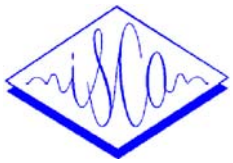
ISCA Archive
http://www.isca-speech.org/archive



INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

A NEW FEATURE EXTRACTION FRONT-END FOR ROBUST SPEECH RECOGNITION USING PROGRESSIVE HISTOGRAM EQUALIZATION AND MULTI-EIGENVECTOR TEMPORAL FILTERING

Shang-nien Tsai and Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University
Taipei, Taiwan.
sntsai@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

10.21437/Interspeech.2004-109

Abstract

In this paper, a new feature extraction front-end for robust speech recognition using progressive histogram equalization and multi-eigenvector temporal filtering is proposed. The progressive histogram equalization (PHEQ) performs the histogram equalization (HEQ) progressively with respect to a reference interval which moves with the present frame to be processed. The multi-eigenvector temporal filtering (m-eigen) uses the linear combination of m eigenvectors corresponding to the largest eigenvalues in the PCA-based temporal filtering approach. The very useful handling of two-stage Wiener filtering (2WF) and SNR-dependent waveform processing (SWP) are first applied to remove the noise and enhance the overall SNR. MFCC parameters are then extracted, followed by the progressive histogram equalization. The multi-eigenvector temporal filtering is finally performed to produce robust feature extraction for speech recognition. Extensive experiments with respect to AURORA2 database and testing conditions verified the effectiveness of each component here and showed that the proposed front-end gives better overall performance when compared to the Advanced Front-End recently announced by ETSI, especially under channel-mismatched conditions.

1. INTRODUCTION

The blueprint for the various applications of the automatic speech recognition (ASR) technologies in the future has been thoroughly laid out and its realization has been highly anticipated by many people [1]. But the recognition accuracy always plays the most dominating role when the realization of real-world applications is considered. It is well known that the recognition accuracy of ASR systems is very often seriously degraded by the mismatch between the acoustic conditions for the training and testing environments and, hence, the robustness for ASR technologies with respect to the varying acoustic environment has always been a key issue in real applications.

One direction towards the above goal is the feature normalization or transformation to alleviate the mismatch. Cepstrum mean subtraction (CMS) and cepstrum normalization (CN) are two widely adopted normalization approaches to produce relatively robust features for this purpose due to their low computational requirements as well as the significant achievable improvements [2, 3]. More recently, an approach for feature transformation based on the concept of histogram equalization (HEQ), which has been widely used in image processing for its capabilities in contrast enhancement, was introduced to speech recognition problems and appears to be superior to the conventional CMS and CN approaches [4].

All transformation approaches mentioned above primarily tried to normalize some statistical characteristics of the feature parameters, i.e., mean, variance, or histogram, over an utterance. In this paper, a progressive histogram equalization (PHEQ) is presented to reshape the time-varying noisy speech feature distribution progressively based on a local reference interval continuously moving with the frame presently being processed. Because the real-world noise is not stationary in general, naturally generating different distributions on the features at different time instants, compensating the features based on the local information instead of the utterance-wise statistics is therefore reasonable.

To further reduce the performance degradation caused by mismatch in acoustic conditions, we further integrate this progressive histogram equalization (PHEQ) with several other novel techniques previously developed to construct a new front-end for ASR systems. The two-stage Wiener filtering is first used to alleviate most of the noise effect [6]; SNR-dependent waveform processing (SWP) is then used to enhance temporally not only the SNR but also the periodicity [7]. The combination of these two approaches creates a more matched environment for the MFCC extraction, which is then followed by the progressive histogram equalization (PHEQ). Finally a data-driven temporal filtering approach based on multi-eigenvectors (m-eigen) [8] is applied to produce the feature parameters with improved robustness for speech recognition. The complete block diagram of the new front-end is shown in Figure 1. Extensive tests with the AURORA2 database verified the robustness of the front-end

The rest of this paper consists of four parts. The various algorithms used in proposed front-end are briefly summarized in section 2. Section 3 gives the experimental conditions, while the experimental results are presented in section 4. Finally, concluding remarks are made in section 5.

