EXPLOITING POLYNOMIAL-FIT HISTOGRAM EQUALIZATION AND TEMPORAL AVERAGE FOR ROBUST SPEECH RECOGNITION

Shih-Hsiang Lin[1], *Yao-Ming Yeh*[1], *Berlin Chen*[2]

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

ABSTRACT

The performance of current automatic speech recognition (ASR) systems radically deteriorates when the input speech is corrupted by various kinds of noise sources. Quite a few of techniques have been proposed to improve ASR robustness in the past several years. Histogram equalization (HEQ) is one of the most efficient techniques that have been used to compensate the nonlinear distortion. In this paper, we explored the use of the data fitting scheme to efficiently approximate the inverse of the cumulative density function of training speech for HEQ, in contrast to the conventional table-lookup or quantile based approaches. Moreover, the temporal average operation was also performed on the feature vector components to alleviate the influence of sharp peaks and valleys that were caused by non-stationary noises. Finally, we also investigated the possibility of combining our approaches with other feature discrimination and decorrelation methods. All experiments were carried out on the Aurora-2 database and task. Encouraging results were initially demonstrated.

Index Terms: histogram equalization, data fitting, temporal average, robustness

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 in the last two decades. In general, 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].

Quite several well-known normalization methods for the feature domain have been developed. For example, 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 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) approaches have been proposed and extensively studied in the recent past, which have also been shown their superiority over the linear compensation approaches, such as CMN and CMVN. A nice feature of the HEQ approaches 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] 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 [5, 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.

Based on the these observations, in this paper, we explored the use of the data fitting scheme to efficiently approximate the inverse of the CDF of training speech for HEQ, in contrast to the conventional table-lookup or quantile based approaches. Moreover, the temporal average operation was also performed on the feature vector components to alleviate the influence of sharp peaks and valleys that were caused by non-stationary noises. Finally, we also investigated the possibility of combining our approaches with other feature discrimination and decorrelation methods.

The rest of this paper is organized as follows. Section 2 describes the basic concept of HEQ and the quantile based histogram equalization (QHEQ), which is an extension of HEQ. Section 3 elucidates our feature normalization approaches. Then, the experimental settings and a series of ASR experiments conducted are presented in Section 4. Finally, conclusions are drawn in Section 5.

2. HISTOGRAM EQUALIZATION

2.1. Basic Formulation

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 [4, 7] or from the cepstral coefficients [8, 9]. Under this assumption, the aim of HEQ is to find a transformation function that can convert the distribution of each feature vector component of the test speech into a predefined target distribution which corresponds to that of the training speech. The formulation is described as follows [9, 10]. Let x be a feature vector component which follows the distribution $p_{Test}(x)$. A transformation function

$F(x)$ converts x to y that follows a reference distribution $p_{Train}(y)$ according to the following expression:

$$p_{Train}(y) = p_{Test}(x)\frac{dx}{dy} = p_{Test}\left(F^{-1}(y)\right)\frac{dF^{-1}(y)}{dy}, \quad (1)$$

where $F^{-1}(y)$ is the inverse function of $F(x)$. Moreover, the relationship between the cumulative probability density functions (CDFs) respectively associated with the test and training speech is governed by:

$$\begin{aligned} C_{test}(x) &= \int_{-\infty}^{x} p_{Test}(x')dx' \\ &= \int_{-\infty}^{F(x)} p_{Test}(F^{-1}(y'))\frac{dF^{-1}(y')}{dy'}dy' \\ &= \int_{-\infty}^{y} p_{Train}(y')dy'\Big|_{y=F(x)} \\ &= C_{Train}(y), \end{aligned} \quad (2)$$

where $C_{Test}(x)$ and $C_{Train}(y)$ are respectively the CDFs for the test and training speech data; y' is the corresponding output of the transformation function $F(x')$; and the transformation function $F(x)$ has the following property:

$$F(x) = C_{Train}^{-1}\left(C_{Test}(x)\right), \quad (3)$$

where C_{Train}^{-1} is the inverse function of C_{Train}. Due to a finite number of speech features being considered, the cumulative histograms are used instead of the cumulative probability density functions for practical implementation. The cumulative histogram of each feature vector component of all training data is computed and divided into a set of equally-probable bins, where the mean $\bar{y}_i$ of each bin i is taken as one of the representative outputs of the transformation function $F(x)$. That is, each feature vector component x 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 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 [11]. Moreover, because a set of cumulative histograms of all speech feature vector dimensions of the training data have to be kept in memory for the table-lookup of restored feature values, such an approach needs huge disk storage consumption and the table-lookup is also time-consuming.

