Noise Robust Front-end for ASR using Spectral Subtraction, Spectral Flooring and Cumulative Distribution Mapping

Eric H. C. Choi

Multimodal User Interaction
National ICT Australia
Eric.Choi@nicta.com.au

Abstract

In this paper, a novel and noise robust front-end based on the combined application of spectral subtraction, spectral flooring and cumulative distribution mapping is proposed. Recognition experiments with the Aurora II connected digits reveal that the proposed front-end achieves an average digit accuracy of 81.46% for a model set trained from clean data and 89.54% for a model set trained from data with various noise conditions. With reference to the ETSI standard Mel-cepstral front-end, the proposed front-end obtains a relative error reduction of around 52% for the clean model set and 14% for the multi-condition model set. Moreover, it is observed that the use of a single fixed parameter to control spectral flooring is beneficial only when cumulative distribution mapping is also applied at a later stage of the front-end processing.

1. Introduction

The state-of-the-art automatic speech recognition (ASR) systems work pretty well if the training and testing conditions are similar and reasonably controlled. However, under the influence of noise, these systems begin to fall apart and their accuracies become unacceptably low in severe environments (e.g. low signal-to-noise ratio). To remedy this noise robustness issue in ASR, various adaptive techniques have been proposed. A common theme of these techniques is the utilisation of some form of compensation to account for the effects of noise on the speech characteristics. In general, a compensation technique can be applied in the signal, feature or model space (Huang et al., 2001) to reduce mismatch between training and usage conditions.

Signal-space methods (Ephraim, 1992; Digalakis et al., 1993; Lee and Jung, 2000) typically try to enhance a noisy speech signal by improving its signal-to-noise ratio (SNR). However, improved SNR does not always contribute to improvement in recognition accuracy. Feature-space methods (Hermansky, 1990; Kim et al., 1999) try to derive some kind of feature representation that is invariant to the change in noise conditions. Typically this is achieved by incorporating some aspects of human auditory modelling. Alternatively, some other feature-space methods (Sankar and Lee, 1996; Tian et al., 2002; Torre et al., 2002) try to understand and compensate the effects of noise on a speech representation and correspondingly reduce the mismatch. Model-space methods (Yao et al., 2001; Ida and Nakamura, 2002; Cerisara et al., 2004; Zhang and Furui, 2004) try to adjust the parameters of recognition models to incorporate the effects of noise on the models. Typically a model of the environment that considers additive and convolutional noise is assumed.

In this work, the main focus is on feature-space compensation for a cepstral based front-end. It is demonstrated that the use of spectral flooring, together with cumulative distribution mapping can be a good alternative to spectral subtraction in compensating the effects of additive noise during the front-end processing. Moreover, additional improvements in recognition accuracy can be achieved by applying all these compensation methods together in a cascade fashion.

The organisation of this paper is as follows. It will describe the details of the proposed front-end processing in Section 2 and some related recognition experiments on the Aurora II digits database in Section 3. Following this is a discussion of the findings in Section 4 and a summary of the conclusions in Section 5.

2. Proposed Front-end Processing

The development of the proposed front-end processing is based on the ETSI standard Mel-frequency cepstral coefficient (MFCC) front-end (ETSI, 2000). In addition

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

to the basic processing blocks (those blocks without underlined labels in Figure 1), three more processing blocks related to additive noise compensation have been added. These additional blocks include noise spectral subtraction (SS), spectral flooring (SF) in log-compression and cumulative distribution mapping (CDM) for the cepstral and log-energy coefficients. A high level diagram of the processing flow is shown in Figure 1, while a more detailed description of the individual compensation blocks can be found in the following sub-sections. Details about the other basic processing blocks can be found in (Hirsch and Pearce, 2000).

