interspeech 2008 incorporating SST 08

# Confusion-Based Entropy-Weighted Decoding for Robust Speech Recognition

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

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

## Abstract

An entropy-based feature parameter weighting scheme was proposed previously [1], in which the scores obtained from different feature parameters are weighted differently in the decoding process according to an entropy measure. In this paper, we propose a more delicate entropy measure for this purpose considering the inherent confusion among different acoustic classes. If a set of acoustic classes are easily confused, those feature parameters which can distinguish them should be emphasized. Extensive experiments with the Aurora 2 testing environment verified that this approach is equally useful for different types of features, and can be easily integrated with typical existing robust speech recognition approaches.

**Index Terms**: speech recognition, robustness.

## 1. Introduction

Various applications of the automatic speech recognition (ASR) technologies in the future have been highly anticipated by many people [2]. But the recognition accuracy always plays the most dominating role when the real-world applications are considered, and many robust speech recognition approaches have been proposed to improve the recognition accuracy under adverse environments. Weighted Viterbi decoding is one of the many approaches proposed to be performed in the back-end decoding process, in which during the Viterbi decoding process different weights can be assigned to the acoustic scores obtained from different frames or even different feature parameters in an utterance [3, 4, 5, 6].

For an acoustic feature vector including many feature parameters, it is possible that some feature parameters carry stronger information to discriminate some acoustic classes against the others, while other feature parameters do not. If we treat all the feature parameters as equally important coefficients as in the conventional Viterbi decoding, the functions of those more discriminating parameters may be smeared out by the functions of other parameters. An entropy-based feature weighting scheme was therefore proposed earlier [1], in which an entropy measure was used to emphasize the scores obtained with these more discriminating feature parameters in back-end Viterbi decoding. However, this may not be the best way to weight the different feature parameters. If a set of acoustic classes are easily confused, those feature parameters which can distinguish them should be weighted more. In this paper, we propose a more delicate entropy measure which considers such confusion information among acoustic classes, and the feature parameters which can distinguish such confusing acoustic classes can be weighted more. Experimental results on the Aurora 2 testing environment verified that the proposed approach is equally useful for different types of features including MFCC, PLP [7] and MVDR [2, 8], and this approach can be easily integrated with typical existing robust speech recognition approaches to offer better performance. Such results are consistent across a wide range of noise types and SNR conditions.

This paper is organized as follows. The proposed confusion-based entropy weighting scheme is described in section 2. In section 3 the experimental results are presented. Section 4 gives our conclusions.

## 2. Proposed Approach

### 2.1. GMM score for each parameter in a testing feature vector

In order to perform entropy-based weighting for different feature parameters, we need to evaluate the scores for each parameter in a testing feature vector from different acoustic classes. We first perform forced alignment on the utterances in the training corpus with the corresponding transcriptions, and collect the feature vectors of the same acoustic class (e.g., the same phone, or any other unit represented by the same acoustic model) together to train a Gaussian mixture model (GMM) with $N$ Gaussian components for each class,

$$G_c(\boldsymbol{x}) = \sum_{n=1}^{N} k_{c,n} N_{c,n}(\boldsymbol{x} \,|\, \theta_{c,n}), \quad c = 1, 2, \ldots, C, \qquad (1)$$

where $c$ is the class (or acoustic model) index, $n$ is the mixture component index of the GMM ($n = 1, 2, \ldots, N$), $k_{c,n}$ is the weight for the $n$-th mixture component $N_{c,n}(\boldsymbol{x}\,|\theta_{c,n})$ for class c, and $\theta_{c,n}$ is the set of parameters of $N_{c,n}(\bullet)$ (i.e., the mean and covariance matrix of the Gaussian distribution).

