Computers and Electrical Engineering 40 (2014) 2139–2149

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Joint variable frame rate and length analysis for speech recognition under adverse conditions ☆

Zheng-Hua Tan [a,*], Ivan Kraljevski [b]

[a] Department of Electronic Systems, Aalborg University, Aalborg 9220, Denmark
[b] voice INTER connect GmbH, Dresden 01067, Germany

ARTICLE INFO

Article history:
Received 9 November 2013
Received in revised form 10 September 2014
Accepted 11 September 2014
Available online 11 October 2014

Keywords:
Frame selection
Noise-robust speech recognition
Variable frame rate
Variable frame length

ABSTRACT

This paper presents a method that combines variable frame length and rate analysis for speech recognition in noisy environments, together with an investigation of the effect of different frame lengths on speech recognition performance. The method adopts frame selection using an *a posteriori* signal-to-noise (SNR) ratio weighted energy distance and increases the length of the selected frames, according to the number of non-selected preceding frames. It assigns a higher frame rate and a normal frame length to a rapidly changing and high SNR region of a speech signal, and a lower frame rate and an increased frame length to a steady or low SNR region. The speech recognition results show that the proposed variable frame rate and length method outperforms fixed frame rate and length analysis, as well as standalone variable frame rate analysis in terms of noise-robustness.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Speech signal analysis is generally performed over short-time frames with a fixed length (FFL) and a fixed frame rate (FFR), based on the assumption that speech signals are non-stationary and, exhibit quasi-stationary behavior in short durations. This fixed frame rate and length (FFRL) analysis is not optimal, since some parts of the signals (e.g. vowels) are stationary over a longer duration compared to others (e.g. consonants and transient speech) that have shorter durations. Consequently, variable frame rate (VFR) and variable frame length (VFL) analysis methods have been proposed for speaker recognition and speech recognition [1,2].

Variable frame rate analysis selects frames according to the signal characteristics. Initially, speech feature vectors (frames) are first extracted at a fixed frame rate and then the decision for the retaining frames is based on distance measures and thresholds [3–5]. The Euclidean distance between the last retained feature vector and the current vector is calculated as the distance measure in [3]. The current frame is discarded if the measure is smaller than the predefined threshold, aimed at reducing the computational load.

Recent research in VFR analysis moves towards finding optimal representation of a speech signal to improve performance in noisy environments. This requires frame analysis in steps smaller than the standard 10 ms, while the average frame rate largely remains unchanged. In [4], an effective VFR method was proposed, that uses a 25 ms frame length with a 2.5 ms frame shift for calculating Mel-frequency cepstral coefficients (MFCCs) and, conducts frame selection based on an energy

☆ Reviews processed and approved for publication by the Editor-in-Chief.

* Corresponding author.
E-mail addresses: zt@es.aau.dk (Z.-H. Tan), ivan.kraljevski@voiceinterconnect.de (I. Kraljevski).

http://dx.doi.org/10.1016/j.compeleceng.2014.09.002
0045-7906/© 2014 Elsevier Ltd. All rights reserved.

weighted cepstral distance. The method significantly improves the recognition accuracy in noisy environments at the cost of degraded performance for clean speech. In [5], an entropy measure instead of a cepstral distance is used, resulting in recognition performance improvement and higher complexity. To provide a fine resolution for rapidly changing events, these methods examine speech signals at much shorter intervals (i.e. 2.5 ms) compared to the normal frame shift of 10 ms. The algorithms extract features such as MFCCs and entropy at a high frame rate for frame selection, which is computationally expensive. An effective energy based frame selection method was proposed in [6] and 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. Evidently, energy based search is more computationally efficient. Speech segments are accounted in speech recognition not only on their characteristics (measured by MFCCs, energy and so on), but also on their reliability. Therefore, a low-complexity VFR method, based on the *a posteriori* signal-to-noise ratio (SNR) weighted energy distance was proposed in [2].

While VFR analysis has been used for improving the noise-robustness of speech recognition – a primary challenge in the field, to the best of our knowledge, VFL analysis has rarely been exploited in dealing with this problem. One exception is a pseudo pitch synchronous analysis method that uses variable frame size and/or frame offset to align frames to natural pitch cycles [7]. Three pitch synchronization methods are presented: depitch, syncpitch and padpitch. On Aurora 2 database, using multi-condition training, all these methods perform worse than the baseline (without pitch synchronization processing) for clean, 20 dB, 15 dB and 10 dB conditions. Depitch is worse than the baseline on all conditions, syncpitch only performs better than the baseline for −5 dB, and padpitch performs better for −5 dB, 0 dB and equally for 5 dB.

For general speech recognition, rather than focusing on noise-robustness, a speaking rate normalization technique that adjusts both the frame rate and frame size (i.e. VFRL) is implemented on a state-of-the-art speech recognition architecture and evaluated on the GALE broadcast transcription tasks [8]. By warping the step size and the window size in the front-end according the speaking rate, the technique shows consistent improvement on all systems and gives the lowest decoding error rates of the corresponding test sets. Instead of using fixed-length frames, a segment-based recognizer represents the observation space as a graph, in which each arc corresponds to a hypothesized variable-length segment [9].

