A Variable Frame Length and Rate Algorithm based on the Spectral Kurtosis Measure for Speaker Verification

Chi-Sang Jung[1], *Kyu J. Han*[2], *Hyunson Seo*[1], *Shrikanth S. Narayanan*[2], *Hong-Goo Kang*[1]

[1]School of Electrical and Electronic Engineering,
Yonsei University, Seoul, Korea
[2]Ming Hsieh Department of Electrical Engineering, Viterbi School of Engineering,
University of Southern California, Los Angeles, CA, USA

jtoctos@dsp.yonsei.ac.kr, kyuhan@usc.edu, hyunson@dsp.yonsei.ac.kr,
shri@sipi.usc.edu, hgkang@yonsei.ac.kr

Abstract

In this paper, we propose a spectral kurtosis based approach to extract features with a variable frame length and rate for speaker verification. Since the speaker-specific information of features in each frame changes depending upon the characteristics of speech, it is important to determine the appropriate frame length and rate to extract the salient feature frames. In order to distinctively represent the characteristics of vowels and consonants both in time and frequency domains, we introduce a variable frame length and rate (VFLR) method based on spectral kurtosis, which provides a local measure of time-frequency concentration. Experimental results verify that the proposed VFLR method improves the performance of the speaker verification system on the NIST SRE-06 database by 9.725% (relative) compared to the feature extraction method with the fixed length and rate.

Index Terms: variable frame length and rate, spectral kurtosis, time-frequency concentration, speaker verification

1. Introduction

In a typical short-time speech signal processing, the frame (i.e., processing window) length is chosen in the range of 15∼30ms with the frame shift typically fixed to half the frame length. This approach has been broadly utilized in automatic speech recognition (ASR) since it ensures capturing phonetic information from speech. The assumption made by this method that such a windowed speech portion is quasi-stationary, however, does not hold all the time because the actual portion of the inherently heterogeneous speech signal that can be deemed stationary depends critically on the underlying speech sound characteristics. For instance, vowels have a long stationary duration, whereas consonants and transient parts of speech occur on a much shorter time scale. Thus, considering a variable frame length could be more effective to accommodate the varying time-frequency acoustic phonetic characteristics. Moreover, it is known that we may not need to repeatedly extract feature frames in steady-state vowel regions, while we need to frequently capture feature frames in dynamically varying regions such as consonants or transitions for speech recognition [1]. Therefore, it is important to investigate the selection of both the frame length and rate.

In the field of speaker recognition, approaches for a variable frame rate have been developed as well [2, 3, 4]. Extensive studies including [5, 6] about the effect of frame size and rate on speaker recognition accuracy have been carried out. However, most of them lack a systematic analysis of the relation between speech (acoustic) characteristics and frame length/rate, especially, how to optimally adjust frame lengths and rates from a speaker recognition perspective as phonetic information changes in speech. In the context of variable frame rate selection, we recently proposed a novel feature frame selection method for speaker verification that not only maximizes mutual information between feature frames and speaker models but also minimizes redundancy of selected feature frames [7]. In that work, we analyzed the relationship between the selection of feature frames and their corresponding phonetic information [8]. Therefore, this paper is a natural extension of our previous works toward considering a variable frame length approach to further improve the performance of speaker verification systems.

For the variable frame length and rate (VFLR) algorithm, we utilize spectral kurtosis as a measure of the variation of phonetic information in speech. Spectral kurtosis is known to be a local measure of time-frequency concentration [9]. It is suitable for obtaining the frame length to represent the characteristics of short-time speech in a spectral domain. In the proposed algorithm, the frame length is determined by the spectral kurtosis measure, and the frame rate is determined from the selected frame length. An iterative frame merging method has also been proposed. The algorithm starts with the consecutive frames having a default frame length, and 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 rate is set to half of the frame length.

We conduct speaker verification experiments using the Gaussian mixture model (GMM) mean supervector based support vector machine (SVM) framework with nuisance attribute projection (NAP) on the NIST SRE-06 database to evaluate the performance of the proposed algorithm compared to the fixed frame length and rate algorithm [10]. In addition, we investigate the individual effects of the variable frame length (VFL) and the variable frame rate (VFR) by considering their modifications independently within the VFLR algorithm. The experimental results show that the proposed algorithms improve the performance of the speaker verification system significantly. Compared to the fixed frame length and rate algorithm, the VFLR reduces error rates by 9.725% relatively when used with both VFL and VFR.

