LETTER

Histogram Equalization Utilizing Window-Based Smoothed CDF Estimation for Feature Compensation

Youngjoo SUH[†a)], *Nonmember*, **Hoirin KIM**[†], *Member, and* **Munchurl KIM**[†], *Nonmember*

SUMMARY In this letter, we propose a new histogram equalization method to compensate for acoustic mismatches mainly caused by corruption of additive noise and channel distortion in speech recognition. The proposed method employs an improved test cumulative distribution function (CDF) by more accurately smoothing the conventional order statistics-based test CDF with the use of window functions for robust feature compensation. Experiments on the AURORA 2 framework confirmed that the proposed method is effective in compensating speech recognition features by reducing the averaged relative error by 13.12% over the order statistics-based conventional histogram equalization method and by 58.02% over the mel-cepstral-based features for the three test sets.

key words: *feature compensation, histogram equalization, robust speech recognition, window-based CDF estimation*

1. Introduction

Speech recognizers trained on clean speech data usually show dramatic degradation of recognition accuracy when they are operated in real-world noisy environments. The main cause of performance degradation is acoustic mismatches between clean training and noisy test environments [1], [2]. The histogram equalization (HEQ) technique is known to be one of the most efficient methods with very comparable effectiveness in compensating for the acoustic mismatch in the feature space of speech recognition [3]–[5]. In the histogram equalization approach, reliable and accurate estimation of cumulative distribution functions (CDFs) is a critical issue for guaranteeing its effectiveness in the feature compensation [5]. We can use a sufficient amount of training data in the reference (or training) CDF estimation. Therefore, the reference CDF can be obtained highly accurately by using the training data. However, the test CDF is apt to be estimated with a limited amount of test data because the basic input unit in most of the current speech recognizers is usually a short utterance or word. As the amount of test data is smaller, the accuracy of the estimated probability distributions tends to be deteriorated further. Therefore, the classical histogram method is not suitable for the test CDF estimation in short utterance-based test environments. As an improved test CDF estimation method, the order statistics-based approach is widely employed in the histogram equalization techniques [3]. However, from the viewpoint of kernel density estimation [6], it is regarded that the order statistics-based method uses a kernel function with a very small bandwidth which covers only a single sample data. As a result, the resulting CDF by the order statistics-based method tends to be extremely under-smoothed. Therefore, by utilizing a reasonable kernel function with a properly chosen bandwidth, we can obtain a more accurate test CDF. Finally, it is surely expected that histogram equalization based on the enhanced test CDF provides better effectiveness in the feature compensation.

In this letter, we propose a new histogram equalization technique which employs a window-based smoothed test CDF estimation method for feature compensation in the noise robust speech recognition. Our method utilizes a window-type smoothing function as the kernel function in the test CDF estimation instead of just using the rank information of each sample data in the conventional order statistics-based method and produces a more accurate test CDF. The experimental results showed that the proposed approach is more effective than the conventional order statistics-based histogram equalization (OS-HEQ) technique in compensating the acoustic features in noisy environments.

2. Order-Statistics-Based Histogram Equalization

The idea of histogram equalization is to convert the probability density function (PDF) of the test variable into its reference PDF. For a given random test variable y, whose PDF is given as $P_Y(y)$, a basic rule for histogram equalization is defined by an inverse transform function $x = F(y)$, which maps $P_Y(y)$ into $P_X(x)$ in [4] as

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

where $C_X^{-1}(x)$ is the inverse of reference CDF $C_X(x)$, and $C_Y(y)$ is the test CDF of random variable y.