Now for each testing feature vector $\boldsymbol{x}(t)$ at time $t$ including $D$ feature parameters, $\boldsymbol{x}(t) = \{x_d(t), d = 1, 2, \ldots, D\}$ (i.e., $d$ is the parameter index and $D$ is the total number of parameters), the score or the probability density value of the $d$-th parameter, $x_d(t)$, for an acoustic class (or acoustic model) $c$ can be evaluated as

$$p_c^{t,d} = \int_{\substack{d'=1,\ldots,D \\ d' \neq d}} \ldots \int G_c(\boldsymbol{x}(t)) dx_1 \ldots dx_{d'} \ldots dx_D |_{x_d(t)}, \qquad (2)$$

where $G_c(\boldsymbol{x}(t))$ is exactly $G_c(\boldsymbol{x})$ in Eq. (1) except replacing $\boldsymbol{x}$ by $\boldsymbol{x}(t)$, and Eq. (2) is evaluated for the $d$-th parameter $x_d(t)$ by an integration of $G_c(\boldsymbol{x}(t))$ over all other feature parameters $x_{d'}(t), d' = 1, 2, \ldots, D$, but $d' \neq d$. The above can be easily simplified when we assume the $D$ feature parameters in $\boldsymbol{x}(t)$ are independent and the covariance matrices of $N_{c,n}(\bullet)$ are diagonal. In that case, $G_c(\boldsymbol{x}(t))$ in Eq. (2) can be simplified into $D$ independent scalar GMMs for the $D$ feature parameters respectively, and $p_c^{t,d}$ in Eq. (2) can be reduced to

$$p_c^{t,d} = \sum_{n=1}^{N} k_{c,n,d} N_{c,n,d}(x_d(t) \,|\theta_{c,n,d}), \qquad (3)$$

This work is supported by the National Taiwan University Advanced Speech Technology Scholarship.

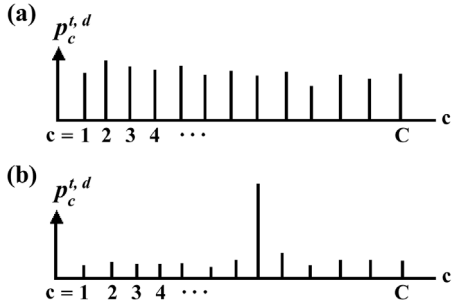


Figure 1: *Distributions of $p_c^{t,d}$ over the $C$ classes give different entropy values: (a) high entropy, and (b) low entropy.*

where $k_{c,n,d}$, $N_{c,n,d}(\bullet)$ and $\theta_{c,n,d}$ are exactly the same as $k_{c,n}$, $N_{c,n}(\bullet)$ and $\theta_{c,n}$ in Eq. (1), except reduced to those for a single feature parameter with index $d$, i.e. $x_d(t)$.

### 2.2. Entropy-based feature weighting

With $p_c^{t,d}$ in Eqs. (2) or (3), the entropy measure for the feature parameter $x_d(t)$ can be defined as follows. We first normalize $p_c^{t,d}$ in Eqs. (2) or (3) into a probability mass function (PMF) $\overline{p}_c^{t,d}$ of $x_d(t)$ across all acoustic classes $c$ as

$$\overline{p}_c^{t,d} = p_c^{t,d} / \sum_{c=1}^{C} p_c^{t,d}, \tag{4}$$

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

$$H^{t,d} = -\sum_{c=1}^{C} \overline{p}_c^{t,d} \cdot \log\left(\overline{p}_c^{t,d}\right). \tag{5}$$

If for a feature parameter $x_d(t)$ the distribution of $p_c^{t,d}$ in Eq. (3) across all classes $c$ looks like the one in Figure 1(a), i.e., the scores for different classes are very similar, the entropy measure $H^{t,d}$ defined in Eq. (5) will be high, which means the discriminating ability of the feature parameter $x_d(t)$ is low. A typical example for such a case is that with $d = 1$ in Figure 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 for $x_d(t)$ the distribution of $p_c^{t,d}$ in Eq. (3) across all classes $c$ looks like the one in Figure 1(b), i.e., one of the classes has a much higher score than all the others, the entropy measure $H^{t,d}$ in Eq. (5) 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 Figure 2. In other words, the distribution of the class with the highest likelihood score is very possibly well separated from those of all other classes. Apparently the recognition should rely more on the latter than on the former.

Based on the entropy measure $H^{t,d}$ in Eq. (5), a reasonable feature parameter weighting function $W^{t,d}$ for $x_d(t)$ can be defined as

$$W^{t,d} = f(H^{t,d}) = \exp\left(-a \cdot H^{t,d}\right), \tag{6}$$

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 Eq. (6) can be a carefully chosen monotonically decreasing function, while here the exponential function is used for simplicity.

