VARIABLE FRAME RATE AND LENGTH ANALYSIS FOR DATA COMPRESSION IN DISTRIBUTED SPEECH RECOGNITION

Ivan Kraljevski[1], Zheng-Hua Tan[2]

[1]voice INTER connect GmbH, Dresden, Germany
[2]Department of Electronic Systems, Aalborg University, Aalborg, Denmark
ivan.kraljevski@voiceinterconnect.de[1], zt@es.aau.dk[2]

Abstract: This paper addresses the issue of data compression in distributed speech recognition on the basis of a variable frame rate and length analysis method. The method first conducts frame selection by using a posteriori signal-to-noise ratio weighted energy distance to find the right time resolution at the signal level, and then increases the length of the selected frame according to the number of non-selected preceding frames to find the right time-frequency resolution at the frame level. It produces high frame rate and small frame length in rapidly changing regions and low frame rate and large frame length for steady regions. The method is applied to scalable source coding in distributed speech recognition where the target bitrate is met by adjusting the frame rate. Speech recognition results show that the proposed approach outperforms other compression methods in terms of recognition accuracy for noisy speech while achieving higher compression rates.

Keywords: Distributed Speech Recognition; Variable Frame Rate; Variable Frame Length

1 Introduction

Speech signals are considered non-stationary and exhibit quasi-stationary behavior in short periods. Therefore, speech analysis is generally done by short-time fixed frame rate (FFR) and fixed frame length (FFL) analysis. On the other hand, the signals have varying time-frequency characteristics and some parts of the speech sequence (e.g., vowels) are stationary over longer periods compared to others (e.g., consonants and transient speech). To respond to these observations, variable frame rate (VFR) and length (VFL) analysis methods have been proposed for automatic speech recognition (ASR) and speaker recognition [1]-[8].

VFR analysis selects frames according to the signal characteristics by first extracting speech feature vectors (frames) at a fixed frame rate, and then determining which frames to retain. The decision on frame selection relies on some distance measures and thresholds [2]-[5]. In [3], an effective VFR method was proposed that uses a 25ms frame length and a 2.5ms frame shift for calculating Mel-frequency cepstral coefficients (MFCCs) and conducts frame selection based on an energy weighted cepstral distance. The method improves the noise-robustness of ASR. In [4], an entropy measure instead of a cepstral distance is used, resulting in further improvement in recognition performance at the cost of higher complexity. An effective energy based frame selection method was proposed in [5] where it uses delta logarithmic energy as the criterion for determining the size of the frame shift on the basis of a sample-by-sample search. A low-complexity VFR method based on the measurement of a posteriori signal-to-noise ratio (SNR) weighted energy was proposed in [2].

VFR and VFL analysis has been used for speaker verification in [6] and speech recognition in [7]. As noted in [6], VFL analysis is more effective than FFL to accommodate the varying time-frequency acoustic phonetic characteristics. For a speech recognition task in laboratory conditions, paper [7] presents a variable frame rate and length algorithm in which frame rate is doubled and frame length is halved in transition regions.

The a posteriori SNR weighted energy distance based VFR method proposed in [2] has shown to be able to assign more frames to fast changing events and less frames to steady regions even for very low SNR signals. This method provides a natural way for dynamically changing frame length: extend the frame length when there are less frames selected. As a result, the frame length is kept as normal in the fast changing regions whereas it is increased in the steady regions. This leads to a joint VFR and VFL analysis (VFRL) for noise robust speech recognition in [8], which is able to further improve ASR performance in noisy environments.

The method in [8] was originally developed to provide good speech recognition performance in adverse conditions. Due to the nature of VFR and VFL analysis, it is considered as suitable also for data compression in distributed speech recognition (DSR). In [2] and [9], VFR analysis is applied for DSR to reduce the transmission bitrate of feature streams by not sending redundant frames that can be reconstructed at the remote server from the received frames. In this paper we present the application of the joint VFR and VFL analysis for DSR and investigates the effects of different compression levels and noise conditions.

The remainder of this paper is organized as follows:

Section II presents the variable frame rate and length algorithm. Section III presents the application of the VFRL method in distributed speech recognition. The experimental results and discussions are given in Section IV. Section V concludes this work.

2 Variable frame rate and length algorithm

This section presents an a posteriori SNR weighted energy distance based VFRL method and shows the results of frame selection and length determination.

2.1 The VFRL analysis algorithm

The variable frame rate and length analysis based on a posteriori SNR weighted energy distance [8] is realized through the iterative algorithm (Figure 1).