2. BRIEF SUMMARIES OF VARIOUS ALGORITHMS USED IN THE NEW FRONT-END

In this section we give brief summaries of the algorithms used in the proposed new front-end for development purposes.

This work is partially supported by the grant number A331XS7X10.

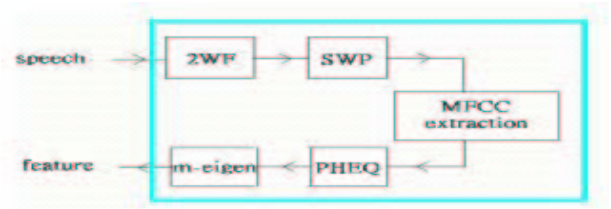


Figure 1: The proposed new front-end.

2.1. Progressive Histogram Equalization (PHEQs)

Histogram equalization (HEQ) [4] has been proved to be able to provide significant improvements in speech recognition under noisy environment, in which the cumulative distribution function (CDF) of the feature parameters, usually evaluated from an utterance, is normalized to a reference distribution assumed for clean speech, usually Gaussian or some distribution obtained in some way. The improvements obtained in this way is often more significant than those simply normalizing the mean or variance of the features [2, 3, 4]. In this paper, different from the conventional utterance-wise normalization approach, we propose to perform the histogram equalization (HEQ) over a reference interval progressively moving with the present frame being considered. This is based on the assumption that very often the additive noise is time-varying, thus equalization with respect to the short interval near the present frame makes better sense than to a whole utterance. This approach is referred to as progressive histogram equalization (PHEQ) here for simplicity.

Let x_t be a feature parameter extracted from a noisy speech frame located at time t, the corresponding reference interval can be defined to be between time indices s and e, where $s \le t \le e$, and we assume $t = (s+e)/2$ here for simplicity. So $[x_s, x_{s+1}, \ldots x_{e-1}, x_e]$ are $n = e - s + 1$ temporally neighboring feature parameters within the reference interval. Histogram equalization (HEQ) can then be performed over this interval rather than over an utterance. It should be pointed out that the length of the reference interval, n, needs to be carefully chosen. If it is too short, the distribution assumed for clean speech may not be applicable. If n is too long, the statistical characteristics of the non-stationary noise may vary substantially within the interval, and the distribution of the feature coefficients can be seriously disturbed. In this study, the reference distribution we chose here is standard Gaussian, which is assumed to be a reasonable model for the clean speech feature distribution.

Note that in this approach we do not need to sort the features within the reference interval each time to obtain the order statistics [5]. When the reference interval moves by one frame, only one feature parameter becomes missing and one new feature is added. So we can simply remove the missing one and insert the new one into the order statistics to obtain the new order statistics to perform the next equalization process. By doing so, the complexity is reduced from $O(n^2)$ to $O(n)$.

There are also advantages of the proposed progressive histogram equalization (PHEQ) over the utterance-wise feature normalization approaches in terms of real-time processing requirements. For the utterance-wise approach, the processing can be performed only when the complete utterance is received. But here with the proposed approach, the processing can be in parallel with the waveform read-in and feature extraction. The delay can in fact be substantially reduced from the length of an utterance to half the reference interval.

2.2. Multi-Eigenvector Temporal Filtering (m-eigen)

The data-driven temporal filtering approaches have been proved to be able to enhance the performance of CN-processed features under various noisy conditions, and a new multi-eigenvector temporal filtering approach was recently proposed and shown to provide very significant improvements in seriously mismatched environments [8, 9]. In this PCA-based approach, the temporal filters are obtained by the linear combination of m eigenvectors corresponding to the m largest eigenvalues, weighted by the eigenvalues, where the eigenvalues and eigenvectors are obtained with the covariance matrix for the segments of the time trajectories of the features obtained with the training corpus. Compared to the conventional PCA temporal filtering which includes only the projection of the time trajectories onto the first principal component, the new multi-eigenvector filter used here tries to include the projections of the time trajectories onto the m major principal components, which also convey important information.