The feature parameter weighting function $W^{t,d}$ in Eq. (6) can then be applied in a weighted Viterbi decoding (WVD) process, in which the log-likelihood score of a given feature vector $\boldsymbol{x}(t)$ for the $j$-th state of a hidden Markov model (HMM)

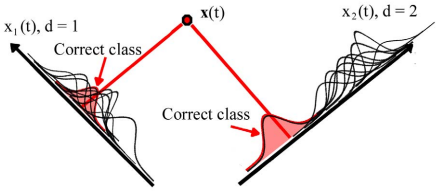


Figure 2: *The situation of $\boldsymbol{x}(t)$ that the entropy $H^{t,d}$ is high for $d = 1$, but low for $d = 2$.*

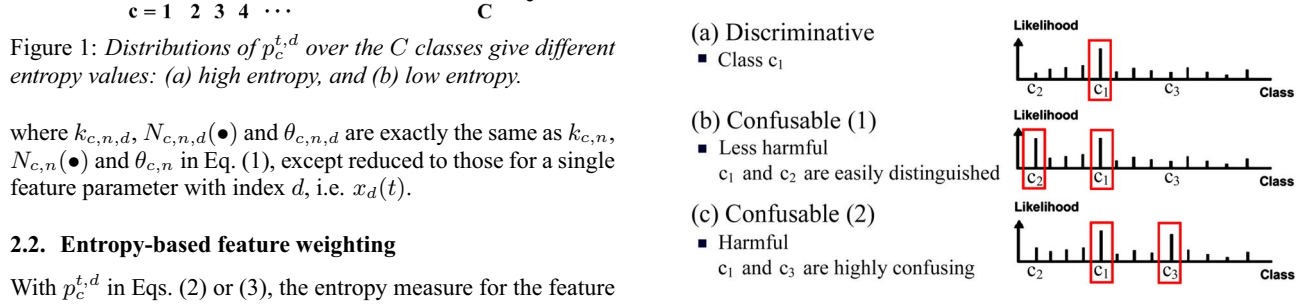


Figure 3: *The situations for a particular feature parameter to be (a) discriminative, (b) confusing but less harmful, and (c) confusing and harmful.*

in the recognizer can be calculated as, assuming a diagonal covariance matrix,

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

where $j$ and $m$ are respectively the indices of the state in the HMM and of the mixture Gaussian component in the state, $c_{jm}$ is the mixture weight, and $\mu_{jmd}$ and $\Sigma_{jmd}$ are the mean and variance of the $m$-th Gaussian component in the $j$-th state of the HMM. So the acoustic scores from the more discriminating feature parameters will be emphasized in the weighted Viterbi decoding process in Eq. (7), and vice versa. This is actually the entropy-based feature weighting approach previously proposed [1].

### 2.3. The concept of confusion-based entropy weighting

In the above initial entropy-based feature weighting scheme, the likelihood scores for all acoustic classes are treated as equally important. This is not necessarily the best way to handle the problem. Consider the three cases in Figure 3. In Figure 3(a), the feature parameter has a good discriminating ability because it has the highest likelihood score for class $c_1$ among all classes, and the score is significantly higher than all the other scores. In Figure 3(b), the feature parameter may have confusion between classes $c_1$ and $c_2$, because they have comparable likelihood scores for this feature parameter. However, this may not cause any problem if we know a priori that classes $c_1$ and $c_2$ are not easily confused with each other, very possibly because these two classes can be correctly distinguished with some other feature parameters in the feature vector. However, in the case shown in Figure 3(c), if $c_3$ is a class known to be easily confused with class $c_1$ a priori, it will be a problem if the feature parameter has comparable scores for them.

This implies that it is the potential confusion among the different classes as discussed in Figure 3 which should be considered more seriously in the parameter weighting, rather than only the entropy measure as shown in Figure 1. For example, in the case of Figure 3(c), the entropy measure for the feature pa-

rameter being considered is low if evaluated for all the classes, but is high if we focus on only the two classes $c_1$ and $c_3$ that are easily confused with each other. A confusion-based entropy weighting scheme is therefore proposed as follows.

