FILTERING ON THE TEMPORAL PROBABILITY SEQUENCE IN HISTOGRAM EQUALIZATION FOR ROBUST SPEECH RECOGNITION

Syu-Siang Wang[1], *Yu Tsao*[1], *Jeih-weih Hung*[2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
[2]Dept. of Electrical Engineering, National Chi Nan University, Nantou, Taiwan

ABSTRACT

In this paper, we propose a filter-based histogram equalization (FHEQ) approach for robust speech recognition. The FHEQ approach first represents the original acoustic feature sequence with statistic probability. Then, a temporal average (TA) filter is applied to smooth the statistic probability sequence. Finally, the filtered statistic probability sequence is transformed to form a new acoustic feature stream. Filtering on statistic probability of a feature sequence is a novel concept that can incorporate the advantages of the conventional histogram equalization (HEQ) and temporal filtering techniques for better noise robustness. Our experimental results on the Aurora-2 and Aurora-4 tasks show that FHEQ outperforms the conventional cepstral mean subtraction (CMS), cepstral mean and variance normalization (CMVN), and HEQ. Furthermore, we conducted a comparison test on TA-HEQ and HEQ-TA, which apply a TA filter to smooth acoustic features before and after the HEQ processing, respectively. The test results show that FHEQ outperforms both TA-HEQ and HEQ-TA, suggesting that filtering in probability is more effective than filtering in acoustic feature.

Index Terms— HEQ, FHEQ, feature normalization, temporal filter, noise robust speech recognition

1. INTRODUCTION

The performance of an automatic speech recognition (ASR) system often degrades dramatically in noisy conditions [1]. To enhance the recognition performance robustness, a lot of approaches have been proposed [2] [3] [4]. A successful category of approaches aims to produce robust features that are less sensitive to environmental mismatch between training and testing conditions. Temporal filtering and feature statistics normalization techniques are two sub-groups of these robustness approaches. Temporal filtering methods focus on designing a filter to suppress noise effects in acoustic features, and the filter is usually designed based on the fact that important speech components for recognition are mainly located around low modulation frequency bands (except for the DC component). Relative spectral (RASTA) [5] [6] is a well-known filtering method, which preserves the informative speech components around 4 Hz while suppresses components at other modulation frequencies. Moving average (MA) [7] and auto-regression moving average (ARMA) [8] are another two notable filtering methods; both produce temporally smoothed acoustic features with reduced noise interferences.

Normalization methods, on the other hand, aim to reduce the mismatch by mapping training and testing acoustic features to make them close to each other in one or more statistical quantities. Among the well-know normalization methods, cepstral mean subtraction (CMS) [9] [10] removes the first moment from the cepstral features; cepstral mean and variance normalization (CMVN) [11] and higher order cepstral moment normalization (HOCMN) [12] perform second and higher-order moment normalizations, respectively, to make the distribution of noisy cepstral features closer to that of the clean ones. Histogram equalization (HEQ) is another powerful normalization approach, which first ranks and converts feature components to probability distribution (PD) values. Then a mapping function is applied to transform the PD values to a pre-defined reference distribution [13] [14]. Based on the type of mapping function, various HEQ approaches have been developed. Representative examples include class-based histogram equalization (CHEQ) [15], quantile-based histogram equalization (QHEQ) [16], and polynomial-fit histogram equalization (PHEQ) [17].

In this paper, we propose a filter-based HEQ (FHEQ) algorithm, which integrates the temporal filtering technique with the HEQ framework. Similar to the conventional HEQ, the proposed FHEQ first converts an acoustic feature sequence into a probability sequence. Then a low-pass filter is applied on the probability sequence with the hope to reduce the noise effect. Finally the filtered probability sequence is transformed to the final acoustic feature sequence. Compared with the conventional temporal filtering methods, the presented FHEQ is a non-linear process for the feature sequence since the filter operates on the probability values associated with the features. Besides, FHEQ differs from HEQ primarily in that FHEQ can alter the relative order of the features among a sequence while HEQ cannot. We evaluate the proposed FHEQ on the Aurora-2 [18] and Aurora-4 [19] tasks. Experimental results confirmed that FHEQ outperforms the conventional CMS, CMVN, and HEQ approaches on both tasks.