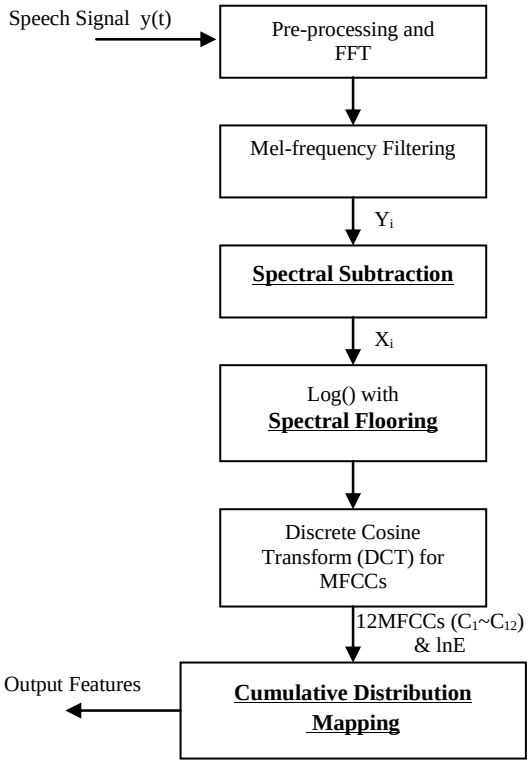

Figure 1: Proposed front-end processing which incorporates spectral subtraction, spectral flooring and cumulative distribution mapping for noise compensation

2.1. Spectral Subtraction

The implementation of the noise spectral subtraction module is based on a commonly used non-linear spectral subtraction algorithm (Vaseghi, 2000). It can be applied to reduce the effect of an additive noise on the magnitude spectrum of a speech signal by subtracting the noise estimate from the noisy signal spectrum. In our case, the spectral subtraction is applied after the Mel-frequency filtering and the subtraction algorithm is given by:

$$X(\omega) = \max\{(Y(\omega) - N(\omega)), \alpha Y(\omega)\} \qquad (1)$$

where $X(\omega)$ is the estimated clean speech magnitude spectrum, $Y(\omega)$ is the magnitude spectrum of the noisy speech signal, $N(\omega)$ is the average magnitude spectrum of the noise and $\alpha \in (0,1)$ is an attenuation constant to prevent $X(\omega)$ from becoming negative due to error in the noise estimate. In this work, the first 10 frames of each utterance are assumed to be noise only and they are used to compute the average noise spectrum. Note that this is a relatively common assumption that would usually hold for practical ASR systems.

2.2. Spectral Flooring

The effect of additive noise on a log filterbank output is nonlinear and it can reduce the dynamic range or variance of the output (Torre et al., 2002). This reduction in variance is particularly significant if the original acoustic models are trained from clean speech data (i.e. speech with high signal-to-noise ratio). The large mismatch between the clean model set and the noisy data can cause recognition accuracy to degrade rapidly. To visualise the effect of noise on log filterbank output, a plot of the output sequences for the clean and noisy version of an example digit string is shown in Figure 2.

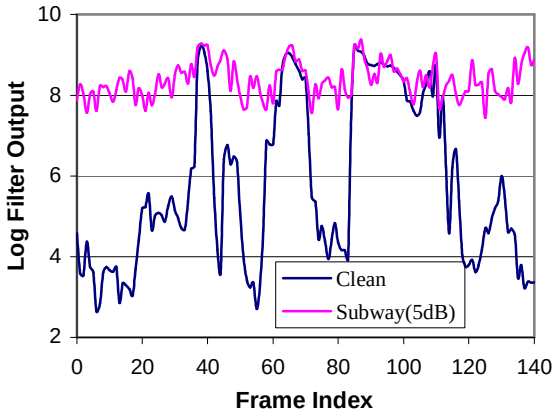

Figure 2: Log Mel-filter output (6th bank) sequences for clean (lower trace) and noisy (upper trace) version of the digit string "Six Three Five", subway noise at 5dB SNR

If a set of acoustic models is trained from clean speech data, spectral flooring can be applied to mask out the potential effect of noise by limiting the lower-bound of a Mel-filterbank output to an appropriate value. By doing that, one can maintain the dynamic range of a feature component in the model set to a desired level and thus help to reduce the potential mismatch between a noisy utterance and the acoustic models. In this work, we adopt the same type of logarithmic transform as

Accepted after full paper review

proposed in (Hermansky et al., 1993) for RASTA processing, given by:

$$\Gamma(y) = \log_e(1+\gamma y) \quad (2)$$

where y is the output of a Mel-filterbank in a particular speech frame (this notation assumes no spectral subtraction for generality) and γ is a flooring factor to be determined. In the literature, the flooring factor has to be adjusted explicitly according to the SNR of the corresponding filterbank output. However, in our case, γ is set to be filterbank-independent and the same single value is used for speech data with different SNR's. This novel approach is possible only when cumulative distribution mapping is used in the front-end processing, as discussed in Section 4. Note that $\Gamma(.)$ is approximately linear for $\gamma y << 1$ and logarithmic for $\gamma y >> 1$.

