ENTROPY-BASED FEATURE PARAMETER WEIGHTING FOR ROBUST SPEECH RECOGNITION

Yi Chen, Chia-yu Wan, Lin-shan Lee

Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan, R. O. C.
chenyi@speech.ee.ntu.edu.tw, chiayui@speech.ee.ntu.edu.tw, lslee@gate.sinica.edu.tw

ABSTRACT

In this work, we propose an entropy-based measure to determine the discriminating ability of a feature parameter in identifying the correct acoustic models, and a feature parameter weighting scheme using this measure during Viterbi decoding. The purpose is to emphasize the scores obtained with more discriminating parameters, and to de-emphasize the scores with less discriminating parameters. Extensive experiments verified that this approach is equally useful for different types of features, and can be easily integrated with typical existing robust speech recognition approaches.

1. INTRODUCTION

Various applications of the automatic speech recognition (ASR) technologies in the future have been highly anticipated by many people [1]. But the recognition accuracy always plays the most dominating role when the real-world applications are considered. Many speech features have been proposed to extract important cues that are helpful for speech recognition in different ways. The Perceptual Linear Predictive (PLP) features [2] differ from the widely used MFCC features in the use of Bark-spaced critical bands followed by cepstral coefficient computation with the autoregressive modeling of the critical band power spectrum. Minimum Variance Distortionless Response (MVDR) spectrum estimation was also proposed [3, 4] for spectral envelope estimation, which can be used to obtain different sets of features including the possibility of warping the frequency before spectrum estimation [1, 5, 6, 7]. It has been well known that some feature parameters are more useful or more important in identifying or distinguishing different acoustic models (e.g. the first several MFCC parameters than others). But in most speech recognition systems such differences in speech features have not been well explored yet. It is natural to believe that when we treat all the feature parameters as equally important coefficients as has been done in conventional Viterbi decoding, the functions of the more discriminating parameters may be smeared out by the functions of other parameters.

In this paper, we propose an entropy-based measure to identify the more discriminating feature parameters, and a feature parameter weighting scheme to emphasize the acoustic scores obtained with these more discriminating feature parameters during recognition. The approach is relatively simple and can be directly applied in the conventional Viterbi decoding process. Experimental results on the AURORA 2 testing environment verified that this approach can be equally useful for different types of features including MFCC, PLP and MVDR, and can be easily integrated with typical existing robust speech recognition approaches to offer even better performance. Such results are consistent across a wide range of noise types and SNR conditions.

This paper is organized as follows. The proposed feature parameter weighting approach is described in section 2. In section 3 the experimental results are presented. Section 4 gives our conclusions.

2. PROPOSED APPROACH

The proposed entropy-based feature parameter weighting scheme is as follows. We construct the feature parameter weighting function using a training corpus. We first perform forced alignment of each utterance in the training corpus with the transcriptions. Then we collect the feature vectors of the same class (or the same acoustic model) together to train a Gaussian mixture model (GMM) with N Gaussian components for each class,

$$G_{\mathrm{c}}(\boldsymbol{x}) = \sum_{\mathrm{n}=1}^{\mathrm{N}} k_{\mathrm{c,\,n}} N_{\mathrm{c,\,n}}(\boldsymbol{x} \mid \theta_{\mathrm{c,\,n}}), \qquad (1)$$

where c is the class or model index (c = 1, 2, ⋯, C), n is the mixture index of the GMM model (n = 1, 2, ⋯, N), $k_{\mathrm{c,\,n}}$ is the weight for the n-th mixture component $N_{\mathrm{c,\,n}}(\mathbf{x} \mid \theta_{\mathrm{c,\,n}})$ for class c, and $\theta_{\mathrm{c,\,n}}$ is the set of parameters (mean and covariance) of $N_{\mathrm{c,\,n}}(\cdot)$.

