Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

An efficient frame selection approach to variable frame rate analysis for noise robust speech recognition

Zheng-Hua Tan and Børge Lindberg

Department of Electronic Systems, Aalborg University, Niels Jernes Vej 12, 9220 Aalborg, Denmark
zt@es.aau.dk

This paper presents a low-complexity, effective variable frame rate (VFR) analysis method that conducts frame selection on the basis of *a posteriori* signal-to-noise ratio (SNR) weighted energy distance. It has two characteristics. First, energy distance (instead of cepstral distance) is used to make it computationally efficient and thus enable a finer granularity in search as compared with cepstral distance criterion. Secondly, SNR weighting is used to emphasize the reliable regions in noisy speech signals. In terms of frame selection, it is experimentally found that 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 VFR method is applied to speech recognition in noisy environments to improve noise robustness. Being a method that takes effect in the time-domain, it is moreover combined with spectral- and cepstral-domain techniques to gain further improvement. Experiments are conducted on the Aurora 2 database, which is the TI digits database artificially distorted by adding different noises, and very encouraging results are obtained.

1 Introduction

Robustness against environmental noises is one of the most important challenges in automatic speech recognition research and development. In general, robustness methods aim at reducing the mismatches between the training and test speech signals through feature-domain or model domain methods. Feature extraction related methods include feature enhancement, distribution normalization and noise robust feature extraction. Feature enhancement attempts to clean noise-corrupted features, for example in spectral subtraction [1]. Distribution normalization reduces the distribution mismatches between training and test speech such as in MVA processing which consists of mean subtraction, variance normalization and auto-regression moving-average (ARMA) based filtering in the cepstral domain [2]. Noise robust features include improved Mel-frequency cepstral coefficients (MFCCs) e.g. root-cepstrum [3] and new features e.g. variance based features [4].

This work, however, investigates a different approach namely variable frame rate (VFR) analysis which is concerned with frame shift and selection, and yet has shown good performance in noise robustness [5].

Fixed frame rate analysis is prevalent in contemporary speech recognition systems. A typical feature extraction processing of such systems computes speech features using a 25 ms frame length and a 10 ms frame shift. This fixed frame rate processing is based on the assumption that speech signals exhibit quasi-stationary behavior in a short time. This assumption is, however, questionable for rapidly changing speech events such as plosives. In addition, fixed frame rate analysis treats noise and speech segments equally.

On the contrary, variable frame rate analysis selects frames according to signal characteristics and has been applied to speech recognition for several purposes [6]. First, it is used for efficient speech recognition by discarding redundant frames while maintaining recognition performance [7], [8]. Given the rapid and steady progress in computing, this is of less interesting today. Secondly, VFR analysis is exploited to improve acoustic modeling by capturing fast changes in the spectral characteristics, and improved performance has been observed on a nasal database [9]. Thirdly, it is used for improving noise robustness [9]-[12] and this will be the focus of the present paper. Finally, a recent work applies VFR to distributed speech recognition (DSR) to increase robustness against transmission errors and compress speech data for low-bit-rate feature transmission [13].

Most of VFR analysis methods calculate cepstral coefficients (used for distance measure) for each frame first and select frames afterwards, as those presented in [7-9]. This kind of process is obviously waste of computing resources. In contrast, the method presented in [11] uses delta logarithmic energy as the criterion for determining the size of the frame step on the basis of a sample-by-sample search. In [12] the authors present *a posteriori* SNR weighted energy based VFR analysis for speech recognition, which also selects frames prior to calculating cepstral features. This paper extensively investigates its use for noise-robust speech recognition.

Since VFR analysis takes effect in the time domain, it has a high potential to be combined with other methods including both feature- and model-domain methods. This work combines VFR analysis with feature enhancement and distribution normalization methods.

This paper is organized as follows. Section 2 presents the *a posteriori* SNR weighted energy based VFR method. Section 3 conducts recognition experiments on noisy speech data and investigates the sensitivity of this method to varying parameters. Sections 4 and 5 present the combination of the method with spectral- and cepstral-domain noise-robustness methods, respectively. Section 6 concludes this paper.

