CDF \(C_{\text{train}}(y_i)\) of the vector component \(y_i\), and \(y_i\) itself, the linear polynomial function \(G_{\theta}(y_i)\) with output \(\hat{y}_i\) can be expressed as:

\[
\hat{y}_i = \sum_{m=0}^{M} \alpha_m (C_{\text{train}}(y_i))^m,
\]

where the coefficients \(\alpha_m\) can be estimated by minimizing the squares error \(E^2\) expressed in the following equation:

\[
E^2 = \sum_{i=1}^{N} \left( y_i - \sum_{m=0}^{M} \alpha_m (C_{\text{train}}(y_i))^m \right)^2,
\]

where \(N\) is the total number of training speech feature vectors. During speech recognition, for each feature vector dimension, the vector components of a test utterance are simply sorted in an ascending order to obtain the corresponding cumulative histogram value of each vector component, which can be then taken as an input to the corresponding inverse function \(G\) to obtain the restored component value.

3.2. Cluster-based PHEQ (CPHEQ)

Instead of merely using a single inverse CDF function for the normalization of each feature component, in this paper we also investigated the use of multiple inverse CDF functions to obtain the restored value, namely the Cluster-based Polynomial-fit Histogram Equalization (CPHEQ). For CPHEQ, not only the clean training utterances but also their corresponding contaminated counterparts that had been corrupted with different kinds of noise types and SNR conditions (the so-called “stereo data”) were used. The contaminated training utterances were first used to train a Gaussian Mixture Model (GMM) whose parameters were estimated by the \(k\)-means algorithm followed by the Expectation Maximization (EM) algorithm. Then, each contaminated training speech vector component \(y_i\) was assigned to a specific cluster (or a mixture component) \(k\) using the following equation:

\[
\delta(k | y_i) = \begin{cases} 1 & \text{arg max } \psi(k | y_i) \\ 0 & \text{otherwise} \end{cases}
\]

where \(\psi(k | y_i)\) is the posterior probability of a specific cluster \(k\). Furthermore, for each cluster \(k\), a polynomial function \(G(k)\), defined in analogy with the function \(G(\cdot)\) stated earlier in Eq.(1), was estimated to restore the contaminated training speech vector components \(y_i\) assigned to \(k\) to their clean counterparts \(\hat{y}_i\). The
The Aurora-2 database is a subset of the TI-DIGITS, which originally defined for the ETSI evaluations [10]. All the experimental results reported below are based on clean-condition (uncontaminated) training utterances.

Experimental results are shown in Table 1. As can be seen, PHEQ significantly boosts the recognition performance of the MFCC-based baseline system, and the average WER of PHEQ is slightly decreased further when the order of the polynomial transformation function becomes larger. The improvements seem to saturate when the order is set to seven.

Then, we evaluate the performance of CPHEQ, for which the polynomial transformation functions were estimated using both the clean and noisy speech feature vectors, while not only the speech feature vectors but also their corresponding distribution characteristics are utilized by CPHEQ. Therefore, CPHEQ is believed to be more general than SPLICE.

4. EXPERIMENTAL RESULTS

4.1. Experimental Setup

The speech recognition experiments were conducted under various noise conditions using the Aurora-2 standard database task [10]. The Aurora-2 database is a subset of the TI-DIGITS, which contains a set of connected digit utterances spoken in English; while the task consists of the recognition of the connected digit utterances interfered with various noise sources at different SNRs, in which Sets A and B are artificially contaminated with eight different types of real world noises (e.g., the subway noise, street noise, etc.) in a wide range of SNRs (-5dB, 0dB, 5dB, 10dB, 15dB, 20dB and Clean) and the channel distortion is additionally included in Set C.

More specifically, the acoustic model for each digit was a left-to-right continuous density HMM with 16 states, and each state has a six-mixture diagonal GMM. Two additional silence models were defined. One had three states with a six-mixture GMM per state for modeling the silence at the beginning and at the end of each utterance. The other one had one state with a six-mixture GMM for modeling the interword short pause. A 39-dimensional feature vector was extracted at each frame, including 12 Mel-Frequency Cepstral Coefficients (MFCCs) and the logarithm of the energy, as well as their corresponding delta and acceleration coefficients. The training and recognition tests used different types of real world noises (e.g., the subway noise, street noise, etc.) in a wide range of SNRs (-5dB, 0dB, 5dB, 10dB, 15dB, 20dB and Clean) and the channel distortion is additionally included in Set C.

Table 1: Average word error rate (WER) results (%) with respect to different polynomial orders and different number of mixture Gaussians which were used in the estimation of the transformation functions of PHEQ.

<table>
<thead>
<tr>
<th>Polynomial Order</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>PHEQ</td>
<td>22.39</td>
<td>21.54</td>
<td>21.08</td>
<td>21.16</td>
</tr>
</tbody>
</table>

Figure 1 shows a schematic description of the CPHEQ approach.