Now for each testing feature vector $\boldsymbol{x}$(t) at time t, the probability density of the d-th feature parameter of $\boldsymbol{x}$(t), x_{d}(t), on the class or acoustic model c can be defined as:

$$p_{\mathrm{c}}(\mathrm{t,\ d}) = \underset{\substack{\mathrm{d'} = 1, \ldots, \mathrm{D} \\ \mathrm{d'} \neq \mathrm{d}}}{\int \ldots \int} G_{\mathrm{c}}(\boldsymbol{x})\ \mathrm{d}x_1 \ldots \mathrm{d}x_{\mathrm{d'}} \ldots \mathrm{d}x_{\mathrm{D}}\ \Big|_{x_{\mathrm{d}}(\mathrm{t})}, \qquad (2)$$

where d is the index for different feature parameters in $\boldsymbol{x}$(t) (d = 1, 2, ⋯, D), and D is the total number of feature parameters in $\boldsymbol{x}$(t). When all the feature parameters in $\boldsymbol{x}$(t) are assumed independent of each other, the Gaussian mixture model (GMM) $G_{\mathrm{c}}(\boldsymbol{x})$ in equation (2) can be simplified into D independent scalar GMMs for the D feature parameters, and p_{c}(t,d) of x_{d}(t) in equation (2) can be reduced to:

$$p_{\mathrm{c}}(\mathrm{t,\ d}) = \sum_{\mathrm{n}=1}^{\mathrm{N}} k_{\mathrm{c,\,n,\,d}} N_{\mathrm{c,\,n,\,d}}(x_{\mathrm{d}}(\mathrm{t}) \mid \theta_{\mathrm{c,\,n,\,d}}), \qquad (3)$$

where $k_{\mathrm{c,\,n,\,d}}$, $N_{\mathrm{c,\,n,\,d}}(\cdot)$ and $\theta_{\mathrm{c,\,n,\,d}}$ are the same as $k_{\mathrm{c,\,n}}$, $N_{\mathrm{c,\,n}}(\cdot)$ and $\theta_{\mathrm{c,\,n}}$ in equation (1), but reduced to those for a simple feature parameter with index d, and x_{d}(t) is the feature parameter in $\boldsymbol{x}$(t) with index d.

To define the entropy measure for the feature parameter x_{d}(t), we convert p_{c}(t, d) into a probability mass function (PMF)-like function $\overline{p}_{\mathrm{c}}(\mathrm{t,\ d})$ of class c as

$$\overline{p}_{\mathrm{c}}(\mathrm{t,\ d}) = p_{\mathrm{c}}(\mathrm{t,\ d}) \,/\, \sum_{\mathrm{c}=1}^{\mathrm{C}} p_{\mathrm{c}}(\mathrm{t,\ d}), \qquad (4)$$

1-4244-0469-X/06/$20.00 ©2006 IEEE

and the entropy for $x_d(t)$ is then defined as

$$H(\mathrm{t}, \mathrm{d}) = -\sum_{c=1}^{C} \bar{p}_c(\mathrm{t}, \mathrm{d})*\log \bar{p}_c(\mathrm{t}, \mathrm{d}). \qquad (5)$$

If the distribution of p_c(t,d) in equation (3) across all classes c looks like the one in Fig. 1(a), i.e., the scores for different classes are very similar, the above entropy H(t,d) will be high, which means the discriminating ability of the feature parameter with index d is low. A typical example for such a case is that with d = 1 in Fig. 2. In other words, even if one of the classes has the highest score, the other competing classes have very similar scores, and therefore this feature parameter is not very reliable. On the other hand, if the distribution of p_c(t,d) in equation (3) across all classes c looks like the one in Fig. 1(b), i.e., one of the classes has much higher score than all the others, the entropy H(t,d) will be low, which means the discriminating ability of this feature parameter is high. A typical example for such a case is that with d = 2 in Fig. 2. In other words, the distribution of the correct class is well separated from those of all other classes. Apparently the recognition should rely more on the latter than on the former.

With the above, the feature parameter weighting function W(t,d) for $x_d(t)$ is defined as

$$W(\mathrm{t}, \mathrm{d}) = f(H(\mathrm{t}, \mathrm{d})) = \exp(-a\, H(\mathrm{t}, \mathrm{d})\,), \qquad (6)$$

which is a function of both time index t and parameter index d, where a is an empirically determined scaling factor. Note that the function $f(\bullet)$ in equation (6) can be a carefully chosen monotonically decreasing function, while here the exponential function is used for simplicity.

