An Energy Search Approach to Variable Frame Rate Front-End Processing for Robust ASR

Julien Epps and Eric H. C. Choi

Interfaces, Machines and Graphic Environments (IMAGEN)
National ICT Australia, Sydney, Australia, 1430
julien.epps@nicta.com.au, eric.choi@nicta.com.au

Abstract

Extensive research has been devoted to robustness in the presence of various types and degrees of environmental noise over the past several years, however this remains one of the main problems facing automatic speech recognition systems. This paper describes a new variable frame rate analysis technique, based upon searching a predefined lookahead interval for the next frame position that maximizes the first-order difference of the log energy (ΔE) between the consecutive frames. The application of this novel technique to noise-robust ASR front-end processing is also reported. In comparison with existing variable frame rate methods in the literature, the proposed energy search approach is simpler and achieves similar recognition accuracy improvements at lower complexity. Experimental work on the Aurora II connected digits database reveals that the proposed front-end, together with cumulative distribution mapping, achieves average digit recognition accuracies of 78.32% for a model set trained from clean data and 89.95% for a model set trained from data with multiple noise conditions, representing 6.1% and 2.3% reductions in word error rates respectively over a cumulative distribution mapping baseline.

1. Introduction

In recent years, numerous improvements to the performance of automatic speech recognition (ASR) systems have been proposed, however these are often undermined in practical conditions, due to the presence of environmental noise. Typically, the deterioration in recognition accuracy as the noise level approaches that of the speech renders all but the smallest vocabulary ASR systems unusable. Various types of approaches have been employed by other researchers to improve speech recognition robustness, including pre-enhancing the noisy speech (e.g. [1]), feature-space compensation of clean/noisy feature mismatch (e.g. [2],) and model-space methods that account for the effects of noise in the speech models (e.g. [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. Furthermore, it seems plausible that the human auditory system allows some degree of flexibility in the analysis period, since no physical mechanism has yet been identified that would support a fixed-duration approach. These considerations have 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 for model sets trained on clean speech in particular, at lower complexity than alternative methods. The organization of this paper is as follows. Details of the proposed energy search VFR analysis 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. Energy Search-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, 8], features have been extracted and then tested against some threshold in order to determine whether they should be retained or not, however in principle any criterion that yields similar discriminating power can be used.

Detailed investigations previously reported in the literature [9] 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 requiring the calculation of any other components of the MFCCs, i.e. without the discrete Fourier transform, Mel filter bank and discrete cosine transform. For this reason, it has significantly lower complexity than previous schemes as a criterion for VFR analysis.

Thus, the criterion employed in this paper is to determine the optimum relative position of the next frame $\hat{k}$ by maximizing the difference in log energy between the current frame and possible next frame, so that

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

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 frame index, and K_{min} and K_{max} are the minimum and maximum admissible values of frame advance in samples.

2.2. Energy Calculation

Here, energy is calculated according to the usual formula, except that the energy of the next frame is dependent upon the candidate frame advance k, i.e.

$$E_{m+1}(k) = \sum_{n=0}^{N} x_m^2(n+k), \quad (2)$$

where k is defined relative to the beginning of the current (mth) frame, as shown in Figure 1.

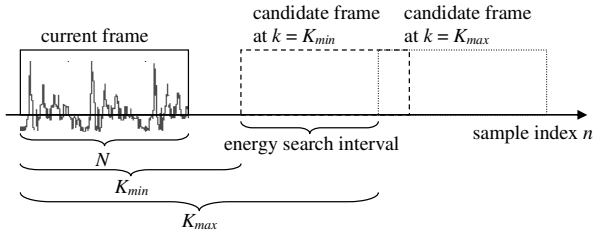

Figure 1: Parameters used in the proposed energy search VFR scheme.

In order to achieve good computational efficiency in calculating the $K_{\max} - K_{\min} + 1$ sample-by-sample next frame candidate energies $\{E_{m+1}(k) \mid K_{\min} \le k \le K_{\max}\}$, these are calculated in a pipelined manner, taking advantage of the fact that

$$E_{m+1}(k+1) = E_{m+1}(k) + x_m^2(N+k+1) - x_m^2(k), \quad (3)$$

Implicitly, equation (3) implies the use of a rectangular window for energy calculation, a window that is susceptible to pitch period artefacts. If the rectangular window is replaced by a tapered window, however, the computational burden of computing the energy on a sample-by-sample basis is greatly increased by this choice.