In recent years, similar efforts also have been made to explore the relationship between the clean training utterances and their contaminated counterparts for speech feature restoration. The SPLICE (Stereo-based Piecewise Linear Compensation for Environments) [9] approach is often considered as a representative of this category. However, there are two main differences between SPLICE and CPHEQ. First, SPLICE simply uses a set of additive biases (or correction vectors) to approximate the true nonlinear relationship between the clean and noisy speech feature vectors, while CPHEQ use a set of polynomial functions to obtain the restored values. Second, SPLCE is only conducted on the speech feature vectors, while not only the speech feature vectors but also their corresponding distribution characteristics are utilized by CPHEQ. Therefore, CPHEQ is believed to be more general than SPLICE.

4.2. Experiments on Proposed Approaches

We first evaluate the performance of the proposed approaches. The average word error rate (WER) obtained by the MFCC-based baseline system is 41.04%, which is an average of the WERs of the test utterances respectively contaminated with eight types of noises under different SNR levels (0dB to 20dB and clean) for three sets (Sets A, B and C). Different orders of the polynomial functions were investigated for PHEQ. The corresponding results are shown in Table 1. As can be seen, PHEQ significantly boosts the recognition performance of the MFCC-based baseline system, and the average WER of PHEQ is slightly decreased further when the order of the polynomial transformation function becomes larger. The improvements seem to saturate when the order is set to seven.

Then, we evaluate the performance of CPHEQ, for which the polynomial transformation functions were estimated using both the clean and noisy speech feature vectors, while not only the speech feature vectors but also their corresponding distribution characteristics are utilized by CPHEQ. Therefore, CPHEQ is believed to be more general than SPLICE.

Figure 2: Average word error rate (WER) results (%) with respect to different polynomial orders and different number of mixture Gaussians which were used in the estimation of the transformation functions of CPHEQ.

4.3. Comparison with Other Approaches

Finally, we compared our proposed PHEQ and CPHEQ approaches with the other conventional approaches, including CMVN, HEQ, QHEQ and SPLICE (where the mixture number of the GMMs used respectively for CPHEQ and SPLICE is similarly set to 512). Table 2 shows the recognition results. Notice here that the results for the HEQ and QHEQ were directly adopted from [5]
Finally, the multiplexer will invoke the appropriate web service mapping module to be converted into a well-formed XML format. The driver or passenger's spoken query is keyword spotting and large vocabulary continuous speech recognition (LVCSR). The driver or passenger can use a spoken natural language query to acquire the specific information provided by the ATIS server. A detailed evaluation of the information retrieval system will return the desired information to the driver or passenger. A detailed evaluation of the information retrieval system has been reported in [12].

5. PROTOTYPE SYSTEM

The proposed robustness technique has also been properly integrated into a prototype system for in-vehicle traffic information retrieval using spoken queries, whose framework is shown in Figure 3. Due to the limited computational capability of the mobile devices, the client-server architecture was adopted. The driver or passenger can use a spoken natural language query to acquire the specific information provided by the ATIS server. When the voice gateway received the spoken query, a series of operations (including speech recognition, information retrieval, and speech synthesis, etc.) will be performed in the server side and then the search results will be returned to the client side either in the form of text, speech, or video.

A. Client Side

The in-vehicle telematics terminals can be any of following devices: personal digital assistant (PDA), laptop, smart phone or onboard unit (OBU). Each kind of the terminals can connect to the server using either the IPv4 or the IPv6 network protocols. The major concern of using IPv6 is the demand for a huge number of IP addresses required for the ITS applications, which cannot be completely satisfied under the current IPv4 protocol, even though there exists an alternative way to alleviate it. For the coexistence of the IPv4 and the IPv6 network protocols, we adopted the simple object access protocol (SOAP) for the communication between the client and server sides, instead of using the traditional socket programming approach, mainly due to the transparency property of the SOAP. That is, the SOAP can provide a unified programming model, and the message transmission using the SOAP will not be affected by the different kinds of the network protocols (e.g., IPv4 or IPv6) that are being used, once the communication channel established.

B. Server Side

The server side is composed of five parts: speech recognition, speech synthesis, message mapping, message multiplexer and information retrieval. The speech recognition system includes both keyword spotting and large vocabulary continuous speech recognition (LVCSR). The driver or passenger’s spoken query is first recognized into words, and is then passed to the message mapping module to be converted into a well-formed XML format. Finally, the multiplexer will invoke the appropriate web service according to the spoken query, and the information retrieval system will return the desired information to the driver or passenger. A detailed evaluation of the information retrieval system has been reported in [12].

6. CONCLUSIONS

In this paper, we have investigated the use of histogram equalization (HEQ) for robust ASR. A novel data fitting scheme has proposed to efficiently approximate the inverse of the cumulative density function of training speech for HEQ, which has the merits of lower storage and time consumption when compared to the conventional HEQ approaches. Moreover, a more elaborate attempt to use multiple inverse functions for different noise conditions was also studied as well. Very encouraging results were obtained. The proposed robustness technique has also been properly integrated into our prototype system for in-vehicle traffic information retrieval using spoken queries.

7. REFERENCES