During the Viterbi decoding process, the log-likelihood of the feature vector $\boldsymbol{x}$(t) in the j-th state of an HMM is calculated as, assuming a diagonal covariance matrix,

$$\log[b_j(\boldsymbol{x}(\mathrm{t}))] = \sum_{d=1}^{D} W(\mathrm{t},\mathrm{d})\ (\ \log[\sum_{m=1}^{M} c_{jm} N(x_d(\mathrm{t});\, \mu_{jmd}, \Sigma_{jmd})]\ ), \quad (7)$$

where j and m are respectively the state and mixture indices of the HMM model, c_{jm} is the mixture weight, and μ_{jmd} and Σ_{jmd} are the mean and variance for the parameter x_d(t) in the m-th mixture of the j-th state. So the more discriminating feature parameters will be emphasized, and vice versa.

3. EXPERIMENTAL RESULTS

The initial experiments reported in this paper were conducted on the AURORA 2 testing environment [8] based on a corpus of English connected digit strings. Only clean-condition training was used, and there are ten different types of noise in the three test sets: sets A, B, and C. In the first set of experiments, three sets of speech features were tested, i.e., MFCCs, MVDR-based cepstral coefficients, and PLP coefficients. In the second set of experiments, we integrated the proposed approach with a set of typical robust speech recognition approaches including a four-stage front-end and a frame selection scheme proposed recently [9]. In these initial experiments, we set N to 1 and D to 39. C is 13 because there are 13 classes (word models) defined for the AURORA 2 task [8].

3.1. Results with Different Features: MFCC, MVDR, and PLP

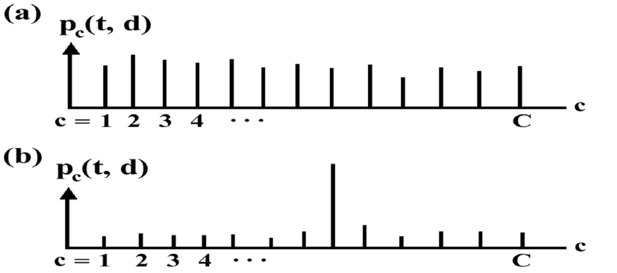

Fig. 1. Distributions of p_c(t,d) over the C classes give different entropy values: (a) high entropy H(t,d), and (b) low entropy H(t,d).

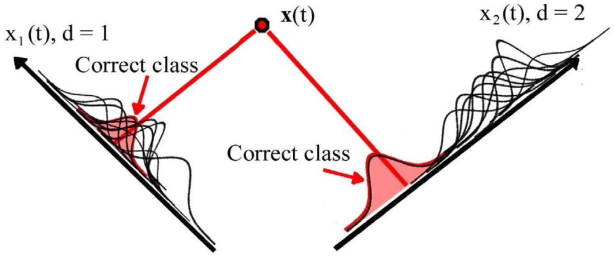

Fig. 2. The situation of $\boldsymbol{x}$(t) that the entropy H(t,d) is high for d = 1, but low for d = 2.

The 13 MFCC parameters (C1~C12 and log-E) were obtained with the WI007 front-end [8]. The MVDR-based features were obtained using the frequency-warped MVDR algorithm with the warping factor λ set to 0.1 for spectrum estimation to replace the conventional FFT. A well-designed filterbank is also used, while everything else is exactly the same as for obtaining MFCC including pre-emphasis filtering, windowing and DCT conversion [7]. For PLP coefficients, the feature vector consists of 12 PLP coefficients and a log-energy term. In all the three cases, 39-dimensional feature vectors including the delta and delta-delta components were used to train the HMM models. During recognition, the log-likelihood scores from different feature parameters were weighted and summed using the proposed method, as shown in equation (7).

The results of applying the proposed feature parameter weighting approach on the MFCC features as compared to the baseline tests without weighting are shown in Fig. 3(a), (b), and (c), respectively for the results averaged over all SNR values (20~0 dB) but separated for different types of noise (Fig. 3(a)), those averaged over all different types of noise but separated for different SNR values (Fig. 3(b)), and those averaged over all different types of noise and SNR values but separated for the three test sets A, B, and C (Fig. 3(c)). The exact accuracies for the three test sets in Fig. 3(c) are also listed in the left part of Table 1. It can be observed that the proposed feature parameter weighting approach consistently offered improved performance across all testing conditions. In Fig. 3(a), the error rate reduction was significant for all types of noise, with good examples including 20.73% of error reduction for street noise in test set B (accuracy from 61.52% to 69.49%), and 19.67% reduction for car noise in test set A (accuracy from 60.60% to 68.35%). In Fig. 3(b), the recognition accuracy was essentially unchanged for clean condition, and slight improved result was obtained for 20 dB SNR. The error rate reduction was 31.20% for SNR of 15 dB (from 85.50% to 90.03%) and 28% for SNR of 10 dB (from 66.95% to 76.21%). In Table 1, the overall average was improved from 61.08% to 67.07%, or an error rate reduction of 15.39%. Parallel results were shown in Fig. 4(a), (b) and (c) and listed in the middle part of Table 1 for MVDR features, and in Fig. 5(a), (b) and (c) and the right part of Table 1 for PLP features. Exactly the same trend as that for MFCC can be found for both MVDR and PLP. All these results

