A Posteriori SNR Weighted Energy Based Variable Frame Rate Analysis for Speech Recognition

Zheng-Hua Tan and Børge Lindberg

Multimedia Information and Signal Processing (MISP), Department of Electronic Systems,
Aalborg University, Denmark
Niels Jernes Vej 12, 9220, Aalborg, Denmark
{zt, bli}@es.aau.dk

Abstract

This paper presents a variable frame rate (VFR) analysis method that uses an *a posteriori* signal-to-noise ratio (SNR) weighted energy distance for frame selection. The novelty of the method consists in the use of energy distance (instead of cepstral distance) to make it computationally efficient and the use of SNR weighting to emphasize the reliable regions in speech signals. The VFR method is applied to speech recognition in two scenarios. First, it is used for improving speech recognition performance in noisy environments. Secondly, the method is used for source coding in distributed speech recognition where the target bit rate is met by adjusting the frame rate, yielding a scalable coding scheme. Prior to recognition in the server, frames are repeated so that the original frame rate is restored. Very encouraging results are obtained for both noise robustness and source coding.

Index Terms: speech recognition, speech analysis, variable frame rate, noise robustness, source coding

1. Introduction

Placed in between input signals and the recognition decoder, the front-end of an automatic speech recognition (ASR) system commonly processes the input signals frame-by-frame at a fixed rate. This processing is based on the two assumptions: that speech signals exhibit quasi-stationary behavior in a short time, and that acoustic models such as hidden Markov models (HMMs) are capable of absorbing the dynamics of variable information rate. However, the two assumptions hold only to some extent as discussed below.

First, an input signal often consists of non-speech parts and speech parts that again consist of steady regions and rapidly changing events. Speech sounds like plosives or speech attributes like transitions can last a very short period, indicating that the use of a fixed frame rate (e.g. at 100 Hz) is insufficient to provide a fine representation for these events. On the other hand, steady regions like vowels can last a relatively long period without significant changes in the spectrum. Over-sampling the spectrum may generate unnecessary frames which can increase insertion errors and computational load. For the non-speech parts the best is no samples at all. Clearly, the fixed frame rate analysis is unsatisfactory [1].

Secondly, HMMs are known to poorly model the variability of sound durations. The variable-duration problem is particularly severe in spontaneous speech, which motivates research interests in duration normalization [2] and speaking-rate dependent decoding [3]. This weakness of HMMs has been demonstrated in [4] as well, yet from a different angle, where it is shown that a mismatch between the frame rate and the number of HMM states may introduce a considerable degradation in recognition performance.

Variable frame rate (VFR) analysis is capable of largely releasing the two assumptions discussed above by providing a fine resolution for rapidly changing events and by normalizing the sound durations. This, however, requires examining the speech signal at a higher rate than 100 Hz, as done in [5]. When the cepstral distance measure that has been widely used in VFR analysis for frame selection is applied, the procedure of extracting cepstral features at a high rate and then discarding the majority of these is waste of computing resources and thus limits the possible high time resolution and the usage of the VFR analysis. However, note that the first-order difference in frame-to-frame energy provides greater discrimination than components of Mel-frequency cepstral coefficients (MFCCs) other than c0 [6] and that the effectiveness of the energy based criterion has been demonstrated in [7]. Evidently, energy based search is much more computationally efficient and thus enables a finer granularity in search.

Moreover, speech segments are accounted in ASR not only on their characteristics, but also on their reliability. The later is important in particular for speech recognition in noisy environments and is pursued in missing data theory and weighted Viterbi decoding methods where low signal-to-noise ratio (SNR) features are either neutralized or less weighted in the ASR decoding process. The SNR information should be exploited for frame selection as well. All these considerations lead us to propose the *a posteriori* SNR weighted energy based selection criterion for VFR.

The paper is organized as follows. First, existing methods and motivations are presented in Section 2. The *a posteriori* SNR weighted energy based VFR is detailed in Section 3. Sections 4 and 5 apply the VFR method to robust speech recognition and to distributed speech recognition (DSR) for data compression, respectively. Finally, we conclude the paper in Section 6.

2. Existing methods and motivations

Variable frame rate analysis has a broad spectrum of applications, ranging from computational reduction in the early days, through improved acoustic modeling and noise robustness, to prolonged speech recognition in singing voice or in spontaneous speech. For these applications, various techniques have been developed. Mostly, VFR analysis extracts speech feature vectors – equivalent to frames – at a fixed-frame-rate first and then uses a certain criterion to retain or omit frames. The frame selection is done by calculating some distance (or similarity) measure and comparing it with a threshold.