2 The *a posteriori* SNR weighted energy based VFR method

The method conducts frame selection based on *a posteriori* SNR weighted energy. Its procedure is as follows [12]:

1. Compute the *a posteriori* SNR weighted energy distance of two consecutive frames as

$$D(t) = |\log E(t) - \log E(t-1)| \cdot SNR_{post}(t) \tag{1}$$

   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

$$T = \overline{D(t)} \cdot f(\log E_{noise}) \tag{2}$$

   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

$$f(\log E_{noise}) = \alpha_1 + \frac{\alpha_2}{1+e^{-2(\log E_{noise}-13)}} \qquad (3)$$

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. Parameters α_1 and α_2 are used to determine average frame rate.

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.

Throughout this work, the $SNR_{post}(t)$ is estimated as the logarithmic ratio of the energy of frame *t*, *E*(*t*), to the energy of noise, E_{noise}. The use of *a posteriori* SNR, rather than *a priori* SNR, avoids the problem of assigning zero or negative weights to frames with $SNR_{prio} \le 0dB$ 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 Experiments on noise robustness

The proposed VFR method is applied to speech recognition in noisy environments.

3.1 Experimental setup

Experiments are conducted on the Aurora 2 database [14], 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 [15]. Each of the whole word digit models has 16 HMM (hidden Markov model) 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.

3.2 Baseline methods

The fixed frame rate (FFR) baseline uses a fixed 10 ms frame shift as implemented in the ETSI DSR standard [16]. The referenced VFR methods include the accumulative, energy weighted cepstral distance VFR [9], the entropy based method [10] and the delta energy based method [11].

In [9] a cumulative, energy weighted cepstral-distance is proposed for frame selection. The distance of adjacent MFCC vectors is calculated as

$$D(t) = D(t,t-1)\cdot(\log E(t) - \overline{\log E(t)}/1.5) \qquad (4)$$

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. Based on the distance, the threshold is then computed as

$$T = \alpha \cdot \overline{D(t)} \qquad (5)$$

where $D(t)$ is the mean of the weighted distance $D(t)$ over a period, and α is a factor that determines the average frame rate. A frame is selected if the distance $A(t) = \sum D(t)$ accumulated since last-selected-frame is greater than the threshold *T*. The method has demonstrated good performance on speech data with low signal-to-noise ratios, and has shown to be superior to the methods presented in [7] and [8] according to the experiments conducted in [6].

With the same motivation, the entropy-based VFR analysis is proposed in [10]. In addition to the MFCCs calculation used in [9], here a 30 ms rectangular-window length and a 15 ms window shift are used for calculating local entropy which is then compared with three different thresholds for frame selection, introducing a high computational cost.

3.3 Experimental results

The word error rate (WER) results for a number of methods are presented in Fig. 1. In the figure, Cep-VFR ($\alpha = 6.8$) refers to the accumulative, energy weighted cepstral distance VFR with the same settings as in [9], while Cep-VFR ($\alpha = 5.0$) uses a smaller α value to select more frames for the purpose of better matching the frame rate with the applied HMMs and in this work this setting gives the best recognition performance. The Cep-VFR method with both settings unfortunately does not give an acceptable performance for clean speech. The reason is that the energy weight $\log E(t) - \overline{\log E(t)}/1.5$ as given in Eq (4) results in no frames output for the first part of speech right after the silence which is often a short-duration consonant.

The results for both Cep-VFR+VAD and Entropy-VFR+VAD presented in Fig. 1 are cited from [10] and they show that the Cep-VFR method combined with VAD (Cep-VFR+VAD) gives a good performance for both clean and noisy speech, and that the entropy method combined with

VAD (Entropy-VFR+VAD) gives a even better recognition performance for noisy speech but a bit worse performance for clean speech.