Equation (1) indicates that histogram equalization uses only two CDFs, the reference and test CDFs. Thus, for accurate histogram equalization, the two CDFs need to be estimated as reliable as possible. Like other statistical estimation, reliable CDF estimation is directly related to the amount of sample data. In developing speech recognizers, training data can be collected large enough that the reference CDF estimated by the classical histogram is regarded highly reliable. However, a short utterance or word is still the major input unit in most current speech recognizers. In this case, the length of each utterance or word may be too short for a reliable estimation of its CDF. Accordingly, the test CDF

Manuscript received March 11, 2008.
Manuscript revised April 21, 2008.
[†]The authors are with School of Engineering, Information and Communications University, Daejeon 305–732, Korea.
a) E-mail: yjsuh@icu.ac.kr
DOI: 10.1093/ietisy/e91–d.8.2199

can be ill-estimated due to the lack of sample data. When the amount of sample data is smaller, it is generally known that the order statistics-based CDF estimation is more accurate than the classical cumulative histogram approach [3]. Therefore, OS-HEQ is a preferred histogram equalization technique in feature compensation. A brief description of OS-HEQ utilizing the order statistics-based CDF estimation is given as follows [3].

Let us define sequence S consisting of N frames of a particular test feature component as

$$S = \{y_1, y_2, \cdots, y_n, \cdots, y_N\}, \tag{2}$$

where y_n is the test feature component at the n-th frame.

The order statistics of (2) is represented as

$$y_{T(1)} \le y_{T(2)} \le \cdots \le y_{T(r)} \le \cdots \le y_{T(N)}, \tag{3}$$

where $T(r)$ denotes the original frame index of feature component $y_{T(r)}$ in which r represents its rank when the elements of sequence S are sorted in ascending order.

From (2) and (3), the order statistics-based test CDF estimate is given as

$$C_Y(y_n) = \frac{R(y_n) - 0.5}{N}, \tag{4}$$

where $R(y_n)$ denotes the rank of y_n, ranging from 1 to N.

Given y_n, an estimate of x_n by OS-HEQ is defined as

$$\begin{aligned} \hat{x}_n &= C_X^{-1}[C_Y(y_n)] \\ &= C_X^{-1}\left[\frac{R(y_n) - 0.5}{N}\right]. \end{aligned} \tag{5}$$

3. Histogram Equalization Based on Smoothed CDF Estimation

In the classical probabilistic definition, PDF is estimated as the number of samples in a given unit bin. Then, CDF is obtained as the accumulation of PDF for the range up to the bin. When the number of samples is not large enough, the estimated PDF is heavily dependent upon the bin width as well as bin position. Therefore, it needs an additional smoothing process for more accurate estimation.

In the order statistics-based estimation, the bin width is infinitesimally small since it contains only a single sample. Therefore, the resulting CDF by the order-statistics-based estimation defined in (4) is prone to be under-smoothed and thus needs to be properly smoothed by introducing a kernel function with an appropriate kernel width according to the kernel density estimation theory [6]. In smoothing PDF estimated by the order statistics-based method, it is important to choose the most appropriate kernel shape as well as kernel width for a given number of samples. In this letter, we propose the window-type kernel-based test CDF estimation method, where the kernel shape is determined by the adopted window functions and kernel width is empirically chosen within the given input utterance to produce an optimally smoothed test CDF in the sense of speech recognition accuracy. It is then utilized in histogram equalization for improving the effectiveness of feature compensation for robust speech recognition. The two approaches for estimating the test CDF with different types of window functions are explained as follows.

3.1 Rectangular Window-Based Test CDF Estimation

One simple choice of a window function is a rectangular window. The test PDF and its CDF estimated by using the rectangular window are obtained as follows

$$P_{Y,rect}(y_n) = \frac{\psi_{rect}(y_n)}{\sum_{m=1}^{N} \psi_{rect}(y_m)}, \tag{6}$$

$$C_{Y,rect}(y_n) = \sum_{r=1}^{R(y_n)} P_{Y,rect}(y_{T(r)}), \tag{7}$$

where relative frequency $\psi_{rect}(y_n)$ at test feature sample y_n obtained by the rectangular window is given as