This paper consists of the following. Section 2 introduces the spectral kurtosis measure and describes the proposed VFLR algorithm based on spectral kurtosis. In Section 3, we compare the performance of the proposed VFLR algorithm with conventional approaches. Conclusions follow in Section 4.

2. A variable frame length and rate algorithm (VFLR) for feature extraction

In this section, we suggest the spectral kurtosis based approach to extract features with a variable frame length and rate for speaker verification systems.

2.1. Spectral kurtosis measure

Spectral kurtosis is known as a measure of local time-frequency concentration for short-time speech signal in the frequency domain [11]. The generation of spectral kurtosis also employs time-frequency analysis and higher-order statistics to detect the existence of transients in a given signal [9]. The fourth order spectral cumulant gives a measure of spectral kurtosis. Then it is normalized by the square of the mean square value. Spectral kurtosis K_X is defined as :

$$K_X \triangleq \frac{C_{4X}(f)}{S_{2X}^2(f)} = \frac{S_{4X}(f)}{S_{2X}^2(f)} - 2, \quad (1)$$

where $S_{2X}(f)$ and $S_{4X}(f)$ denote the second and fourth order spectral moment. $C_{4X}(f)$ which denotes the fourth order spectral cumulant is written as :

$$C_{4X}(f) = S_{4X}(f) - 2S_{2X}^2(f). \quad (2)$$

Spectral kurtosis can be computed by the following equation.

$$K(x_w) = \frac{\sum_k \left|\sum_n x_w[n] e^{-j2\pi kn/N}\right|^4}{\left(\sum_k \left|\sum_n x_w[n] e^{-j2\pi kn/N}\right|^2\right)^2}, \quad (3)$$

where $x_w[n]$ indexes the windowed short-time speech signal in a frame. k and N denote a frequency index and FFT size, respectively. The constant value in equation (1) is eliminated because the relative comparison between spectral kurtosis values of different frames is used in this algorithm.

The spectral kurtosis values vary depending on the characteristics of the speech signal in a short-time frame [11]. Maximizing the local time-frequency concentration favors the short durations that have most of the energy in the smallest region of the time-frequency plane. In particular, time-localized speech, such as stop consonants, produces the most concentrated energy distribution in a short time frame. On the other hand, vowels are spread over time but localized in a frequency domain. Thus, the spectral kurtosis value in a frame which has a periodic signal would increase if the frame length became longer. Considering these properties of spectral kurtosis, we utilize the spectral kurtosis measure to determine the frame length and rate of each frame depending on the characteristics of the speech.

2.2. The proposed VFLR algorithm based on the spectral kurtosis measure

In this subsection, we describe the VFLR algorithm based on the spectral kurtosis measure.

Figure 1 describes the proposed variable frame length and rate scheme based on the spectral kurtosis measure in equation (3). At first, we introduce an iterative frame-merging algorithm

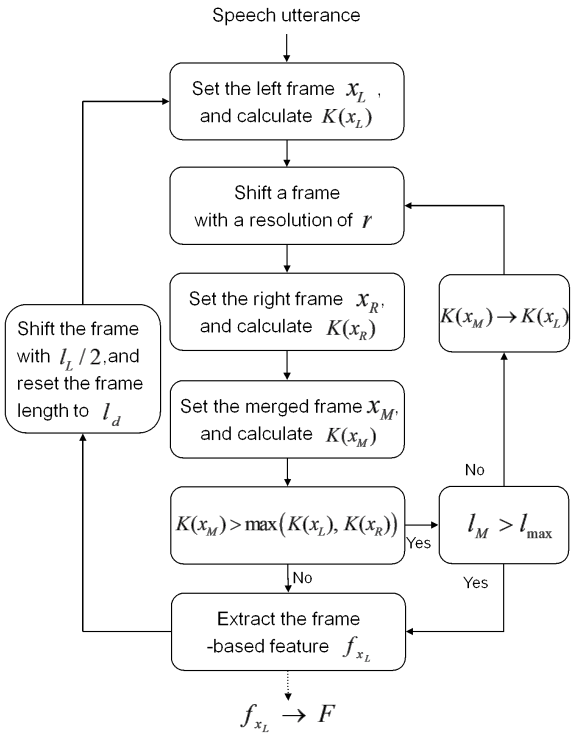

Figure 1: *A flowchart of the proposed variable frame length and rate (VFLR) algorithm*