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 or low SNR regions, even for very low SNR signals, thus significantly improving noise-robustness. The method can be combined with VFL analysis through a natural way of determining frame length: Extend the frame length when less frames are selected. Specifically, the lengths of the selected frames are extended when their preceding frames are not selected, for which motivations and details are presented in Section 2. As a result, the frame length is kept as normal in the fast changing regions, whereas it is increased in the steady or low SNR regions. The proposed VFRL method is applied to speech recognition in noisy environments.

As the VFRL method operates in the time domain in the sense that it decides which frame to retain, it has a good potential to be combined with other robustness methods which in general operate in the feature or model domain, to reduce the mismatch between the training and test speech signals. Feature based methods include feature enhancement, distribution normalization and noise robust feature extraction. Feature enhancement attempts to remove the noise from the signal, such as in spectral subtraction (SS) [10], non-local means de-noising [11] and vector Taylor series (VTS) [12]. Distribution normalization reduces the distribution mismatches between training and test speech, for example in cepstral mean and variance normalization (CMVN) [13]. Noise robust features include improved MFCCs [14], and the newly proposed features called power-normalized cepstral coefficients [15]. Acoustic modelling approach called deep neural networks [16] has recently attracted a significant amount of attention in the field of noise robust speech recognition. In this work, the VFRL analysis is combined with minimum statistics noise estimation based SS [10,17].

The remainder of this paper is organized as follows: Section 2 presents the proposed variable frame rate and length algorithm. The experimental results and discussions are given in Section 3. Section 4 investigates the effect of frame length on speech recognition performance. Finally, Section 5 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 illustrative results of frame selection and length determination.

2.1. Motivations

In general, VFRL analysis methods determine one of the frame analysis parameters (length or rate) first, and then use it as the basis for calculating the other in a relatively straightforward way.

Aiming at improved modelling of transition segments for speech recognition [18] presents a method where the frame shift is increased during stationary regions, while frame shift and frame length are decreased for non-stationary regions. Specifically, it uses MFCC based measures to determine local non-stationary, and then doubles the frame rate at transition regions and halves the frame size. In [19], if a transient frame is detected, the frame is segmented into two – each having the half of a normal frame length, which has shown improved recognition accuracy on TIMIT database. Ref. [8] presents a technique that adjusts both the frame rate and frame length according to the detected speaking rate, achieving impressive speech recognition performance. A pseudo pitch synchronous analysis method uses variable frame size and/or frame offset

to align frames to natural pitch cycles for speech and speaker recognition [7]. In [1], for speaker verification, fixed frame rate and length analysis is applied first, and then the frame length is iteratively expanded until the spectral kurtosis of the merged frame is less than the maximum value of the spectral kurtosis of two consecutive frames. After obtaining the frame length, the frame shift is simply set to half of the frame length.

Variable frame rate and length analysis methods commonly assign higher frame rates and smaller frame lengths to fast changing regions, and lower frame rates and larger frame lengths to steady regions. In this work, the frame length is expanded (to a maximal frame length, 32 ms unless otherwise stated) until the accumulative *a posteriori* SNR weighted distance is larger than a threshold and at the same time a new frame is selected.

The motivations are: (1) the first-order difference in frame-to-frame energy provides greater discrimination than the components of MFCCs other than c_0 [20]; (2) speech segments, besides their characteristics, are accounted also on their reliability e.g., measured by SNR; (3) the *a posteriori* SNR for noise-only segments will be theoretically equal to 0 dB, so that it acts as a soft voice activity detection; (4) both energy and *a posteriori* SNR are easy to estimate, resulting in a low complexity; (5) the accumulative distance measure provides a natural way for expanding the frame length to better represent the time–frequency characteristics of the segment, as larger frame lengths are assigned to steady regions; (6) the accumulative distance uses multiple frames, rather than only two frames, for the frame rate and size determination. Details about the variable frame rate analysis are available in [2].

2.2. The VFRL analysis algorithm

The flowchart of the VFRL algorithm is shown in Fig. 1. It operates iteratively from the beginning to the end of a speech file as follows:

Step 1. Load the first frame into a superframe and set the accumulative *a posteriori* SNR weighted energy distance $A(0) = 0$.
Step 2. Load the next frame and merge it with the superframe (i.e. expand the superframe to include the samples of the frame shift). If the length of the superframe is greater than the predefined maximal length, remove the samples from the beginning of the superframe to match the maximal length.
Step 3. Calculate the *a posteriori* SNR weighted energy distance of the two consecutive frames:

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

where $E(t)$ is the energy of frame t, and $\mathrm{SNR}_{post}(t)$ 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 $E_{noise}(t)$. Logarithmic energy distance is used due to its power for discrimination and SNR is used to take into account the reliability of speech segments. In addition, the computational complexity of this measure is much lower than some other calculations such as MFCC distance or entropy.

Step 4. Update the accumulative *a posteriori* SNR weighted energy distance:

$$A(t) = A(t-1) + D(t) \tag{2}$$

The accumulative distance is used instead of the distance between only two frames in order to take into account more frames for frame selection decision.

Step 5. Compare the accumulative distance with a threshold $T(t)$ (to be described below).

