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.

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/

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