INVESTIGATING THE USE OF SPEECH FEATURES AND THEIR CORRESPONDING DISTRIBUTION CHARACTERISTICS FOR ROBUST SPEECH RECOGNITION

Shih-Hsiang Lin, Yao-Ming Yeh, Berlin Chen

Department of Computer Science & Information Engineering
National Taiwan Normal University, Taipei, Taiwan
{ shlin, ymyeh, berlin }@csie.ntnu.edu.tw

ABSTRACT

The performance of current automatic speech recognition (ASR) systems often deteriorates radically when the input speech is corrupted by various kinds of noise sources. Quite a few of techniques have been proposed to improve ASR robustness over the last few decades. Related work reported in the literature can be generally divided into two aspects according to whether the orientation of the methods is either from the feature domain or from the corresponding probability distributions. In this paper, we present a polynomial regression approach which has the merit of directly characterizing the relationship between the speech features and their corresponding probability distributions to compensate the noise effects. Two variants of the proposed approach are also extensively investigated as well. All experiments are conducted on the Aurora-2 database and task. Experimental results show that for clean-condition training, our approaches achieve considerable word error rate reductions over the baseline system, and also significantly outperform other conventional methods.

Index Terms **—** speech recognition, robustness, histogram equalization, polynomial regression, clustering

1. INTRODUCTION

Most of the current state-of-the-art ASR systems can achieve quite high recognition performance levels in controlled laboratory environments. However, as the systems are moved out of the laboratory environments and deployed into real-world applications, the performance of the systems often degrade dramatically due to the reason that varying environmental effects will lead to mismatch and uncertainty between the acoustic conditions of the training and test speech data. Therefore, the development of robustness methods has attracted a great deal of attention in recent years. Related work reported in the literature can be generally divided into two aspects according to whether the orientation of the methods is either from the feature domain or from the corresponding probability distributions. Each of them has their own advantages and limitations.

Methods conducted from the feature domain, such as feature compensation [1-3], feature transformation [4-6], or feature reconstruction [7-8], can usually achieve higher performance by having a prior knowledge about the actual distortions caused by various kinds of noises, or by assuming that there exist a fixed (or known) relationship between the clean speech and the corresponding noisy one. Nevertheless, some of them are only effective in tackling the linear distortions but sometimes fails in handling the non-linear distortions. This might be explained by the fact that noise corruptions do not always appear with a one-to-one linear relationship. On the other hand, it is difficult to enumerate all possible noise conditions in real-world scenarios, so their effectiveness is restricted. Another research line is relied on exploring noise-resistant distribution characteristics of speech features. Representative methods, include, but not limited to, CMS [9], CMVN [10], HOCMN [11], HEQ [12-15], etc. Although these methods have already been demonstrated their capability in preventing performance degradation of speech recognition systems under various noisy environments and also have been proven their effectiveness in compensating the environmental mismatch between the training and test speech data, they to some extent have their inherent limitation. Most of the approaches still have room for improvement. For example, some of them must assume the speech features follow a predefined distribution (e.g., Gaussian), but such an assumption is not entirely correct. The other problem is that noises will not only modify the distributions of the speech features but also inject uncertainties into the speech features due to the random behavior of them. However, most of these methods can only deal with the mismatch between the training and test conditions but few with such uncertainties.

Based on the above observations, we believe that these two research orientations could complement each other, and it might be possible to inherit the individual merits from them to overcome their inherent limitations. As an example, in our previous work, we proposed a cluster-based polynomial-fit histogram equalization (CPHEQ) approach [16], which makes use of both the speech features and their corresponding distribution characteristics for speech feature compensation. CPHEQ inherits the merits of above two orientations and uses the data fitting technique in a purely data-driven manner to approximate the actual distributions without the need of unrealistic assumptions about the speech feature distributions. Experimental results have shown that CPHEQ could achieve a considerable word error

rate reduction over the baseline MFCC-based system. In this paper we attempt to elucidate the theoretical foundation of CPHEQ based on a more rigorous mathematical treatment. Moreover, two variants derived from CPHEQ by making additional assumptions and extensions are presented as well, namely, the polynomial-fit histogram equalization (PHEQ) [15] and the selective cluster-based polynomial-fit histogram equalization (SCPHEQ).