If $A(t) < T(t)$, check whether there are more frame(s) to be dealt with. If yes, go to Step 2; if no, the process terminates. Otherwise, output the superframe (since maximal frame length control is done in Step 2 when each new frame is loaded, there is no need to do it again here). If the current frame is not the last one, load the next frame into a superframe, set the accumulative *a posteriori* SNR weighted energy distance $A(0) = 0$, and go to Step 3; otherwise, the process terminates.

The threshold $T(t)$ for frame selection is computed as:

$$T(t) = \overline{D(t)} \cdot f(\log\ E_{noise}(t)) \tag{3}$$

where $\overline{D(t)}$ is the average weighted distance over a certain period, which can be calculated over one utterance for simplicity. In practice, $\overline{D(t)}$ is calculated over preceding frames. The function $f(\log E_{noise}(t))$ is a sigmoid function of $\log E_{noise}(t)$ to allow a smaller threshold and thus a higher frame rate for clean speech. The sigmoid function is defined by the following equation:

$$f(\log\ E_{noise}(t)) = \alpha + \frac{\beta}{1 + e^{-2(\log\ E_{noise}(t) - \gamma)}} \tag{4}$$

where $\alpha = 9.0$, $\beta = 2.5$, $\gamma = 13$. The constant $\gamma = 13$ is chosen so that the turning point of the sigmoid function is at an *a posteriori* SNR value of between 15 dB and 20 dB. The motivation is to select more frames for clean speech and relatively less for noisy speech (as more frames for noisy speech can result in high insertion errors).

In this work, for the VFRL algorithm, unless otherwise stated, the frame shift is 1 ms and the initial and maximal frame lengths are 25 ms and 32 ms, respectively.

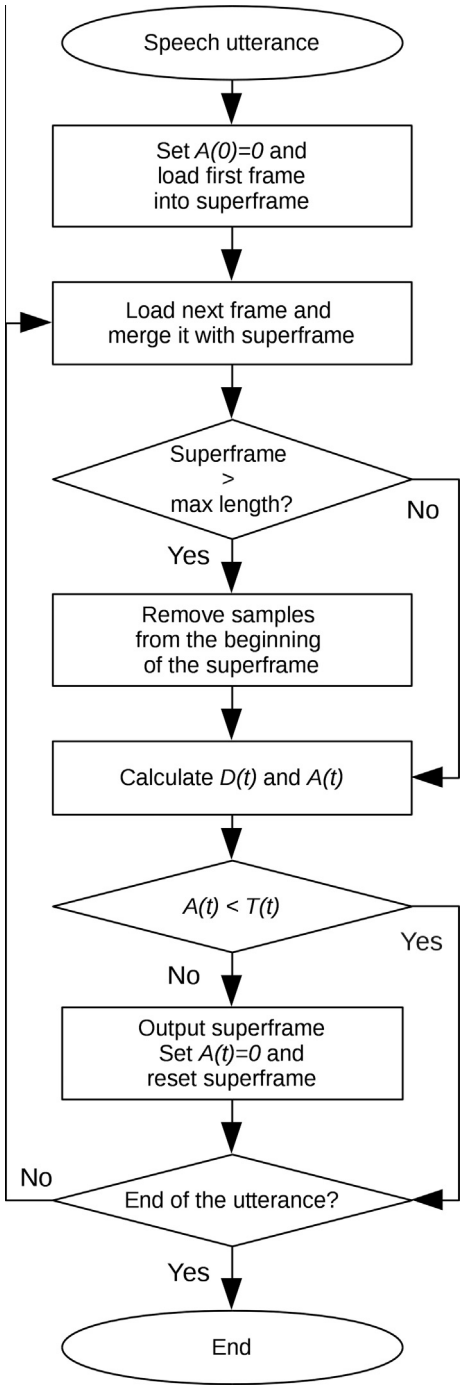

Fig. 1. Flowchart of the proposed VFRL method.

2.3. Frame selection and length determination results

Fig. 2 depicts the results of the *a posteriori* SNR weighted energy distance based VFRL for (a) clean speech and (b) 5 dB noisy speech, where the panels of each sub-figure show the spectrogram (the first panel), the selected frames and their lengths (the second panel, with the dashed line showing the initial length of 25 ms), the $D(t)$ in Eq. (1) (the third panel), and $A(t)$ in Eq. (2) (the fourth panel, with the dashed line showing $T(t)$ in Eq. (3)), respectively. From Fig. 2(a) it can be observed that more frames with normal or slightly greater than normal frame lengths are selected in regions with rapidly

