A NEW APPROACH TO VARIABLE FRAME RATE FRONT-END PROCESSING FOR ROBUST SPEECH RECOGNITION

Julien Epps
IMAGEN Program, National ICT Australia
Locked Bag 9013, Alexandria NSW 1435 Australia
julien.epps@nicta.com.au

ABSTRACT

Robustness in the presence of various types and severities of environmental noise has been researched extensively over the past several years, however this remains one of the main problems facing automatic speech recognition systems. This paper describes a noise-robust ASR front-end that employs a new variable frame rate analysis, based upon the first-order difference of the log energy for each frame (ΔE). Compared with previous variable frame rate methods, this delta energy approach is simpler and achieves similar recognition accuracy improvements but at reduced complexity. Recognition experiments on the Aurora II connected digits database reveal that the proposed front-end achieves an average digit recognition accuracy of 68.69% for a model set trained from clean data and 85.82% for a model set trained from data with multiple noise conditions. Compared with the ETSI standard Mel-cepstral front-end, the proposed front-end obtains a relative error rate reduction of around 20% for the clean model set, achieved consistently across nearly all signal-to-noise ratios and noise conditions tested.

1. INTRODUCTION

Substantial improvements to the performance of automatic speech recognition (ASR) systems have been seen in recent years, however these are often undermined in practical conditions, due to the presence of environmental noise. In this case, the deterioration in recognition accuracy as the signal-to-noise ratio (SNR) approaches 0 dB renders all but the smallest vocabulary ASR systems unusable. Numerous approaches have been employed by other researchers to improve ASR robustness, including pre-enhancing the noisy speech [1], feature-space compensation of clean/noisy feature mismatch [2], and model-space methods that account for the effects of noise in the speech models [3].

One approach, variable frame rate analysis [4], is based upon the assumption that the fixed frame rate employed by nearly all ASR systems is merely a convenient approximation, the legacy of fixed-frame speech processing applications such as speech coding. The unlikely use of such a fixed-duration approach by the human auditory system has motivated various variable frame rate (VFR) ASR front-ends, all of which report improvements over their fixed-frame counterparts.

Pointing and Peeling [4] used a Euclidean distance between consecutive feature vectors, an approach that has subsequently been shown [5] to outperform the more recently proposed feature vector time-derivative [6]. To date the most effective approach appears to be one exploiting the entropy of the feature vector [7], however all these approaches come at the cost of requiring that feature vectors be pre-computed before any variation to the frame rate is applied.

In this work, a new method for determining the instantaneous frame rate (or frame advance) is introduced based upon the change in log energy. It is demonstrated that this method can produce good front-end compensation for the effects of additive noise at lower complexity than alternative methods.

The organization of this paper is as follows. Details of the proposed VFR front-end processing are given in Section 2, and related recognition experiments on the Aurora II digits database are described in Section 3. Following this is a discussion of the findings in Section 4 and a summary of the conclusions in Section 5.

2. DELTA ENERGY-BASED VARIABLE FRAME RATE ANALYSIS

2.1. A New Criterion for Frame Rate Variation

Variable frame rate analysis relies upon some criterion to determine at what point a new feature should be extracted. Intuitively, new features should be extracted only after sufficient changes have occurred within the speech signal to warrant their extraction. Previously [4, 5, 6, 7, 10], some kind of feature-based measure has been compared against a pre-determined threshold in order to determine whether a particular feature should be retained or not, however in principle any criterion that yields similar discriminating power can be used.

Previous detailed investigations [8] into the properties of components of the standard Mel-cepstral front-end have shown that the first-order difference in frame-to-frame energy, ΔE, provides greater discriminative power than any other component of the Mel-frequency cepstral coefficients (MFCCs). Conveniently, it can also be computed without needing to calculate any other components of the MFCCs, i.e. without the discrete Fourier transform, Mel filter bank and discrete cosine transform. For this reason, it is expected to have significantly lower complexity than previous VFR schemes.

0-7803-9243-4/05/$20.00 ©2005 IEEE

Thus, the criterion employed in this paper is to retain the current frame if the change in energy ΔE is greater than a fixed threshold T, and discard it if $\Delta E < T$, where

$$\Delta E = E_m - E_{m-1}, \quad \text{and} \quad E_m = \log\left[\sum_{n=0}^{N} x_m^2(n)\right], \qquad (1)$$

m is the frame number, N is the frame length and $x_m(n)$ is the n'th sample of speech in the m'th frame.