The efficient computation of all possible candidate next frame energies allows a very fine-grained search, which cannot easily be replicated by existing Euclidean MFCC distance [8] and entropy-based MFCC [7] VFR methods. Since in this scheme, the frame advance $\hat{k}$ is selected by search, the main design consideration is the setting of the predefined search interval limits K_{min} and K_{max}.

3. Experimental Results

3.1. Experimental Setup

The proposed front-end was evaluated on the Aurora II database, test set A, with various configurations described in this section. This test set contains noisy connected digits created by adding subway, babble, car, and exhibition noise at different SNRs to the original clean utterances. 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. 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 are the mean accuracies over the 0 dB to 20 dB conditions.

All the pre-processing and Mel filtering of speech signals followed the ETSI standard MFCC front-end. The Hidden Markov Model (HMM) Toolkit (HTK) 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 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. Sample-by-Sample Energy Variation

An example of the sample-by-sample energy variation in the speech signal is shown in Figure 2, for which the optimum frame advance according to (1) occurs at around 12ms. In this example, the delta energy contour is $\left\{\frac{1}{k}\left(\log\left(E_{m+1}(k)\right) - \log\left(E_m\right)\right) \,\middle|\, 60 \le k \le 200\right\}$, where the current energy E_m is calculated from a 25ms frame centred around 12.5ms, and $E_{m+1}(k)$ is calculated from a 25ms frame centred between 7.5 and 25ms ahead relative to the current frame centre. Here, because the peak in delta energy in Figure 2(c) occurs at about 24.5ms, the optimum next frame would be centred at 24.5 ms.

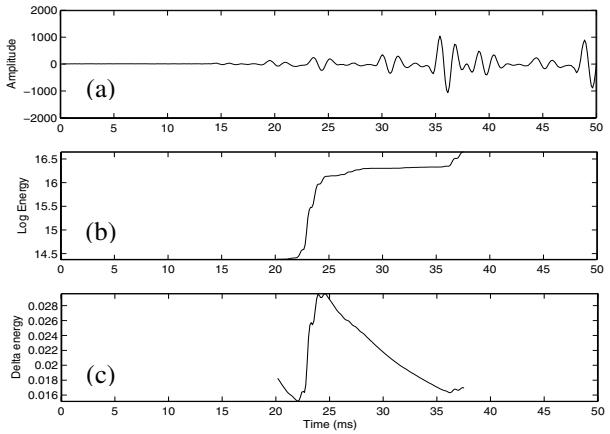

Figure 2: (a) A 50 ms speech signal excerpt, (b) its log energy contour, and (c) its delta energy contour.

3.3. Results for Different Search Intervals

In this experiment, both K_{min} and K_{max} (the minimum and maximum frame advance) were varied, and model sets were

trained on clean speech data. The trained model sets were then applied to the entire test set A, and the resulting accuracies are shown in Figure 3, where each curve represents the variation in accuracy with different values of K_{max}, for constant K_{min}. The highest accuracy found in this experiment was for K_{min} and K_{max} set to 8.75 and 16.75 ms respectively.

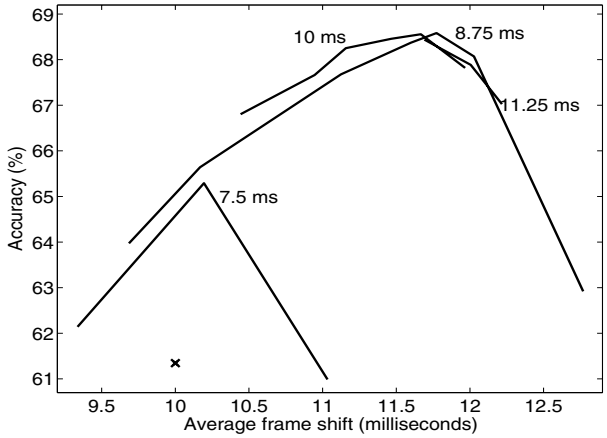

Figure 3: Recognition results from Aurora test set A (clean models), showing the accuracies resulting from different search interval limits (K_{min} is marked on each curve) against the average frame shift, compared with the ETSI standard front-end, marked 'x'.