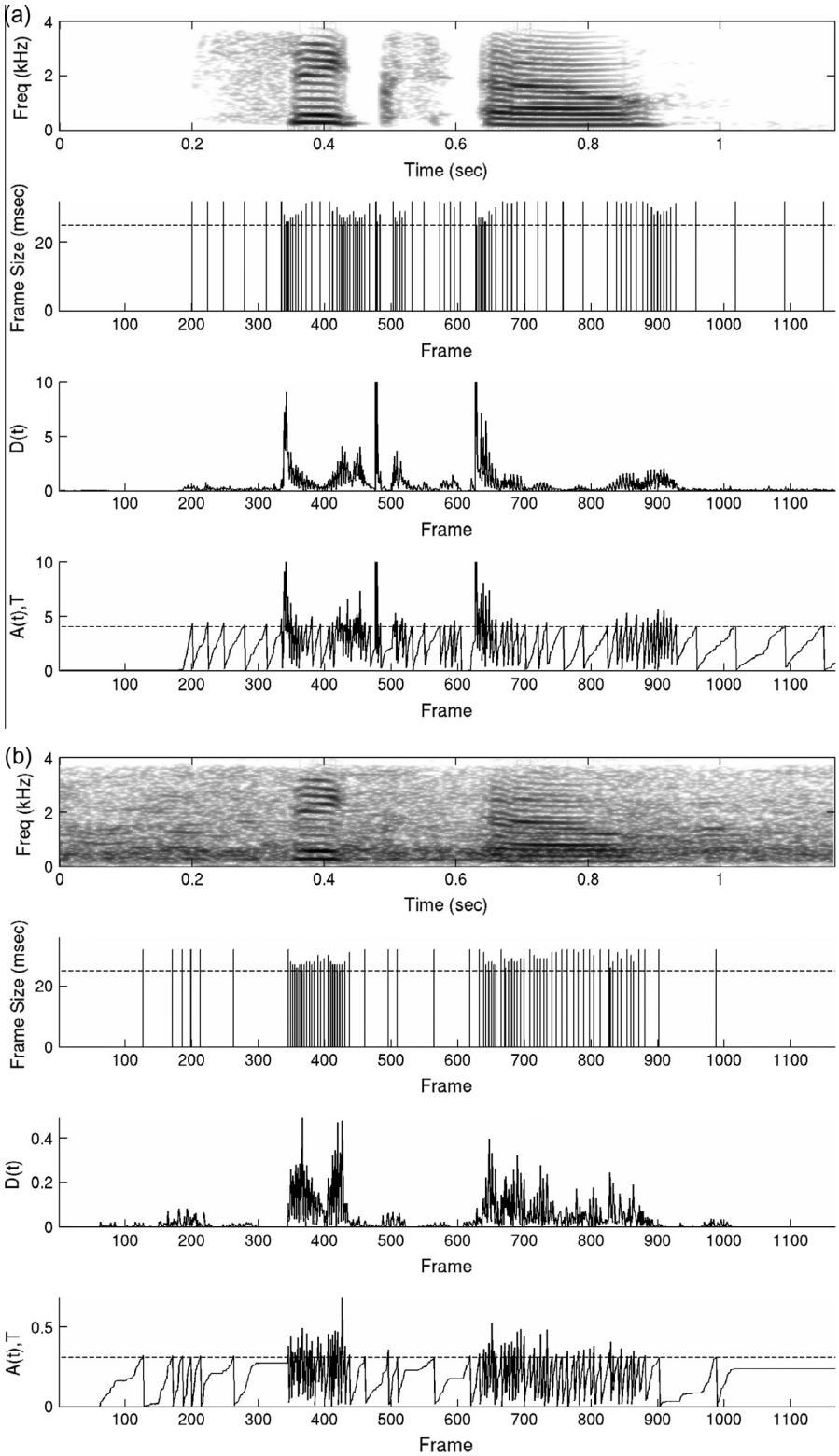

Fig. 2. Frame selection and frame length determination results: (a) for clean speech: spectrogram (the first panel) and the selected frames and their length (the second panel, with the dashed line showing the initial length of 25 ms), $D(t)$ in Eq. (1) (the third panel) and $A(t)$ in Eq. (2) (the fourth panel, with the dashed line showing $T(t)$ in Eq. (3)); (b) for 5 dB noisy speech with the same order of panels as in (a), respectively. In (a) there are a few strikes (large values) for $D(t)$ and $A(t)$ that have been cut off to better show the details.

changing characteristics for the clean speech signal. In steady regions, fewer frames with increased frame lengths are selected. In silence regions, almost no frames are selected. From Fig. 2(b) it can be observed that more frames with normal or slightly greater than normal frame lengths are selected in fast changing and high SNR regions for noisy signals. Fewer or no frames with maximal frame length are selected in non-speech regions. These results are desirable for VFRL analysis.

3. Speech recognition experiments and discussions

This section evaluates the proposed method through a number of experiments. Further we combine it with spectral-domain method and present experimental results.

3.1. Database and experimental setups

Experiments in this work were conducted with the Aurora 2 database [21], which is the TI digits database artificially distorted by adding noise and using a simulated channel distortion. The sampling rate is 8 kHz. Whole word models were created for all digits using the HTK recognizer [22] 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, 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. Each noise type and condition has 1001 test utterances, which gives 24,024 utterances in total for testing.

Throughout this work, MFCCs are used as the speech features for recognition. The features include 12 MFCCs (without c_0), logarithmic energy, and their corresponding velocity and acceleration components.

The FFRL baseline method is the ETSI Distributed Speech Recognition (DSR) Standard [23] with a frame length of 25 ms and a frame rate of 100 Hz.

3.2. Comparison with other methods

Table 1 compares the word error rate (WER) for FFRL, SNR-LogE-VFR and SNR-LogE-VFRL for various noise types and SNR values.

