Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 78 (2016) 62–71

www.elsevier.com/locate/specom

A mode-shape classification technique for robust speech rate estimation and syllable nuclei detection

Chiranjeevi Yarra[a], Om D. Deshmukh[b], Prasanta Kumar Ghosh[a,*]

[a] Department of Electrical Engineering, Indian Institute of Science (IISc), Bangalore 560012, India
[b] Xerox Research Center India, Bangalore 560103, India

Received 30 April 2015; received in revised form 30 December 2015; accepted 23 January 2016
Available online 3 February 2016

Abstract

Acoustic feature based speech (syllable) rate estimation and syllable nuclei detection are important problems in automatic speech recognition (ASR), computer assisted language learning (CALL) and fluency analysis. A typical solution for both the problems consists of two stages. The first stage involves computing a short-time feature contour such that most of the peaks of the contour correspond to the syllabic nuclei. In the second stage, the peaks corresponding to the syllable nuclei are detected. In this work, instead of the peak detection, we perform a mode-shape classification, which is formulated as a supervised binary classification problem – mode-shapes representing the syllabic nuclei as one class and remaining as the other. We use the temporal correlation and selected sub-band correlation (TCSSBC) feature contour and the mode-shapes in the TCSSBC feature contour are converted into a set of feature vectors using an interpolation technique. A support vector machine classifier is used for the classification. Experiments are performed separately using Switchboard, TIMIT and CTIMIT corpora in a five-fold cross validation setup. The average correlation coefficients for the syllable rate estimation turn out to be 0.6761, 0.6928 and 0.3604 for three corpora respectively, which outperform those obtained by the best of the existing peak detection techniques. Similarly, the average *F*-scores (syllable level) for the syllable nuclei detection are 0.8917, 0.8200 and 0.7637 for three corpora respectively.

Keywords: Speech rate estimation; Syllable nuclei detection; TCSSBC; Mode-shape classification; Mode-shape feature vectors; Peak detection.

1. Introduction

Speech rate estimation and syllable nuclei detection are important problems in the areas of automatic speech recognition (ASR), computer assisted language learning (CALL) and fluency analysis. The ASR accuracy has been shown to improve by using the speech rate and syllable nuclei information in the recognition model (Bartels and Bilmes, 2007; Morgan et al., 1997) . In CALL, the features used for fluency analysis (Cucchiarini et al., 2000) or non-nativeness analysis (Hönig et al., 2012) are based on one or more combinations of speech rate and syllable nuclei locations. The problems of speech rate and syllable nuclei detection are closely related. The speech rate is typically estimated by counting the number of speech units per second. Most of the existing works in the literature use syllable as the speech unit (Heinrich and Schiel, 2011; Morgan et al., 1997; Wang and Narayanan, 2007). The speech rate estimation typically involves identification of the syllable nuclei locations followed by syllable rate computation (Reddy et al., 2013). Generally the approaches for the speech rate estimation and the syllable nuclei detection are based on either acoustic features (Heinrich and Schiel, 2011; Morgan et al., 1997; Reddy et al., 2013; Wang and Narayanan, 2007) or hidden Markov model (HMM) based recognition systems (Cincarek et al., 2009; Cucchiarini et al., 2000; Hönig et al., 2012; Yuan and Liberman, 2010).

The HMM based methods involve the identification of the phoneme/syllable boundaries using an ASR system. The estimated boundaries are then used to compute the syllable rate. HMM based approaches are used in the applications related to CALL where a good quality speech rate estimation is

* Corresponding author. Tel.: +91 80 2293 2694; fax: +91 80 2360 0444.
E-mail addresses: chiranjeevi.yarra@ee.iisc.ernet.in, chiranjeevi.yarra@gmail.com (C. Yarra), prasantg@ee.iisc.ernet.in (P.K. Ghosh).

http://dx.doi.org/10.1016/j.specom.2016.01.004
0167-6393/© 2016 Elsevier B.V. All rights reserved.

essential (Cucchiarini et al., 2000; Deshmukh et al., 2008; Hönig et al., 2012; Witt, 1999). However for accurate speech rate estimation, methods based on HMM are time consuming particularly when the reference transcription is not available and, hence, often not useful in real time applications (Wang and Narayanan, 2007). In contrast to the HMM based methods, the acoustic feature based methods are computationally less expensive (Morgan and Fosler-Lussier, 1998). The acoustic feature based methods are typically developed using acoustic properties of the vowels, which in general correspond to the syllable nuclei. Therefore, the vowel rate corresponds directly to the syllable rate (Pfau and Ruske, 1998; Yuan and Liberman, 2010).