3.4. Results for Different Noise Types

In order to investigate the results of section 3.3 a little more deeply, energy search VFR (ES-VFR) with K_{min} and K_{max} equivalent to 8.75 and 16.75 ms respectively was compared with the ETSI standard front-end by noise type for both clean and multi-condition model sets, as seen in Figure 4.

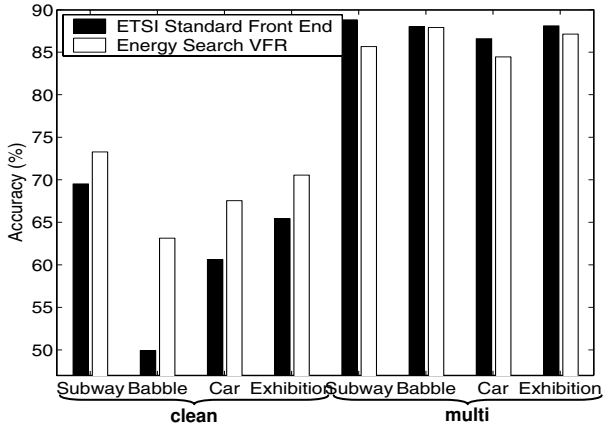

Figure 4: Recognition results from Aurora test set A comparing ES-VFR with the ETSI standard front-end by noise type, for clean and multi-condition models.

3.5. Results for Different Signal to Noise Ratios

For the purposes of SNR comparisons, ES-VFR was compared against two baselines: the ETSI standard front end and the standard front end with cumulative distribution mapping (CDM) applied [10]. The motivation for using a second baseline was both to test the hypothesis that ES-VFR should produce additive improvements to other robust front-end methods, and in recognition of the fact that contemporary front-ends have significantly better performance than the standard front-end. Figures 5(a) and (b) thus show the recognition accuracies for all four schemes: the ETSI standard front-end, the standard front-end using energy search VFR, the standard front end using CDM, and finally the standard front end using both CDM and ES-VFR. The recognition results averaged across all SNRs and noise conditions are summarized in Table 1.

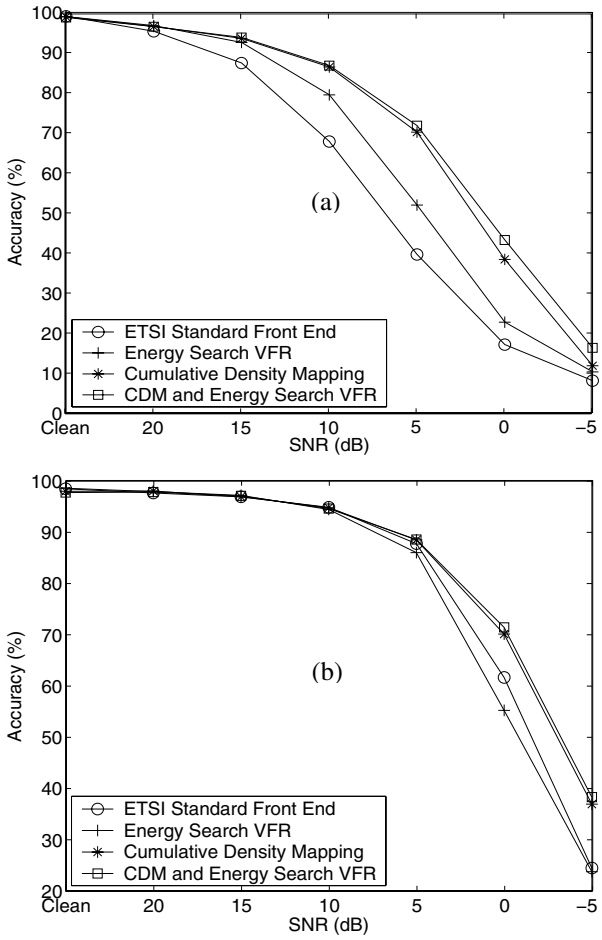

Figure 5: Recognition results from Aurora test set A showing the proposed ES-VFR based front-end compared with the ETSI standard front-end for (a) clean- and (b) multi-condition model sets.