It can be noticed that for all noise types, SNR-LogE-VFRL outperforms SNR-LogE-VFR that again outperforms FFRL. SNR-LogE-VFRL significantly outperforms SNR-LogE-VFR for 0 dB, 5 dB and 10 dB. For 15 dB and 20 dB, SNR-LogE-VFR marginally outperforms SNR-LogE-VFRL on average, but not for all noise types. For clean speech, FFRL outperforms the others. In practice, the VFRL analysis can be switched off for clean speech and therefore the recognition performance will be the same as

Table 1
Percent WER across the methods.

		FFRL baseline (ETSI standard)	SNR-LogE-VFR	SNR-LogE-VFRL
Noisy speech	Average	38.7	28.7	25.8
Subway	Average	30.5	28.4	26.6
	20 dB	2.9	5.2	4.7
	15 dB	6.5	10.1	8.5
	10 dB	21.3	20.2	17.0
	5 dB	47.8	38.3	34.8
	0 dB	74.0	67.9	67.9
Babble	Average	50.1	27.8	26.0
	20 dB	9.8	3.8	4.9
	15 dB	26.2	7.3	9.0
	10 dB	50.6	18.2	17.1
	5 dB	73.2	40.0	34.1
	0 dB	90.7	69.8	64.9
Car	Average	39.4	29.2	24.9
	20 dB	2.6	4.1	5.3
	15 dB	10.0	7.9	8.9
	10 dB	33.0	18.5	15.8
	5 dB	65.9	40.9	32.3
	0 dB	85.5	74.6	62.3
Exhibition	Average	34.6	29.6	25.8
	20 dB	3.6	4.2	5.1
	15 dB	8.0	9.8	9.1
	10 dB	24.3	19.9	16.2
	5 dB	55.2	41.0	33.3
	0 dB	81.9	73.3	65.1
Clean speech	Average	1.0	1.4	1.7

that of FFRL, at the cost of an increased system complexity. Figs. 3 and 4 further show the performance trends for these methods. The proposed method constantly demonstrates the superior performance across the SNR values and noise types.

Further experiments were conducted to investigate the behavior of VFRL through the analysis of recognition error types. Table 2 shows the number of correctly recognized words, the number of recognition errors in different types and the WER results for speech corrupted by car noise at 10 dB, which was chosen as an example. It should be noted that the improvement comes from all types of recognition errors and from the correctly recognized words.

The recognition results for a number of methods are presented in Table 3. Additionally the table includes the results of FFRL with a frame length of 32 ms as a reference. Cep-VFR refers to the energy weighted cepstral distance based VFR [4]. Cep-VFR + VAD is the combination of the Cep-VFR method with voice activity detection and the results for this method are cited from [5]. LogE-VFR is the energy-based VFR presented in [6] 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]. Finally SNR-LogE-VFRL is the proposed VFRL method.

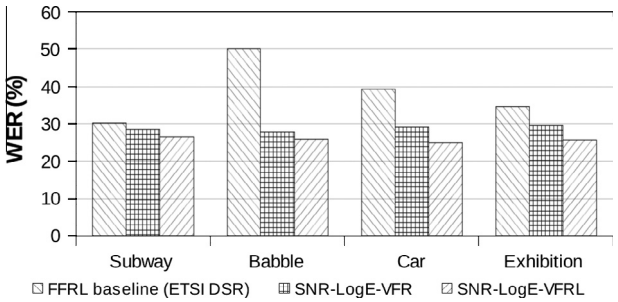

Fig. 3. Average WER performance of different methods across different noise types (as also given in Table 1).

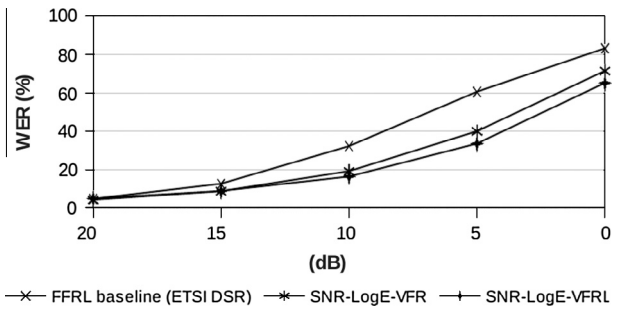

Fig. 4. Average WER performance of different methods across different SNR values (as also given in Table 1).

Table 2
Number of correctly recognized words, number of errors for different recognition error types and percent WER (car noise at 10 dB, in total 3353 words).

	Correct	Deletion	Substitution	Insertion	WER (%)
FFRL baseline	2739	154	460	492	33.0
SNR-LogE-VFR	2760	167	426	28	18.5
SNR-LogE-VFRL	2837	129	387	13	15.8

Table 3
Percent WER across the methods (for noisy speech, the WER is averaged over 0–20 dB and four noise conditions). The results for Cep-VFR + VAD are cited from [5] and the results for LogE-VFR are cited from [6].

	FFRL baseline (ETSI)	FFRL 32 ms	Cep-VFR	Cep-VFR + VAD	LogE-VFR	SNR-LogE-VFR	SNR-LogE-VFRL	MSNE-SS	SNR-LogE-VFRL + MSNE-SS
Noisy speech	38.7	36.8	29.5	30.0	31.4	28.7	25.8	33.7	20.2
Clean speech	1.0	1.0	3.5	1.4	1.1	1.4	1.7	1.5	1.3

Table 4
Percent WER for VFRL with different parameter settings (for noisy speech, the WER is averaged over 0–20 dB and four noise conditions).