2.2. Quantile-Based Histogram Equalization (QHEQ)

In [5, 12], a parametric type of histogram normalization, which was 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-match of the cumulative histogram as that done by the table-lookup approach described above. A transformation function $H(x)$ is applied to each feature vector component x to make the CDF of the equalized feature match that observed in training:

$$H(x) = Q_K\left(\alpha\left(\frac{x}{Q_K}\right)^{\gamma} + (1-\alpha)\frac{x}{Q_K}\right), \quad (4)$$

where K is the total number of quantiles; Q_K is the K-th quantile of a specific feature vector dimension calculated from the training data; and α and γ are transformation coefficients. For each feature vector dimension, α and γ are optimized using the following equation:

$$\{\alpha,\gamma\} = \arg\min_{\{\alpha,\gamma\}}\left(\sum_{k=1}^{K-1}\left(H(Q_k) - Q_k^{train}\right)^2\right), \quad (5)$$

where Q_k^{train} is the k-th quantile of the same feature vector dimension calculated from the training speech. It allows the estimation of the transformation function performing 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.

3. IMPROVED APPROACHES

3.1. Polynomial-Fit Histogram Equalization (PHEQ)

Least squares regression 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 $G(u_i)$ whose output $\tilde{v}_i$ is closely approximates v_i. That is, it minimizes the sum of the squares error (or the squares of the ordinate differences) between the points $(u_i,\tilde{v}_i)$ and their corresponding points (u_i,v_i) in the data. The function $G(u_i)$ to be estimated can be either linear or nonlinear in its coefficients. For example, if $G(u_i)$ is a linear M-order polynomial function:

$$G(u_i) = a_0 + a_1u_i + a_2u_i^2 + \ldots + a_Mu_i^M, \quad (6)$$

where $a_0, a_1, \ldots, a_M$ are the coefficients, then its corresponding squares error E^2 can be defined as:

$$E^2 = \sum_{i=1}^{N}\left(v_i - \sum_{m=0}^{M} a_m u_i^m\right)^2. \quad (7)$$

In this paper, we presented a polynomial-fit histogram equalization approaches (PHEQ) that uses least squares regression to fit the inverse function of the CDF of the training speech. For each speech feature vector dimension of the training data, given the pair of the CDF $C_{Train}(y_i)$ of the vector component y_i and y_i itself, the linear polynomial function $G(C_{Train}(y_i))$ with output $\tilde{y}_i$ can be expressed as:

$$G(C_{Train}(y_i)) = \tilde{y}_i = \sum_{m=0}^{M} a_m (C_{Train}(y_i))^m, \quad (8)$$

where the coefficients a_m can be estimated by minimizing the squares error E'^2 expressed in the following equation:

$$E'^2 = \sum_{i=1}^{N}\left(y_i - \sum_{m=0}^{M} a_m (C_{Train}(y_i))^m\right)^2, \quad (9)$$

where N is the total number of training speech feature vectors. During speech recognition, for each feature vector dimension, the vector components of the 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.

A quantile-based CDF matching (QCM) approach using least squares regression for CDF matching between the training and test speech data was also proposed recently [6], in which a set of pairs of the means $(\mu_{Test,i}, \mu_{Train,i})$ in the corresponding bins i of the histograms of each feature vector dimension of the test and training data were used for estimating a linear polynomial function that transforms $\mu_{Test,i}$ to $\mu_{Train,i}$. Once the transformation function of each feature vector dimension is obtained, it can be used to transform the corresponding feature vector component in the test data. However, the corresponding experimental results reported in

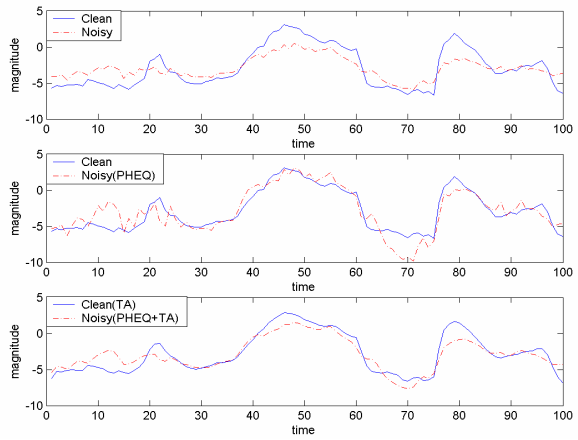

Figure 1: The 2-*th* cepstral feature component sequence of an utterance.

[6] were not conducted on the standard evaluation task, so they can not be directly adopted here for comparison.