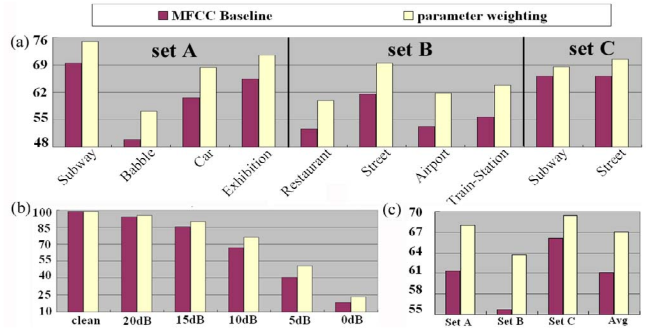

Fig. 3. Accuracies for the proposed feature parameter weighting approach as compared to the baseline tests without weighting, for *MFCC features:* (a) averaged over all SNR values, but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all SNR values and noise types but separated for sets A, B, C.

	MFCC		MVDR		PLP	
	Original	Parameter Weighting	Original	Parameter Weighting	Original	Parameter Weighting
Set A	61.34	68.00	63.86	68.25	63.49	68.99
Set B	55.75	63.74	62.20	68.72	58.40	64.41
Set C	66.14	69.46	64.78	65.57	68.45	71.83
Average	61.08	67.07	63.61	67.51	63.45	68.41

Table 1. Averaged accuracies (%) for sets A, B, C in Fig. 3(c), 4(c) and 5(c).

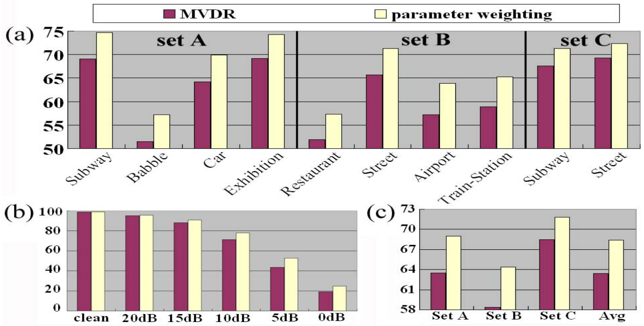

Fig. 4. Accuracies for the proposed feature parameter weighting approach as compared to the baseline tests without weighting, for *MVDR features:* (a) averaged over all SNR values, but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all SNR values and noise types but separated for sets A, B, C.

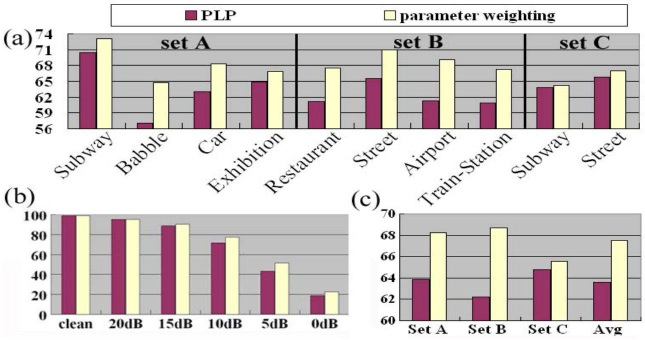

Fig. 5. Accuracies for the proposed feature parameter weighting approach as compared to the baseline tests without weighting, for *PLP features:* (a) averaged over all SNR values, but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all SNR values and noise types but separated for sets A, B, C.