2.3. Cumulative Distribution Mapping

The cumulative distribution mapping method described here can be traced back to the use of histogram equalisation (HE) in image processing (Russ, 1995). The use of the HE method for additive noise compensation in front-end processing of speech can also be found in (Torre et al., 2002; Pelecanos, 2003). The main idea of this method is to map the distribution of the noisy speech features into a target distribution with a pre-defined probability density function (PDF). In our case, it is assumed that for a given feature value v_o, the mapping relationship would be:

$$\int_{v=-\infty}^{v_o} f(v)dv = \int_{z=-\infty}^{z_o} h(z)dz \quad (3)$$

or

$$F_v(v_o) = F_z(z_o) \quad (4)$$

where $F_v(v)$ is the corresponding cumulative distribution function (CDF) of a given set of speech features and $F_z(z)$ is the target CDF, $f(v)$ and $h(z)$ are the respective PDF's. From equation (4), we have

$$z_o = F_z^{-1}[F_v(v_o)] \quad (5)$$

Therefore the required mapping from a given speech feature v_o into the corresponding target feature z_o is represented by equation (5). In our work, the target PDF of z is assumed to be a Gaussian with zero mean and unity variance. That is:

$$h(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (6)$$

We also use the following formula to estimate $F_v(v_o)$:

$$F_v(v_o) = \frac{K + 0.5}{N}; \quad (7)$$

$$K = \text{Count}\{v < v_o\} \quad (8)$$

i.e. K is the number of frames whose corresponding feature values are less than a particular v_o in an utterance and N is the total number of frames in the utterance.

In our work, the target $F_z(z)$ is divided into 100 bins and the corresponding z values are stored in a table. Note that since a Gaussian distribution is symmetric, in practice we need only to store 50 entries. In the experiments, cumulative distribution mapping is applied only to the basic feature vector which consists of the 12 MFCC's and the log-energy coefficient.

3. Experimental Results

The proposed front-end has been evaluated on the Aurora II database (Hirsch and Pearce, 2000) with various configurations. This database contains noisy connected digits (spoken by American adults), which were created by adding various types of noises at different SNR's to the original clean utterances. The types of noises include subway, babble, car, and exhibition noise. There are three test sets in the database, but we have performed the evaluation using test set *A* only. The SNR's of the test data range from -5dB to more than 20dB, while those of the training data can have SNR's ranging from 5dB to more than 20dB.

3.1. Experimental Setup

The basic feature vector of our front-end consisted of 12 MFCC's and log-energy. This basic feature vector was appended with their corresponding 1st-order and 2nd-order time derivatives to form a resultant vector with 39 coefficients for speech recognition at the backend, as per the Aurora evaluation framework. The Hidden Markov Model (HMM) Toolkit (Young et al., 1997) was used for the speech recognition experiments. Each model was represented by a continuous density HMM with left-to-right configuration. Digit models had 16 states with 3 Gaussians per state, while the noise model had 3 states with 6 Gaussians per state. An inter-digit silence model with 1 state was also used, and it was tied with the middle state of the 3-state silence model.

Two sets of HMM's were trained for the evaluation. The clean model set was trained from clean speech data only and the multi-condition model set was trained from the noise-added version of the same training data. Note that the training and test data are disjoint sets and comprise the original data from the Aurora CDs without end-point detection. Each of the two training sets contains 8440

utterances, while the test set *A* contains about 28K utterances in total.