2.3. Two-Stage Wiener Filter (2WF)

The two-stage Wiener filter (2WF) has been found very useful in handling the problem of additive noise [6, 10]. This technique effectively removes the noise on a frame-by-frame basis in two stages, each includes a Wiener filter. The first stage includes a VAD component for noise spectrum estimation and evaluation of the coefficients of the first Wiener filter. After the first stage of the Wiener filtering, the residual noise is assumed to be white and is to be processed by the second stage. In both stages, the frequency domain Wiener filter coefficients are calculated using the estimation of the current noisy speech spectrum and the noise spectrum. In both stages, the linear Wiener filter coefficients are smoothed along the frequency axis by a Mel filter bank, resulting in a Mel-warped frequency domain Wiener filter. The impulse response of this Mel-warped Wiener filter is then obtained by applying a Mel IDCT (Mel-warped Inverse Discrete Cosine Transform). The input waveform signal is thus convolved with this impulse response. More detailed description of this algorithm can be found in the literature [10].

2.4. SNR-Dependent Waveform Processing

The SNR-dependent waveform processing has also been found useful in handling the noise in speech signals [7, 10]. It tries to enhance the high SNR portion and attenuate the low SNR portion of the noisy signal in the time domain. The overall SNR of the noisy speech can thus be increased, and the periodicity of voiced speech can also be enhanced. The estimation of SNR is based on Teager energy. The peaks of temporally smoothed Teager energies are first identified for locating the high and low SNR portions of the input waveform. The waveform segments in high SNR portions are amplified by a factor and the segments that fall in low SNR portions are attenuated by a factor. More detailed description of this algorithm can be found in the literature [10].

Testing set	Set A	Set B	Set C	average
MFCC+HEQ	80.38	81.43	80.81	80.87
MFCC+PHEQ	82.02	82.96	82.87	82.62
Error rate reduction (%)	8.36	8.24	10.73	9.15

1(a)

SNR	20dB	15dB	10dB	5dB	0dB	average
MFCC+HEQ	96.36	93.91	88.24	75.87	50.00	80.87
MFCC+PHEQ	96.40	94.40	89.24	78.07	54.96	82.62
Error rate reduction (%)	1.10	8.05	8.50	9.12	9.92	9.15

1(b)

Table 1: Word accuracy for the proposed PHEQ as compared to HEQ for (a) different testing sets averaged over all conditions and (b) different SNR's averaged over all types of noise.

car	20dB	15dB	10dB	5dB	0dB	average
MFCC+HEQ	97.46	95.65	90.13	77.75	52.58	82.71
MFCC+PHEQ	97.32	95.79	91.23	81.66	56.40	84.48
Error rate reduction (%)	-5.51	3.22	11.14	17.57	8.06	10.24

2(a)

street+ch.	20dB	15dB	10dB	5dB	0dB	average
MFCC+HEQ	97.01	95.01	89.48	77.03	50.33	81.77
MFCC+PHEQ	97.13	95.28	90.66	79.59	57.89	84.11
Error rate reduction (%)	4.01	5.41	11.22	11.14	15.22	12.84

2(b)

Table 2: Word accuracy for the proposed PHEQ as compared to HEQ for typical noise types: (a) car (set A), (b) street + channel effect (set C).

3. EXPERIMENTAL CONDITIONS

The experiments reported in this paper were performed on the English connected-digit string database as well as the various testing environments defined by AURORA2. Ten different types of noises, as representatives of the real-world noise, were considered in three sets of testing conditions, each with different mismatched environments, i.e., set A (subway, babble, car, and exhibition noises), set B (restaurant, street, airport, and train station noises), and set C (subway and street noises, both with channel effect). The MFCC were obtained using the AURORA2 WI007 Front-end, which gives 13 coefficients (C1 ~ C12 + log energy). The transformations were applied to the 13 static parameters only; first and second

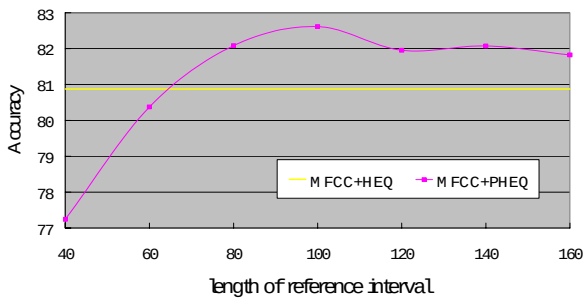


Figure 2: Average word accuracies as a function of the reference interval length over all testing sets and all SNR's for MFCC normalized with PHEQ or HEQ.