Copyright © 2008 ISCA
Accepted after peer review of full paper

In [8], the distance measure is computed as the Euclidean distance between the last retained feature vector and the current vector. The decision criterion is to discard the current frame if the measure is smaller than a defined threshold. In [9], it is based on the norm of the first derivative cepstrum vector. The current frame is discarded if the norm is smaller than a threshold. In this way, neighboring frames of the current frame, rather than only two frames as in [8], are used in the decision making. Due to the reduced number of feature vectors, computation time for decoding is saved.

Lately, there has been a growing interest in applying VFR to deal with additive noise in the time domain [5], [7], [10]. In [5], Zhu and Alwan proposed an effective VFR method that uses a 25 ms frame length and a 2.5 ms frame shift for calculating MFCCs and conducts frame selection as follows. First, the energy weighted Euclidean distance of adjacent MFCC vectors is calculated as

$$D(t) = D(t,t-1)\cdot(\log E(t) - \overline{\log E(t)} / \beta) \qquad (1)$$

where $D(t,t-1)$ is the Euclidean distance between frame t and frame t-1, $\log E(t)$ is the logarithmic energy of frame t and $\overline{\log E(t)}$ is the mean of $\log E(t)$ over a certain period, for example, an utterance. β was set to be 1.5 both in [5] and in this work. Based on the distance, the threshold is then computed as

$$T = \alpha \cdot \overline{D(t)} \qquad (2)$$

where $\overline{D(t)}$ is the mean of the weighted distance $D(t)$ over a period, and α is a factor that determines the average frame rate. α was set to be 6.8 in [5]. Finally, a frame is selected if the distance $A(t) = \sum D(t)$ accumulated since last-selected-frame is greater than the threshold *T*.

A thorough comparison of the VFR methods referred above was conducted in [11] and the one in [5] was found to outperform the others for both frame selection and speech recognition, but it did not show improvement in recognition accuracy over an FFR analysis on a general database.

A few observations are obtained from analyzing the existing methods. First, it is noted from Eq. 1 that $D(t)$ is not guaranteed to be a non-negative value. For clean speech, due to the significant difference in energy between silence and speech regions, the weights will be negative for a silence region and the resulting negative values will accumulate and thus influence the frame selection for the speech right after the silence region. This is likely to be the reason why it performs well for low SNR speech, but shows no improvement on a general database.

Next, the procedure of extracting cepstral features at a high rate and then discarding the majority of these is waste of computing resources. The entropy-based VFR analysis proposed in [10] introduces even higher computational cost though with improved recognition accuracy. Given that energy provides a good discrimination, energy based search can potentially enable a determination of frame shift without pre-computing feature vectors at a fixed rate.

Finally, speech segments are accounted in ASR not only on their characteristics, but also on their reliability. SNR is a good measure for reliability and thus can be exploited for frame selection.

All these considerations lead us to propose the *a posteriori* SNR weighted energy selection criterion for VFR.

3. *A posteriori* SNR weighted energy based VFR

The proposed method conducts frame selection on the basis of an accumulative, *a posteriori* SNR weighted energy distance. *A posteriori* SNR is defined as the logarithmic ratio of the energy of noisy speech to the energy of noise; in contrast, *a priori* SNR is the logarithmic ratio of the energy of speech to the energy of noise. Calculating *a posteriori* SNR is rather straightforward, while calculating *a priori* SNR requires estimating the energy of clean speech which is a challenging task in itself.

3.1. The proposed VFR method

The algorithm of the method is as follows:

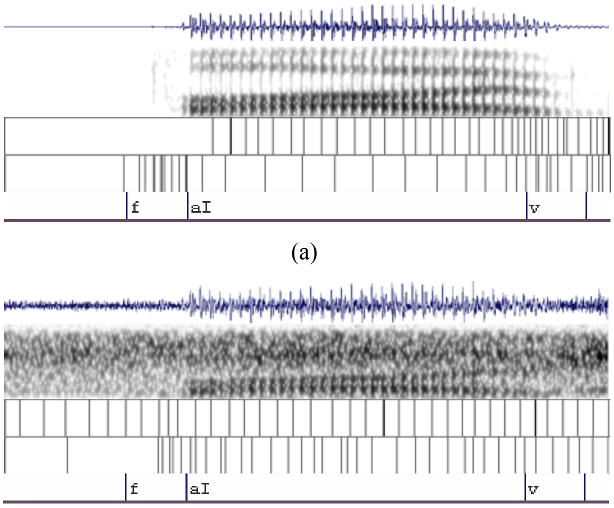
1. Compute the *a posteriori* SNR weighted energy distance of two consecutive frames as
$$D(t) = |\log E(t) - \log E(t-1)| \cdot SNR_{post}(t) \qquad (3)$$
where $\log E(t)$ is the logarithmic energy of frame *t*, and $SNR_{post}(t)$ is the estimated *a posteriori* SNR value of frame *t* by using a 1 ms frame shift and a 25 ms frame length.
2. Compute the threshold T for frame selection as
$$T = \overline{D(t)} \cdot f(\log E_{noise}) \qquad (4)$$
where $\overline{D(t)}$ is the average weighted distance over a certain period and $f(\log E_{noise})$ is a sigmoid function of $\log E_{noise}$ to allow a smaller threshold and thus a higher frame rate for clean speech. The sigmoid function is defined as $f(\log E_{noise}) = 9.0 + \frac{2.5}{1+e^{-2(\log E_{noise}-13)}}$ where the constant of 13 is chosen so that the turning point of the sigmoid function is at *a posteriori* SNR of between 15 and 20 dB. The choice of sigmoid parameters and their influence on ASR performance are detailed in [17].
3. Update the accumulative distance: $A(t) += D(t)$ on a frame-by-frame basis and compare it against the threshold T : If $A(t) > T$, the current frame is selected and $A(t)$ is reset to zero; otherwise, the current frame is discarded. If the current frame is not the last one, the search continues, that is, go back to step 1.

The use of *a posteriori* SNR, rather than *a priori* SNR, avoids the problem of assigning zero or negative weights to frames with $SNR_{prio} \leq 0dB$ and subsequently discarding them due to their non-positive weights. As such, the *a posteriori* SNR weight for noise-only frames will be theoretically equal to 0 dB, which serves as an implicit, soft VAD; negative *a posteriori* SNR values may still appear in practice and are then set to zero to prevent negative weights. In this work E_{noise} for calculating $SNR_{post}(t)$ and $\log E_{noise}$ for calculating T are both simply estimated by averaging the first 10 frames of an utterance which are considered noise only. The average weighted distance $\overline{D(t)}$ is calculated over one utterance; in practice, $\overline{D(t)}$ calculated over preceding segments can be used and it is then updated frame-by-frame based on a forgetting factor.

As only the logarithmic energy and the *a posteriori* SNR value are calculated for each frame, the VFR method has a very low complexity as compared with the existing methods described.

3.2. Frame selection

The comparison study in [11] showed that the VFR method in [5] outperformed a few other methods for both frame selection and speech recognition. Figure 1(a) illustrates a comparison between the proposed method and the method in [5] in terms of frame selection for the clean speech of the English digit “five”. The five panels in Fig. 1(a), sequentially, illustrate the waveform, the spectrogram, the frames selected by the referenced method with $\alpha = 5.0$, the frames selected by the proposed one and the phoneme annotation. Figure 1(b) shows the same comparison for 0 dB speech. In this work, it has been experimentally found that $\alpha = 5.0$, rather than $\alpha = 6.8$, gives the best recognition results.



(a)

(b)

Fig. 1. Frame selection for the English digit “five”: (a) For clean speech: waveform (the 1st panel), spectrogram (the 2nd panel), frames selected by the referenced method [5] with $\alpha = 5.0$ (the 3rd panel), frames selected by the proposed method (the 4th panel), phoneme annotation (the 5th panel); (b) for 0 dB speech with the same order of panels as in (a).

Figure 1(a) shows that the proposed VFR assigns a higher frame rate to fast changing events such as consonants, lower frame rate to steady regions like vowels and no frames to silence, which exactly represents the objective of applying VFR analysis. In contrast, the referenced method also performs well but with one weakness namely eliminating the first part of speech following a silence due to the negative weights resulting from $\log E(t) - \overline{\log E(t)} / \beta$. Figure 1(b) shows that the proposed VFR method realizes an implicit VAD very well even for a 0 dB signal as there is only one frame output for the silence part, while the referenced method results in almost evenly distributed frames.

4. Noise robust speech recognition

The proposed VFR method is applied to noise robust speech recognition. Experiments are conducted on the Aurora 2 database [12], which is the TI digits database artificially distorted by adding noise and using a simulated channel distortion. Whole word models are created for all digits using the HTK recognizer. Each of the whole word digit models has 16 HMM states with three Gaussian mixtures per state. The silence model has three HMM states with six Gaussian mixtures per state. A one state short pause model is tied to the second state of the silence model.