The remainder of this paper is organized as follows. Section 2 first elucidates the theoretical foundation of CPHEQ. Moreover, the two extensions of CPHEQ, i.e., PHEQ and SCPHEQ, are described in detail in Section 3. Then, Section 4 presents the experimental settings, as well as the experimental results and discussions. Finally, conclusions and future work are drawn in Section 5.

2. THEORETICAL FOUNDATION OF CPHEQ

The basic idea of CPHEQ is inspired from two diverse approaches. The first one is SPLICE (Stereo-based Piecewise Linear Compensation for Environments) [1], which attempts to use a Gaussian mixture model (GMM) to model the noisy feature space, and each Gaussian component represents one specific distortion condition, or it can be treated as the condition that a certain phoneme class interfered with a particular kind of noise. In addition, each Gaussian component has one corresponding correction vector (or bias) for compensating the noisy speech. The estimation of the correction vector is generally done by utilizing the MMSE (Minimum Mean Squares Error) criterion with a set of stereo data. However, the main drawback of SPLICE is that it simply uses a set of linear additive biases to approximate the true nonlinear relationship between the clean and noisy speech for each distortion condition. In order to overcome this shortcoming, we additionally take the idea from HEQ [12-14]. HEQ uses non-linear transformation functions to compensate the nonlinear distortions by utilizing the relationship between the cumulative distribution functions (CDFs) of the test speech and those of the corresponding training (or reference) one. The critical success factor behind HEQ is strongly relied on the assumption that the distribution of test data should be identical to the one of the training data. Nevertheless, this assumption is sometimes invalid due to the fact that the data distribution will be affected by different noise corruptions and phonetic characteristics, especially when the test utterance becomes much shorter. For this reason, the use of only a single global transformation (or inverse) function seems inadequate to compensate the noisy feature vector component. Therefore, we purpose the use of CPHEQ to combine the merits of both SPLICE and HEQ, as well as to overcome their shortcomings.

For CPHEQ, we first use the noisy speech data to train a GMM whose parameters are estimated by the K-means algorithm followed by the expectation maximization (EM) algorithm. The GMM is expressed as follows:

$$p(Y_t)=\sum_{k=1}^{K}p(k)p(Y_t\mid k)=\sum_{k=1}^{K}p(k)N(Y_t;\mu_k,\Sigma_k), \quad (1)$$

where K is the total mixture number used in GMM; $p(Y_t\mid k)$ and $p(k)$ are the likelihood of feature vector Y_t being generated by the k-th mixture and the corresponding weight of the k-th mixture, respectively; and each Gaussian is associated with a mean vector μ_k and a diagonal covariance matrix Σ_k. Furthermore, we assume the compensated feature vector can be derived by:

$$\mathbf{E}[X_t\mid Y_t]=\mathbf{E}[\mathbf{E}[X_t\mid Y_t,k]]=\sum_{k=1}^{K}p(k\mid Y_t)\mathbf{E}[X_t\mid Y_t,k], \quad (2)$$

where $\mathbf{E}[\bullet]$ is the expectation operation and $p(k\mid Y_t)$ is the posterior probability given by:

$$p(k\mid Y_t)=\frac{p(Y_t\mid k)p(k)}{\sum_{k'=1}^{K}p(Y_t\mid k')p(k')}. \quad (3)$$

Then, we further assume that the feature vector component y_t of Y_t is independent of each other given the k-th mixture, and the restored value of y_t can be obtained by utilizing a polynomial regression model and taking its corresponding CDF value as the explanatory variable [13]. Therefore, the restore value of y_t for the k-th mixture is defined as:

$$\tilde{x}_{t,k}=\mathbf{E}[x_t\mid y_t,k]=G_k(CDF(y_t))=\sum_{m=0}^{M}a_{km}(CDF(y_t))^m, \quad (4)$$