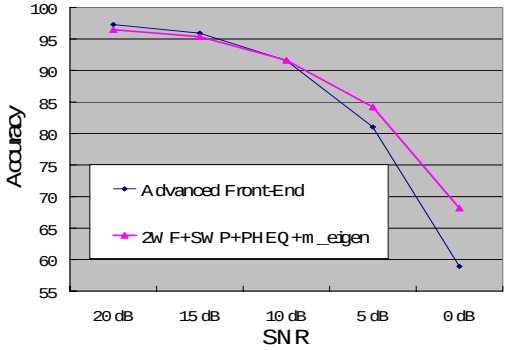


Figure 3: Accuracies for the proposed new front-end as compared to the Advanced Front-End of ETSI for the street noise in set C of different SNR's.

Accuracy	Set A	Set B	Set C	average
pure MFCC	61.34	55.75	66.14	61.08
2WF+MFCC	84.09	82.55	80.66	82.43
2WF+MFCC+PHEQ	82.55	84.34	82.50	83.13
2WF+SWP+MFCC+PHEQ	84.33	85.56	84.43	84.77
2WF+SWP+MFCC+PHEQ+m-eigen	85.11	85.57	86.46	85.71
Advanced Front-End	86.17	85.24	84.72	85.38

Table 3: Word accuracy averaged over all conditions in different testing sets for the integration of various techniques as compared to the Advanced Front-End.

derivatives were then calculated afterward. To evaluate the robustness against mismatched conditions, we adopted the clean speech training scheme only, in which all training data are noise-free speech and testing data are noise-corrupted at different SNR's.

4. EXPERIMENTAL RESULTS

In the following, we first show that the progressive histogram equalization (PHEQ) alone is a better feature normalization technique than HEQ. Experiments were then conducted to verify the individual effectiveness of the several other approaches when integrated with PHEQ. Finally we compare

the results of the integrated new front-end with the Advanced Front-End recently announced by ETSI[10].

41. Progressive Histogram Equalization

We assume that the coefficient to be transformed is located at the center of the reference interval, i.e. t = (s + e)/2 as mentioned in 2.1. The length of the reference interval n is first determined empirically in preliminary tests shown in Figure 2, in which the word accuracy averaged over all conditions and all SNR's are plotted for different length n. As can be found, a good value of the reference interval length n was chosen to be 100 frames. This value of n was used in all the following tests in 4.1.

The test results for HEQ/PHEQ alone applied on MFCC are listed in Table 1. From Table 1(a) and (b), we find that both HEQ and PHEQ give better performance than pure MFCC in AURORA2 tests, the improvements that PHEQ provides over HEQ is consistent for each test set and each SNR condition, and these improvements are more significant for low SNR's. In fact, there is a 9.15% relative error rate reduction when averaged over all types of noise and all SNR's in all testing sets.

To gain more insight than the average accuracies can provide, we list the accuracies for some typical examples of noise types in Table 2, i.e. car (set a, stationary low-pass) and street with channel effect (set c, non-stationary). From Table 2(a) and (b), we can clearly observe the consistent improvements that PHEQ can offer over HEQ.