	Noisy speech	Clean speech
FFRL baseline	38.7	1.0
($\alpha_0 = 9.0, \beta_0 = 2.5$)	25.8	1.7
($\alpha_0 + 0.5, \beta_0$)	26.6	1.8
($\alpha_0 - 0.5, \beta_0$)	25.8	1.7
($\alpha_0, \beta_0 + 0.5$)	25.6	1.7
($\alpha_0, \beta_0 - 0.5$)	26.0	1.7
($\alpha_0 - 1, \beta_0 + 1$)	25.4	2.0
($\alpha_0 + 1, \beta_0 - 1$)	27.2	1.8
($\gamma_0 - 1 = 12$)	25.7	1.7
($\gamma_0 + 1 = 14$)	25.9	1.7

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

It should be noted that in the implementation of VFRL, the noise energy value is first calculated as an average from the frames at the beginning of an utterance and is then replaced when a frame appears to have a lower energy. This implementation improves noise estimation especially when there are non-representative noise samples in the beginning, and it boosts the VFR method performance up from 28.7% to 28.3% for noisy speech and introduces no difference for clean speech.

SNR-LogE-VFRL is applied in combination with multi-condition training and compared with SNR-LogE-VFR and FFRL. The WER results of Test Set A for SNR-LogE-VFRL, SNR-LogE-VFR and FFRL are 16.2%, 15.2% and 12.2%, respectively. These results show that VFRL and VFR methods do not work well with multi-condition training, which is consistent with the findings in [6] (LogE-VFR) and [7] (pseudo pitch synchronous analysis).

3.3. Parameter settings

Regarding the threshold setting for frame selection, a number of parameters are used. Experiments have been conducted to investigate how much the performance varies according to the change of parameters. The experimental results are shown in Table 4. It can be noted that changing the parameters in a number of ways, does not result in dramatic changes in the performance.

3.4. VFRL in combination with spectral-domain method

VFRL analysis relies on some distance measures for frame selection. These measures, however, can be largely affected by noises that corrupt the speech signal. If the noisy speech signal is first de-noised by a speech enhancement method and thereafter analyzed by the VFRL method, it is expected that applying the speech enhancement method will both enhance the speech signal and improve the frame selection. The speech enhancement method adopted in this work, is the minimum statistics noise estimation (MSNE) based SS [10,17]. MSNE assumes that speech cannot occupy a frequency bin all the time and thus 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. The WERs for the MSNE-SS are 33.7% for noisy speech and 1.5% for clean speech. The combination of the VFRL and the SS achieves a performance of 20.2% and 1.3% for noisy speech and clean speech respectively, in comparison with the performance of 25.8% and 1.7% obtained by VFRL alone. This significant improvement in recognition performance indicates the VFRL method and the enhancement method compensates each other very well. This verifies that the enhancement improves the frame selection while at the same time de-noising the speech. These results are included in Table 3 for ease of comparison.

4. Analysis of the effect of frame length

An interesting and important question is how the different frame lengths have impact on the performance of speech recognition. In general, frame length for speech analysis is determined so that it is short enough to keep unchanged the speech properties of interest roughly within the frame and long enough to be able to estimate the desired parameters [24]. In [25], it is shown that longer speech segments can be recognized more accurately from noise compared to shorter ones in the context of speaker recognition. In this section, we evaluate how it influences speech recognition performance.

4.1. The effect of the frame length range of VFRL

In order to investigate how the maximal frame length influences the performance of the VFRL method, we conducted a number of experiments with the initial/minimal frame length being 25 ms as previously used and the maximal length varying from 32 ms to 64 ms (32 ms is the default maximal length and has been used for all previous experiments for the VFRL). Speech recognition results presented in Columns 3–6 of Table 5 show that 32 ms maximal length gives the best performance in almost all conditions. It can be noticed that the performances of 35 ms VFRL and 37.5 ms VFRL are worse than that of 32 ms VFRL, but they are still better than that of SNR-LogE-VFR for noisy speech (28.7% WER as shown in Table 1).

Furthermore, the performances for VFLR with a smaller initial frame of 20 ms and different maximal frame lengths (25 ms and 32 ms) were investigated. It can be seen that smaller initial frame lengths perform worse than 25 ms frame length due to the decreased resolution for the low frequencies. These results can well be explained by the fact that formant patterns are best exhibited when the frames are of a certain length around 25–30 ms.

4.2. The effect of the frame length of FFRL

We further investigate the effect of different frame lengths on FFRL based speech recognition. A number of experiments were conducted, with the frame shift fixed at 10 ms and the different fixed frame length ranging from 25 ms to 64 ms. Speech recognition results are presented in Table 6. The results show that increasing the frame length (up to 37.5 ms) improves speech recognition performance in noisy environments, while the performance for clean speech is maintained. The 32 ms FFRL outperforms the 25 ms FFRL for almost all conditions while equally performing for clean one. This indicates that longer speech segments in noisy environments can be more accurately recognized than short ones. The performance for 64 ms is slightly worse than others, which is reasonable as 64 ms is apparently too large frame length.