### 2.4. Modified class confusion matrix

We first define a modified class confusion matrix $\{v_c(i)\}_{c,i=1}^{C}$ containing elements $v_c(i)$ defined as follows, which is an indicator regarding how frequently the class $i$ is misclassified as a given class $c$,

$$v_c(i) = \begin{cases} \log(count_c(i)+1)/\log(count_c(i_c^*)+1), & \forall i \neq c, \\ 1, & i = c, \end{cases} \quad (8)$$

where $count_c(i)$ is the number of frames in the training corpus belonging to class $i$ which are mis-classified as belonging to the given class $c$. The class $i_c^*$ is the most confusing class with respect to the given class $c$,

$$i_c^* = \arg\max_{i \neq c} (count_c(i)). \quad (9)$$

As a result, we have $v_c(i_c^*) = 1$ and $v_c(c) = 1$ from Eqs. (8) and (9), and $0 \leq v_c(i) \leq 1$ for all other $i$, $i \neq i_c^*$, $i \neq c$. The additions of one in the first expression in Eq. (8) are simply to avoid taking logarithm of zero for those classes $i$ with zero frames misclassified as class $c$. This matrix $\{v_c(i)\}$ can be pre-trained with a training corpus serving as the *a priori* knowledge about the recognition task.

### 2.5. Confusion-based entropy weighting

With the modified class confusion matrix $\{v_c(i)\}$ defined above, for a given feature parameter $x_d(t)$, the probability density values $p_i^{t,d}$ for all classes $\{i, i = 1, 2, \ldots, C\}$ can first be obtained using Eqs. (2) or (3) above. They are then normalized as in Eq. (10) below to give a probability distribution $\{\bar{p}_i^{t,d}, i = 1, 2, \ldots, C\}$,

$$\bar{p}_i^{t,d} = p_i^{t,d} / \sum_{i=1}^{C} p_i^{t,d}, \; i = 1, 2, ..., C. \quad (10)$$

Therefore, with the modified class confusion matrix defined in Eqs. (8) and (9) above, the entropy measure defined in Eq. (11) below over the distribution $\{\bar{p}_i^{t,d}, i = 1, 2, \ldots, C\}$ for a given class $c$ considers both the probabilities $\bar{p}_i^{t,d}$ for the classes $i$ and $v_c(i)$ representing the confusions from the classes $i$ to the given class $c$,

$$H_c^{t,d} = -\sum_{i=1}^{C} v_c(i) \cdot \bar{p}_i^{t,d} \cdot \log(\bar{p}_i^{t,d}). \quad (11)$$

The entropy measure $H_c^{t,d}$ naturally tells the discriminating ability of the feature parameter $x_d(t)$ to classify the class $c$ with respect to those other classes which either have higher probabilities $\bar{p}_i^{t,d}$ or are very often confused with the class $c$. But this entropy measure is for a given class $c$. For a given feature parameter $x_d(t)$, the above entropy measure $H_c^{t,d}$ should be averaged over all classes $c$ ($c = 1, 2, \ldots, C$), weighted by the corresponding probability densities $p_c^{t,d}$,

$$H_{confusion}^{t,d} = (\sum_{c=1}^{C} p_c^{t,d} \cdot H_c^{t,d}) / \sum_{c=1}^{C} p_c^{t,d}. \quad (12)$$

The difference between the entropy measure $H^{t,d}$ in Eq. (5) and the entropy measure $H_{confusion}^{t,d}$ in Eq. (12) is that for the same feature parameter the former puts equal emphasis on the confusions between all classes, but the latter uses $v_c(i)$ to automatically ignore some confusions that may not cause serious problems to the final classification, and to put more emphasis on the discriminating ability of the parameter to classify a given class with its most confusing classes.