4.2. The New Front-End by Integrating Various Techniques

The baseline experiments using MFCC only is listed in the first row of Table 3. We first apply alone the two-stage Wiener filter (2WF) as mentioned previously before MFCC to reduce the noise, and then add the proposed PHEQ to the MFCC thus obtained to reduce the residual mismatch. The resulted average accuracy for every testing set is listed in the next two rows of Table 3. We can see from these results that 2WF can effectively remove the noise with relatively stationary characteristics; therefore adding PHEQ actually slightly degrades the average word accuracy of set A. But the function of PHEQ becomes apparent when there is some non-stationary noise (set B) or channel distortion (Set C). We then further applied the SNR-dependent waveform processing (SWP) algorithm after 2WF before MFCC extraction and PHEQ, and the results are listed in the fourth row of Table 3. The effectiveness of SWP here is clearly shown by the consistent improvements. Finally, we added the temporal filtering based on multi-eigenvector (m-eigen) before the speech recognition, and the results are listed in the fifth row of Table 3. It can be found that the improvements are especially evident for the channel-mismatched conditions, i.e. Set C, and this temporal filtering technique in fact provides an additional 6.17% relative error rate reduction on average. This integration of four different techniques (2WF + SWP + PHEQ +m-eigen) is the new front-end proposed here in this paper. This proposed front-end is then compared with the state-of-the-art Advanced Front-End recently announced by ETSI [10, 11], whose results are listed in the last row of Table 3. By comparing the last two rows of Table 3, we can see that the Advanced Front-End gives slightly higher performance for set A, while the new front-end proposed in this paper is slightly better for set B and much more better in set C. The overall accuracy averaged over all cases of the new front-end proposed here is also slightly higher. In fact, the proposed new front-end provides the most significant improvement when the noise is non-stationary and channel condition is mismatched. Typical experimental results for the non-stationary street noise in the channel-mismatched set C are plotted in Figure 3, in which significant improvements can be observed when the environment is very adverse, although at the cost of slight degradation in high SNR cases. Such a trend actually holds true for all types of noise in the Aurora 2 database.

5. CONCLUSIONS

We presented a new front-end which can reduce the effect of mismatch by integrating several novel techniques including progressive histogram equalization, two-stage Wiener filtering, SNR-dependent waveform processing, and the temporal filtering based on the multi-eigenvectors. Extensive evaluation was conducted with respect to the AURORA2 database, and the effectiveness of each individual component as well as the integrated front-end was experimentally verified. Better robustness for channel-mismatched noisy conditions for the new front-end proposed here as compared to the recently announced Advanced Front-End was also demonstrated.

6. REFERENCES

[1] L.-s. Lee and Y. Lee, "Voice Access of Global Information for Broad-band Wireless: Technologies of Today and Challenges of Tomorrow", Proceedings of the IEEE, Jan 2001.

[2] A. E. Rosenberg, C.-H. Lee, and F. K. Soong, "Cepstral Channel Normalization Techniques for HMM-based Speaker Verification", ICSLP, 1992.

[3] O. Viikki and K. Laurila, "Noise Robust HMM-based Speech Recognition Using Segmental Cepstral Feature Vector Normalization", in ESCA NATO Workshop Robust Speech Recognition Unknown Communication Channels, Pont-a-Mousson, France 1997.

[4] Á. de la Torre, J. C. Segura, C. Benítez, A. M. Peinado, and A. J. Rubio, "Non-linear Transformations of the Feature Space for Robust Speech Recognition", ICASSP, 2002.

[5] N. Balakrishnan and C.R. Rao, *Order Statistics: Theory and Methods*, Elsevier Science B. V., The Netherlands, p. 1, 1998.

[6] Agarwal, A., Cheng, and Y.M., "Two-Stage Mel-Warped Wiener Filter for Robust Speech Recognition", *Proc. ASRU'99*, 1999.

[7] Macho, D., Cheng, Y.M., "SNR-dependent Waveform Processing for Robust Speech Recognition", *Proc. ICASSP'01*, 2001.

[8] N.-c. Wang, J.-h. Hung, and L.-s. Lee, "Data-Driven Temporal Filters Based on Multi-Eigenvectors for Robust Features in Speech Recognition", ICASSP, 2003.

[9]I. T. Jolliffe, "Principal Components Analysis", Berlin, Germany: Springer-Verlag, 1986.

[10] "Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms" ETSI ES 202 050 v1.1.3, ETSI standard, 2003.

[11] D. Macho et al. "Evaluation of a Noise-Robust DSR Front-End on Aurora Databases", ICSLP '02.