Table 1: *Recognition accuracy results from Aurora test set A. '% impr.' denotes the relative reduction in error rate over the corresponding baseline (i.e. ETSI or CDM).*

Front-end	clean	% impr.	multi	% impr.
ETSI	61.34	N/A	87.82	N/A
ES-VFR	68.58	18.7	86.21	-13.2
CDM	76.92	N/A	89.71	N/A
CDM+ES-VFR	78.32	6.1	89.95	2.3

3.6. Complexity Comparison

In order to estimate the complexity of the proposed variable frame rate scheme, the ETSI standard and ES-VFR front-ends were run on the clean and multi-condition training data and the test set A data (a total of 45000 files), and the duration was recorded. Again, the optimum front-end configuration from previous experiments (K_{min} and K_{max} equivalent to 8.75 and 16.75 ms respectively) was used, and no other processes were running on the processor (2.66 GHz CPU, 2 GB RAM) at the time. The ETSI standard was found to require 778s of processing time, while the proposed ES-VFR took 888s.

4. Discussion

From the results in section 3.3, it becomes evident that there is an optimum average frame shift for the proposed ES-VFR approach, around 11.75 ms in the case of the test set used in section 3. Compared with earlier VFR approaches using model sets trained on clean Aurora II data, ES-VFR (68.58%) yields comparable, but slightly poorer performance averaged over all SNRs than the energy-weighted Euclidean MFCC distance [8] (70%) and entropy-based MFCC [7] (71.54%) VFR schemes. The energy-weighted Euclidean MFCC distance [8] and entropy-based MFCC [7] VFR schemes both produce significantly poorer accuracy under clean conditions, with their word error rates degrading by 42% and 89% [7] respectively relative to the ETSI front-end. By contrast, the new ES-VFR produces a more acceptable degradation of 10% in relative error rate for the clean condition.

From sections 3.4 and 3.5, it appears that the ES-VFR accuracy improvements over the ETSI front-end for a clean model set are consistent across all SNRs except clean, and across all noise types. Interestingly, for the multi-condition model set ES-VFR yielded poorer accuracies than the ETSI front-end, with the schemes, 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. A remarkable feature of the ES-VFR scheme is that it seems to perform well in babble noise relative to the ETSI standard front-end, producing a 26% reduction in word error rate for a clean model set and barely degrading the accuracy at all in the multi-condition case.

The most encouraging result from section 3 is the consistent improvement over the CDM baseline found for both clean and multi-condition model sets and over all SNRs. Since the ES-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 ES-VFR with pre-enhancement, feature-space or model-space techniques has good prospects for further improvements in recognition accuracy.

The 14% increase in processing time relative to the ETSI standard front-end required by ES-VFR qualitatively compares well with previous VFR schemes [4, 5, 6, 7, 8], which have required all candidate MFCCs to be pre-computed, presumably resulting in complexities significantly greater than that of the ETSI standard front-end.

5. Conclusions

A noise-robust variable frame rate ASR front-end based upon an energy search that maximizes the first-order difference of the log energy between consecutive frames has been presented. From experiments with the Aurora II database, ES-VFR shows promise for improving recognition of speech in the presence of all types of noise, particularly where the test data are less well-matched to 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. When combined with cumulative distribution mapping, the proposed front-end obtains a relative error rate reduction of 6.1% for the clean model set, and 2.3% for the multi-condition model set. Future research will focus on investigating means for improving the multi-condition performance of the proposed scheme, and on the possibilities for combining ES-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, April 1992, pp. 725-735.

[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, May 1996, pp. 190–202.

[3] Zhang, Z. and Furui, S., "Piecewise-linear Transformation-based HMM Adaptation for Noisy Speech", *Speech Communication*, vol. 42, iss. 1, Jan. 2004, pp. 43-58.

[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*, 2003, pp. 1809–1812.

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

[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, 2004, pp. 529-552.

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

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

[10] 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.