With the confusion-based entropy measure in Eq. (12), the proposed confusion-based entropy weighting function $W_{confusion}^{t,d}$ for $x_d(t)$ is, similar to Eq. (6),

$$W_{confusion}^{t,d} = f(H_{confusion}^{t,d}) = \exp(-a \cdot H_{confusion}^{t,d}), \quad (13)$$

where $a$ is an empirically determined scaling factor. This weighting factor $W_{confusion}^{t,d}$ is also a function of both time index $t$ and parameter index $d$. Similar to Eq. (7) above, the weighted Viterbi decoding process can be equally performed,

$$\log[b_j(\boldsymbol{x}(t))] = \sum_{d=1}^{D} W_{confusion}^{t,d} \cdot \log[\sum_{m=1}^{M} c_{jm} N(x_d(t);\, \mu_{jmd}, \Sigma_{jmd})]. \quad (14)$$

## 3. Experiments

### 3.1. Experimental conditions

The initial experiments reported in this paper were conducted on the Aurora 2 testing environment [9] 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, PLP coefficients, and Minimum Variance Distortionless Response (MVDR)-based cepstral coefficients. In the second set of experiments, we integrated the proposed approach with an example existing robust speech recognition approach, i.e., Cepstral Mean and Variance Normalization (CMVN) [10]. In these experiments, we define the acoustic classes as the 13 word models ($one$ to $nine$, $zero$, $oh$, $silence$ and $short\ pause$ models) used in the Aurora 2 task [9].

### 3.2. Recognition results with different features: MFCC, PLP and MVDR

In the first set of experiments, the previously proposed entropy-based feature weighting [1] and the presently proposed confusion-based entropy weighting schemes were tested on three sets of feature parameters, i.e., MFCCs, PLP coefficients and MVDR-based cepstral coefficients. The 13 MFCC parameters ($C1$–$C12$ and log-E) were obtained with the WI007 front-end [9]. For PLP coefficients, the feature vector consists of 12 PLP coefficients and a log-energy term. The MVDR-based features were obtained using the frequency-warped MVDR algorithm with the warping factor set to 0.1 for spectrum estimation to replace the conventional Fourier spectrum for MFCC [8]. In all the three cases, 39-dimensional feature vectors including the delta and delta-delta components were used.

The recognition results on MFCC features are shown in Figure 4, where the three bars are respectively for the baseline, the previously proposed entropy-based weighting, and the confusion-based entropy weighting proposed in this paper, for different types of noise but averaged over all SNR values (0–20 dB) in Figure 4(a), for different SNR values but averaged over all types of noise in Figure 4(b), and for the three testing sets in Figure 4(c). Significant improvements were obtained by the new approach proposed in this paper in all cases consistently across all testing conditions. In Figure 4(a), the error rate reductions were significant for all types of noise, with good examples including 32.20% of relative error reduction for car

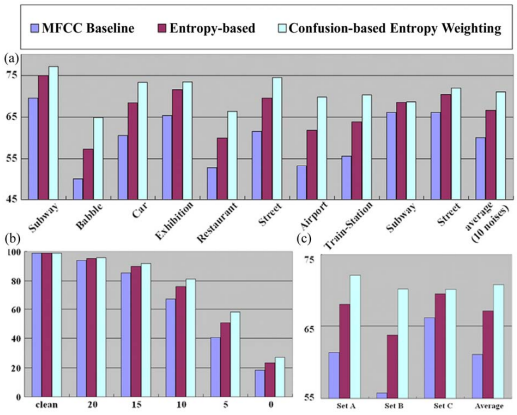


Figure 4: *Accuracies for baseline, previously proposed entropy-based weighting, and the presently proposed confusion-based entropy weighting for MFCC features: (a) averaged over all SNR values but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all SNR values and noise types but separated for sets A, B, C.*

noise in test set A (60.60% to 73.29%), and 35.45% reduction for airport noise in test set B (53.25% to 69.82%). In Figure 4(b), the recognition accuracies were essentially unchanged for tests in clean condition, and example of significant improvements were error rate reductions of 44.31% for SNR of 15 dB (85.50% to 91.93%) and 44.05% for SNR of 10 dB (66.95% to 81.51%). As shown in Table 1, the total error reductions of the final average results are 25.22% for MFCC (from 61.08% to 70.89%), 15.74% for PLP (from 63.61% to 69.34%), and 25.18% for MVDR (from 63.45% to 72.65%), respectively. The detailed results and relative improvements for PLP and MVDR parameters are similar but left out for space limitation.