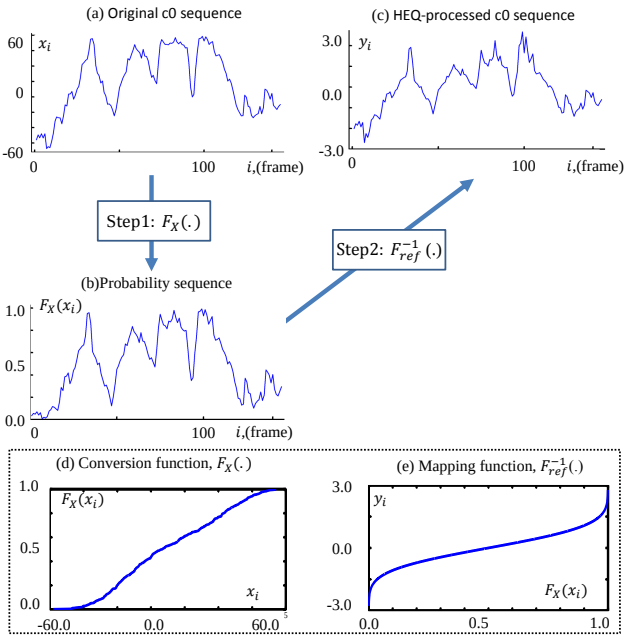

Fig. 1. Procedure of the conventional HEQ algorithm

2. HISTOGRAM EQUALIZATION

The technique of histogram equalization (HEQ) normalizes the feature sequences in the training and testing sets so that they can approximately match a common probability distribution function (PDF), known as reference PDF or target PDF. The reference PDF can be obtained from the features in the clean training set, or any non-negative and monotonically non-decreasing function. In HEQ, an arbitrary feature sequence, denoted by $\{x_1, ..., x_N\}$, N is the total number of frames, is viewed as the sample set of a random variable X with PDF, $F_X(x)$. Then, applying the mapping process:

$$y_i = F_{ref}^{-1}(F_X(x_i)), \quad 1 \le i \le N \tag{1}$$

can make the PDF of the new feature sequence $\{y_1, ..., y_N\}$ approach the reference PDF, $F_{ref}(y)$. Fig. 1 shows the overall HEQ procedure. In Fig. 1, (a), (b), and (c), respectively, show the original feature (c_0 in MFCC) sequence, probability distribution (PD) sequence, and new feature sequence; $F_X(.)$ and $F_{ref}^{-1}(.)$ in Eq. (1), are also illustrated as (d) and (e) in the figure, where we adopt the standard normal distribution function as the reference PDF, $F_{ref}(.)$. According to Fig. 1, each feature in the sequence is first converted to a PD value by $F_X(.)$. Then, the PD sequence is transformed to form the new feature sequence by $F_{ref}^{-1}(.)$.

3. FILTER-BASED HISTROGRAM EQUALIZATION

HEQ compensates the distortion of the statistics (i.e., the mean, variance, and any higher-order moments) caused by noise and effectively reduces the histogram mismatch between the features of the training and testing sets. However, HEQ cannot recover the loss of the (size) ordering information of each noise-free feature in the sequence due to the random effects of the noise. In more detail, the functions F_X, F_{ref} and $F_{ref}^{-1}(F_X)$ in Eq. (1) are always monotonically non-decreasing functions, and thus the (size) ordering of the original sequence $\{x_i\}$ is preserved in the new sequence $\{y_i\}$. That is,

If $x_i \le x_j$,
then $F_X(x_i) \le F_X(x_j)$,
and
$y_i = F_{ref}^{-1}(F_X(x_i)) \le F_{ref}^{-1}(F_X(x_j)) = y_j$.