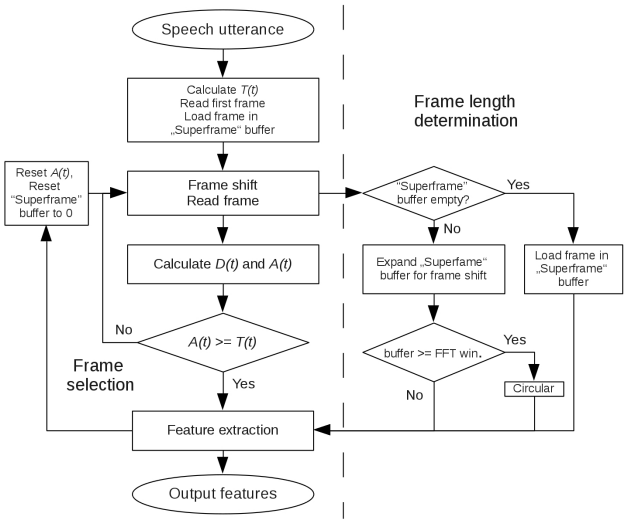

Figure 1 Diagram of the VFRL analysis method

Step1. The number of frames are determined and the value for frame selection threshold *T(t)* is calculated from the equation:

$$T(t) = \overline{D(t)} \cdot f(\log E_{noise}(t)) \tag{1}$$

where $\overline{D(t)}$ is the weighted energy distance *D(t)* averaged over a certain period. It can be calculated over one utterance for simplicity, but in practice, usually $\overline{D(t)}$ is estimated over the preceding segment. The a posteriori SNR weighted energy distance *D(t)* of two consecutive frames is calculated as:

$$D(t) = |logE(t) - logE(t-1)| \cdot SNR_{post}(t) \tag{2}$$

where *E(t)* is the energy of frame *t*, and $SNR_{post}(t)$ is the a posteriori SNR value of the frame that is defined as the logarithmic ratio of the energy of noisy speech *E(t)* to the energy of noise $E_{noise}(t)$. The function $f(logE_{noise}(t))$ is a sigmoid function of $logE_{noise}(t)$ to allow a smaller threshold and thus a higher frame rate for clean speech. The sigmoid function is defined with the following equation, where $\alpha = 9.0$, $\beta = 2.5$ and $\gamma = 13$. The constant $\gamma = 13$ is chosen so that the turning point of the sigmoid function is at an a posteriori SNR value of between 15dB and 20dB:

$$f(logE_{noise}(t)) = \alpha + \frac{\beta}{1 + e^{2 \cdot logE_{noise}(t) - \gamma}} \tag{3}$$

Since the algorithm is used in offline mode, the value for *T* is estimated over all available frames and it is not changed over time. However, in live mode the value of *T(t)* can be estimated over preceding segments and updated on a frame-by-frame basis. Afterwards, the first frame is loaded in the input buffer and a separate "Superframe" linear buffer is established with the same frame content. The initial size of the buffer is set on 25ms (200 samples for the 8kHz sampled speech).

Step 2. The frame is moved with the defined time shift (1ms) and the content of the following frame is loaded into the input buffer. In the same time the "Superframe" buffer is extended with the number of samples of the frame shift.

Step 3. The accumulative distance is updated per frame basis: *A(t)* += *D(t)* and it is compared against the threshold *T(t)*. If *A(t)* < *T(t)* the process continues again from Step 2. The "Superframe" buffer length is determined by the frequency of the frame selection events and the initial frame size is expanded gradually by the number of samples from the last frame selection point. If there are no significant changes in the frame energy, the function *D(t)* will grow slowly and the "Superframe" length will reach the size of the pre-defined FFT analysis window (unless specified otherwise, it is 32ms or 256 samples).

When the "Superframe" length equal the FFT window, the input buffer is changed from linear into circular and its length is kept fixed until the moment of frame selection. That means, the selected frame will have a length between the initial defined size (25ms) and the length of the FFT analysis window (32ms).

Step 4. When the condition *A(t)* ≥ *T(t)* is met, the current frame is selected and the feature extraction process is carried on the "Superframe" buffer and the resulting frame based feature vector is extracted and stored.

A(t) is reset back on zero and the "Superframe" buffer is re-initialized again on the initial length and the consecutive frame content and the execution continues again from Step 2. The process is iteratively repeated until the end of the input speech utterance.