$$\psi_{rect}(y_n) = \sum_{m=1}^{N} W_{rect}(y_n - y_m), \tag{8}$$

in which rectangular window function $W_{rect}(\lambda)$ is defined as

$$W_{rect}(\lambda) = \begin{cases} 1, & \text{if } |\lambda| \le B \\ 0, & \text{otherwise} \end{cases}, \tag{9}$$

where B is the window kernel width, obtained as

$$B = \frac{y_{\max} - y_{\min}}{D}, \tag{10}$$

where $y_{\max}$ and $y_{\min}$ are the maximum and minimum values of y_n in the sequence, respectively, and D is an empirically chosen constant.

3.2 Triangular Window-Based Test CDF Estimation

Another choice of the window function is a triangular window, which takes into account the decaying contributions to the relative frequency in a given bin as the neighboring samples are apart from the center of the window. Similar to the rectangular window case, the test PDF and resulting CDF estimated by using the triangular window are obtained as follows

$$P_{Y,tri}(y_n) = \frac{\psi_{tri}(y_n)}{\sum_{m=1}^{N} \psi_{tri}(y_m)}, \tag{11}$$

$$C_{Y,tri}(y_n) = \sum_{r=1}^{R(y_n)} P_{Y,tri}\,(y_{T(r)}), \tag{12}$$

where relative frequency $\psi_{tri}(y_n)$ at test feature sample y_n obtained by the triangular window is given as

$$\psi_{tri}(y_n) = \sum_{m=1}^{N} W_{tri}(y_n - y_m), \tag{13}$$

in which triangular function $W_{tri}(\lambda)$ is defined as

$$W_{tri}(\lambda) = \begin{cases} 1 - \dfrac{|\lambda|}{B}, & \text{if } |\lambda| \le B \\ 0, & \text{otherwise} \end{cases}. \tag{14}$$

4. Experimental Results

In the performance evaluation, we used the AURORA 2 database, which is based on the TI-DIGITS database. In our experiments, we employed only clean training-conditions which indicate that the training data for training the speech recognizer are composed of clean speech only. We examined the effectiveness of the proposed approaches on the three test sets, test sets A, B, and C, where test sets A and B are corrupted by different four kinds of additive noise, respectively, and test set C is contaminated by two kinds of additive noise used in test sets A and B, and channel distortion together. The number of connected digits in each utterance of the AURORA 2 test sets ranges from one to seven, which makes utterance lengths of the test sets quite variable. Figure 1 depicts the utterance length distribution of the AURORA 2 test sets. From the figure, we see that the test data contains speech utterances of various lengths with the higher population at the shorter lengths, which makes it especially suitable to evaluate the effectiveness of the proposed test CDF estimation method in small as well as various numbers of test samples.

We employed the ETSI AURORA 2 framework in conducting the feature extraction and determining the architecture of speech recognizers used in the experiments as follows [7]. Speech signals are blocked into a sequence of frames, each 25 ms in length with a 10 ms interval. Speech frames are pre-emphasized using a factor of 0.97, and a Hamming window is then applied. From a set of 23 mel-scaled filter-bank log energies, a 39-dimensional mel-frequency cepstral coefficient(MFCC)-based feature vector consisting of twelve MFCCs, the log energy, and their first and second derivatives is extracted. The baseline speech recognizer employs whole-word models and its dictionary consists of 13 words, composed of 11 digits, a silence, and a short-pause. Each digit-based hidden Markov model (HMM) consists of sixteen states while silence and short-pause HMMs has three states and one state, respectively. Each state in the word models has three mixture components and that in the silence and short-pause models has six mixture components. Each mixture component is modeled with a diagonal covariance matrix. The number of histogram bins in the reference CDFs was chosen as 64 in all of the histogram equalization techniques. Due to the use of a linear interpolation within each histogram bin in the computation of reference CDF estimate, further increasing the number of histogram bins did not show any meaningful performance improvement. The histogram equalization was conducted on all of the 39 components of the MFCC feature vector for the training and test data with utterance-by-utterance estimation of the test CDFs. The parameter D is empirically set to 340 and 210 for the rectangular and triangular windows, respectively.

