

A Generalized Framework for Compensation of Mel-filterbank Outputs in Feature Extraction for Robust ASR

Eric H. C. Choi

Interfaces, Machines and Graphic Environments (IMAGEN)
National ICT Australia, Sydney, Australia
Eric.Choi@nicta.com.au

Abstract

This paper describes a novel and efficient noise-robust front-end that utilizes a set of Mel-filterbank output compensation methods, together with cumulative distribution mapping of cepstral coefficients, for noisy speech recognition. The proposed compensation framework includes the use of noise spectral subtraction, spectral flooring and log Mel-filterbank output weighting. Recognition experiments on the Aurora II connected digit database have revealed that the proposed front-end achieves an average digit recognition accuracy of 83.46% for a model set trained from clean data. Compared with the recognition results obtained by using the ETSI standard Mel-cepstral front-end, these results represent a relative error reduction of around 58%.

1. Introduction

State-of-the-art automatic speech recognition (ASR) systems work pretty well if the training and usage 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. To remedy this noise robustness issue in ASR, various adaptive techniques have been proposed. A common theme of these techniques is the utilization 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 to reduce mismatch between training and usage conditions.

Signal-space methods, e.g. [1], 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, e.g. [2], try to derive some kind of feature representation that is invariant to the change in noise conditions. This is often achieved by incorporating some aspects of human auditory modeling. Alternatively, some other feature-space methods [3, 4] try to understand and compensate the effects of noise on a speech representation and correspondingly reduce the mismatch. Model-space methods, e.g. [5, 6], try to adjust the recognition model parameters to incorporate the effects of noise on the models.

In this work, the main focus is on feature-space compensation [4, 7, 8] for a cepstral based front-end. It is demonstrated that a general framework of Mel-filterbank output compensation can be used together with cumulative distribution mapping to compensate the effects of noises. This novel framework includes the use of noise spectral

subtraction, spectral flooring and log Mel-filterbank output weighting.

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

2. Front-end Processing

The development of the proposed front-end processing is based on the ETSI standard Mel-frequency cepstral coefficient (MFCC) front-end [9]. Typically, the MFCC's (C_i) of a frame of speech data are given by:

$$C_i = \sum_{j=1}^M m_j \cos\left[\frac{i\pi}{M}(j-0.5)\right]; \quad m_j = \log_e(Y_j); \quad (1)$$
$$i = 0, 1, 2, \dots, N; \quad N < M$$

where Y_j is the output magnitude of the j -th Mel-filterbank and M is the total number of Mel-filters in the filterbank analysis.

In this work, three more processing blocks related to additive noise compensation have been added to the ETSI standard MFCC front-end. These additional blocks include noise spectral subtraction (SS), spectral flooring (SF) in log-compression and log Mel-filterbank output weighting (LMW) based on signal-to-noise ratio. Moreover, noise robustness is further enhanced by applying cumulative distribution mapping (CDM) to the resultant cepstral coefficients. A detailed description of this novel noise compensation framework is presented as in the following sub-sections.

2.1. Mel-filterbank Output Compensation

The noise robustness of the proposed front-end is enhanced by compensating the Mel-filterbank outputs based on the noise spectral characteristics. In this work, an enhanced log Mel-filterbank output is given by:

$$m_j = \alpha_j \log_e \{1 + \beta_j \text{MAX}[Y_j - \hat{N}_j], \gamma_j Y_j\} \quad (2)$$

where α_j , β_j , γ_j all $\in (0, 1)$ are parameters to adjust the noise compensation, \hat{N}_j is the noise estimate of the j -th Mel-filterbank output and $\text{MAX}[\cdot]$ is a function which returns the maximum value of its arguments.

Note that γ_j is used to control the degree of noise spectral subtraction [10] and β_j is used to control the degree of

spectral flooring [7]. Here, both γ_j and β_j are assumed to be independent of the Mel-filterbank index j as we are more interested in the log Mel-filterbank output weighting and this assumption can simplify the formulation. Also these two parameters are applied globally in that they have the same values for all the speech utterances.