2.2 Frame selection and length determination results

Figure 2 shows the results of the VFRL method for (a) clean speech, and (b) 5dB noisy speech, where the first panel of each sub-figure shows the waveform and selected frames and the second panel shows the length of selected frames with dashed line illustrating the initial length of 25ms. From the figure it can be seen that more frames were selected in regions with higher SNR values

and rapidly changing characteristics for both clean and noisy speech. The length of these high-dense frames is close to the initial frame length.

For the low SNR or steady regions, fewer frames are selected and the length of these frames is larger and limited by the maximal defined value. These characteristics are desirable for VFRL analysis in DSR.

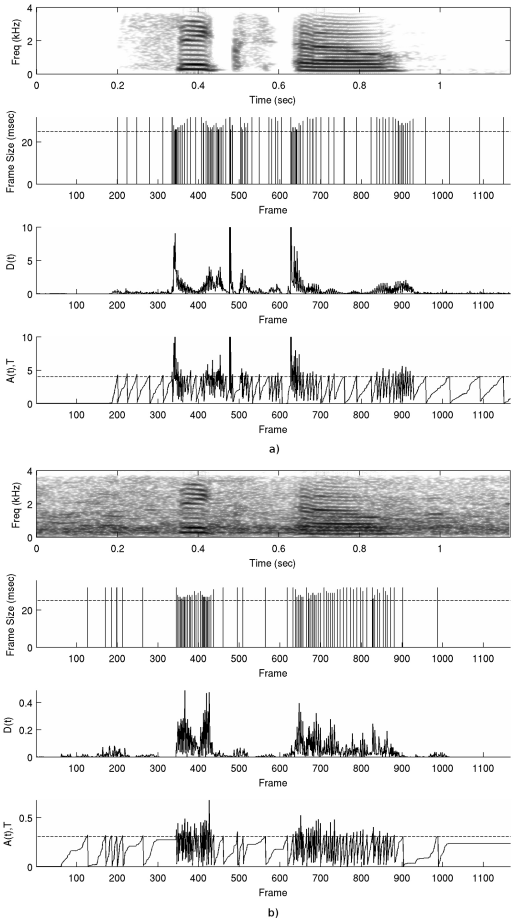

Figure 2 Frame selection and length determination results: (a) For clean speech: spectrogram (the first panel) and the selected frames and their length (the second panel, with the dashed line showing the initial length of 25ms), *D(t)* in Equation (1) (the third panel) and *A(t)* in Equation (4) (the fourth panel, with the dashed line showing *T* in Equation (2)); (b) for 5dB noisy speech with the same order of panels as in (a), respectively.

3 Scalable speech compression in distributed speech recognition

To benefit from the resources available over networks, DSR employs the client-server architecture and submits the computation-intensive ASR decoding task to a powerful server [10]. Specifically, speech recognition features are compressed and transmitted through networks to a server, where they are decoded and used in the recognition phase. The ETSI-DSR standard [11] uses split vector quantization to compress MFCC features. An efficient compression method applied in DSR is the two-dimensional discrete cosine transform (2D-DCT) based code [12]. In [13] the group of pictures concept (GoP) from video coding was applied to DSR to achieve a variable bitrate inter-frame compression scheme.

As shown in Figure 2, the VFRL implementation in DSR provides a high time resolution for fast changing events and a low time resolution for steady regions. Moreover, the extended frame lengths give large overlaps between consecutive frames. As a result, there is a big potential for scalable data compression by simply reducing the frame rate. It is important to note that, as shown in [14], there is a strong correlation between the number of states of the back-end hidden Markov models (HMMs) and the frame rate used in the front-end and a mismatch between the two significantly degrades the speech recognition performance. After reducing the frame rate at the client side, it is necessary to restore the proper frame rate at the server side by e.g. repetition or interpolation.

4 Speech recognition experiments

To evaluate the VFRL analysis based scheme for DSR, speech recognition experiments were conducted on Test Set A of the Aurora 2 database across different methods and compression levels.

4.1 Database

Experiments in this work were conducted on the Aurora 2 database [15], which is the TI digits database artificially distorted by adding noise and using a simulated channel distortion. The sampling rate is 8kHz. Whole word models were created for all digits using the HTK recognizer [16] and trained on clean speech data. For testing, Test Set A was used. The four noise types in Test Set A are ”subway”, ”babble”, ”car”, and ”exhibition” and the testing conditions include Clean, 20dB, 15dB, 10dB, 5dB and 0dB. MFCCs are used as the speech features with 12 coefficients (without c0) and logarithmic energy and their delta and delta-delta features, resulting in 39 dimensions. The FFR and FFL baseline method is the ETSI Distributed Speech Recognition (DSR) [11] with a frame length of 25ms and a frame rate of 100Hz.