3.2. Results with Various Front-End Configurations

We followed the official Aurora evaluation framework in that average recognition accuracy for each test set is calculated from the recognition results for those test data with SNR's from 0 dB to 20dB only. When the ETSI standard MFCC front-end was used, the average digit accuracy for the test set *A* was found to be 61.34% for the clean HMM set and 87.82% for the multi-condition HMM set. We also modified the calculation of log-energy to make use of Mel-filterbank outputs instead of calculating it from a frame of raw speech signal. In our case, the log-energy (lnE) is calculated as:

$$\mathrm{lnE} = \log_e\left(\sum_{i=1}^{M} X_i^{\,2}\right) \qquad (9)$$

where X_i is the output amplitude of the i-th Mel-filterbank after spectral subtraction and M is the total number of Mel-filterbanks (M=23). With this modification, the average digit accuracy was found to be 65.01% for the clean HMM set and 86.21% for the multi-condition HMM set. Based on this modification, various recognition experiments were performed using different front-end configurations, and the results are summarised as shown in Table 1. Note that the 1st-order and the 2nd-order time derivatives of a basic feature vector were generated after those basic features had been compensated.

Table 1: Average digit accuracies (%) for Aurora test set *A* with various front-end configurations

Front-end Configuration	Clean HMM Set	Multi-condition HMM Set
No compensation	65.01	86.21
SS only	75.64	89.07
SF only	53.75	86.61
CDM only	76.92	89.71
SS/CDM	79.61	**90.23**
SF/CDM	79.46	89.31
SS/SF	61.15	87.63
SS/SF/CDM	**81.46**	89.54
ETSI standard	61.34	87.82

SS: Spectral subtraction (α=0.4)
SF: Spectral flooring (γ=0.001)
CDM: Cumulative distribution mapping

From the above table, it is found that the front-end configuration SS/SF/CDM, which combines the three noise compensation methods, achieves the best recognition accuracy for the clean HMM set (81.46%). On the other hand, the spectral flooring method is found to have no further improvement on the accuracy of the SS/CDM front-end for the multi-condition HMM set.

For easy comparison with the baseline results, a break-down of the recognition results for the final front-end configuration (SS/SF/CDM) according to individual SNR levels is shown in Figure 3. From the figure, it is observed that this front-end configuration achieves better recognition accuracy than that of the ETSI standard front-end in every noise level (SNR). Moreover, the difference in recognition accuracy for the clean and multi-condition HMM sets is found to be much reduced using the proposed front-end.

ETSI_clean: standard front-end, clean HMM set
ETSI_multi: standard front-end, multi-condition HMM set
Proposed_clean: SS/SF/CDM, clean HMM set
Proposed_multi: SS/SF/CDM, multi-condition HMM set

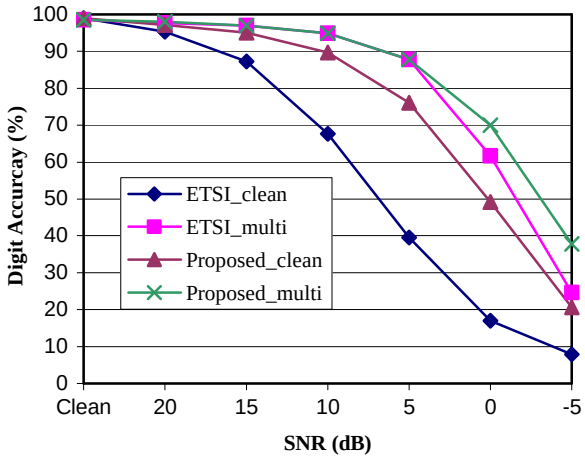

Figure 3: Recognition results for Aurora test set *A*, SS/SF/CDM front-end compared with ETSI standard front-end

3.3. Results with Various Degrees of Spectral Flooring

We also investigated the effect of various degrees of spectral flooring on the recognition accuracy for using the SS/SF/CDM front-end. The experimental results are shown in Table 2.

Table 2: Average digit accuracies (%) for Aurora test set *A* with various degrees of spectral flooring, SS/SF/CDM front-end; α = 0.4

Spectral Flooring (γ)	Clean HMM Set	Multi-condition HMM Set
0.01	79.88	90.01
0.005	79.06	89.71
0.001	81.46	89.54
0.0005	81.31	88.96

Accepted after full paper review

From the previous table, it can be observed that although there is an optimal value of γ for the clean HMM set, the change in recognition accuracy is barely significant if γ is sufficiently small (i.e. a reasonably higher degree of flooring). Also the results here further confirm that spectral flooring has limited benefit for multi-condition training.