### 3.3. Integration with CMVN

The popularly used Cepstral Mean and Variance Normalization (CMVN) technique [10] is taken as a typical example of robust speech recognition techniques to be integrated with the feature weighting scheme proposed here. The results are shown in Figure 5. Again reasonable improvements were obtained in all cases, although the achievable improvements were relatively smaller obviously because CMVN itself is very powerful. As typical examples, in Figure 5(a) the relative error rate reductions were 6.60% for babble noise (70.17% to 72.14%) in set A, 6.96% for train-station noise (68.52% to 70.71%) and 6.75% for street noise (72.02% to 73.91%) in set B. Similar improvements can be obtained in Figure 5(b) for different SNR conditions and in Figure 5(c) for the three sets A, B, and C. The improvements are consistent for all three sets, with overall average accuracy improved from 69.13% to 70.86%, which implied a total relative error reduction of 5.61%.

## 4. Conclusions

In this work, we propose a confusion-based entropy weighting scheme for emphasizing the acoustic scores obtained with the feature parameters with better discriminating ability for confusing acoustic classes. Significant improvements were obtained with the proposed scheme applied on different types of features, or integrated with existing robustness techniques, in extensive experiments with the Aurora 2 testing environment under a wide range of noise types and SNR conditions.

| | MFCC | | PLP | | MVDR | |
|---|---|---|---|---|---|---|
| | Original | Entropy Weighting | Original | Entropy Weighting | Original | Entropy Weighting |
| Set A | 61.34 | 72.19 | 63.86 | 70.15 | 63.49 | 73.32 |
| Set B | 55.75 | 70.26 | 62.20 | 71.99 | 58.40 | 71.67 |
| Set C | 66.14 | 70.24 | 64.78 | 65.88 | 68.45 | 72.97 |
| Average | 61.08 | 70.89 | 63.61 | 69.34 | 63.45 | 72.65 |

**Table 1.** Averaged accuracies (%) for sets A, B, C for experiments with respectively MFCC, PLP and MVDR features.

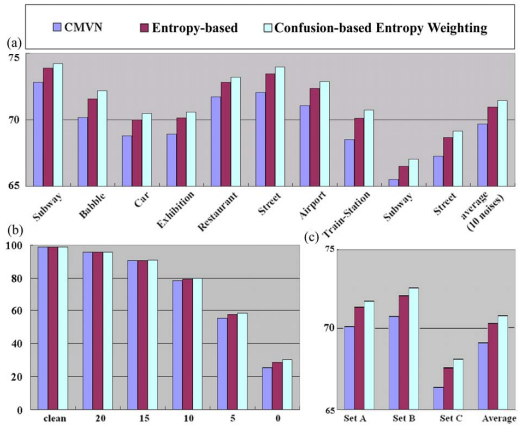


Figure 5: *Accuracies for baseline, previously proposed entropy-based weighting, and the presently proposed confusion-based entropy weighting for CMVN: (a) averaged over all SNR values but separated for different types of noise, (b) averaged over all types of noise but separated for different SNR values, and (c) averaged over all SNR values and noise types but separated for sets A, B, C.*

## 5. Acknowledgment

The authors would like to thank the reviewers for their extensive and valuable comments.

## 6. References

[1] Y. Chen, C.-Y. Wan, L.-S. Lee, "Entropy-Based Feature Parameter Weighting for Robust Speech Recognition," ICASSP 2006.

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

[3] N. B. Yoma, M. Villar, "Speaker Verification in Noise Using a Stochastic Version of the Weighted Viterbi Algorithm," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 3, pp. 158-166, Mar. 2002.

[4] N. B. Yoma, I. Brito, C. Molina, "The Stochastic Weighted Viterbi Algorithm: A Frame Work to Compensate Additive Noise and Low-Bit Rate Coding Distortion," InterSpeech 2004.

[5] A. Bernard, A. Alwan, "Low-Bitrate Distributed Speech Recognition for Packet-Based and Wireless Communication," IEEE Transactions on Speech and Audio Processing, vol. 10, no. 8, pp. 570-579, Nov. 2002.

[6] X. Cui, A. Alwan, "Noise Robust Speech Recognition Using Feature Compensation Based on Polynomial Regression of Utterance SNR," IEEE Transactions on Speech and Audio Processing, vol. 13, no. 6, pp. 1161-1172, May 2006.

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

[8] 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.

[9] 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.

[10] O. Viikki, K. Laurila, "Cepstral Domain Segmental Feature Vector Normalization for Noise Robust Speech Recognition," Speech Communication, 1998.