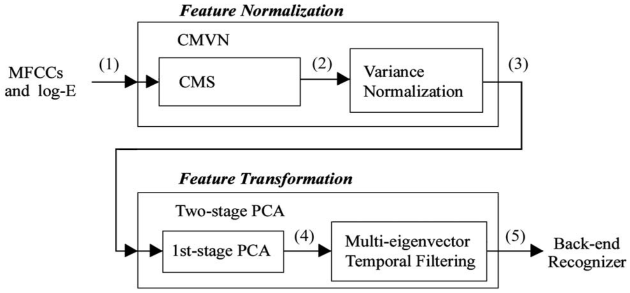

Fig. 6. The four-stage feature enhancement front-end [9]**.** The results in each column of Table 2 are obtained at the output of the corresponding stage as labeled.

verified that the entropy-based feature parameter weighting approach proposed here can successfully identify the more discriminating feature parameters and improve the recognition accuracies by properly emphasizing the scores of such parameters.

3.2. Integration with Typical Robust Speech Recognition Approaches

Here a four-stage feature enhancement front-end and an energy-based frame selection algorithm recently proposed [9] were taken as typical examples of robust speech recognition approaches to be integrated with the feature parameter weighting approach proposed in this paper. The four-stage front-end is shown in Fig. 6. CMS and CMVN form the first part of this front-end. The second part is then a two-stage Principal Component Analysis (PCA) process, consisting of a first stage PCA which transforms 14-dimensional MFCC features (C0~C12 and log-E) obtained with the WI007 front-end into vectors of 13 principal components, and a multi-eigenvector (M-eigen) temporal filtering [10] applied on the temporal trajectories of the 13 resulting feature parameters obtained in the first PCA stage. The energy-based frame selection algorithm, on the other hand, uses the local order statistics of the smoothed instantaneous energy of signal samples to select the reliable frames in a noisy utterance. These reliable frames are then used to estimate the mean, variance and various principal components which are used in the above four stages [9].

The results obtained with the integration of the above four stages and frame selection with the feature parameter weighting approach proposed here, averaged over all types of noise and all SNR values in sets A, B, C, are summarized in Table 2. Column (1) of Table 2 are those for the original MFCC without any processing, while column (2)(3)(4)(5) are those obtained at the outputs of the four stages respectively, as shown in Fig. 6. The first row of the table shows the results without feature parameter weighting, and it can be found that the performance was improved when each of the four stages was applied one after one. The second row shows the corresponding results when the parameter weighting approach was applied in addition, and it can be found that reasonable improvements were obtained at all stages. The third row shows the corresponding results when frame selection was further performed, and very good improvements can be obtained at each stage. The last row is the total error rate reduction.

Processing Stages	(1) MFCC	(2) with CMS	(3) with CMVN	(4) plus PCA	(5) plus M-eigen
Original[9]	61.08	67.29	69.13	79.77	84.10
plus Feature Parameter Weighting	66.84	67.93	70.38	80.89	85.26
plus Frame Selection	67.16	74.40	78.97	82.38	86.31
Total Relative Error Reduction (%)	15.62	21.74	31.88	12.90	13.90

Table 2. Accuracies (%) averaged over all types of noise and all SNR values in sets A, B, C, obtained at the outputs of the four stages in Fig. 6, with baseline MFCC (original), the feature parameter weighting proposed here, and the frame selection algorithm.

The accuracies for a few typical cases in Table 2, i.e. 61.08% of the original MFCC baseline (column (1)), 84.10% with the four stages, 85.26% when feature parameter weighting was applied in addition, and 86.31% when frame selection was applied in addition (all in column (5)), are further analyzed in Fig. 7(a)(b)(c), with accuracies averaged over all SNR values but separated for different types of noise (Fig. 7(a)), averaged over all types of noise but separated for different SNR values (Fig. 7(b)), and averaged over all types of noise and all SNR values but separated for sets A, B, C and their average (Fig. 7(c)), respectively.