where $CDF(y_t)$ is the CDF value of the feature vector component y_t, which can be obtained by using the cumulative histogram [12] or order statistics [13]; and $G_k(\bullet)$ is the transformation function which maps CDF values onto their corresponding predefined feature values for the k-th mixture. In contrast to the conventional time-consuming table-lookup based HEQ [12], we use the polynomial functions to approximate the inverse function of CDF. The reason why we choose the polynomial function here is mainly because that it has a simple form, without the need of a complicated computation procedure, and it has moderate flexibility in controlling the shape of the function. Though the polynomial function is efficient to delineate the transformation function, it is worth mentioning that the polynomial function, to some extent, has its inherent limitations. For example, high order polynomial functions might lead to over-fitting of the training data. Moreover, the polynomial function would provide good fits for the input data points that are located within the range of values of the training data, but would also probably have rapid deteriorations when the input data points are located outside

the range of values of the training data when the order becomes much higher.

During the training phase, the coefficients a_{km} of the polynomial function $G_k(\bullet)$ for the k -th mixture can be estimated by using a set of stereo data and by minimizing the squares error defined by:

$$E_k^2 = \sum_{t=0}^{T-1}\left(p(k \mid Y_t)\times\left(x_t - \sum_{m=0}^{M} a_{km}(CDF(y_t))^m \right)\right)^2, \qquad (5)$$

where T is the total number of frames in training data; y_t and x_t are respectively the feature vector component for noisy speech and its corresponding clean one. During the speech recognition process, each feature vector component of test speech y_t is first used to estimate its corresponding CDF value, and then the restored value $\tilde{x}_t$ of y_t hence can be expressed by:

$$\tilde{x}_t = \sum_{k=1}^{K}\left(p(k \mid Y_t)\times\left(\sum_{m=0}^{M} a_{km}(CDF(y_t))^m \right)\right). \qquad (6)$$

As can be seen from Eq. (6), the restored value $\tilde{x}_t$ is obtained by a weighted sum of the output of each transformation function $G_k(\bullet)$. Obviously, the computation time will increase when the number of mixtures becomes larger. In order to reduce the computation time, we therefore use the maximum a posteriori probability (MAP) criterion and redefine Eq. (5) and (6) as follows:

$$E_k^2 = \sum_{t=0}^{T-1}\left(\delta(k \mid Y_t)\times\left(x_t - \sum_{m=0}^{M} a_{km}(C_{Train}(y_t))^m \right)\right)^2, \qquad (7)$$

$$\tilde{x}_t = \sum_{k=1}^{K}\left(\delta(k \mid Y_t)\times\left(\sum_{m=0}^{M} a_{km}(CDF(y_t))^m \right)\right), \qquad (8)$$

where

$$\delta(k \mid Y_t) = \begin{cases} 1 & \text{if } k = \arg\max_{k'} p(k' \mid Y_i) \\ 0 & \text{otherwise} \end{cases}. \qquad (9)$$

That is, each feature vector Y_t will be assigned to a specific mixture k . The main difference between Eq. (5) and (7) is the amount of data pairs being considered for obtaining the polynomial functions. The estimation using Eq. (5) can be viewed as a soft-decision approach, where the error contributed by each data pair is weighted by the corresponding posterior probability of the mixture it probably belongs to. On the other hand, a hard-decision approach is used in Eq. (7), where each frame is exactly associated with one mixture. During the recognition phase, each feature vector Y_t is first assigned to a specific mixture k by using Eq. (9) and then each of its feature vector component y_t is replaced by a restored value $\tilde{x}_t$ using Eq. (8).

3. TWO VARIANTS OF CPHEQ

In the previous section, we have described the theoretical details of CPHEQ. In this paper, we also present two different extensions of CPHEQ. The first extension, named the polynomial histogram equalization (PHEQ) [15], assumes that only a global transformation function was used for restoring the noisy component values and the transformation function is instead delineated from the clean speech data without the use of the noisy speech data. This extension in fact is in analogy with the conventional HEQ approaches. The other extension, named the selective cluster-based polynomial-fit histogram equalization (SCPHEQ), combines the missing feature theory (MST) [7-8] and the prediction capabilities of polynomial functions to reconstruct the unreliable feature vector components.