In order to arrive at candidate values of ΔE, some initial fixed spacing between frames, or 'base' frame rate, is required. In the ETSI standard front-end [9], this is 100 frames per second, i.e. a 10 ms frame shift. Section 3.2 investigates the effect of different base frame shifts upon the accuracy of the proposed delta-energy VFR approach. For the different base frame rates arising from this approach, it is also reasonable to consider whether the frame length N needs to be adjusted according to the base frame rate, and this question is examined experimentally in section 3.3.

2.2. Delta Energy Calculation

Two methods for calculating ΔE are possible for a VFR approach:

§ ΔE is based on the log energy difference between consecutive frames with a fixed spacing.

§ ΔE is based on the log energy difference between the last retained frame and the current frame.

The recognition performance of each of these two methods is investigated in section 3.4.

3. EXPERIMENTAL RESULTS

The proposed front-end was evaluated on the Aurora II database, which contains noisy connected digits created by adding various types and intensities of noises to the original clean utterances. The types of noises include subway, babble, car, and exhibition noise. Model sets can alternatively be trained on clean or multi-condition data, so that evaluation can be performed under mismatched and matched training/testing conditions respectively. There are three test sets in the database, and in this work, evaluation is performed using test set A. The SNRs of the test data range from -5 dB to more than 20 dB, while the training data SNRs range from 5 dB to more than 20 dB. In keeping with the conventional reporting of Aurora II results, average recognition accuracies throughout this section represent the mean accuracies over the 0 dB to 20 dB conditions.

3.1. Experimental Setup

All the pre-processing and Mel filtering of speech signals followed the ETSI standard MFCC front-end [9]. The Hidden Markov Model (HMM) Toolkit (HTK) was used for the speech recognition experiments. Each model was represented by a continuous density HMM with a 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 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 the noise-added version of the same training data. All the test and training data were obtained from the original Aurora II CDs without end-point detection.

3.2. Results for Different Base Frame Rates

In this experiment, both the base frame rate (or equivalently, the base frame shift) and the threshold T were varied, and model sets were trained on clean speech data. The resulting model sets were then applied to the entire test set A to determine the accuracies. For each base frame shift tested (5, 7.5, 8.75, 10 and 11.25 ms) the optimum threshold T was estimated. The resulting accuracies are shown in figure 1, plotted against the threshold T.

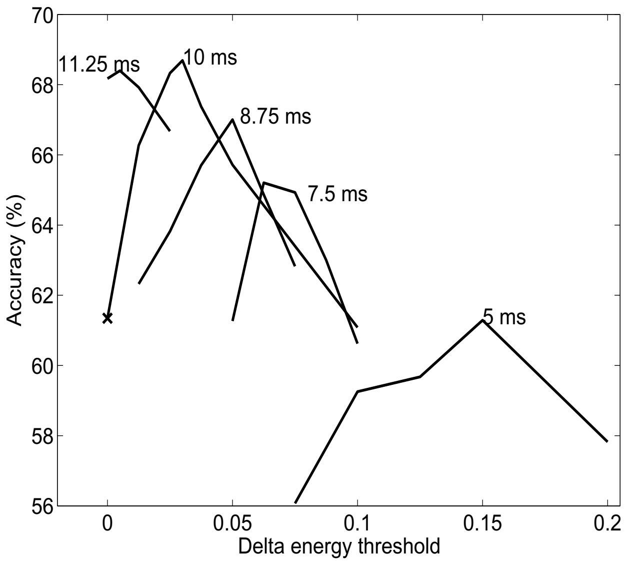

Figure 1. Recognition results from Aurora test set A (clean models), showing the accuracies resulting from different base frame shifts in the range 5 to 11.25 ms as marked vs. the threshold T, compared with the ETSI standard front-end, marked 'x'.

The same accuracy results are shown again in figure 2, plotted against the resulting average frame shift. Here there is a clear relationship between the optimum threshold T and the resulting average frame shift, irrespective of the base frame rate. In general, better recognition accuracies were obtained for base frame rates similar to the ETSI standard of 10 ms, but when thresholding to increase the average frame shift from 10 ms to the range 10.5 to 11.5 ms was applied, significant improvements in accuracy were attained. In both figures 1 and 2, ΔE is based on the log energy difference between the last retained frame and the current frame.