Note that FFRL gives the same WERs for clean speech regardless of the frame length while VFRL generates different WERs when different maximal frame lengths are used. The main reason for the different behaviors between the VFRL and FFRL as shown in Tables 5 and 6 is because FFRL has the frame length fixed (even though the lengths are different for different settings), but VFRL has the frame lengths varying within a range (e.g. 25–32 ms, 25–35 ms). What is more important for VFRL is the range, not only the maximal length. This is further justified by the same performance of VFRL on clean speech given by 25–32 ms and 20–25 ms as shown in Table 5.

Table 5
Percent WER for VFRL with different ranges of frame length.

Range of frame length		25–32 ms	25–35 ms	25–37.5 ms	25–64 ms	20–25 ms	20–32 ms
Noisy speech	Average	25.8	26.8	27.5	37.2	27.1	28.4
Subway	Average	26.6	28.0	28.7	37.8	27.9	29.2
	20 dB	4.7	6.4	6.2	14.1	5.6	7.9
	15 dB	8.5	9.3	10.5	19.9	9.5	12.0
	10 dB	17.0	18.3	19.2	29.9	17.1	20.6
	5 dB	34.8	36.7	38.6	51.1	37.6	38.3
	0 dB	67.9	69.4	69.1	74.8	69.6	67.0
Babble	Average	26.0	27.0	27.7	37.0	26.5	29.4
	20 dB	4.9	5.3	5.7	11.6	4.9	7.6
	15 dB	9.0	9.9	10.1	16.9	8.8	11.8
	10 dB	17.1	18.0	18.9	29.2	17.0	21.1
	5 dB	34.1	35.4	36.6	51.0	35.8	38.1
	0 dB	64.9	66.3	67.3	76.2	66.0	68.4
Car	Average	24.9	26.2	26.7	37.3	26.6	27.2
	20 dB	5.3	6.3	7.0	12.8	5.0	7.4
	15 dB	8.9	9.9	10.3	18.2	8.8	11.3
	10 dB	15.8	16.8	17.5	29.2	17.0	18.7
	5 dB	32.3	32.6	32.7	50.6	34.9	33.7
	0 dB	62.3	65.6	65.8	75.5	67.0	65.1
Exhibition	Average	25.8	25.9	27.5	36.8	27.3	27.7
	20 dB	5.1	5.8	6.2	11.8	5.4	6.8
	15 dB	9.1	8.8	9.5	17.7	9.2	11.0
	10 dB	16.2	15.8	17.0	28.0	17.1	18.1
	5 dB	33.3	33.9	34.8	50.9	36.4	36.3
	0 dB	65.1	65.5	65.9	75.4	68.5	65.5
Clean speech	Average	1.7	1.9	2.1	4.7	1.7	2.1

Table 6
Percent WER for FFRL with different frame lengths.

		25 ms	32 ms	35 ms	37.5 ms	64 ms
Noisy speech	Average	38.7	36.8	36.3	37.0	39.6
Subway	Average	30.5	28.4	28.5	30.4	34.6
	20 dB	2.9	3.1	3.1	2.9	3.4
	15 dB	6.5	6.2	6.6	6.8	9.4
	10 dB	21.3	18.6	18.8	20.8	27.5
	5 dB	47.8	42.8	43.1	46.9	54.7
	0 dB	74.0	71.5	71.0	74.3	78.0
Babble	Average	50.1	47.4	45.8	46.0	46.0
	20 dB	9.8	7.6	7.7	7.6	7.3
	15 dB	26.2	22.2	21.9	22.1	21.7
	10 dB	50.6	47.4	45.5	46.1	44.9
	5 dB	73.2	72.1	69.1	69.0	68.0
	0 dB	90.7	87.9	85.0	86.0	88.3
Car	Average	39.4	37.8	37.6	37.7	41.0
	20 dB	2.6	2.5	2.5	2.6	2.7
	15 dB	10.0	7.7	7.3	7.9	10.4
	10 dB	33.0	28.9	28.1	28.5	34.0
	5 dB	65.9	63.1	62.5	62.3	69.6
	0 dB	85.5	86.8	87.7	87.2	88.2
Exhibition	Average	34.6	33.5	33.3	33.8	37.0
	20 dB	3.6	3.3	3.2	3.3	3.3
	15 dB	8.0	7.3	7.2	7.1	9.3
	10 dB	24.3	21.0	21.9	21.5	28.1
	5 dB	55.2	53.4	51.8	53.6	59.5
	0 dB	81.9	82.3	83.5	83.8	84.7
Clean speech	Average	1.0	1.0	1.0	1.0	1.0

5. Conclusions

This paper has shown that the proposed variable frame length and rate method, using accumulative *a posteriori* SNR weighted energy distance, is able to assign more frames with normal lengths to fast changing events and, fewer frames with larger frame lengths to steady regions. The variable frame rate analysis targets at finding the right time resolution at the signal level while the variable frame length analysis targets at the right time–frequency resolution at the frame level for noisy speech. Speech recognition experiments verify that the proposed variable frame rate and length method improves speech recognition performance in noisy environments. The method was combined with the minimum statistics noise estimation based spectral subtraction method and good recognition performance was achieved. The effect of frame lengths on speech recognition was investigated to explain the behavior of the proposed variable frame length and rate method. It was found that setting the right range which frame length can vary between is as important as setting the maximal length that is allowed.