To understand under what noise conditions spectral flooring can further improve the SS/CDM front-end, a comparison of recognition results between the SS/CDM and the SS/SF/CDM front-ends at different noise levels was undertaken (shown in Figure 4). As observed from Figure 4, the improvement in the average digit accuracy is mainly obtained from the noisier conditions (5 and 0 dB SNR).

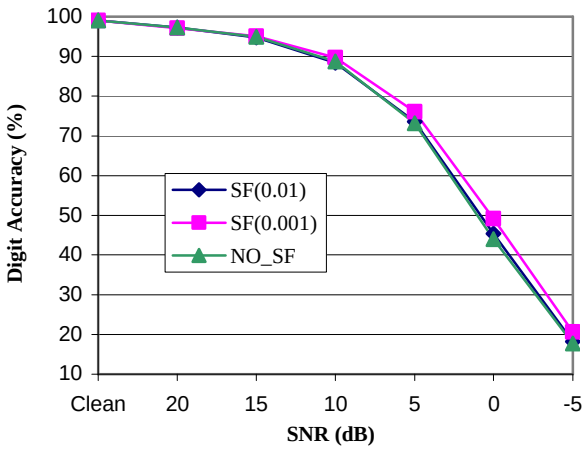

Figure 4: Recognition results for Aurora test set *A* with clean HMM set, SS/SF/CDM front-end (γ = 0.01 or 0.001), and SS/CDM front-end (NO_SF); α = 0.4

4. Discussion

As far as individual compensation methods are concerned, the cumulative distribution mapping was found to provide the greatest improvement in recognition accuracy, independent of the type of HMM set being used. It is interesting to note that although log-energy is used as a component of the basic feature vector, one can actually get rid of the logarithm and just use the energy for the cumulative distribution mapping method. This will not affect the recognition results since the CDF of the energy values is invariant to any monotonically increasing transformation, such as logarithmic operation.

It can be observed from Table 1 that the use of spectral flooring alone with a single fixed parameter is not very workable. This agrees with other findings in the literature (Hermansky, 1993; Lieb and Fischer, 2001) which observed that γ should be made dependent on the filterbank output SNR. Nevertheless, when spectral flooring is used with cumulative distribution mapping, reasonable improvement in recognition accuracy can be obtained. In fact, it is observed that the recognition accuracies with the clean HMM set are almost the same for both the SS/CDM and the SF/CDM front-ends. The implicit SNR normalisation of each Mel-filterbank output by the cumulative distribution mapping on the cepstral coefficients may be a possible reason why a single flooring parameter can be used in this case.

With reference to the ETSI standard front-end configuration, the novel SS/SF/CDM front-end configuration achieves a relative error reduction of around 52% for the clean HMM set and 14% for the multi-condition HMM set. These results compare favourably with those reported in (Cui et al., 2002) which utilised more than seven different compensation algorithms, including explicit end-point detection, spectral subtraction and RASTA filtering, in obtaining the results for the same test set.

5. Conclusions

A new and noise robust front-end based on the combined application of spectral subtraction, spectral flooring and cumulative distribution mapping has been proposed. Experimental results on the Aurora II speech database have revealed the effectiveness of the novel combination of these three compensation methods. The proposed front-end achieves an average digit accuracy of 81.46% for the test *A* with the clean HMM set and 89.54% for the multi-condition HMM set. Moreover, it is observed that the use of a single fixed parameter (γ) to control spectral flooring is beneficial only when cumulative distribution mapping is also applied at a later stage of the front-end processing. Based on the proposed front-end configuration, we have also investigated the effect of varying the degree of spectral flooring on the recognition results and found that higher degree of flooring is required for the Aurora database. Possible future extension work includes the use of dynamic noise estimates to handle non-stationary noises, the replacement of the simple spectral flooring with a more advanced temporal masking algorithm, and the use of a different target CDF for the cumulative distribution mapping method.

6. References

Cerisara, C., Rigazio, L. and Junqua, J.C. (2004). *α-Jacobian Environmental Adaptation*. Speech Communication, Vol. 42, Issue 1, Jan., pp. 25-41.