The word models used in the experiments are trained on clean speech data. The test data is Test Set A including clean speech and noisy speech corrupted by four noise types: “Subway”, “Babble”, “Car”, and “Exhibition” with SNR ranging from 0 to 20 dB. The speech features are 12 MFCC coefficients, logarithmic energy as well as their corresponding velocity and acceleration components.

4.1. Experimental results

The word error rate (WER) results for a number of methods are presented in Table 1. The fixed frame rate (FFR) baseline uses a fixed 10 ms frame shift. VFR ($\alpha = 5.0$) is the VFR in [5]. The referenced method does not give an acceptable performance for clean speech. The reason is that the energy weight $\log E(t) - \overline{\log E(t)} / \beta$ results in no frames output for the first part of speech right after the silence which is often a short-duration consonant, as exemplified in Fig. 1(a).

The energy based VFR (LogE-VFR) [7] also gives a good performance on noisy speech, though worse than that of [5]. The proposed method (SNR-LogE-VFR) is superior to the cited methods and has lower complexity.

Table 1. Percent WER across the methods for Test Set A. The results for LogE-VFR are cited from [7].

Methods	0 ~ 20 dB					Clean
	Subway	Babble	Car	Exhibit.	Average	
FFR	30.5	50.1	39.4	34.6	38.7	1.0
VFR ($\alpha = 5.0$)	28.9	29.0	28.9	31.1	29.5	3.5
LogE-VFR	N/A	N/A	N/A	N/A	31.4	1.1
SNR-LogE-VFR	28.3	27.8	29.2	29.6	**28.7**	**1.4**

4.2. Combination with spectral subtraction

Variable frame rate analysis relies on distance measures for frame selection; however, these measures can be largely affected by noises that corrupt the speech. On the other hand, as the VFR method operates in the time domain, it has a good potential to be combined with other methods, e.g. spectral subtraction. The idea of the following experiment is to first use spectral subtraction to de-noise the speech and then apply a VFR analysis.

Table 2 shows the results of combining the VFR with the minimum statistics noise estimation (MSNE) [13] based spectral subtraction (SS). The constant of 13 in the sigmoid function is optimized to be 10 due to the use of SS. It is observed that the combination achieves a 17.1% absolute WER reduction over the FFR baseline. Interestingly, the improvement of the combined method is greater than the summation of the gains obtained by applying the two methods individually. This justifies the dual contributions of spectral subtraction when combined with the VFR method, i.e. improving frame selection and enhancing speech.

Table 2. Percent WER for SS and its combination with the VFR for Test Set A.

Methods	0 ~ 20 dB					Clean
	Subway	Babble	Car	Exhibit.	Average	
MSNE-SS	31.9	43.0	25.6	34.1	33.7	1.5
MSNE-SS +SNR-LogE-VFR	19.7	26.4	18.6	21.7	**21.6**	**1.3**

5. Source coding in DSR

Distributed speech recognition employs the client-server architecture by placing the front-end in the client and the computation-intensive back-end in the server. This architecture relieves the burden of computation, memory and energy consumption from mobile devices. One issue induced by the distributed solution is the requirement of data compression.

The VFR method aims at a high time resolution for fast changing events and a low time resolution for steady regions. The same philosophy is applied in source coding as well. Frame allocation in the feature extraction process optimized over a certain period in the VFR analysis is likely of benefit to the source coding which follows right after the feature extraction.

An efficient compression method in DSR is the two-dimensional Discrete Cosine Transform (2D-DCT) based code [14]. More recently, the group of pictures concept (GoP) from video coding was applied to DSR to achieve a variable-bit-rate interframe compression scheme [15]. The results for these methods are cited and presented in Table 3. The ETSI-DSR standard, however, uses a split vector quantization for data compression without exploiting interframe information [16].

In this work, we use the VFR method for data compression where the target bit rate is simply realized by choosing a proper frame rate. For comparison purpose, we optimized the SNR-LogE-VFR, by constraining the range of the frame selection search, to give a comparable performance on clean speech to the ETSI-DSR baseline. After applying split vector quantization, this gives a DSR front-end with a bit rate of approximately 3.5 kbps (SNR-LogE-VFR-DSR) and its recognition results are shown in Table 3. A bit rate of approximately 1.5 kbps is implemented as well and to restore the original frame rate for the match between the frame rate and the applied HMMs, frame repetition is applied in the server. The mismatch can as well be removed by using a smaller number of HMM states, at the expense of additional acoustic model sets. Experimental results in Table 3 show that the VFR based source coding is significantly superior to the 2D-DCT method and the GoP one.