3.1. Polynomial-Fit Histogram Equalization (PHEQ)

For PHEQ, we simply assume that only a single global transformation function is utilized to obtain the restored value $\tilde{x}_t$ of the noisy feature vector component y_t , and therefore Eq. (6) can be rewritten as:

$$\tilde{x}_t = \sum_{m=0}^{M} a_m(CDF(y_t))^m, \qquad (10)$$

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

$$E^2 = \sum_{t=0}^{T-1}\left(x_t - \sum_{m=0}^{M} a_m(CDF(x_t))^m \right)^2. \qquad (11)$$

Notice that only the clean speech data x_t is used for estimating the coefficients. During speech recognition, for each feature vector dimension, the CDF value of each feature vector component is first estimated and then taken as an input to the polynomial function to obtain its restored value. The advantage of PHEQ is it can 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 based HEQ (THEQ) [12] or quantile-based HEQ (QHEQ) approaches [14].

3.2. Selective Cluster-based Polynomial-Fit Histogram Equalization (SCPHEQ)

Generally, there are two essential steps in using the missing feature theory to compensate the corrupted spectral vector. The first one is to identify which spectral vector component is unreliable or missing, and the second one is to either reconstruct the unreliable components for recognition [8] or ignore them during the decoding process [7]. In this paper, we only focus on using the same concept of CPHEQ to reconstruct the unreliable spectral vector components given that the oracle data mask of reliable or unreliable components is known in advance. We simply assume that the unreliable components can be reconstructed by Eq. (6)

or (8). It is worth noting that the feature vectors used here for SCPHEQ is represented in the spectral domain, in comparison to that of CPHEQ in the cepstral domain. The discrepancy is because the noise that corrupts the speech may only occur in some particular frequency bands (or spectral vector components), so the noise corruptions can be easily identified in the spectral domain, but much more difficult in the cepstral domain since each cepstral feature vector component will encompass the information from all spectral bands. On the other hand, because of the range of spectral values varying dramatically, the use of cumulative histograms or order statistics to estimate the corresponding CDF value is no longer applicable. Therefore, we use the Gaussian error function to estimate the CDF values of each feature vector. If the distribution of a univariate random variable x belongs to a Gaussian, i.e. $x \sim N(\mu, \sigma^2)$, then its CDF is thus defined by [17]:

$$CDF(x) = \Phi(x) = \frac{1}{2}\left(1 + erf\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right), \quad (12)$$

where μ and σ are the sample mean and standard deviation of x, respectively; and the Gaussian error function $erf(v)$ is defined as

$$erf(v) = \frac{2}{\sqrt{\pi}}\int_0^v e^{-t^2}\,dt. \quad (13)$$

Nevertheless, as mentioned earlier, it would be inappropriate by simply assuming the feature vector component follows a Gaussian distortion. Thus, GMM is instead used to approximate the true distribution of the feature vector component. Consequently, $CDF(x_t)$ in Eq. (12) can be further expressed by using GMM:

$$CDF(x_t) = \sum_{j=1}^{J} c_j \times \Phi_j(x_t) = \frac{1}{2}\sum_{j=1}^{J} c_j\left(1 + erf\left(\frac{x_t-\mu_j}{\sigma_j\sqrt{2}}\right)\right), \quad (14)$$

where J is the total number of Gaussian distributions used in GMM; μ_j and σ_j are the mean and standard deviation of the j-th Gaussian distribution, respectively; and c_j is the mixture weight. Therefore, there are two GMMs being used in SCPHEQ: one for modeling the noisy feature space and the other for estimating the CDF value. The training procedure of SCPHEQ is almost as the same as that of CPHEQ except for the way to estimate the CDF value $CDF(x_t)$. The main difference between SCPHEQ and CPHEQ lies in their applications in the recognition phase. Due to the additive property in the spectral domain, the noisy spectral value must be greater than the clean one. Thus, for SCPHEQ, we can not directly take the restored value of the unreliable component from the output of the polynomial functions, since the restored value might be abnormal. Therefore, a bounded function is applied to ensure that restored value must be no greater than the corresponding noisy one:

$$\hat{x}_t = \begin{cases} \min\{y_t, \tilde{x}_t\} & \text{, for unreliable components} \\ y_t & \text{, for reliable components} \end{cases} \quad (15)$$