to determine the frame length. From the input utterance, we set the current left frame x_L with a default frame length l_d and calculate the spectral kurtosis $K(x_L)$ of the current left frame. Then, we move a frame with a shift resolution of r. The current right frame x_L is also set with the default frame length, and the spectral kurtosis $K(x_R)$ of the current right frame is also computed. Finally, the spectral kurtosis of the merged frame x_M with the length l_M is calculated. When we calculate the spectral kurtosis, frame length (i.e., a window length) is variable, but the FFT length which denotes N in equation (3) is fixed to 512. If the spectral kurtosis of merged frame $K(x_M)$ exceeds the maximum value of the spectral kurtosis of the left and right frames, denoted by $K(x_L)$ and $K(x_R)$ respectively, then we merge two frames to one expanded frame which becomes a new left frame in the next iteration. At the iterative step, the frame length expands until the spectral kurtosis of the merged frame is less than the maximum value of the spectral kurtosis of two consecutive frames. In this step, the maximum frame length l_{max} is limited to prevent the frame length from expanding too long. After deciding the current frame length, the frame-based feature is extracted in the determined frame, and the frame rate is set to half of the selected frame length. Then, the process to determine the next frame length iteratively repeats until the given utterance ends.

In addition, we modify the VFLR algorithm to evaluate the performance of the variable frame length (VFL) and the variable frame rate (VFR) algorithms. For the VFR, at first, the frame length is fixed to the conventional frame length 20ms. And then, we only use the frame rate result obtained by the VFLR algorithm. In a similar way, we also use the VFL algorithm with the fixed frame rate 10ms and the result of the variable frame length determined by the VFLR algorithm.

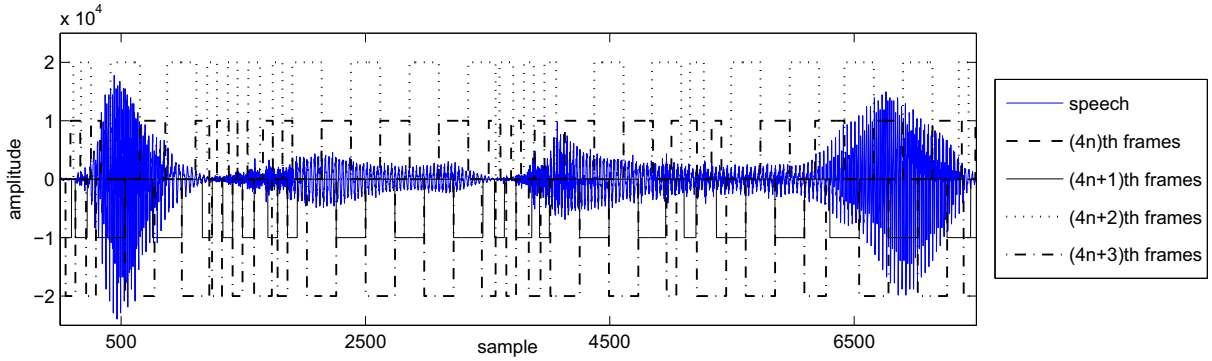

Figure 2: *An example of frame length and rate determined by the VFLR algorithm*

In the proposed algorithm, the maximum value of time-frequency concentration favors frames that have most of the energy in the time-frequency plane. In particular, short-length frames are selected around time-localized transitions and consonants, since these produce the most concentrated energy distribution in the short-time duration. On the other hand, vowels are spread over time but localized in a frequency domain. Thus, the frame length of vowels tends to be much longer.

Figure 2 represents the selected frame length and rate results obtained by the VFLR algorithm. The speech signal with an 8kHz sampling rate is shown in this figure. While the different types of lines and vertical scales of each box are used for distinguishing the concatenated frames visually, the horizontal values of each box denote the selected frame lengths. Figure 2 illustrates that frame lengths in the vowel-like periodic speech are mostly longer than frame lengths in the aperiodic or transient speech. From the results of selected frame length, the frame rates are also long in the vowel-like speech since they are set to half of the selected frame length. Thus, the redundant frames from periodic speech parts such as vowels are relatively reduced. On the other hand, consonants and transitions favor the short frame lengths and rates because they contain most of the energy in a short frame. From these results, we can confirm that the spectral kurtosis based VFLR algorithm determines the appropriate frame length and rate to describe the various characteristics of frame-based speech in a time-frequency plane. Based on these properties of variable frame length and rate algorithm using spectral kurtosis, we next apply the proposed VFLR algorithm to speaker verification systems.

