A Novel Front-end Based on Variable Frame Rate Analysis and Mel-filterbank Output Compensation for Robust ASR

Eric H.C. Choi and Julien Epps

ATP Research Laboratory
National ICT Australia, Sydney, Australia
Eric.Choi@nicta.com.au, Julien.Epps@nicta.com.au

Abstract

For automatic speech recognition (ASR) systems, robustness in the presence of various types and levels of environmental noise remains an important issue, despite the various advances of recent years. This paper describes a new noise-robust ASR front-end employing a combination of variable frame rate processing based on the sample-by-sample delta energy parameter, Mel-filterbank output compensation and cumulative distribution mapping. Recognition experiments on the Aurora II connected digits database reveal that the proposed front-end achieves an average digit recognition accuracy of 84.3% for a model set trained from clean speech data. Compared with the ETSI standard Mel-cepstral front-end, the proposed front-end is found to obtain a relative error rate reduction of around 60%. Moreover, the proposed front-end can provide almost comparable recognition accuracy with the ETSI advanced front-end, at roughly half the computational complexity.

1. Introduction

Under the influence of noise, the performance of even state-of-the-art automatic speech recognition (ASR) system begins to degrade, and its accuracy can become unacceptably low in severe environments (i.e. low signal-to-noise ratio). Various adaptive techniques have been proposed to remedy this noise robustness issue in ASR, often using of some form of compensation to account for the effects of noise on the speech characteristics. Typical approaches employed by other researchers to improve ASR robustness include pre-enhancing the noisy speech (Ephraim, 1992), feature-space compensation of clean/noisy feature mismatch (Hermansky, 1990; Sankar and Lee, 1996), and model-space methods that account for the effects of noise in the speech models (Yao, Paliwal and Nakamura, 2001; Zhang and Furui, 2004).

In this work, the main focus is on feature-space compensation for a cepstral based front-end. Previous research efforts into varying the frame shift between extracting consecutive features has shown promise (Pointing and Peeling, 1991), and here we present a new, computationally efficient approach. While variable frame rate processing addresses the issue of speech signal information redundancy in the temporal domain, the proposed front-end also integrates a novel framework for emphasizing speech frequency content that is less susceptible to noise effects. Compared with previous work (Segura et. al, 2002), this generalized Mel-filterbank output compensation framework provides a more flexible and effective means for compensating noisy speech.

2. Proposed Front-end

The proposed front-end employs a series of processing steps that aim to reduce the effects of noise. These steps include variable frame rate (VFR) processing, Mel-filterbank output compensation (MOC) and cumulative distribution mapping (CDM). The motivation for this scheme is the belief that VFR can remove noisy speech frames that are made similar to each other due to the noise flooring effect and MOC can reduce the effects of noise in different frequency bands within a frame of speech. In addition, CDM normalizes cepstral coefficients, to help compensate other noise effects not accounted for by the previous two steps.

2.1. Variable Frame Rate Processing

The concept of a variable frame rate for speech processing has been proposed previously (Pointing and

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

Peeling, 1991), and can be motivated by the following considerations: (i) To date, no physical mechanism has yet been identified that would support a fixed-duration analysis frame in the human auditory system; (ii) Any improvements that can be gained by adjusting the analysis frame positions will be highly independent of other front-end enhancements, and thus should provide a cumulative advantage; and (iii) The non-stationary nature of many noises means that small adjustments in frame position may result in different signal to noise ratios. Early (Pointing and Peeling, 1991) and subsequent VFR analysis techniques have employed a Euclidean distance between consecutive feature vectors, however this comes at the cost of needing to pre-compute a number of feature vectors in order to determine the frame shift.