3.2. Temporal Average (TA)

Though the above HEQ approaches are very effective in matching the global feature statistics of the test (or noisy) speech to that of the training (or reference) speech, the undesired sharp peaks or valleys of the feature vector component sequence of a noisy speech utterance that are caused by some non-stationary noise can not be restored well to that of the original clean speech utterance, as illustrated in the upper and middle parts of Figure 1. Therefore, in this paper, a finite impulse response moving average filtering operation was performed on the time trajectory of the PHEO restored feature vector component sequence, and each feature vector component is then replaced by its corresponding temporal average (TA):

$$\hat{y}[t] = \frac{1}{2L+1} \sum_{l=-L}^{L} \tilde{y}[t+l] \qquad (10)$$

where $\tilde{y}[t]$ is the PHEQ restored feature vector component at time t and $\hat{y}[t]$ is the corresponding one after the TA operation; and the span order L is empirically set to 2. The feature vector component sequence after the TA operation is shown in the lower part of Figure 1. Notice that the TA operation also can be implemented with an exponential moving average filter [13].

4. EXPERIMENTAL RESULTS

4.1. Experimental 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 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 signal-to-noise ratios (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. For the baseline system, the training and recognition tests used the HTK recognition toolkit [15], which followed the setup originally defined for the ETSI AURORA evaluations.

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 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].

		Polynomial Order			
		3	5	7	9
Clean-Condition Training	All Training Data	22.39	21.54	21.08	21.30
	1000 Quantiles	21.80	21.46	21.13	21.16
	100 Quantiles	22.68	21.31	20.75	20.55
	10 Quantiles	23.42	22.20	22.54	23.42
Multi-Condition Training	All Training Data	10.80	10.34	10.43	10.54
	1000 Quantiles	10.48	10.32	10.40	10.45
	100 Quantiles	10.73	10.45	10.36	10.45
	10 Quantiles	11.65	10.61	10.79	11.58

Table 1: Average word error rates (WERs) with respect to different numbers of training data and different polynomial orders which were used in the estimation of the transformation function of PHEQ.

4.2. Experiments on PHEQ and TA

We first evaluated the performance of the PHEQ and TA approaches. For PHEQ, either all feature vector components or the means of the histogram quantile bins of the training data were used for estimating the linear polynomial function (i.e., the inverse function of the CDF). Different numbers of histogram quantile bins (1000, 100 and 10) and different orders of the polynomial regression were extensively investigated. The results are shown in Table 1, which are averaged for three sets (Sets A, B and C) and SNR levels between 0 dB and 20 dB. As can be seen, for both clean- and multi-condition training, the word error rate (WER) is slightly improved when the order of the polynomial regression becomes higher. However, the improvements seem to saturate for most cases when the order is set to seven. Moreover, the estimation of the linear polynomial function using 100 histogram quantile bins yields the best performance for most cases. Therefore, in the following experiments, the linear polynomial function was set with a regression order of seven and was estimated using 100 histogram quantile bins. On the other hand, the TA operation was additionally performed on the resultant feature vector component of the PHEQ approach and the results are respectively shown in the ninth rows (PHEQ-TA) of Tables 2 and 3 for clean- and multi-condition training. As the results indicate, TA is very effective for clean-condition training, and it provides an average WER reduction of about 4.0%. While, for multi-condition training, the WER reduction provided by TA is almost negligible, which is mainly because that multi-condition training to some extent can model very well the sharp peaks or valleys of the feature

Clean-Condition Training				
Method	Set A	Set B	Set C	Average
MFCC (Baseline)	41.06	41.52	40.03	41.04
ETSI	38.69	44.25	28.76	38.93
CMVN	27.73	24.60	27.17	26.37
HEQ	19.72	18.57	19.24	19.16
QHEQ	23.53	21.90	22.36	22.64
PHEQ	20.98	20.17	21.43	20.75
PHEQ-TA	16.83	15.10	20.02	16.78
HLDA-MLLT+CMVN	21.63	21.37	21.59	21.52
HLDA-MLLT+PHEQ-TA	15.98	15.96	15.91	15.96

Table 2: Comparison between the WER results of the baseline and various approaches under clean-condition training.

Multi-Condition Training				
Method	Set A	Set B	Set C	Average
MFCC (Baseline)	14.78	16.01	19.33	16.18
ETSI	10.64	10.76	12.85	11.13
CMVN	12.70	12.45	14.52	12.98
HEQ	10.02	10.41	10.34	10.24
QHEQ	10.20	10.75	10.76	10.53
PHEQ	9.91	9.41	13.14	10.36
PHEQ-TA	9.41	9.53	11.21	9.82
HLDA-MLLT+CMVN	9.49	9.51	10.40	9.68
HLDA-MLLT+PHEQ-TA	9.06	8.87	8.55	8.88

Table 3: Comparison between the WER results of the baseline and various approaches under multi-condition training.

vector component sequence that was caused by various noise sources at different SNRs.