Figure 2 shows the recognition results at various signal-to-noise ratio (SNR) conditions when OS-HEQ, rectangular window-based HEQ (RW-HEQ), and triangular window-based HEQ (TW-HEQ) techniques as well as the MFCC features are used. The results are represented in terms of the overall averaged word accuracy for the three test sets. In this figure, we observe that the proposed techniques meaningfully outperform OS-HEQ. Of the two proposed techniques, TW-HEQ provides slightly better performance. As the SNR level is lowered, the corresponding absolute error reduction is proportionally increasing, which indicates that the consistency of relative error reduction is well kept across various SNR levels. From these recognition results and the utterance length distribution in Fig. 1, we also note that the proposed histogram equalization techniques are substantially effective in test utterances with short as well as various lengths.

Table 1 shows the recognition results for test sets A, B, and C obtained by MFCC, OS-HEQ, RW-HEQ, and TW-HEQ, respectively. Each result is obtained as the averaged value between 0 and 20 dB SNRs. For sets A and B, the proposed RW-HEQ technique shows outstanding im-

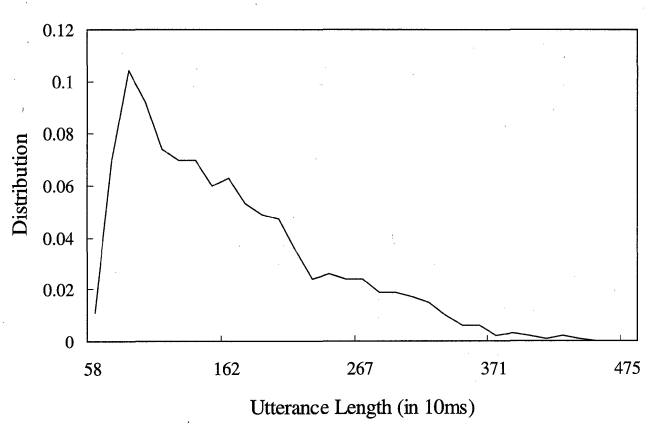

Fig. 1 Utterance length distribution of the AURORA 2 test sets.

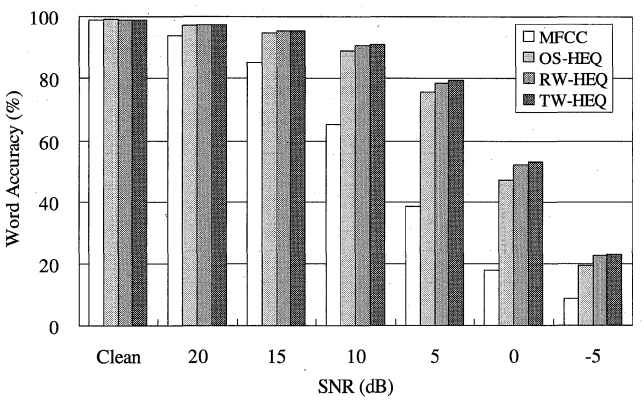

Fig. 2 Recognition results of MFCC, OS-HEQ, RW-HEQ, and TW-HEQ at various SNR conditions.

Table 1 Word accuracy of the baseline MFCC, OS-HEQ, RW-HEQ, and TW-HEQ at the AURORA 2 task (Averaged between 0 and 20 dB SNRs).