Recently, we proposed a new variable frame rate-based front end that employs the delta energy parameter as a criterion for determining the size of the frame shift (Epps and Choi, 2005). This offers two advantages: (i) delta energy can be calculated very efficiently relative to the full Mel-frequency cepstral coefficient (MFCC) vector, and (ii) this permits a fine, sample-by-sample search for the optimum frame shift estimate. In this approach, the relative position of the next frame $\hat{k}$ (in samples) is found by maximizing the difference in log energy between the current frame and possible next frame, so that

$$\hat{k} = \underset{K_{\min} \le k \le K_{\max}}{\arg\max} \frac{\left|\log\left(E_{m+1}(k)\right) - \log\left(E_m\right)\right|}{k}, \qquad (1)$$

where k is the candidate frame advance relative to the current frame position in samples, E_m is the energy of the current frame, $E_{m+1}(k)$ is the energy of the next frame, m is the current frame index, and K_{min} and K_{max} are respectively the minimum and maximum admissible values of frame advance in samples. The divisor k arises from the first-order approximation to the energy derivative. More perceptually relevant formulations of the search criterion (1) could be contemplated. A fast method for generating the sample-by-sample energy $E_{m+1}(k)$ was also given in (Epps and Choi, 2005).

In section 3.2, we provide more insight into the effect that noise has on the estimated frame shift $\hat{k}$, and also investigate the use of variable frame rate processing together with other noise compensation methods, to improve the overall front-end noise robustness.

2.2. Mel-filterbank Output Compensation

The noise robustness of the proposed front-end is enhanced by compensating the Mel-filterbank outputs based on the speech and noise spectral characteristics. In this work, an enhanced log Mel-filterbank output (L_j) is given by (Choi, 2005):

$$L_j = \alpha_j \log_e\{1 + \beta_j MAX[(Y_j - \hat{N}_j), \gamma_j Y_j]\}; \quad 1 \le j \le M \qquad (2)$$

where α_j, β_j, $\gamma_j \in (0,1)$ are parameters to adjust the noise compensation, Y_j is the magnitude of the j-th Mel-filterbank output, $\hat{N}_j$ is the corresponding noise estimate, M is the total number of Mel-filters and $MAX[.]$ is a function that returns the maximum value of its arguments.

In our case, γ_j and β_j are determined empirically (optimum values are given in Section 3.3) and they are assumed to have the same values for all the noise conditions. Here, it is assumed that γ_j and β_j are independent of the Mel-filter index j, as we are more interested in the log Mel-filterbank output weighting and this assumption can simplify the formulation.

The motivation for incorporating log Mel-filterbank output weighting is to emphasize those filterbank outputs that 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 perspective of psychoacoustics (Stevens, 1957), these weighing factors (α_j) are related to the spectral compression process that converts sound intensity into the loudness as perceived by humans. In the literature to date, each of the weighting factors has been assumed to be dependent on its individual output SNR only. However, in our case, the weighting factors are also dependent on the SNRs of other filterbank outputs, and are given by:

$$\alpha_j = \frac{\log_e(1 + \frac{Y_j}{\hat{N}_j})}{\sum_{r=1}^{M} \log_e(1 + \frac{Y_r}{\hat{N}_r})}; \qquad \sum_{j=1}^{M} \alpha_j = 1 \qquad (3)$$

In essence, α_j is calculated as the ratio of the SNR of a particular filterbank output to the sum of the SNRs of all the filterbank outputs. Moreover, in this case, all the weighing factors are calculated dynamically frame-by-frame, 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 of the dynamic ranges of the compensated cepstral coefficients would be beneficial. For this purpose, we choose to apply cumulative distribution mapping (CDM) to the cepstral coefficients (C_0~C_{12}) after noise compensation, as detailed in (Choi, 2004).

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

3. Experimental Results

3.1. Experimental Setup

Various configurations of the proposed front-end were evaluated on the Aurora II database. All pre-processing and Mel filtering of speech signals followed the ETSI standard MFCC front-end, except that C_0 was used here instead of log-energy. Each model was represented by a continuous density hidden Markov model (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. Two sets of HMMs 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 noise-added version of the same training data.

3.2. Investigation of Variable Frame Rate Processing

Although previous work has examined the recognition accuracy of the delta energy-based variable frame rate processing approach (Epps and Choi, 2005), it is instructive to investigate the variation in frame shift produced under different noise conditions. Since the frame shift calculated from equation (1) will vary with any modification to the speech data (e.g. addition of noise), in our experiments, the average frame shift for a single utterance is used as the basic measurement.