The motivation to incorporate log Mel-filterbank output weighting is to emphasize those filterbank outputs which are found to be more reliable and less affected by the actual noise spectral characteristics. One way to measure the reliability of a filterbank output is the signal-to-noise ratio (SNR). From the viewpoint of psychoacoustics [11], these weighing factors (α_j) are related to the spectral compression process that converts sound intensity into perceived loudness by human. So far in the literature, each of the weighting factors has been assumed to be dependent on its individual output SNR only. However, in our case, a weighting factor is also dependent on the SNR's of other filterbank outputs and it is given by:

$$\alpha_j = \frac{\log_e(1 + \frac{Y_j}{\hat{N}_j})}{\sum_{k=1}^M \log_e(1 + \frac{Y_k}{\hat{N}_k})}; \quad \sum_{j=1}^M \alpha_j = 1 \quad (3)$$

The constant "1" is added to the log function to prevent it from having negative values since there may be errors in the noise estimates. In essence, α_j is basically calculated as the ratio of the SNR of a particular filterbank output to the sum of the SNR's of all the filterbank outputs. Moreover, in this case, all the weighing factors are calculated frame-by-frame dynamically based on the noise estimates from the first 10 frames of each speech utterance.

While equation (2) provides a general framework to perform the noise compensation, it is anticipated that some kind of normalization to the dynamic ranges of the compensated cepstral coefficients would be beneficial. For this purpose, we choose to apply cumulative distribution mapping to the cepstral coefficients after noise compensation.

2.2. Cumulative Distribution Mapping

The cumulative distribution mapping method described here can be traced back to the use of histogram equalisation (HE) in image processing. The use of the HE method for additive noise compensation in front-end processing of speech can also be found in [4, 7, 8]. 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:

$$\int_{v=-\infty}^{v_o} f(v)dv = \int_{z=-\infty}^{z_o} h(z)dz; \text{ or } F_v(v_o) = F_z(z_o) \quad (4)$$

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 this work, the target PDF of z is assumed to be a Gaussian with zero mean and unity variance. In the experiments, CDM is applied only to the individual static feature vector which consists of 13 MFCC's ($C_0 \sim C_{12}$).

3. Experimental Results

The proposed front-end has been evaluated on the Aurora II database [12] with various configurations. This database contains noisy connected digits, which were created by adding various types of noises at different SNR's to the original clean utterances. There are three test sets in the database and they contain 8 types of additive noises. The test set C includes channel distortion as well.

3.1. Experimental Setup

The static feature vector of our front-end consisted of 13 MFCC's ($C_0 \sim C_{12}$). This static 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. Two sets of hidden Markov models (HMM) were trained following the setups provided in the database. 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. Both the training and test data comprise the original data from the Aurora CDs without end-point detection.

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. 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.

Various recognition experiments were performed using different configurations of our proposed front-end, and the results for using only one of the Mel-filterbank output compensation methods are summarized as shown in Table 1.

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
Baseline	58.89	86.81
CDM only	81.67	90.90
SS+CDM	80.90	90.91
SF+CDM	81.21	90.32
LMW+CDM	81.07	90.55
ETSI standard (logE)	61.34	87.82

SS: Spectral subtraction ($\gamma=0.4$)

SF: Spectral flooring ($\beta=0.001$)

LMW: Log Mel-filterbank output weighting

CDM: Cumulative distribution mapping (100 bins)

Note that the 1st-order and the 2nd-order time derivatives of a static feature vector were generated after the static features had been compensated. Also the baseline front-end configuration is basically the same as the ETSI standard MFCC front-end, except that the baseline front-end uses C_0 , instead of log-energy. C_0 was used in the experiments because it works better than log-energy when CDM is applied [7]. In implementation, the target $F_z(z)$ of the CDM was divided into 100 bins and the corresponding z values were stored in a lookup table. Furthermore, both β_j and γ_j were determined empirically based on some preliminary experiments.

Table 2 summarizes the recognition results for the cases where the compensation methods were applied in combination according to the generalized framework. All front-end settings were the same as before.

Table 2: Average digit accuracies (%) for Aurora test set A with two or more Mel-filterbank output enhancements

Front-end Configuration	Clean HMM Set	Multi-condition HMM Set
SS+SF+CDM	82.67	90.20
SF+LMW+CDM	82.53	89.64
SS+LMW+CDM	81.42	90.76
SS+SF+LMW+CDM	83.65	89.82

To get an insight on how the proposed front-end is performing in different noise conditions, a break-down of the recognition results for the front-end configuration SS+SF+LMW+CDM according to individual SNR levels is shown in Figure 1.