In other words, the rank mismatch existing in the original feature sequence $\{x_i\}$ is left unprocessed.

Based on the aforementioned observation, here we propose a novel method, termed filter-based HEQ (FHEQ), to enhance noise robustness of speech features. Briefly speaking, FHEQ applies a filter to the PD sequence $\{F_X(x_i)\}$ in Eq. (1) during HEQ to alleviate the rank mismatch. The input-output relationship of the FHEQ process is

$$y_i = F_{ref}^{-1}(\sum_k h_k F_X(x_{i-k})), \quad 1 \le i \le N \tag{2}$$

where $\{h_k\}$ denotes the filter coefficients. In this study, we select a simple two-point low-pass FIR filter, i.e. $h[k] = \alpha\delta[k] + (1-\alpha)\delta[k-1]$, $0 < \alpha < 1$, for FHEQ in Eq. (2). Thus in FHEQ, the new PD sequence is a smoothed version of the original one, and each new PD point is the weighted sum of two original neighboring PD points. The idea of FHEQ is partially motivated by the concept of temporal filtering methods like RASTA [5], MA [7], and ARMA [8]. In general, a noise-corrupted feature sequence reveals a more oscillating characteristic than the clean counterpart, implying that noise introduces relatively high modulation frequency distortion. RASTA, MA, and ARMA suppress the high modulation frequency portion in a feature sequence and achieve better noise robustness. Similarly, due to the monotonic relationship between the feature sequence and PD sequence, the PD sequence for a noise-corrupted feature sequence appears more fluctuating. As a result, we adopt a low-pass filter to smooth the PD sequence as it should be in the clean case.

Fig. 2 shows the procedure of the FHEQ algorithm. In Fig. 2, (a), (b), (c), and (d), respectively, show the original feature sequence, PD sequence, filtered PD sequence, and new feature sequence, where (e) illustrates the frequence response of the low-pass filter, $h[k] = 0.25\delta[k] + 0.75\delta[k-1]$. Similar to the conventional HEQ as shown in Fig. 1, each feature in the sequence is first converted to a probability value. The resulting probability sequence is then processed by a moving-average filter. Finally, the filtered probability sequence is transformed back to form the new feature sequence.

Compared with HEQ, the proposed FHEQ has two potential advantages:

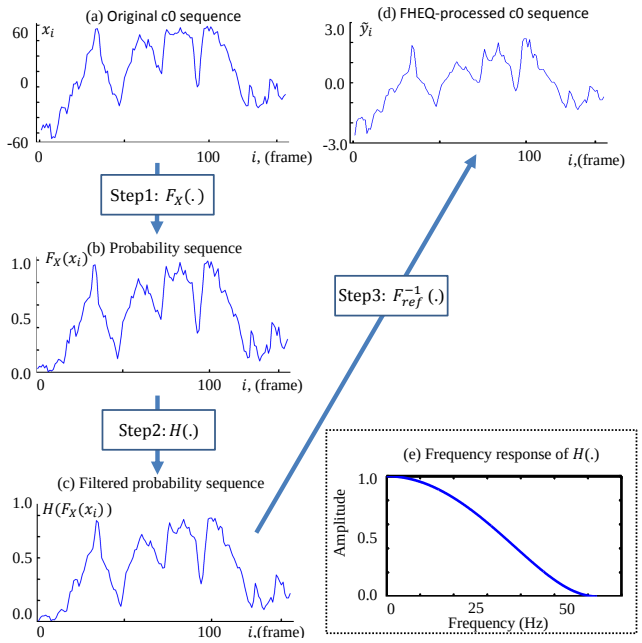

Fig. 2. Procedure of the FHEQ algorithm

1. FHEQ considers the inter-frame information to determine the new feature sequence, while HEQ does not. By properly selecting the filter coefficients $\{h_k\}$ in Eq. (2), FHEQ can likely reduce the rank mismatch among the features that caused by noise.