The energy based VFR (LogE-VFR) [11] also gives a good performance on noisy speech, though worse than the previous two. Finally the proposed method (SNR-LogE-VFR) with $\alpha_1 = 6.5$ and $\alpha_2 = 4.5$ demonstrates a slightly improved performance over the Entropy-VFR method combined with VAD, yet the proposed method has substantially lower complexity, no support from VAD (need for a rough estimation of E_{noise} but no explicit need for a VAD) and less parameters to tune (for example, the entropy method compares against three thresholds).

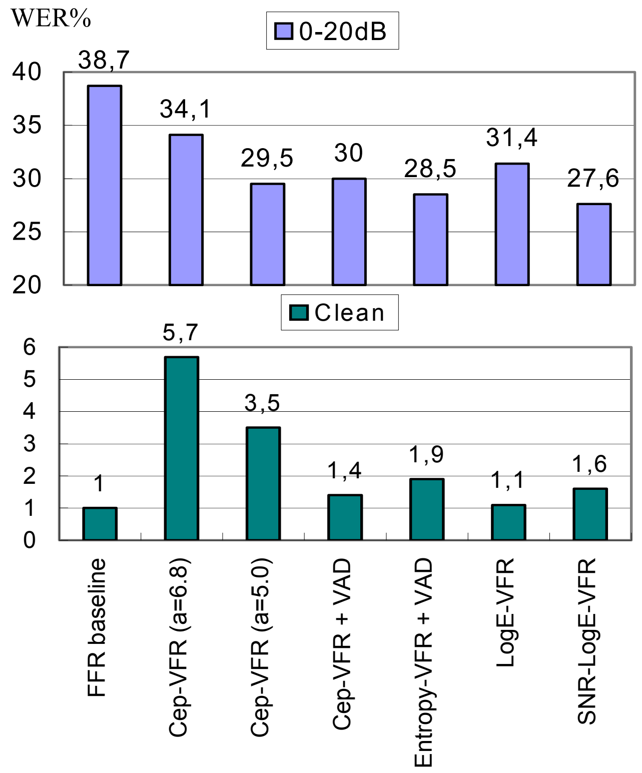

Fig. 1. Percent WER across the methods for Test Set A: upper figure for 0-20 dB speech and lower figure for clean speech.

3.4 Parameters for adjusting frame rate

The key parameters in the *a posteriori* SNR weighted energy based VFR method are α_1 and α_2, which are used to control average frame rate of the front-end. To investigate the sensitivity of the method with regard to the changes of α_1 and α_2, a number of experiments have been conducted and recognition results are shown in Fig. 2.

The results verify that changing the two parameters around their default settings only has a marginal effect on the recognition performance. It has even less influence on clean speech. The WERs are the same for clean speech (i.e. 1.6%) when changing α_2 from 4.0 to 5.0 with a fixed value of $\alpha_1 = 6.5$; they are also the same when changing α_1 from 6.3 to 6.7 with a fixed value of $\alpha_2 = 4.5$, while it is 1.5% for $\alpha_1 = 7.0$ and 2.0% for $\alpha_1 = 6.0$.