To obtain an idea of the variation in average frame shift across a number of different utterances, the distribution of average frame shift from the 1001 clean utterances for the 'Car' condition in the Aurora test set *A* was determined, and this is shown in Figure 1(a). Here the average over all utterances was 11.85 ms, and the variance was narrow relative to the range of the upper and lower bounds K_{min} and K_{max}, which were set to 8.75 and 16.75 ms respectively.

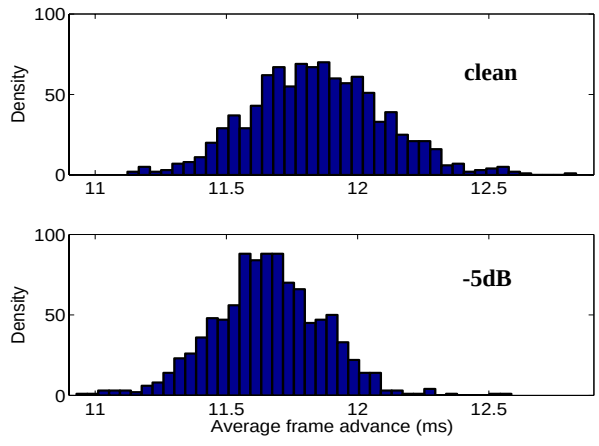

Figure 1: Distribution of average frame shift for (a) the 1001 clean utterances and (b) the same 1001 utterances at -5 dB SNR from the 'Car' condition of Aurora test set *A*.

Within each utterance, the frame shift was generally observed to be biased towards the lower bound, an effect produced by the 1/*k* factor in equation (1). Without the 1/*k* factor, the frame shift was generally observed to be biased towards the upper bound instead. Experiments on VFR without the 1/*k* factor showed that recognition accuracies close to those arising from the use of equation (1) could be obtained, providing that new values for the upper and lower bounds K_{min} and K_{max} were determined empirically.

The distribution for the same utterances but with -5 dB SNR car noise added is shown in Figure 1(b). Here the size of the frame shift has been reduced from the clean case (Fig. 1(a)) by the addition of severe noise, however only by about 0.2 ms overall, to a new average of 11.66 ms. Similar comparisons were made for all four noise types and all six SNRs of the test set *A*, however the average difference in average frame shift was consistently less than 0.2 ms, with similar distributions to that of Figure 1(b).

3.3. Results with Various Front-End Configurations

In keeping with the conventional reporting of Aurora II results, the average recognition accuracies to be reported throughout this section are the mean accuracies over the 0 dB to 20 dB SNR conditions. Results for various front-end combinations evaluated on the Aurora test set *A* are shown in Table 1. Note that CDM was applied on a per-utterance basis and the 1st and 2nd order time derivatives of static feature vectors were generated after the static features had been compensated and normalized.

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
CDM only	81.67	**90.90**
VFR+CDM	82.12	90.80
MOC+CDM	83.65	89.82
VFR+MOC+CDM	**84.41**	89.57
ETSI standard (logE)	61.34	87.82

CDM: Cumulative distribution mapping (100 bins)
VFR: Variable frame rate (K_{min}:K_{max} = 8.75:16.75ms)
MOC: Mel-filterbank output compensation (β_j=0.001, γ_j=0.4)

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

As can be observed from Table 1, the front-end configuration VFR+MOC+CDM achieved the best accuracy with the clean HMM set, while for multi-condition HMM set, the CDM only configuration provided the best accuracy. It can also be observed that for using the clean HMM set, applying either VFR or MOC individually before CDM improved the accuracy over the use of CDM only.

To gain an insight into how the proposed front-end performs in different noise conditions, a break-down of the recognition results for the front-end configuration VFR+MOC+CDM according to individual SNR levels is shown in Figure 2. From this figure, it can be observed that the proposed front-end achieved better recognition accuracy than that of the ETSI standard MFCC front-end at almost every SNR level. In addition, it was found to be far more robust than the ETSI front-end in mismatched training/testing conditions.