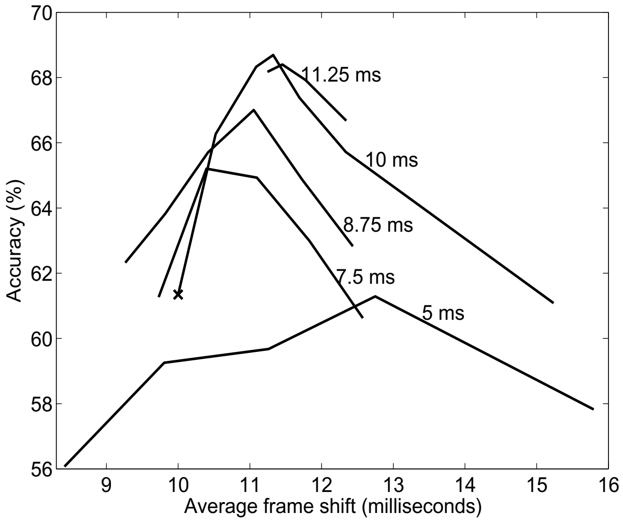

Figure 2. Recognition results from Aurora test set A (clean models), showing the accuracies resulting from different base frame shifts in the range 5 to 11.25 ms vs. the frame shift (averaged over all frames in the test set), compared with the ETSI standard front-end, marked ‘x’.

3.3. Results for Different Frame Lengths

The recognition accuracies reported in section 3.2 are all based on a frame length of 25 ms, resulting in a varying ratio of frame shift to frame length. In order to determine whether the frame length needs to be varied with the frame shift, the ratio of frame shift to frame length was held constant (at 0.4), and the optimum thresholds for three different frame lengths were determined, as seen in Table 1. Note that here again *ΔE* is based on the log energy difference between the last retained frame and the current frame.

Table 1. Average digit recognition accuracies (%) over Aurora test set A (clean models) for various frame lengths.

Frame shift (ms)	7.5	10	11.25
Frame length (ms)	18.75	25	28.13
Optimum T	0.0625	0.03	0.025
Accuracy (%)	64.21	68.69	64.24

3.4. Results for Different Delta Energy Calculation Methods

The two different delta energy calculation methods described in section 2.2 were compared, as shown in Table 2, for the optimum experimental configuration found from section 3.2, i.e. for $T = 0.03$ and a base frame shift of 10 ms.

3.5. Results for Different Signal to Noise Ratios

For the optimum front-end configuration obtained in the previous experiments ($T = 0.03$, base frame shift 10 ms), model sets were then trained on both clean and multi-condition data, and the resulting average accuracies are shown for each SNR in figure 3.

Table 2. Average digit recognition accuracies (%) over Aurora test set A (clean models) for two different delta energy calculation methods.

Method	Diff. between consecutive frames	Diff. between last retained and current frames
Accuracy (%)	68.53	68.69

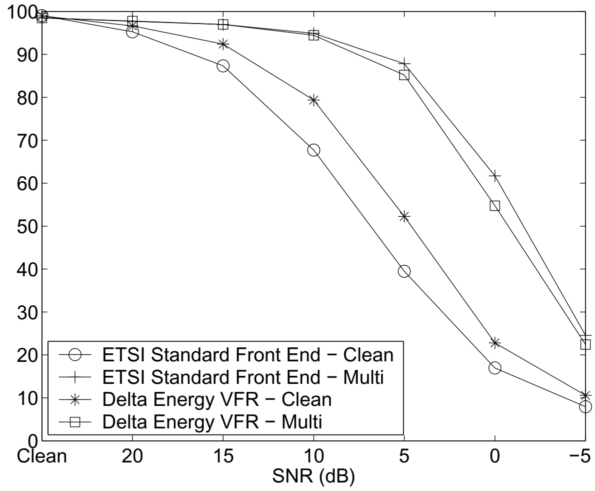

Figure 3. Recognition results from Aurora test set A (clean models), showing the proposed *ΔE* VFR-based front-end compared with the ETSI standard front-end.

3.6. Complexity Comparison

In order to estimate the complexity of the proposed variable frame rate scheme, the ETSI standard and proposed front-ends were run on the clean and multi-condition training data and the test data (a total of 45000 files), and the duration was recorded. Again, the optimum front-end configuration from previous experiments ($T = 0.03$, base frame shift 10 ms) was used, and no other processes were running on the processor at the time.

Table 3. Comparison of running times for front-end processing of both the training data and test set A from Aurora II, on a 2.66 GHz processor with 2 GB RAM.

Front-end	ETSI standard	Proposed *ΔE* VFR
Time (s)	778	761

4. DISCUSSION

From the results in section 3.2, it becomes evident that there is an optimum base frame rate (10 ms in the case of Aurora II) for the proposed *ΔE* VFR approach. It is suggested that below this rate (e.g. for an 11.25ms frame shift), recognition accuracy drops off due to insufficient data, while above this rate (e.g. for a 5ms shift) larger thresholds are needed that appear to discard some features containing important information. Further, from figure 2, there is also an optimum average frame shift, around 11 ms in these experiments, for which nearly any given *ΔE* VFR configuration achieves its maximum accuracy.