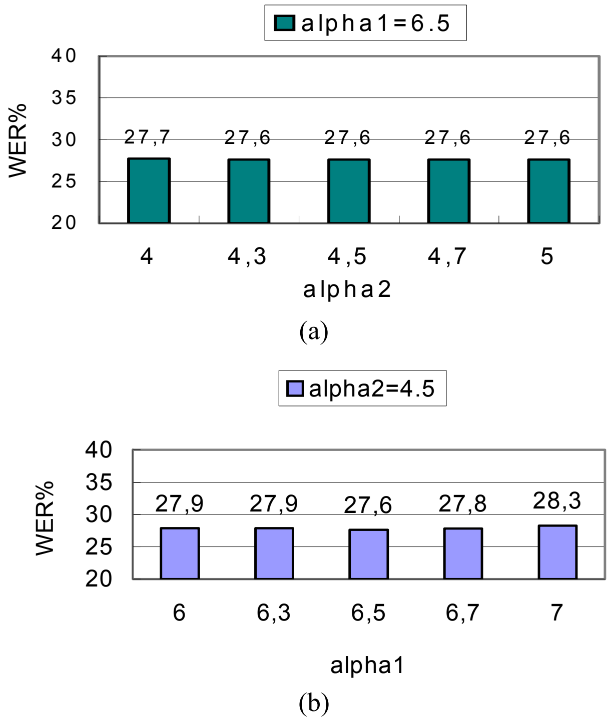

Fig. 2. Percent WER for SNR-LogE-VFR for 0 ~ 20 dB speech in Test Set A: (a) α_1 has a fixed value of 6.5 and the value of α_2 changes; (b) α_2 has a fixed value of 4.5 and the value of α_1 changes.

3.5 Frame selection

Frame selection is in the core of VFR analysis. Figure 3 compares the *a posteriori* SNR weighted energy based VFR method with the accumulative, energy weighted cepstral distance VFR in terms of frame selection. The testing utterance is English digit string "five nine four" and has an *a priori* SNR of 0 dB.

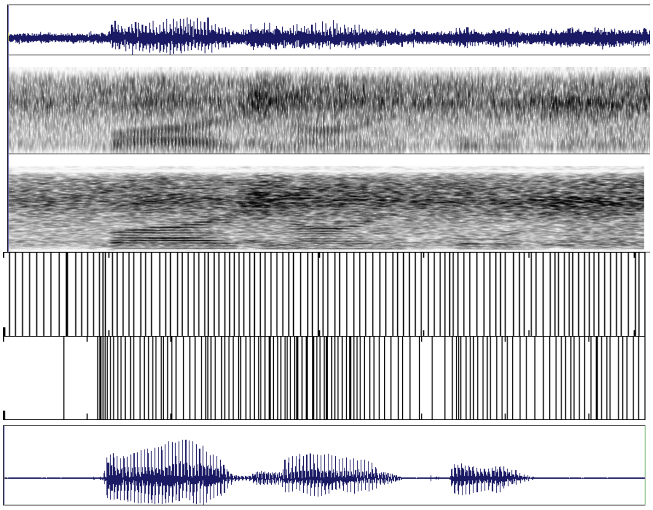

Fig. 3. Frame selection for English digits "five nine four" (0 dB): waveform (the first panel), wideband spectrogram (the second panel), narrowband spectrogram (the third panel), frames selected by the referenced method [9] with

$\alpha = 5.0$ (the fourth panel), frames selected by the proposed method (the fifth panel) and the corresponding clean speech waveform as a reference (the last panel).

From the figure, it is evident that the *a posteriori* SNR weighted energy based 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.

4 Combination with spectral-domain method

Since VFR analysis takes effect in the time domain, it has a high potential to be combined with other methods [5]. VFR analysis emphasizes speech transitions and deemphasizes silence and vowel regions based on distance measures. However, for noisy speech the measurement can be largely affected by additive noise. We propose to use speech enhancement methods to de-noise speech first and apply VFR analysis secondly. The purpose of applying speech enhancement methods is to both improve the frame selection and to enhance the speech.

Based on the assumption that speech cannot occupy a frequency bin all the time, the minimum statistics noise estimation (MSNE) method [17] treats the minimum value of each frequency bin in the power spectral density domain within a long-enough window as the noise estimate of the current frame. This method gets rid of VAD and is capable of tracking noise changes even within speech segments.

4.1 Recognition results

Table 1 shows the results for the MSNE based spectral subtraction (MSNE-SS) and its combination with the *a posteriori* SNR weighted energy based VFR. It is observed that the combination of the proposed SNR-LogE-VFR and MSNE-SS achieves a 17.1% absolute WER reduction. Interestingly, the improvement of the combined method is greater than the summation of the gains obtained by applying the two methods individually (11.1% for SNR-LogE-VFR and 5% for MSNE-SS) – it is often the opposite way when combining two methods. This justifies the dual contributions of speech enhancement when combined with the VFR method.