3. Performance evaluation

3.1. Experimental setup

We conduct speaker verification experiments to evaluate the proposed VFLR algorithm on the NIST SRE-06 database [12]. We focus on the single-side 1 conversation train and test. This setup results in 1,799 true trials and 18,985 false trials.

We build the GMM supervector SVM based speaker verification system. The universal background model (UBM) includes 256 Gaussians, and the single-side 1 conversation on the NIST SRE-04 database is used for training the UBM [13]. We produce one GMM supervector per utterance using maximum a posteriori (MAP) adaptation. The supervectors, which are concatenated vectors of GMM means, are used as the input of SVM. Another part of the single-side 1 conversation on the NIST SRE-04 database, which is not used for training the

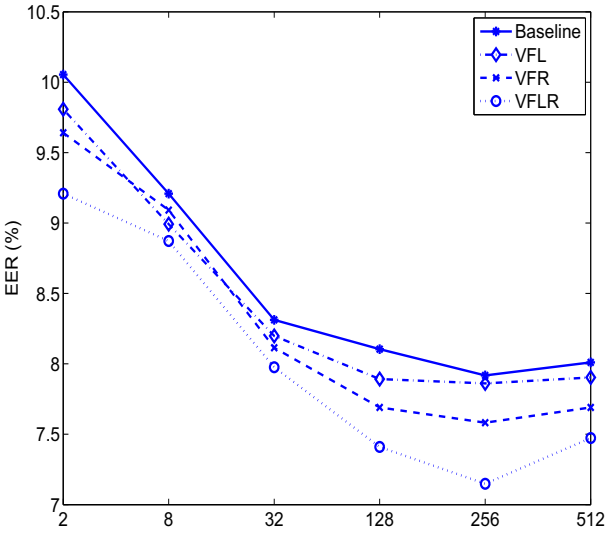

Figure 3: *A variation of EERs (%) according to the corank for each algorithm*

UBM, is used for training imposter models of SVM. Then, nuisance attribute projection (NAP) is applied to remove the session variability, and the single-side 8 conversation on the NIST SRE-04 database is used for NAP [10].

Voice activity detection (VAD) is applied to remove the silent frames. And then, the fifteenth order mel-scaled frequency cesptral coefficients (MFCCs) and their delta-MFCCs are used as feature vectors. Cepstral mean subtraction (CMS) and variance normalization are applied to eliminate the linear channel effects. For the baseline system, the frame length and rate are fixed to 20ms and 10ms, respectively. For the VFLR algorithm, we set the resolution of the variation of frame length and rate to 2ms and 1ms, respectively. The initial and maximum frame lengths are set to 10ms and 30ms, respectively, because the stationary speech duration is generally known as around 15∼30ms. A frame shift is determined by the VFLR algorithm as half of the calculated frame length.

In order to evaluate the effect of the variable frame length (VFL) and variable frame rate (VFR) algorithms, respectively, we conduct four types of experiments-baseline, VFL, VFR, and

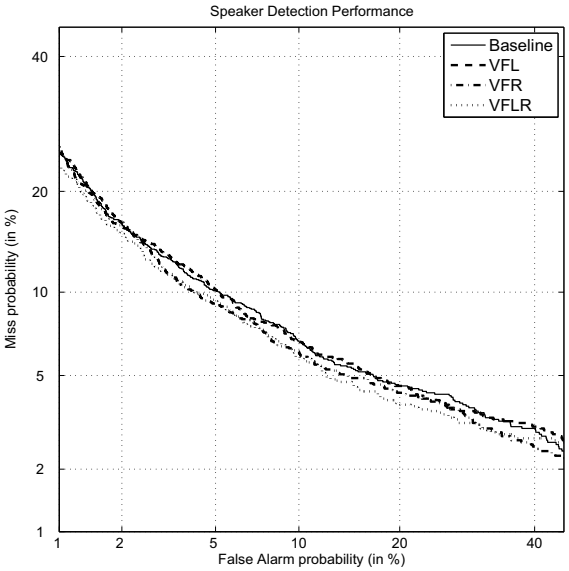

Figure 4: *DET curves for each algorithm (corank=256)*

Table 1: *Comparison of EERs (%) for each algorithm (corank=256)*

	EER (%)
Baseline	7.918
VFL	7.861
VFR	7.582
VFLR	7.148

VFLR-on the speaker verification system.