4. EXPERIMENTAL SETUP AND RESULTS

4.1 Experimental Setup

The speech recognition experiments were conducted under various noise conditions using the Aurora-2 database and task [18]. 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 the Test 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 (-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB and Clean) and the Test Set C additionally includes the channel distortion. For the baseline system, the training and recognition tests used the HTK recognition toolkit [19], which followed the setup originally defined for the ETSI evaluations [18]. All the experimental results reported below are based on clean-condition training, i.e., the acoustic models were trained only with the clean training utterances.

4.2 Experimental Results

The average WER result obtained by the MFCC-based baseline system is 41.04%, which is an average of the WER results of the test utterances respectively contaminated with eight types of noises under different SNR levels (0 dB to 20 dB) for the three test sets (Sets A, B and C). We first evaluate the performance of CPHEQ when different criteria were used to obtain the polynomial transformation functions. The number of GMM mixtures is set from 32 to 1,024, and the order of the polynomial is initially set to 3. The associated results are shown in Table 1. It can be found that CPHEQ provides significant performance boosts over the MFCC-baseline system, especially when the number of mixtures becomes much larger (e.g., 512 or 1,024), and there is no significant difference between soft-decision and hard-decision approaches. This might be due to the fact that when the transformation functions are estimated using Eq. (5), the error contributions are prone to be dominated by the mixture with the highest posterior probability for each training speech feature vector component, which would make the estimation of the transformation functions using Eq. (5) have the same effect as that using Eq. (7). Accordingly, this may also suggest that using Eq. (7) to derive the polynomial functions for CPHEQ is enough and it can also simplify the computation of CPHEQ.

In the next set of experiments, we evaluated the performance of CPHEQ with respect to different number of mixtures and different orders of the polynomial function. The corresponding WER results are illustrated in Figure 1.

	Number of Mixtures					
	32	64	128	256	512	1024
Hard	19.84	19.49	18.24	17.33	16.36	15.41
Soft	19.88	19.46	18.23	17.31	16.33	15.40

Table 1: Comparison of the average WER results (%) between hard/soft decision approaches used for deriving the polynomial transformation functions of CPHEQ

	Number of Polynomial Orders					
	1-th	3-th	5-th	7-th	9-th	11-th
PHEQ	23.25	21.80	21.46	21.13	21.16	22.14

Table 2: Average WER results (%) of PHEQ with respect to different orders of the polynomial transformation functions.

As can be seen from Figure 1, the WER is slightly improved when the order of the polynomial regression becomes higher, but in the case of large number (e.g., 512 or 1,024) of mixtures, the performance seems to degrade substantially if the order of the polynomial functions becomes too large. These results may be explained by the facts that the limited training data was used in this study (the fact of the curse of dimensionality), and the use of higher order polynomial functions might led to oscillations between the exact-fit values. As we further compare the best result obtained from Table 1 with the result of the MFCC-based baseline system, it can be found that CPHEQ can provide a relative WER reduction of about 62% over the MFCC-based baseline system.

In the third set of experiments, we evaluate the performance of PHEQ with respect to different polynomial orders and the associated results are presented in Table 2. Due to the end behavior property of polynomial functions, the even order polynomials are either "up" on both ends or "down" on both ends which is not appropriate to characterize the behavior of a cumulative distribution. Therefore, only odd-order polynomials are utilized in this study for PHEQ. As evidenced by the results shown in Table 2, the average WER results of PHEQ are slightly improved when the order of the polynomial functions become higher. However, as the order increases, the polynomial functions might sometimes tend to over-fit the training data and further degrade the performance. As it is indicated, PHEQ yields about a relative WER reduction of about 48.51% as compared to the MFCC-based baseline system. The performance of PHEQ does not outperform the CPHEQ, but it only requires clean speech data to estimate the polynomial coefficients without the needs of stereo data.