4.2 Recognition results

In this work, we implemented two compression schemes after applying split vector quantization and the resulting data streams have bitrates of around 1.8kbps and 1.2kbps, respectively. The original frame rate is restored by frame repetition before HMM decoding in the server. Both the speech for training acoustic models and the speech for testing were quantized.

Table I. Percent WER across the methods, averaged over 0-20dB and four noise conditions.

Methods	Bitrate (kbps)	Noisy speech	Clean speech
FFR baseline (ETSI-DSR)	4.4	39.8	1.0
2D-DCT	1.4	40.5	1.6
GOP	2.6	N/A	2.5
GOP	1.3	N/A	2.6
SNR-LogE-VFR	~1.5	32.8	1.2
SNR-LogE-VFRL	~1.8	27.1	2.1
SNR-LogE-VFRL	~1.2	29.3	2.5

The recognition results are shown in Table I, where the results for 2D-DCT and GoP are cited from [12] and [13], respectively. Since there could be a mismatch in the train/test sets between the various simulation systems in the references, the comparisons are indicative only.

The table shows that the proposed SNR-LogE-VFRL approach gives 12.7% absolute improvement in recognition accuracy over the FFR baseline at a bitrate of 40.9% (1.8 kbps vs. 4.4 kbps and 10.5% absolute recognition improvement at a bitrate of 27.3% (1.2kbps vs. 4.4 kbps). Further, the proposed method achieves high compression rates at reduced computational cost since there are fewer frames (MFCC features) to be calculated due to the low bitrate/frame rate.

Table II compares the proposed approach with SNR-LogE-VFR for 1.5kbps, 1.8kbps and 1.2kbps for various noise types and SNR levels. It is noticed that the improvement for SNR-LogE-VFRL 1.8kbps and 1.2kbps is achieved for all noise types. SNR-LogE-VFRL 1.8kbps significantly outperforms the one with 1.5kbps also for all noise levels, while SNR-LogE-VFRL 1.2kbps outperforms the one with 1.5kbps for all noise levels except for 20dB.

Table II. Percent WER across the methods and bitrates, averaged over 0-20dB and four noise conditions.

	SNR-LogE-VFR	SNR-LogE-VFRL	SNR-LogE-VFRL
	~1.5 kbps	~1.8 kbps	~1.2 kbps
Average	32.8	27.1	29.3
Subway	34.3	26.7	31.0
Babble	30.9	26.9	29.8
Car	33.0	26.0	27.9
Exhibition	33.0	26.9	28.6
20dB	6.4	6.1	7.4
15dB	13.4	10.0	11.6
10dB	25.4	18.0	20.4
5dB	45.7	35.5	38.4
0dB	73.1	66.0	68.9

Further experiments were conducted to investigate the behavior of VFR and VFRL with different compression rates through the analysis of recognition error types. Table III shows the number of correctly recognized words, the number of recognition errors in different types and the percent WER for speech corrupted by "exhibition" noise at 10dB. It should be noted that the improvement comes from all types of recognition errors and from the correctly recognized words.

Table III. Number of correctly recognized words, substitutions, deletions, insertions, and percent WER ("Exibition" noise at 10dB, in total 3241 words).

	Corr.	Del.	Sub.	Ins.	WER (%)
SNR-LogE-VFR 1.5k	2473	170	598	53	25.3
SNR-LogE-VFRL 1.8k	2712	126	403	37	17.5
SNR-LogE-VFRL 1.2k	2631	155	455	42	20.1

To provide more insight about the noise robustness brought by the SNR-LogE-VFR, Table IV provides the recognition results across different frame analysis methods without data compression (i.e. standard ASR where no data compression is applied).

The ETSI-DSR is used as the standard method to represent the FFR baseline with a FFL of 25ms. Cep-VFR refers to the energy weighted cepstral distance based VFR [3]. Cep-VFR+VAD is the combination of the Cep-VFR method with voice activity detection and the results for this method are cited from [4]. LogE-VFR is the energy-based VFR presented in [5] and the results are cited from this reference as well. The SNR-LogE-VFR is the a posteriori SNR weighted energy distance based VFR [2].