3.2. Results and analysis

Figure 3 represents the comparison of the equal error rates (EERs) for each algorithm by varying the corank of NAP. The corank means how many dimensions are projected out from the total dimensions of the supervector. In Figure 3, the VFL and VFR algorithms show a better performance than the fixed frame length and rate algorithm (baseline) across all coranks. Moreover, the VFR algorithm is better than the VFL and the fixed algorithm. As we verified in [8], the VFR reduces the error rate of the speaker verification system compared to the fixed frame rate algorithm. In the proposed algorithm, the frame rate is determined from the frame length obtained by the VFLR algorithm. Thus, the redundant frames for vowels are reduced since the frame length is mostly long in vowels. On the other hand, the VFL shows a little improved or similar performance compared to the performance of the fixed length algorithm. However, when the VFL is combined with the VFR, which is the VFLR, the performance becomes the best among all algorithms over all coranks. In other words, the performance of speaker verification systems is improved when both the frame length and rate are controlled by the spectral characteristics of speech.

Figure 4 plots the detection error trade-off (DET) curves of each algorithm when the corank is set to 256, where all the algorithms compared show the best performance (Figure 3). In addition, Table 1 indicates EERs of each algorithm. Figure 4 and Table 1 show that the performance of the VFLR algorithm is the best. It reduces EER by 9.725% compared to the conventional fixed frame length and rate algorithm. From these results, we confirm that the spectral kurtosis based variable frame length and rate (VFLR) algorithm improves the performance of speaker verification system.

4. Conclusions

In this paper, we suggest the variable frame length and rate (VFLR) algorithm based on the spectral kurtosis measure for speaker verification systems. The characteristics of speech signals dynamically vary depending on the phonetic information. Thus, we control the frame length and rate based on the spectral kurtosis to extract the speaker-specific information depending on the nature of the speech. In the proposed algorithm, the VFL is used for extracting the signal-dependant speaker information in each frame, and the VFR removes the frames containing the redundant information. Experimental results show the superiority of the proposed VFLR algorithm that combines the VFL and VFR.

References

[1] P. Scanlon, D. Ellis, and R. Reilly, "Using broad phonetic group experts for improved speech recognition," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 15, pp. 803-812, 2007.

[2] H. You, Q. Zhu and A. Alwan, "Entropy-based variable frame rate analysis of speech signals and its application to ASR," *in Proc. Internat. Conf. Acoust. Speech Signal Processing*, pp. 549-552, 2004.

[3] J. Epps and E. Choi, "An Energy Search Approach to Variable Frame Rate Front-End Processing for Robust ASR," *in Proc. Interspeech*, pp. 2613-2616, 2005.

[4] T. Wu, D. Compernolle, J. Duchateau, and H. Hamme, "Maximum likelihood based temporal frame selection," *in Proc. Internat. Conf. Acoust. Speech Signal Processing*, pp. 349-352, 2006.

[5] D. Impedovo and M. Refice, "Frame length selection in speaker verification task," *WSEAS Trans. Systems*, vol. 7, no. 10, pp. 1028-1037, 2008.

[6] H. S. Jayanna and S. R. Mahadeva Prasanna, "Multiple frame size and rate analysis for speaker recognition under limited data condition," *IET Signal Processing*, vol. 3, no. 3, pp. 189-204, 2009.

[7] C. Jung, M. Kim, and H. Kang, "Normalized minimum-redundancy and maximum-relevancy based feature selection for speaker verification systems," *in Proc. Internat. Conf. Acoust. Speech Signal Processing*, pp. 4549-4552, 2009.

[8] C. Jung, M. Kim, and H. Kang, "Selecting feature frames for automatic speaker recognition using mutual information," *IEEE Trans. on Audio, Speech and Language Signal Processing, in press*.

[9] D. Jones and R. Baraniuk, "A simple scheme for adapting time-frequency representations," *IEEE Transactions on Signal Processing*, vol. 42, no. 12, pp. 3530-3535, 1994.

[10] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM based speaker verification using a GMM supervector kernel and NAP variability compensation," *in Proc. Internat. Conf. Acoust. Speech Signal Processing*, pp. 97-100, 2006.

[11] D. Rudoy, P. Basu, T. Quatieri, R. Dunn, and P. Wolfe, "Adaptive short-time analysis-syntheis for speech enhancement," *in Proc. Internat. Conf. Acoust. Speech Signal Processing*, pp. 4905-4908, 2008.

[12] The NIST Speaker Recognition Evaluation. Available: http://www.itl.nist.gov/iad/mig/tests/sre/

[13] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using Adapted Gaussian Mixture Models," *Digital Signak Processing*, vol. 10, no. 1-3, pp. 19-41, 2000.