Then, we evaluate the performance of SCPHEQ. Since the objective of this paper is to deal with the unreliable feature vector component, we preliminarily used the oracle mask to identify whether a spectral vector component belong to reliable or unreliable component for our experiments. Moreover, we found that four mixtures are sufficient to estimate the CDF value expressed in Eq. (14), and this setting was thus used for the following experiments

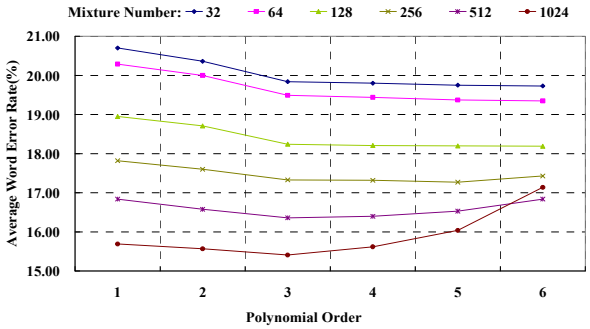

Figure 1: Average WER results (%) of CPHEQ with respect to different number of mixtures and different orders of the polynomial transformation functions.

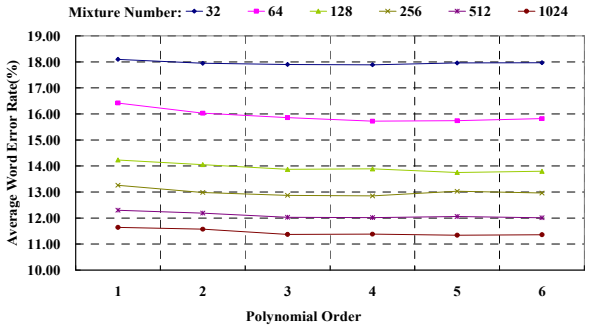

Figure 2: Average WER results (%) of SCPHEQ with respect to different numbers of the mixtures and different orders of the polynomial transformation functions.

of SCPHEQ. Figure 2 shows the average WER results with respect to different number of mixtures for modeling the noisy speech characteristics and different orders of the polynomial functions. As it is illustrated, increasing the mixtures can have a steady improvement in WER results, but the performance seems not to degrade even we used a large number of mixtures together with a higher order of the polynomial functions. This may be probably explained by the reason that a bounded function is applied for ensuring the reconstructed spectral feature vector components must be lower than noisy ones. Even though the value of reconstructed component is abnormal, the bounded function can alleviate, to some extent, the effect caused by it. The best result achieved with a relative WER reduction of about 72% over the MFCC-based baseline system.

To go a step further, SCPHEQ can be thought as a spectral subtraction operation that removes noise effects in the spectral domain, but sometimes the residual noise may still present in the resulting cepstral feature vectors. It may be possible to use HEQ approaches to reduce the residual noise in the cepstral domain. Therefore, we evaluate the ASR performance when combining SCPHEQ with PHEQ. The experiment result shows that the combination of SCPHEQ with PHEQ can further provide a very significant improvement as compared to the results obtained by using

each of them individually. It yields an average WER result of about 4.30%.

Finally, we compare our proposed approaches with the other conventional approaches. Table 3 shows the average WER results obtained by various conventional approaches and our proposed approaches. The number of mixtures used in SPLICE was set to 1,024, which was the same as that used in CPHEQ and SCPHEQ. As can be seen from Table 3, our proposed approaches are considerably better or competitive to all the other conventional approaches.

Method	Test A	Test B	Test C	Average
Baseline(MFCC)	41.06	41.52	40.03	41.04
CMS	32.40	27.16	34.15	30.65
CMVN	22.73	19.60	31.70	23.27
THEQ	18.37	16.92	19.51	18.02
QHEQ	23.08	22.03	24.08	22.86
PHEQ	20.92	18.12	25.68	21.73
SPLICE	17.03	17.13	26.90	19.04
CPHEQ	14.35	14.04	20.28	15.41
SCPHEQ	10.21	15.09	6.30	11.38
SCPHEQ+PHEQ	4.29	4.11	4.70	4.30