Table IV. Percent WER across the methods, averaged over 0-20dB and four noise conditions.

	Noisy speech	Clean speech
FFR baseline (ETSI-DSR)	38.7	1.0
Cep-VFR	29.5	3.5
Cep-VFR+VAD	30.0	1.4
LogE-VFR	31.4	1.1
SNR-LogE-VFR	28.7	1.4
SNR-LogE-VFRL	25.8	1.7

It is noticed that all VFR methods outperform the baseline FFR in noisy conditions. SNR-LogE-VFR has both a lower complexity and a better recognition performance as compared with the other VFR methods. SNR-LogE-VFRL further introduces 2.9% absolute improvement over SNR-LogE-VFR for noisy speech. As compared with the FFR-FFL baseline, the improvement is very significant from 38.7% to 25.8% in WER. The performance on clean speech decreases moderately.

5 Conclusions

In this paper we presented a data compression approach for distributed speech recognition. The approach is based on a computationally efficient variable frame length and rate (VFLR) method that uses a posteriori SNR weighted energy distance for frame selection. It was demonstrated that using variable frame rate (VFR) analysis for DSR provides higher data compression rates and more efficient data transmission to the server side. In the same time, introducing variable frame length (VFL) analysis and selecting relatively longer frames has positive impact on the noise-robustness of DSR. Speech recognition experiments confirmed that the proposed variable frame rate and length method improves the performance for low bitrate source coding compared to the other methods commonly used in DSR. The combination of the variable frame rate and length introduces twofold benefit: lower bitrate source coding and improved noise robustness, effectively improving the overall performance of distributed speech recognition systems, in the terms of computational and transmission efficiency as well as noise robustness.

References

[1] T. Wu, Feature Selection in Speech and Speaker Recognition. PhD Thesis, Katholieke Universiteit Leuven, 2009.

[2] Z-H. Tan and B. Lindberg, Low-complexity variable

frame rate analysis for speech recognition and voice activity detection, IEEE Journal of Selected Topics in Signal Processing, 2010, 4, (5).

[3] Q. Zhu, A. Alwan, On the use of variable frame rate analysis in speech recognition. ICASSP-2000, Istanbul, Turkey, June 2000.

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

[5] J. Epps, E. Choi, An energy search approach to variable frame rate front-end processing for robust ASR. Eurospeech-2005, Lisbon, Portugal, September 2005.

[6] C-S. Jung, KJ. Han, H. Seo, S.S. Narayanan, and H.G. Kang, A variable frame length and rate algorithm based on the spectral Kurtosis measure for speaker verification. INTERSPEECH, Makuhari, Japan, September 2010.

[7] K. Samudravijaya, Variable frame size analysis for speech recognition. Int. Conf. Natural Language Processing, New Delhi, India, 2004.

[8] Z-H. Tan and I. Kraljevski, Joint variable frame rate and length analysisf or speech recognition under adverse conditions, Manuscript submitted for publication.

[9] H. Deng, D. O'Shaughnessy, J. Dahan, W. F. Ganong, Interpolative variable frame rate transmission of speech features for distributed speech recognition, 2007 IEEE Automatic Speech Recognition and Understanding Workshop, Kyoto, Japan, December 2007.

[10] Z-H. Tan, B. Lindberg (eds.). Automatic Speech Recognition on Mobile Devices and Over Communication Networks. Springer-Verlag, London, 2008.

[11] ETSI. Speech processing, transmission and quality aspects (STQ), distributed speech recognition, advanced front-end feature extraction algorithm, compression algorithm. ES 202 050 v1.1.1, 2002.

[12] 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 ICASSP04, Montreal, Canada, pp. 69-72, 2004.

[13] B. J. Borgstrom and A. Alwan, A packetization and variable bitrate interframe compression scheme for vector quantizer-based distributed speech recognition. In Proc. Interspeech07, Antwerp, Belgium, 2007.

[14] Z.-H. Tan, P. Dalsgaard, and B. Lindberg, Exploiting temporal correlation of speech for error-robust and bandwidth-exible distributed speech recognition. IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 4, pp. 1391-1403, May 2007.

[15] H.G. Hirsch, D. Pearce, The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy condition. ISCA ITRW ASR, Paris, France, September 2000.

[16] S.J. Young et al, HTK: Hidden Markov Model Toolkit V3.2.1, Reference Manual. Cambridge, U.K.: Cambridge Univ. Speech Group, 2004.