	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	19.8	26.3	18.3	21.9	21.6	1.3

Table 1 Percent WER for MSNE-SS and its combination with the VFR for Test Set A

5 Combination with cepstral-domain method

The joint time- and spectral-domain method presented in the previous section is further combined with the MVA (cepstral mean subtraction, variance normalization and ARMA filtering) method [2]. Here, the MVA processing is applied to the static MFCC features only.

5.1 Recognition results

The results for combining SNR-LogE-VFR, MSNE-SS and MVA are given in Table 2. The performance for the MVA is cited from [2]. The results show that the combination with MVA further improves the performance and suggest that the VFR method is orthogonal to other methods. The method is expected to benefit from combination with other advanced methods as well.

	0 ~ 20 dB					Clean
	Subway	Babble	Car	Exhibition	Average	
MVA	-	-	-	-	24.8	1.0
MSNE-SS + MVA+ SNR-LogE-VFR	20.3	19.2	16.6	20.1	**19.0**	**1.4**

Table 2. Percent WER across the methods for Test Set A

5.2 Analysis of recognition error types

In [5] it was revealed, by analyzing recognition error types, that VFR analysis reduces insertion errors significantly. Table 3 shows the same analysis for the *a posteriori* SNR weighted energy based VFR and its combination with MSNE and MVA methods.

5 dB				
	H	D	S	I
Baseline	1982	260	1066	1095
SNR-LogE-VFR	2202	359	747	134
MSNE-SS+SNR-LogE-VFR	2279	289	740	216
MSNE-SS + MVA+ SNR-LogE-VFR	2551	411	346	42
Clean				
	H	D	S	I
Baseline	3285	10	13	10
SNR-LogE-VFR	3268	6	34	18
MSNE-SS+SNR-LogE-VFR	3265	12	31	4
MSNE-SS + MVA+ SNR-LogE-VFR	3260	14	34	5

Table 3. Number of correct words (H), deletions (D), substitutions (S) and insertions (I) on clean speech and speech corrupted by “Babble” noise (in total 3308 words).

It is particularly interesting to study the 5 dB case. Number of correct words steadily increases and number of substitutions steadily decreases after applying VFR, MSNE-SS and MVA. Number of insertions decreases significantly and number of deletions increases a bit after applying VFR and MVA.

6 Conclusion

This paper has studied the 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 applied to noise-robust speech recognition and was further combined with spectral- and cepstral-domain methods. Encouraging results were obtained.

References

[1] S. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2), 113-120 (1979)

[2] C.-P. Chen and J. A. Bilmes, “MVA Processing of Speech Features,” *IEEE Transactions on Audio, Speech and Language Processing*, 15(1), 257-270 (2007)

[3] R. Sarikaya and J.H.L. Hansen, “Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition,” in Proc. Eurospeech, Aalborg, Denmark, Sepember 2001

[4] H. Xu, Z.-H. Tan, P. Dalsgaard and B. Lindberg, “Exploitation of spectral variance to improve robustness in speech recognition,” *IEE Electronics Letters*, 42(5), 312-314 (2006)

[5] Z.-H. Tan, “Variable frame rate analysis for automatic speech recognition,” in Proc. SPIE Multimedia systems and Applications X, Boston, MA, USA, September 2007.

[6] 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.

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

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

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

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

[12] Z.-H. Tan and B. Lindberg, “A posteriori SNR weighted energy based variable frame rate analysis for speech recognition”, submitted to *Interspeech 2008*, Brisbane, Australia, September 2008.

[13] Z.-H. Tan and B. Lindberg, "A variable frame rate method for distributed speech recognition over wireless networks,” The 10th International Symposium on Wireless Personal Multimedia Communications, Jaipur, India, December 2007.

[14] 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.

[15] Young, S. J. et al., “HTK: Hidden Markov Model Toolkit V3.2.1, Reference Manual”, Cambridge Univ. Speech Group, Mar. 2004.

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