Noises		Word Accuracy (%)			
		MFCC	OS-HEQ	RW-HEQ	TW-HEQ
A	Subway	69.86	81.33	83.20	83.68
	Babble	50.24	81.46	82.14	82.37
	Car	59.87	81.58	84.48	85.04
	Exhibition	64.53	78.28	79.94	79.96
	Average	**61.12**	**80.66**	**82.44**	**82.76**
B	Restaurant	51.51	82.01	82.79	83.06
	Street	61.52	81.84	83.72	84.17
	Airport	53.33	82.50	84.82	85.14
	Station	55.92	80.69	82.92	83.62
	Average	**55.57**	**81.76**	**83.56**	**84.00**
C	Subway	67.23	77.72	81.07	81.85
	Street	66.13	79.36	82.69	83.23
	Average	**66.68**	**78.54**	**81.88**	**83.24**
Overall Average		**60.01**	**80.68**	**82.78**	**83.21**

provements over MFCC by producing error reductions of 54.83% and 62.99% over MFCC and substantial improvements over OS-HEQ with error reductions of 9.20% and 9.87%, respectively. For the same test sets, TW-HEQ provides better improvements with error reductions of 55.66% and 63.99% over MFCC and 10.86% and 12.28% over OS-HEQ. The results for set C also confirm the superiority of the two proposed techniques to the conventional OS-HEQ by yielding error reductions of 45.62% and 47.60% over MFCC and 15.56% and 18.64% over OS-HEQ, respectively. From the results for test sets A, B, and C, we notice that the proposed window-based approaches are more effective in compensating for the acoustic mismatch caused by both additive noise and channel distortion together than additive noise only. When clean speech is corrupted by both additive noise and channel distortion, the resulting acoustic mismatch behaves as a more complex nonlinear function, because additive noise, channel distortion, and clean speech affect on the function independently each other. We also note from the results that the proposed histogram equalization techniques are more effective in the noise types of Car and Airport than those of Babble and Restaurant. The former noise types are known to be more stationary. In the stationary noise environments, the acoustic mismatch leads more closely to the monotonic transformation, a transformation that does not cause any change of rank information in (3). When this major requirement is satisfied, the histogram equalization technique can work more effectively on the feature compensation by more faithfully meeting other requirements for its full performance such as the accurate test CDF estimation.

5. Conclusion

For more effective feature compensation in noisy speech recognition environments, the histogram equalization technique requires more accurate test CDF estimation. The order statistics-based CDF estimation is the most frequently employed approach in the conventional histogram equalization techniques, but it still has some room for further improvement in its accuracy. The window-based smoothed CDF estimation method can obtain more accurate test CDF by smoothing the order statistics-based CDF with the use of window functions. The proposed histogram equalization technique then utilizes the improved test CDF estimated by the window-based smoothed estimation method. Experimental results showed the effectiveness of the proposed histogram equalization technique for feature compensation in robust speech recognition.

References

[1] X. Huang, A. Acero, and H.-W. Hon, Spoken Language Processing, Prentice-Hall, 2001.
[2] A. Sankar and C.-H. Lee, "A maximum-likelihood approach to stochastic matching for robust speech recognition," IEEE Trans. Speech Audio Process., vol.4, no.3, pp.190–202, May 1996.
[3] J.C. Segura, C. Benítez, Á. de la Torre, A.J. Rubio, and J. Ramírez, "Cepstral domain segmental nonlinear feature transformations for robust speech recognition," IEEE Signal Process. Lett., vol.11, no.5, pp.517–520, May 2004.
[4] Á. de la Torre, A.M. Peinado, J.C. Segura, J.L. Pérez-Córdoba, M.C. Benítez, and A.J. Rubio, "Histogram equalization of speech representation for robust speech recognition," IEEE Trans. Speech Audio Process., vol.13, no.3, pp.355–366, May 2005.
[5] Y. Suh, M. Ji, and H. Kim, "Probabilistic class histogram equalization for robust speech recognition," IEEE Signal Process. Lett., vol.14. no.4, pp.287–290, April 2007.
[6] R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, Wiley Interscience, 2001.
[7] H.-G. Hirsch and D. Pearce, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," Proc. ICSLP, pp.16–20, Oct. 2000.