Cui, X., Iseli, M., Zhu, Q. and Alwan, A. (2002). *Evaluation of Noise Robust Features on the Aurora Databases*. Int. Conf. on Spoken Language Processing, Sept, pp. 481-484.

Digalakis, V., Rohlicek, J.R. and Ostendorf, M. (1993). *ML Estimation of a Stochastic Linear System with the EM Algorithm and Its Application to Speech Recognition*. IEEE Trans. Speech and Audio Processing, Vol. 1, No.4, October, pp. 431-442.

Ephraim, Y. (1992). *A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models*. IEEE Trans. Signal Processing, Vol. 40, No. 4, April, pp. 725-735.

ETSI (2000). *Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms*. ETSI standard document ES 201 108, April.

Hermansky, H. (1990). *Perceptual Linear Predictive (PLP) Analysis of Speech*. J. Acoustical Soc. Amer., Vol. 87 (4), pp. 1738-1752.

Hermansky, H., Morgan, N. and Hirsch, G. (1993). *Recognition of Speech in Additive and Convolutional Noise Based on RASTA Spectral Processing*. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Vol. 2, April, pp. 83-86.

Hirsch, H.G. and Pearce, D. (2000). *The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noise Conditions*. Proc. ISCA ITRW ASR2000, Sept., pp. 181-188.

Huang, C.S., Wang, H.C. and Lee, C.H. (2001). *An SNR-incremental Stochastic Matching Algorithm for Noisy Speech Recognition*. IEEE Trans. Speech and Audio Processing, Vol. 9, Issue 8, Nov., pp. 866 – 873.

Ida, M. and Nakamura, S. (2002). *HMM Composition-Based Rapid Model Adaptation using A Priori Noise GMM Adaptation Evaluation on AURORA2 Corpus*. Proc. Int. Conf. on Spoken Language Processing, Sept., pp. 437-440.

Kim, D.S., Lee, S.Y. and Kil, R.M. (1999). *Auditory Processing of Speech Signals for Robust Speech Recognition in Real-World Noisy Environments*. IEEE Trans. Speech and Audio Processing, Vol. 7(1), pp. 55-69.

Lee, K.Y. and Jung, S. (2000). *Time-Domain Approach using Multiple Kalman Filters and EM Algorithm to Speech Enhancement with Nonstationary Noise*. IEEE Trans. Speech and Audio Processing, Vol. 8, No. 3, May, pp. 282-291.

Lieb, M. and Fischer, A. (2001). *Experiments with the Philips continuous ASR system on the AURORA noisy digits database*. Proc. European Conf. on Speech Communication and Technology, Sept., pp. 625-628.

Pelecanos, J.W. (2003). *Robust Automatic Speaker Recognition*. PhD Thesis, Queensland University of Technology, Jan. 2003.

Russ, J.C. (1995). *The Image Processing Handbook*. CRC Press.

Sankar, A. and Lee, C.H. (1996). *A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition*. IEEE Trans. Speech and Audio Processing, Vol. 4, May, pp. 190–202.

Tian, Y., Wu, J., Wang, Z. and Lu, D. (2002). *Robust Noisy Speech Recognition with Adaptive Frequency Bank Selection*. Proc. IEEE Int. Conf. Multimodal Interfaces, Oct., pp. 75 – 80.

Torre, A., Segura, J., Benitez, C., Peinado, A.M. and Rubio, A.J. (2002). *Non-Linear Transformations of the Feature Space for Robust Speech Recognition*. Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, Vol. 1, May, pp. 401-404.

Vaseghi, S.V. (2000). *Advanced Digital Signal Processing and Noise Reduction*. Wiley Press.

Yao, K., Paliwal, K.K. and Nakamura, S. (2001). *Sequential Noise Compensation by a Sequential Kullback Proximal Algorithm*. Proc. European Conf. on Speech Communication and Technology, Sept., pp. 1139-1142.

Young, S., Odell, J., Ollason, D., Valtchev, V. and Woodland, P. (1997). *The HTK Book*. Cambridge University.

Zhang, Z. and Furui, S. (2004). *Piecewise-linear Transformation-based HMM Adaptation for Noisy Speech*. Speech Communication, Vol. 42, Issue 1, Jan., pp. 43-58.

Accepted after full paper review