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

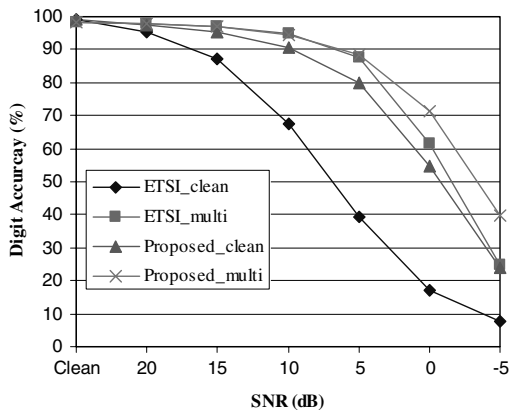


Figure 1: Recognition results for Aurora test set A, SS+SF+LMW+CDM front-end compared with ETSI standard MFCC front-end by SNR

Figure 2 shows the corresponding recognition results for the test set A according to the different noise types. It can be observed that overall the biggest improvement is obtained for the babble-noise type speech, while the best average digit accuracy is obtained for the car-noise type speech by using the proposed front-end.

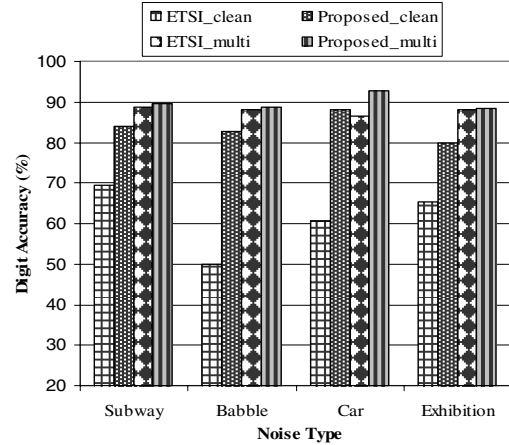


Figure 2: Recognition results for Aurora test set A, SS+SF+LMW+CDM front-end compared with ETSI standard MFCC front-end by noise type

Further recognition results for all the test sets in the Aurora database are shown in Table 3. Again the results were obtained by using the front-end configuration SS+SF+LMW+CDM with the same settings as before.

Table 3: Average digit accuracies (%) for Aurora test sets, SS+SF+LMW+CDM front-end compared with ETSI standard MFCC front-end, clean HMM set

Front-end	Test A	Test B	Test C	Avg.
ETSI	61.34	55.75	66.14	61.08
Proposed	83.65	84.00	82.74	83.46

A detailed break-down of the recognition results by SNR for the test sets B and C is shown in Figure 3. It can be observed that the proposed front-end is much more noise robust than the ETSI standard MFCC front-end.

ETSI_testB: standard front-end, test set B
 ETSI_testC: standard front-end, test set C
 Proposed_testB: SS+SF+LMW+CDM, test set B
 Proposed_testC: SS+SF+LMW+CDM, test set C

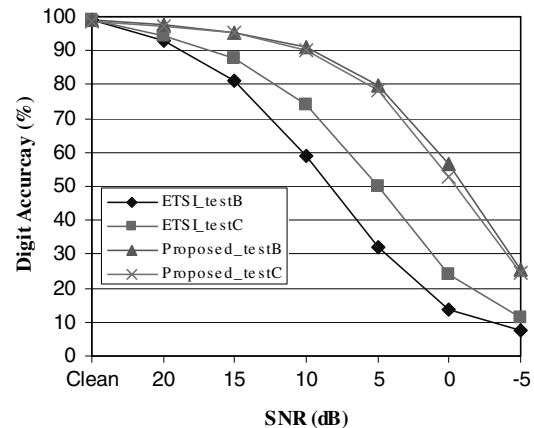


Figure 3: Recognition results for Aurora test sets B and C, SS+SF+LMW+CDM front-end compared with ETSI standard MFCC front-end by SNR

4. Discussion

From Table 1 and Table 2, it can be observed that the front-end configuration SS+SF+LMW+CDM achieves the best recognition accuracy for the clean HMM set (83.65%), while the SS+CDM configuration is the best for the multi-condition HMM set (90.91%). Nevertheless, the CDM only configuration is as good as the SS+CDM one for the multi-condition HMM set. For the results with the clean HMM set, none of the compensation methods alone (SS, SF or LMW) can improve the recognition accuracy further when it is applied individually after the CDM. On the other hand, further improvement in accuracy is found to be possible (from Table 2) if two or more of the compensation methods are added to the front-end with CDM.