Table 3: Average WER results (%) obtained by the MFCC-based baseline system and various approaches.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we have investigated the use of speech features and their corresponding distribution characteristics for robust speech recognition. Three variants of polynomial regression approaches were proposed to increase the robustness of the ASR system. Their performance has been extensively tested and verified by comparison with the other conventional approaches. Very encouraging results on the Aurora-2 database have been obtained. Finally, we list below some possible future extensions of the proposed polynomial regression approaches: 1) The stereo data sometimes is difficult to collected, and therefore a possible future work is using mono data (either clean speech or noisy speech) alone to estimate the parameters of the transformation functions. 2) The data-fitting technique is prone to be affected by abnormal values. Therefore, another possible future work is outlier detection/elimination, or the so-called robust regression. 3) Speech signal is slowly time-varying, so the contextual information between consecutive speech feature vectors might be an important cue that can help in improving the ASR robustness.

6. ACKNOWLEDGEMENTS

This work was supported in part by the National Science Council, Taiwan, under Grants: NSC96-2628-E-003-015-MY3 and NSC95-2221-E-003-014-MY3.

7. REFERENCES

[1] L. Deng, et al., “Large Vocabulary Speech Recognition under Adverse Acoustic Environments," in *Proc. ICSLP 2000*.

[2] J. Wu, et al., “An Environment-Compensated Minimum Classification Error Training Approach Based on Stochastic Vector Mapping,” *IEEE Trans. on Audio, Speech and Language Processing*, 14(6), 2006.

[3] L. Buera, et al., “Cepstral Vector Normalization Based on Stereo Data for Robust Speech Recognition,” *IEEE Trans. on Audio, Speech and Language Processing* 15(3), 2007.

[4] R. O. Duda, et al., *Pattern Classification and Scene Analysis*, New York, John Wiley and Sons, 1973.

[5] G. Saon, et al., “Maximum Likelihood Discriminant Feature Spaces,” in *Proc. ICASSP 2000*.

[6] M. J. F. Gales, “Maximum Likelihood Multiple Subspace Projections for Hidden Markov Models,” *IEEE Trans .on Speech and Audio Processing* 10(2), 2002.

[7] M. P. Cooke, et al., ”Robust Automatic Speech Recognition with Missing and Uncertain Acoustic Data,” *Speech Communication* 34, 2001

[8] B. Raj, et al., “Missing-feature Approaches in Speech Recognition,” *Signal Processing Magazine* 22(5), 2005.

[9] B. S. Atal, “Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification,” *Journal of the Acoustical Society of America* 55, 1974.

[10] O. Vikki, et al., “Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition,” *Speech Communication* 25, 1998.

[11] C. W. Hsu, et al., “Extension and Further Analysis of Higher Order Cepstral Moment Normalization (HOCMN) for Robust Features in Speech Recognition,” in *Proc. ICSLP 2006*.

[12] S. Dharanipragada, et al., “A Nonlinear Unsupervised Adaptation Technique for Speech Recognition,” in *Proc. ICSLP 2000*.

[13] A. Torre, et al., “Histogram Equalization of Speech Representation for Robust Speech Recognition,” *IEEE Trans. on Speech and Audio Processing* 13(3), 2005.

[14] F. Hilger, et al., “Quantile Based Histogram Equalization for Noise Robust Large Vocabulary Speech Recognition,” *IEEE Trans. on Audio, Speech and Language Processing* 14(3), 2006.

[15] S.H. Lin, et al., “Exploiting Polynomial-Fit Histogram Equalization and Temporal Average for Robust Speech Recognition,” in *Proc. ICSLP 2006*.

[16] S.H. Lin, et al., “Cluster-based Polynomial-Fit Histogram Equalization (CPHEQ) for Robust Speech Recognition,” in *Proc. Eurospeech 2007*.

[17] M. Abramowitz, et al, “Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables,” New York: Dover, 1972.

[18] H. G. Hirsch, et al., “The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions,” in *Proc. ICSLP 2000*.

[19] S. Young, et al., “The HTK Book (for HTK Version 3.3),” Cambridge University Engineering Department, Cambridge, UK, 2005.