2. FHEQ emphasizes the slow-varying portions of the PD sequence, and thus the FHEQ-processed feature sequence is smoother than the HEQ-processed one. As a result, FHEQ acts like a low-pass temporal filter and can preserve the important modulation frequency components for speech recognition.

4. EXPERIMENT RESULTS AND ANALYSES

This section introduces our experimental setup and provides recognition results and discussions.

4.1. Experimental Setup

First, we describe the feature extraction procedure and briefly introduce the Aurora-2 [18] and Aurora-4 [19] databases that are used to test performance in this study.

4.1.1. Feature extraction

Each utterance in the training and testing sets was converted into a sequence of Mel-frequency cepstral coefficients (MFCC) vectors. Each vector included 13 static components plus their first- and second- order time derivatives. The frame length and shift were set to 32 ms and 10 ms, respectively.

4.1.2. Database: Aurora-2

Aurora-2 is a standardized database for connected digit speech recognition under noisy conditions [18]. The original clean speech utterances in Aurora-2 were acquired from the TIDIGITs corpus [20]; then different noises were artificially added into the clean speech to generate noisy speech data. Aurora-2 includes three test sets: Sets A, B, and C. Speech signals in test Sets A and B were distorted by four additive noise individually, and speech signals in test Set C were distorted by two additive noise and channel interferences; each noise instance was added to the clean speech at six SNR levels (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, and 20 dB); accordingly, Aurora-2 provides 70 test conditions in total. There are two training sets in the Aurora-2 database: clean- and multi- condition training sets. The clean-condition training set includes 8440 speech utterances, all recorded from a clean condition. The multi-condition training set includes the same 8440 utterances with artificially affected by the same four types of additive noise as those in test Set A, at five SNR levels: 5 dB, 10 dB, 15 dB, 20 dB, and clean. In this paper, we adopted the multi-condition training set and a complex back-end model topology suggested in [21] to train acoustic models. The acoustic models include 11 digit models with silence and short pause models. Each digit model contains 16 states and 20 Gaussian mixtures per state. Silence and short pause models include three and one states, respectively, both with 36 Gaussian mixtures per state.

4.1.3. Database: Aurora-4

Aurora-4 is a standardized database for large vocabulary continuous speech recognition (LVCSR) under noisy conditions [19]. The clean speech utterances in Aurora-4 were acquired from the Wall Street Journal (WSJ0) corpus [22] and then contaminated by different noises artificially to generate noisy speech data. Two sampling rates, 8k Hz and 16k Hz, were provided in Aurora-4, and we chose 8k Hz data for both training and testing processes. Aurora-4 also includes clean- and multi- condition training sets, both consisting of 7138 utterances. The Aurora-4 database comprises 14 test sets with different noise and channel interferences. These 14 sets were further categorized into four sets: Set A (clean speech in the same channel condition as the training data; set 1), Set B (noisy speech in the same channel condition as the training data; sets 2-7), Set C (clean speech in a different channel condition to the training data; set 8), and Set D (noisy speech in a different channel condition to the training data; sets 9-14). The multi-condition training set was used to train acoustic models. In this study, we used context-dependent triphone acoustic models, where each triphone was characterized by an HMM. Each HMM consists of 3 states, with 8 Gaussian mixtures per state. A tri-gram language model was prepared based on the reference transcription of the training utterances.

4.2. Recognition Results

In the following experiments, we reported the word error rate (WER) as the performance measure. For Aurora-2, we present the average WERs of the three test sets and an overall average result (denoted as Avg). For Aurora-4, we present the WERs of the four test sets and an overall average result (denoted as Avg). The low-pass filter in Eq. (2) for FHEQ is preliminarily set to be $h[k] = 0.25\delta[k] + 0.75\delta[k-1]$.