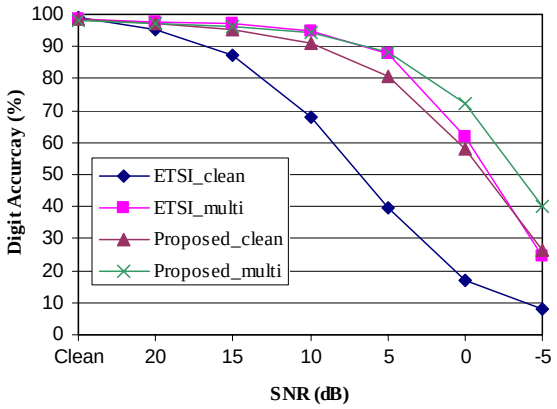

Figure 2: Recognition results for Aurora test set *A* grouped by SNR, showing the proposed front-end (VFR+MOC+CDM) compared with the ETSI standard MFCC front-end, clean HMM set (_clean) and multi-condition HMM set (_multi).

Figure 3 shows the corresponding recognition results for the test set *A* broken down according to the different noise types. It can be observed that overall the greatest improvement was obtained for the babble-noise type speech, while the best average digit accuracy was obtained for the car-noise type speech using the proposed front-end. Interestingly, for the car noise conditions, the proposed front-end with a clean HMM set was able to provide better accuracy than the ETSI standard front-end with a multi-condition HMM set. This better accuracy (88.39% vs. 86.52%) is found to be statistically significant (z=5.59, p<0.001, two tailed).

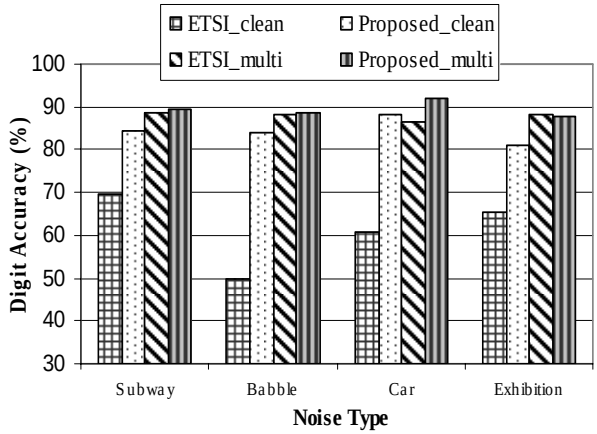

Figure 3: Recognition accuracy averaged over 0 to 20 dB SNRs from Aurora test set *A*, showing the proposed front-end compared with the ETSI standard MFCC front-end by noise type.

Further recognition results for the three Aurora test sets with the use of a clean HMM set are shown in Table 2. For the proposed front-end, the VFR+MOC+CDM configuration with the same settings as previously was used. Overall, the proposed front-end was found to be much more robust than the ETSI standard MFCC front-end in various noisy conditions. The difference in accuracy (84.29% vs. 61.08%) is found to be statistically significant (z=149.74, p<0.001, two tailed). Compared with the more contemporary ETSI advanced front-end (ETSI, 2002), the average accuracy of the proposed front-end was found to be marginally lower. In this case, the difference in accuracy (84.29% vs. 85.39%) is also statistically significant (z=8.80, p<0.001, two tailed).

Table 2: Average digit accuracies (%) for Aurora test sets, comparing the proposed front-end with the ETSI MFCC front-ends, clean HMM set.

Front-end	Test *A*	Test *B*	Test *C*	Avg.	% * Improv
ETSI std.	61.34	55.75	66.14	61.08	0.0
ETSI adv.	86.20	85.24	84.72	85.39	62.5
Proposed	84.41	84.89	83.58	84.29	59.6

* Improvement measured in % of relative error rate reduction reference to the ETSI standard front-end

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review

A detailed break-down of the average recognition results across all three test sets by SNR level is shown in Figure 4. Again it can be observed that the accuracy of the proposed front-end closely matches that of the ETSI advanced front-end at every SNR level.