In Fig. 7(a), the most significant improvements with the proposed approach are obtained for babble, car, restaurant, airport, and train-station cases. For example, in the case of non-stationary airport noise, the relative error rate reduction is about 9.45% (from 85.78% to 87.12%) with feature parameter weighting applied on the four stages, and 23.05% (to 89.05) with frame selection further applied. This verified that the proposed entropy-based feature parameter weighting approach works well under non-stationary noise, and it can be well integrated with the four stages and frame selection. In Fig. 7(b), slight degradation occurred in the clean speech case, but when SNR value goes down from 20 dB all the way to 0 dB, it is clear that the accuracy was improved with the four stages, feature parameter weighting, and frame selection applied one after one. In the case of 5 dB SNR, for example, the accuracy obtained after applying the feature parameter weighting approach was as high as 81.58%, while the result for the MFCC baseline is only 40.55%, which implied a relative error reduction of 69.02%. Further integration with the frame selection approach gives an accuracy of 83.75%. Similar improvement can be observed in the case of 0 dB SNR, in which the improvement was from 18.27% to 61.02% with the four stages and feature parameter weighting, and 62.28% after applying frame selection further. Apparently these different approaches can be successfully integrated. Because feature parameter weighting gives different weights to the scores of different parameters within a feature vector to help the recognizer, while the frame selection scheme selects the frames that are more reliable among all frames in an utterance to help the four stages, their respective merits are complementary and thus additive. From the overall average accuracies shown in column (5) of Table 2 or in the right-most part of Fig. 7(c), significant improvements can be obtained as compared to the MFCC baseline result of 61.08%.

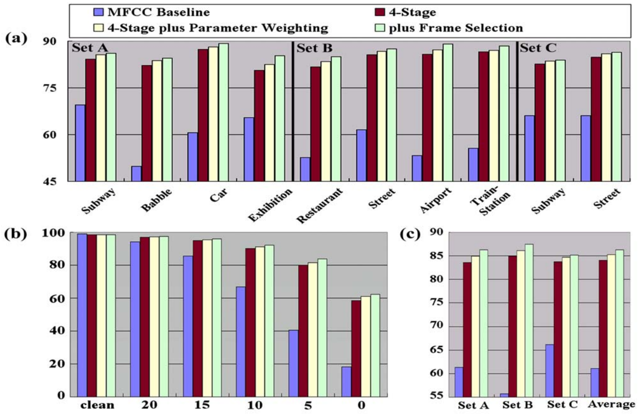

Fig. 7. Incremental improvements in the accuracies obtained with the MFCC baseline, at the outputs of the four stages, plus the feature parameter weighting proposed here, and plus the frame selection algorithm, for (a) averaged over all SNR values but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all SNR values and noise types but separated for sets A, B, C.

4. CONCLUSIONS

In this paper, we propose an entropy-based feature parameter weighting scheme to emphasize the acoustic log-likelihood scores of the more discriminating feature parameters during the Viterbi decoding process. The proposed approach can be equally useful with different kinds of speech features, and can be easily integrated with typical existing robustness techniques to offer improved robustness for the recognition process. The concept of the feature parameter weighting scheme is very simple, and its effectiveness is well verified by extensive experimental results.

5. REFERENCES

[1] Special section on “Speech Technology in Human-Machine Communication,” IEEE Signal Processing Magazine, vol. 22, no. 5, Sep. 2005.

[2] H. Hermansky. “Perceptual linear predictive (PLP) analysis of speech, ” J. Acoust. Soc. Am. 87 (4), 1990.

[3] M. N. Murthi, B. D. Rao, “All-pole Model Parameter Estimation for Voiced Speech,” 1997 IEEE Workshop on Speech Coding for Telecommunications Proceeding, 1997.

[4] M. N. Murthi, B. D. Rao, “All-Pole Modeling of Speech Based on the Minimum Variance Distortionless Response Spectrum,” IEEE Transactions on Speech and Audio Processing, vol. 8, no. 3, May 2000.

[5] S. Dharanipragada, B. D. Rao, “MVDR Based Feature Extraction for Robust Speech Recognition,” ICASSP 2001.

[6] S. Dharanipragada, “Feature Extraction for Robust Speech Recognition,” IEEE ISCAS 2002, May 2002.

[7] Y. Chen, L.-S. Lee, “Robust features for speech recognition using minimum variance distortionless response (MVDR) spectrum estimation and feature normalization techniques,” IEEE ISCSLP 2004, Dec. 2004.

[8] H.-G. Hirsch, D. Pearce, "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems under Noisy Conditions," ISCA ITRW ASR2000, Paris, France, Sep. 2000.

[9] Y. Chen, L.-S. Lee, "Energy-Based Frame Selection for Reliable Feature Normalization and Transformation in Robust Speech Recognition," Proc. InterSpeech 2005 - Eurospeech, Sep. 2005.

[10] N.-C. Wang, J.-W. Hung, L.-S. Lee, “Data-driven Temporal Filters based on Multi-eigenvectors for Robust Features in Speech Recognition,” ICASSP 2003.