From Figure 1 and Figure 3, it can be found that the proposed front-end configuration achieves better recognition accuracy than that of the ETSI standard front-end at almost every SNR level. 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, as shown in Figure 1. For the results with multi-condition training in this figure, we can observe the superiority of the proposed end in handling unseen noise conditions. When the noise conditions of the test data are not encountered in the multi-condition training (i.e. the 0 and -5 dB SNR noises), the degradation in accuracy for the ETSI standard front-end is observed to be higher.

With reference to the ETSI standard MFCC front-end, our novel front-end achieves a relative error reduction of around 58% across all the three test data sets (83.46% vs. 61.08%) with the clean HMM training. These results compare favourably with those reported in [13] which utilized more than seven different compensation methods, including explicit end-point detection, spectral subtraction and RASTA filtering, in obtaining the results for the same test sets.

Another comparison can be made with the more noise robust ETSI advanced front-end. As reported in [8], the ETSI advanced front-end achieved an average digit accuracy of 85.38% across the three test sets. Although there is a performance gap of around 2% digit accuracy, the computation demand of our proposed front-end is much lower. An informal evaluation on processing speed carried out by us has revealed that, on average, the ETSI advanced front-end is about three times slower than our front-end in processing an utterance on a computer with 2.66 GHz CPU and 2 GB RAM. The much lighter computation requirement of our front-end can be a distinguished advantage for applications running on handheld devices. Moreover, the proposed front-end is easier to be implemented on fixed-point processors.

5. Conclusions

A new and noise robust front-end based on the combined application of spectral subtraction, spectral flooring, log Mel-filterbank output weighting 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 compensation methods. The proposed front-end achieves an average digit accuracy of 83.46% for the three test sets with clean HMM training. 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.

6. References

- [1] Ephraim, Y., "A Bayesian Estimation Approach for Speech Enhancement Using Hidden Markov Models", *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725-735, April 1992.
- [2] Hermansky, H., "Perceptual Linear Predictive (PLP) Analysis of Speech", *JASA*, vol. 87 (4) , pp. 1738-1752, 1990.
- [3] Sankar, A. and Lee, C.H., "A Maximum Likelihood Approach to Stochastic Matching for Robust Speech Recognition", *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 190-202, May 1996.
- [4] Torre, A., Segura, J., Benitez, C., Peinado, A.M. and Rubio, A.J., "Non-Linear Transformations of the Feature Space for Robust Speech Recognition", *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Vol. 1, May 2002, pp. 401-404.
- [5] Yao, K., Paliwal, K.K. and Nakamura, S., "Sequential Noise Compensation by a Sequential Kullback Proximal Algorithm", *Proc. EUROSPEECH*, pp. 1139-1142, 2001.
- [6] Zhang, Z. and Furui, S., "Piecewise-linear Transformation-based HMM Adaptation for Noisy Speech", *Speech Communication*, vol. 42, iss. 1, pp. 43-58, Jan. 2004.
- [7] Choi, E., "Noise Robust Front-end for ASR using Spectral Subtraction, Spectral Flooring and Cumulative Distribution Mapping", *Proc. 10th Australian Int. Conf. on Speech Science and Technology*, Dec. 2004, pp. 451-456.
- [8] Tsai, S. and Lee, L., "A New Feature Extraction Front-end for Robust Speech Recognition Using Progressive Histogram Equalization and Multi-eigenvector Temporal Filtering", *Proc. Int. Conf. on Spoken Language Processing*, Oct. 2004, pp. 165-168.
- [9] ETSI, "Speech Processing, Transmission and Quality Aspects (STQ); Distributed Speech Recognition; Front-end Feature Extraction Algorithm; Compression Algorithms", *ETSI standard document ES 201 108*, April 2000.
- [10] Vaseghi, S.V., *Advanced Digital Signal Processing and Noise Reduction*, Wiley Press, 2000.
- [11] Stevens, S.S., "On the Psychological Law", *Psychological Review*, Vol. 64, 1957, pp. 153-181.
- [12] Hirsch, H.G. and Pearce, D., "The AURORA Experimental Framework for the Performance Evaluation of Speech Recognition Systems Under Noise Conditions", *Proc. ISCA ITRW ASR2000*, Sept. 2000, pp. 181-188.
- [13] Cui, X., Iseli, M., Zhu, Q. and Alwan, A., "Evaluation of Noise Robust Features on the Aurora Databases", *Proc. Int. Conf. on Spoken Language Processing*, Sept. 2002, pp. 481-484.