4.2.1. Aurora-2 Results

Table 1 presents the results of the conventional HEQ and the proposed FHEQ, along with the MFCC baseline, CMS, and CMVN on the Aurora-2 task. Each WER value in Table 1 is an average result over five SNR levels (0 dB, 5 dB, 10 dB, 15 dB, and 20 dB). From Table 1, we first notice that HEQ outperformed the MFCC baseline, CMS, and CMVN. Next, we observe that FHEQ achieved better performance than HEQ. This set of results confirms that the integration of a filter can effectively improve the conventional HEQ method.

In a previous study, a PHEQ with temporal average (TA) approach (PHEQ-TA) was proposed to suppress noise components of the PHEQ-processed acoustic features [17]. Here, we tested the HEQ-TA performance to compare with the proposed FHEQ. In a similar manner, we designed a new algorithm that performed TA on acoustic features before feeding them into the HEQ process; we named this algorithm TA-HEQ and also tested its performance on Aurora-2. Fig. 3 illustrates the WERs of HEQ, TA-HEQ, HEQ-TA, and the proposed FHEQ on different test sets of Aurora-2; here a same filter was applied for TA-HEQ, HEQ-TA, and FHEQ. From Fig. 3, we observe that both TA-HEQ and HEQ-TA provided lower WERs than the conventional HEQ. This set of results suggests that applying a TA filter before and after the HEQ processing can produce more robust acoustic features. In addition, we found that FHEQ outperformed TA-HEQ and HEQ-TA consistently over the four test sets. The results indicate that applying the filter on the PD sequences is more effective than applying the same filter on acoustic feature sequences either before or after the HEQ processing.

Table 1. WER (%) of three test sets of Aurora-2

Set	Set A	Set B	Set C	Avg
Baseline	8.29	9.86	10.74	9.41
CMS	7.29	7.45	6.87	7.27
CMVN	6.87	7.50	7.31	7.21
HEQ	7.00	7.42	7.07	7.18
FHEQ	**6.61**	**7.12**	**6.77**	**6.84**

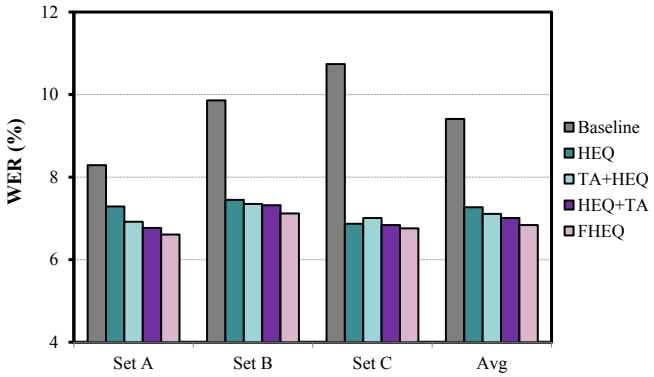

Fig. 3. WER (%) of the MFCC baseline, HEQ, TA+HEQ, HEQ+TA, and FHEQ on the Aurora-2 task.

4.2.2. Aurora-4 Results

In this section, we present the testing results of FHEQ on the Aurora-4 task. Table 2 lists the recognition results of the MFCC baseline, HEQ, and FHEQ on the four test sets in Aurora-4. From Table 2, we observe that FHEQ outperformed the conventional HEQ consistently over different test sets. When compared with the MFCC baseline, FHEQ achieved a significant 18.65% (from 22.95% to 18.67%) average WER reduction over the 14 test sets on Aurora-4.

5. CONCLUSION

In this study, we proposed an FHEQ algorithm for robust speech recognition. Similar to the conventional HEQ, FHEQ first converts a feature sequence into a probability sequence. Then, a temporal moving average filter is applied on the probability sequence. Finally, the filtered probability sequence is transformed to form a new feature stream. Applying the filter on the probability sequence enables FHEQ to effectively reduce noise interferences and preserve important speech components. From the experimental results on Aurora-2 and Aurora-4, we verified that FHEQ can provide better performance than the conventional HEQ for both connected digit recognition and LVCSR under noisy conditions. Moreover from a comparison experiment, we found that FHEQ outperformed TA-HEQ and HEQ-TA, suggesting that applying the filter on the probability sequences is more effective than applying the same filter on acoustic feature sequences.