Table 3. Percent WER across the data compression methods for Test Set A. The results for 2D-DCT and GOP are cited from [14] and [15], respectively.

Methods	kbps (pay load)	0 ~ 20 dB					Clean
		Subway	Babble	Car	Exhibit.	Aver-age	
ETSI-DSR	4.40	32.3	50.4	40.6	36.1	39.8	1.0
2D-DCT	1.45	N/A	N/A	N/A	N/A	40.5	1.6
GOP	2.57	N/A	N/A	N/A	N/A	N/A	2.5
	1.27	N/A	N/A	N/A	N/A	N/A	2.6
SNR-LogE-VFR-DSR	3.50	34.0	31.6	34.8	34.6	33.7	1.0
SNR-LogE-VFR-DSR	**1.50**	34.3	30.9	33.0	33.0	**32.8**	**1.2**

6. Conclusions

This paper has presented a new variable frame rate analysis method that relies on the accumulative, *a posteriori* SNR weighted energy distance for frame selection. In terms of frame selection, the method is able to assign a higher frame rate to fast changing events such as consonants, a lower frame rate to steady regions like vowels and no frames to silence, even for very low SNR signals. The method was then applied to noise-robust speech recognition and was further combined with a spectral-domain method. Encouraging results were obtained. The VFR was moreover applied to distributed speech recognition for source coding, resulting in an efficient, scalable coding scheme. The advantage of the proposed method lies in its low complexity and improved performance.

7. References

[1] S.J. Young and D. Rainton, “Optimal frame rate analysis for speech recognition,” IEE Colloquium on Techniques for Speech Processing, Dec 1990.

[2] J.P. Nedel and R.M. Stern, “Duration normalization for improved recognition of spontaneous and read speech via missing feature methods,” in Proc. IEEE ICASSP, Salt Lake City, USA, 2001.

[3] H. Nanjo and T. Kawahara, “Language model and speaking rate adaptation for spontaneous presentation speech recognition,” IEEE Trans. Speech and Audio Processing, 12(4), 2004.

[4] Z.-H. Tan, P. Dalsgaard, and B. Lindberg, “Exploiting temporal correlation of speech for error-robust and bandwidth-flexible distributed speech recognition," IEEE Transactions on Audio, Speech and Language Processing, 15(4), pp. 1391-1403, 2007.

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

[6] E. L. Bocchieri and J. G. Wilpon, “Discriminative analysis for feature reduction in automatic speech recognition,” in Proc. IEEE ICASSP, 1992.

[7] J. Epps and E. Choi, “An energy search approach to variable frame rate front-end processing for robust ASR,” in Proc. Eurospeech, Lisbon, 2005.

[8] K. M. Pointing and S. M. Peeling, “The use of variable frame rate analysis in speech recognition,” Computer Speech and Language, 5(2), pp. 169–179, 1991.

[9] P. Le Cerf and D. Van Compernolle, “A new variable frame rate analysis method for speech recognition,” IEEE Signal Processing Letters, 1(12), pp. 185–187 1994.

[10] H. You, Q. Zhu, and A. Alwan, “Entropy-based variable frame rate analysis of speech signals and its application to ASR”, in Proc. IEEE ICASSP, 2004.

[11] J. Macias-Guarasa, J. Ordonez, J. M. Montero, J. Ferreiros, R. Cordoba and L. F. D. Haro, “Revisiting scenarios and methods for variable frame rate analysis in automatic speech recognition,” in Proc. Eurospeech, 2003.

[12] H. G. Hirsch and D. Pearce, “The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions,” in Proc. ISCA ITRW ASR, 2000.

[13] R. Martin, “Noise power spectral density estimation based on optimal smoothing and minimum statistics”, IEEE Trans. on Speech and Audio Processing, 9(5), pp. 504-512, 2001.

[14] W.-H Hsu and L.-S. Lee, “Efficient and robust distributed speech recognition (DSR) over wireless fading channels: 2D-DCT compression, iterative bit allocation, short BCH code and interleaving”, in Proc. IEEE ICASSP, Canada, 2004.

[15] B.J. Borgstrom and A. Alwan, “A packetization and variable bitrate interframe compression scheme for vector quantizer-based distributed speech recognition”, in Proc. Interspeech, Antwerp, Belgium, 2007.

[16] ETSI Standard ES 202 212 (2003) Distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithm, back-end speech reconstruction algorithm.

[17] Z.-H. Tan and B. Lindberg, “An efficient frame selection approach to variable frame rate analysis for noise robust speech recognition,” in Proc. Acoustics, Paris, France, 2008.