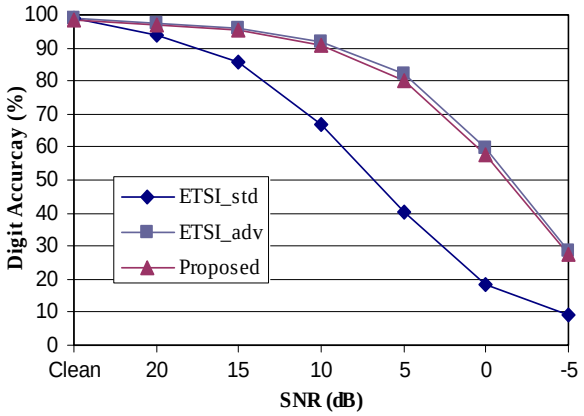

Figure 4: Recognition accuracy averaged across all 3 test sets, showing the proposed front-end compared with the ETSI front-ends by SNR, clean HMM set.

3.4. Complexity Comparison

In order to estimate the complexity of the proposed front-end processing, the ETSI standard, the ETSI advanced and the proposed front-end were run on the clean training data (8440 utterances), and the duration was recorded. No other processes were running on the processor at the time. On average, the computational load of the proposed front-end (running time 171s) was found to be about 30% more than that of the ETSI standard MFCC front-end (133s), but only about half that of the ETSI advanced front-end (334s). The higher computational load of the ETSI advanced front-end is expected, as the advanced front-end applies Wiener filtering twice to a speech signal based on time-domain convolution.

4. Conclusion

A new noise-robust ASR front-end that combines a variable frame rate approach, Mel-filterbank output compensation and a cumulative distribution mapping technique has been presented. Based on experiments with the Aurora II database, this combination shows promise for improving recognition of speech in the presence of a range of different noise types and levels, although cumulative distribution mapping remains the most effective component of the overall front end. For a clean model set, the proposed front end was shown to achieve recognition accuracies close to those of the ETSI advanced front-end, at roughly half of the computational complexity. Future research will focus on the use of dynamic noise estimates to handle non-stationary noise and the development of a robust voice activity detector.

5. References

Choi, E. (2004). 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*, pp. 451-456.

Choi, E. (2005). A generalized framework for compensation of Mel-filterbank outputs in feature extraction for robust ASR. *Proc. 9th European Conference on Speech Communication and Technology*, Lisbon, pp. 933-936.

Ephraim, Y. (1992). A Bayesian estimation approach for speech enhancement using hidden Markov models. *IEEE Trans. Signal Processing*, vol. 40, no. 4, pp. 725-735.

Epps, J., & Choi, E. (2005). An energy search approach to variable frame rate front-end processing for robust ASR. *Proc. 9th European Conference on Speech Communication and Technology*, Lisbon, pp. 2613-2616.

ETSI (2002). Speech processing, transmission and quality aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithm. *ETSI standard document ES 202 050*.

Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. *Journal of the Acoustical Society of America*, vol. 87, pp. 1738-1752.

Pointing, K. M., & Peeling, S. M. (1991). The use of variable frame rate analysis in speech recognition. *Computer Speech and Language*, vol. 5, no. 2, pp. 169–179.

Sankar, A. & Lee, C. H. (1996). A maximum likelihood approach to stochastic matching for robust speech recognition. *IEEE Trans. Speech and Audio Processing*, vol. 4, pp. 190–202.

Segura, J. C., Benitez, M. C., Torre, A. & Rubio, A. J. (2002). Feature extraction combining spectral noise reduction and cepstral histogram equalization for robust ASR. *Proc. International Conference on Spoken Language Processing*, vol. 1, pp. 225-228.

Stevens, S. S. (1957). On the psychological law. *Psychological Review*, vol. 64, pp. 153-181.

Yao, K., Paliwal, K. K. & Nakamura, S. (2001). Sequential noise compensation by a sequential Kullback proximal algorithm. *Proc. 7th European Conference on Speech Communication and Technology*, pp. 1139-1142.

Zhang, Z. & Furui, S. (2004). Piecewise-linear transformation-based HMM adaptation for noisy speech. *Speech Communication*, vol. 42, issue 1, pp. 43-58.

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review