Table 2. WERs (%) of four test sets on Aurora-4

Set	A	B	C	D	Avg
Baseline	10.83	20.66	16.57	28.34	22.95
HEQ	10.46	17.23	13.33	22.92	18.91
FHEQ	**9.39**	**17.21**	**12.67**	**22.69**	**18.67**

6. REFERENCES

[1] Y. Gong, "Speech recognition in noisy environments: A survey," *Speech Communication*, vol. 16, pp. 261–291, 1995.

[2] S. Molau, D. Keysers, and H. Ney, "Matching training and test data distributions for robust speech recognition," *Speech Communication*, vol. 41, pp. 579–601, 2003.

[3] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm and System Development*. New Jersy: Prentice Hall PTR, 2001.

[4] L.-C. Sun and L.-S. Lee, "Modulation spectrum equalization for improved robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 20, pp. 828–843, 2012.

[5] H. Hermansky and N. Morgan, "Rasta processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 587–589, 1994.

[6] S. Tibrewala and H. Hermansky, "Multi-band and adaptation approaches to robust speech recognition," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2619–2622, 1997.

[7] C.-P. Chen, J. A. Bilmes, and K. Kirchhoff, "Low-resource noise-robust feature post-processing on aurora 2.0," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 2445–2448, 2002.

[8] C.-P. Chen, K. Filali, and J. A. Bilmes, "Frontend post-processing and backend model enhancement on the aurora 2.0/3.0 databases," in *Proc. International Conference on Speech and Language Processing (ICSLP)*, pp. 241–244, 2002.

[9] D. B. Olli Viikki and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," in *Proc. International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 733 –736, 1998.

[10] H. K. Kim and R. C. Rose, "Cepstrum-domain acoustic feature compensation based on decomposition of speech and noise for asr in noisy environments," *IEEE Transactions on Speech and Audio Processing*, vol. 11, pp. 435–446, 2003.

[11] S. Tibrewala and H. Hermansky, "Multi-band and adaptation approaches to robust speech recognition," in *Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 2619–2622, 1997.

[12] C.-W. Hsu and L.-S. Lee, "Higher order cepstral moment normalization for improved robust speech recognition," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 17, pp. 205–220, 2009.

[13] D. P. Ibm, S. Dharanipragada, and M. Padmanabhan, "A nonlinear unsupervised adaptation technique for speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 556–559, 2000.

[14] A. de la Torre, A. M. Peinado, J. C. Segura, J. L. Prez-Crdoba, M. C. Bentez, and A. J. Rubio, "Histogram equalization of speech representation for robust speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 13, pp. 355–366, 2005.

[15] S. K. Youngjoo Suh and H. Kim, "Compensating acoustic mismatch using class-based histogram equalization for robust speech recognition," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, 2007.

[16] F. Hilger, H. Ney, and L. F. I. Vi, "Quantile based histogram equalization for noise robust speech recognition," in *Proc. European Conference on Speech Communication and Technology (Eurospeech)*, pp. 1135–1138, 2001.

[17] S.-H. Lin, Y.-M. Yeh, and B. Chen, "Exploiting polynomial-fit histogram equalization and temporal average for robust speech recognition," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, pp. 1135–1138, 2006.

[18] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *Proc. International Conference on Spoken Language Processing (Interspeech)*, pp. 29–32, 2000.

[19] N. Parihar and J. Picone, "Aurora working group: Dsr front end lvcsr evaluation au/384/02," in *Institutefor Signal and Information Processing Report*, 2002.

[20] R. Leonard, "A database for speaker-independent digit recognition," in *Proc. International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, pp. 328–331, 1984.

[21] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, 2005.

[22] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proc. International Conference on Spoken Language Processing (ICSLP)*, 1992.