4.3. Comparison with Other Compensation Approaches

Then, we compared our presented feature normalization approaches with the conventional approaches. The WER results for the baseline MFCC system and the ETSI standard system, as well as CMVN, HEQ and QHEQ, for clean- and multi-condition training are respectively shown in Tables 2 and 3. Notice that the results for the ETSI system, HEQ and QHEQ were directly adopted from [14], [4] and [5], respectively. As compared with the results of PHEQ shown in the eighth rows of Tables 2 and 3 (which were obtained with the best setting indicated by Table 1), it can be found that PHEQ provides significant performance boosts for the baseline MFCC system, and it is also better than CMVN, and competitive to HEQ and QHEQ. If TA is further applied after the PHEQ operation (PHEQ-TA), the recognition results of PHEQ-TA will be considerably better than that of HEQ and QHEQ for clean-condition training.

4.4. Integration with the Discriminative Feature Transformation and Feature Normalization Approaches

Finally, we integrated our proposed feature normalization approach, i.e. PHEQ-TA, with the conventional discriminative feature transformation and feature decorrelation approaches, in which HLDA (Heteroscedastic Linear Discriminant Analysis) was used for discriminative feature transformation and MLLT (Maximum Likelihood Linear Transform) for feature decorrelation. HLDA and MLLT were conducted directly on the Mel-frequency filter bank outputs before PHEQ-TA. For HLDA and MLLT, the states of each HMM were taken as the unit for class assignment. The front-end processing was conducted with HLDA and then with MLLT. The feature vectors from every nine successive frames were spliced together to form the supervectors for the construction of the HLDA transformation matrix. The dimension of the resultant vectors was set to 39. As can be seen from the last rows of Tables 2 and 3, HLDA-MLLT can provide significant performance gains when combined with PHEQ-TA. It is also the same situation when HLDA-MLLT is combined with CMVN, as shown in the tenth rows of Tables 2 and 3.

5. CONCLUSIONS

In this paper, we have investigated the use of data fitting schemes to efficiently approximate the inverse of the CDF of the training speech for HEQ, in contrast to the conventional table-lookup or quantile based approaches. Moreover, the temporal average operation also has been performed on the feature vector components to alleviate the influence of sharp peaks and valleys that were caused by non-stationary noises. Finally, the presented feature normalization approaches have been integrated with the conventional feature discrimination and decorrelation approaches. Significant performance gains have been initially demonstrated.

6. REFERENCE

[1] Y. Gong. “Speech recognition in noisy environments: A survey,” *Speech Communication*, Vol. 16, 1995.

[2] A. Viikki, K. Laurila, “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication*, Vol. 25, 1998.

[3] S. Dharanipargda, M. Padmanabhan, "A Nonlinear Unsupervised Adaptation Technique for Speech Recognition", in *Proc. ICSLP 2000*.

[4] S. Molau et al., “Histogram Based Normalization in the Acoustic Feature Space,” in *Proc. ASRU 2001*.

[5] F. Hilger, H. Ney, “Quantile based histogram equalization for noise robust speech recognition,” in *Proc. EUROPSEECH 2001*.

[6] S. Prasad, S. A. Zahorian, “Nonlinear and Linear Transformations of Speech Features to Compensate for Channel and Noise Effects,” in *Proc. EUROPSEECH 2005*.

[7] S. Molau et al., “Feature Space Normalization in Adverse Acoustic Conditions,” in *Proc. ICASSP 2003*.

[8] A. de la Torre et al., “Non-linear Transformation of the Feature Space for Robust Speech Recognition”, in *Proc. ICASSP 2002*.

[9] J. C. Segura et al., “Cepstral domain segmental nonlinear feature transformations for robust speech recognition,” *IEEE Signal Processing Letters*, Vol. 11(5), 2004.

[10] A. de la Torre et al., “Histogram equalization of the speech representation for robust speech recognition,” *IEEE Trans. Speech and Audio Processing*, Vol. 13(3), 2005.

[11] S. Molau et al., “Histogram Normalization in the Acoustic Feature Space,” in *Proc. ICASSP 2002*.

[12] F. Hilger et al., “Quantile Based Histogram Equalization for Online Applications,” in *Proc. ICSLP 2002*.

[13] M.M. Kantard. *Data Mining: Concepts, Models, Methods and Algorithms*. Wiley-IEEE Press, 2002.

[14] H. G. Hirsch, D. Pearce, “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,” in *Proc. ISCA ITRW ASR 2000*.

[15] S. Young et al., “The HTK Book Version 3.3,” 2005.