Future work is focused on applying the variable frame rate and length method to speaker identification and verification task.

References

[1] Jung C-S, Han KJ, Seo H, Narayanan SS, Kang HG. A variable frame length and rate algorithm based on the spectral Kurtosis measure for speaker verification. In: Proceedings of INTERSPEECH-2010, Makuhari, Japan; September 2010.
[2] Tan Z-H, Lindberg B. Low-complexity variable frame rate analysis for speech recognition and voice activity detection. IEEE J Sel Top Signal Proc 2010;4(5):798–807.
[3] Pointing KM, Peeling SM. The use of variable frame rate analysis in speech recognition. Comput Speech Lang 1991;5(2):169–79.
[4] Zhu Q, Alwan A. On the use of variable frame rate analysis in speech recognition. In: Proceedings of ICASSP-2000, Istanbul, Turkey; June 2000.
[5] You H, Zhu Q, Alwan A. Entropy-based variable frame rate analysis of speech signals and its application to ASR. In: Proceedings of IEEE ICASSP, Montreal, Quebec, Canada; 2004.
[6] Epps J, Choi E. An energy search approach to variable frame rate front-end processing for robust ASR. In: Proceedings of Eurospeech, Lisbon, Portugal; September 2005.
[7] Zilca RD, Kingsbury B, Navrátil J, Ramaswamy GN. Pseudo pitch synchronous analysis of speech with applications to speaker recognition. IEEE Trans Audio Speech Lang Process 2006;14(2):467–78.
[8] Chu S, Povey D. Speaking rate adaptation using continuous frame rate normalization. In: Proceedings of ICASSP 2010, Dallas, TX, USA; March 2010.
[9] Glass JR. A probabilistic framework for segment-based speech recognition. Comput Speech Lang 2003;17(2):137–52.
[10] Martin R. Spectral Subtraction based on Minimum Statistics. In: Proceedings of EUSIPCO, Edinburgh, Scotland, UK; September 1994.
[11] Xu H, Tan Z-H, Dalsgaard P, Lindberg B. Robust speech recognition by non-local means de-noising processing. IEEE Signal Process Lett 2008;15:701–4.
[12] Droppo J, Deng L, Acero A. A comparison of three non-linear observation models for noisy speech features. In: Proceedings of Eurospeech 2003, Geneva, Switzerland; September 2003.
[13] Viikki O, Laurila K. Cepstral domain segmental feature vector normalization for noise robust speech recognition. Speech Commun 1998;25(1–3):133–47.

[14] Sarikaya R, Hansen JHL. Analysis of the root-cepstrum for acoustic modeling and fast decoding in speech recognition. In: Proceedings of Eurospeech 2001, Aalborg, Denmark; September 2001.
[15] Kim C, Stern RM. Power-normalized cepstral coefficients (PNCC) for robust speech recognition. In: Proceedings of IEEE ICASSP. Japan: Kyoto; 2012.
[16] Seltzer ML, Yu D, Wang Y. An investigation of deep neural networks for noise robust speech recognition. In: Proceedings of IEEE ICASSP. Canada: Vancouver; 2013.
[17] Martin R. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans Speech Audio Process 2001;9(5):504–12.
[18] Samudravijaya K. Variable frame size analysis for speech recognition. In: Proceedings of int conf natural language processing; 2004.
[19] Sam K, He Q. The use of adaptive frame for speech recognition. EURASIP J Adv Sig Process 2001;2:82–8.
[20] Bocchieri EL, Wilpon JG. Discriminative analysis for feature reduction in automatic speech recognition. In: Proceedings of IEEE ICASSP, San Francesco, USA; 1992.
[21] Hirsch HG, Pearce D. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy condition. In: Proceedings of ISCA ITRW ASR, Paris, France; September 2000.
[22] Young S, Evermann G, Gales M, Hain T, Kershaw D, Liu XA, et al. Hidden Markov model toolkit v3. 4. Cambridge University; 2006.
[23] Tan Z-H, Lindberg B, editors. Automatic speech recognition on mobile devices and over communication networks. London: Springer-Verlag; 2008.
[24] O'Shaughnessy D. Speech communications: human and machine. 2nd ed. Wiley-IEEE Press; 1999.
[25] Jafari A, Srinivasan R, Crookes D, Ming J. Exploiting long-range temporal dynamics of speech for noise-robust speaker recognition. In: Proceedings of EUSIPCO, Barcelona, Spain; August 2011.

Zheng-Hua Tan is an associate professor at Aalborg University, Denmark. His research interests include speech processing, multimodal sensing, human–robot interaction, and machine learning. He has a PhD in electronic engineering from Shanghai Jiao Tong University, China. He was a Visiting Scientist at MIT, USA, an Associate Professor at Shanghai Jiao Tong University, and a postdoctoral fellow at KAIST, Korea.

Ivan Kraljevski obtained his PhD degree at the Faculty of Electrical Engineering and Information Technology, University "St. Cyril and Methodius", Skopje, Macedonia. His scientific and professional interests include: Speech and Audio Signal Processing, Speech Recognition, Speech Synthesis, Speaker Identification, Noise Robust Speech Recognition, Pattern Recognition and Artificial Neural Networks. Current position is Speech Communication Engineer at VoiceINTERConnect GmbH, Dresden, Germany.