From the experiments of section 3.3, varying the frame size according to the frame shift does not appear to produce any advantage. The different *ΔE* calculation methods listed in section 2.2 appear to provide extremely similar performance. The complexity improvement of 2% of *ΔE* VFR over the ETSI standard front-end compares well with previous VFR schemes [4, 5, 6, 7, 10], which have required candidate all MFCCs to be pre-computed, presumably resulting in complexities significantly greater than that of the ETSI standard front-end.

Compared with earlier VFR approaches using model sets trained on clean Aurora II data, *ΔE* VFR (68.69%) yields comparable, but slightly poorer performance averaged over all SNRs than the energy-weighted Euclidean MFCC distance [10] (70%) and entropy-based MFCC [7] (71.54%) VFR schemes. The energy-weighted Euclidean MFCC distance [10] and entropy-based MFCC [7] VFR schemes both produce poor accuracy under clean conditions, their word error rates degrading by 42% and 89% [7] respectively relative to the ETSI front-end. By contrast, *ΔE* VFR produces a more acceptable degradation of 11% in relative error rate for the clean condition.

The *ΔE* VFR accuracy improvements over the ETSI front-end for a clean model set are consistent across all SNRs except clean (cf. section 3.5), and further investigation showed that improvements in accuracy were found consistently across all noise types. Interestingly, for the multi-condition model set *ΔE* VFR yielded poorer accuracies than the ETSI front-end, with the schemes averaging 85.82% and 87.82% across all SNRs respectively. This is possibly due to the *ΔE* VFR algorithm rejecting some frames that contain relevant spectral shape information even when there is very little change in inter-frame energy, and it would be interesting to know whether this result generalizes to other VFR approaches. Although You *et al.* [7] also performed Aurora II evaluations on their VFR schemes, they did not give results for a multi-condition model set.

While the exact parameter values are expected to vary from one corpus to the next, the characteristics observed in these experiments are conjectured to hold for other corpora. Since the new *ΔE* VFR approach provides an improvement based on the temporal characteristics of the speech signal, it is suggested that the performance gains observed in section 3 would be largely independent of performance gains from many other front-end improvements in the literature. This implies that the combination of *ΔE* VFR with any pre-enhancement, feature-space and/or model-space techniques (as mentioned in section 1) has good prospects for further improvements in recognition accuracy.

5. CONCLUSION

A noise-robust variable frame rate (VFR) speech recognition front-end based upon the delta energy component of the Mel frequency cepstral coefficients has been presented. Based on experiments with the Aurora II database, *ΔE* VFR analysis shows promise for improving recognition of speech in the presence of babble, car or exhibition noise, particularly where the test data are noisier and more diverse than the training data. Further, this approach is faster than existing methods for VFR speech analysis as it does not require candidate features to be pre-computed. For the Aurora II test set A, the proposed front-end obtains a relative error rate reduction of around 20% for the clean model set, while a degradation of around 15% in relative error rate was found for the multi-condition model set. Future research will focus on exploiting the ease of computing *ΔE* on a sample-by-sample basis for a search-based VFR approach, on investigating means for improving the multi-condition performance of the proposed scheme, and on the possibilities for combining *ΔE* VFR with other robust front-end techniques.

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, 1992.

[2] 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.

[3] 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.

[4] Pointing, K. M., and Peeling, S. M., “The use of variable frame rate analysis in speech recognition,” *Computer Speech and Language*, vol. 5, no. 2, pp. 169–179, April 1991.

[5] Macias-Guarasa, J., Ordonez, J., Montero, J. M., Ferreiros, J., Cordoba, R., and Haro, L. F. D, “Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition,” in *Proc. EUROSPEECH*, pp. 1809–1812, 2003.

[6] Le Cerf, P., and Van Compernolle, D., “A new variable frame rate analysis method for speech recognition,” *IEEE Sig. Proc. Letter*s, vol. 1, no. 12, pp. 185–187, December 1994.

[7] You, H., Zhu, Q., and Alwan, A., “Entropy-based variable frame rate analysis of speech signals and its application to ASR”, in *Proc. Int. Conf. on Acoust., Sp. And Sig. Proc.*, vol. 1, pp. 529-552, 2004.

[8] Bocchieri, E. L., and Wilpon, J. G., “Discriminative analysis for feature reduction in automatic speech recognition”, in *Proc. IEEE ICASSP*, vol. 1, pp. 501-504, March 1992.

[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] Zhu, Q., and Alwan, A., “On the use of variable frame rate analysis in speech recognition,” in *Proc. IEEE ICASSP*, pp. 3264–3267, 2000.