A typical approach for estimating syllable nuclei locations, which is also useful for estimating syllable rate, involves two steps – (1) computing a short-time feature contour such that most of the peaks corresponding to the syllable nuclei locations, (2) detecting the peaks belonging to the syllable nuclei. Pfau and Ruske (1998) estimated the vowel locations based on prominent peaks in smoothed loudness contour. They proposed a peak identification strategy based on the steepness information around the local maxima. Zhang and Glass (2009) proposed a contour based on Hilbert envelope and used a rhythm guided peak counting to estimate the syllable nuclei. De Jong and Wempe (2009) used intensity based envelope with simple peak counting based on voicing decisions to estimate speaking rate. Landsiedel et al. (2011) proposed a contour based on long short term memory neural networks and identified peaks based on the region based selection above a threshold limit. Wang and Narayanan (2005, 2007) introduced a method by proposing a feature contour “temporal correlation and selected sub-band correlation (TCSSBC)”, which involves computing a spectral and a temporal correlation; they also proposed a peak detection strategy which involves smoothing and a thresholding mechanism. A comprehensive comparative study of eight different methods for speech rate estimation has been summarized by Dekens et al. (2007), who found that the TCSSBC method performs the best for speaking rate estimation.

The methods addressed in the literature for both the problems focus on the feature computation as well as on the peak detection strategies. Wang et al improved the speech rate estimation accuracy by optimizing parameters in the TCSSBC feature contour computation and using a robust peak detection strategy (Wang and Narayanan, 2007). A modified version of the peak detection strategy is used by Reddy et al. (2013) along with perceptually motivated features. A neural network based syllabic peak detection was proposed by Howitt (2000). Most of the existing peak detection strategies are typically heuristic and rule based. A generic formulation for syllabic peak detection is necessary to overcome the limitations of the rule based approaches. We observe that the rule based peak detection strategies often fail to detect target peaks mainly because the target peaks do not always satisfy the heuristically designed rules. In that direction Jiao et al. (2015) proposed a convex optimization based speech rate estimation to avoid dependency on heuristic peak detection strategy. Faltlhauser et al. (2000) used the Gaussian mixture model (GMM) for classification of speaking rate into three categories – slow, medium and fast. Following this, they used the class probabilities to estimate speaking rate with the help of Neural Networks.

We, in this work, use TCSSBC as a short-time feature contour and perform mode-shape classification. In the vicinity of syllable nuclei locations TCSSBC contour typically has local maxima (Dekens et al., 2007; Howitt, 2000; Wang and Narayanan, 2007). Therefore, almost all syllables correspond to the peaks in the TCSSBC feature contour. However, some of the peaks corresponding to the syllable nuclei (referred to as syllabic peaks) are often less prominent compared to the peaks that do not correspond to any syllable (non-syllabic peaks). We hypothesize that the contour shape around each mode of the TCSSBC contour could be used for robust detection of target TCSSBC peaks.

We use a support vector machine (SVM) based binary classification method for distinguishing the syllabic mode-shapes from the non-syllabic ones. Note that although one mode-shape carries information about only one peak, we use the term ‘mode-shape’ instead of ‘peak’ because we exploit the shape of the TCSSBC feature contour around the peak for the binary classification. We propose different feature vectors spanning across multiple modes to represent each mode-shape of the TCSSBC feature contour. We also propose an automatic way of labeling each mode-shape – syllabic and non-syllabic – for training the SVM classifier. The effectiveness of the proposed mode-shape classification (MSC) approach is demonstrated using three large corpora, namely, Switchboard, TIMIT and CTIMIT. Experiments for both speech rate estimation and syllabic nuclei detection are performed on each corpus. The proposed MSC based syllabic peak detection approach achieves better performance in comparison to the best of the existing methods for both speech rate estimation and syllable nuclei detection.

The rest of the paper is organized as follows: Section 2 describes the corpora details, Section 3 discusses the details of the proposed MSC approach including TCSSBC feature contour computation, smoothing, mode-shape feature vector computation, labeling and classification procedures. Section 4 includes the experimental setup, results on various corpora and discussions. The conclusions are summarized in Section 5.

2. Database

We use ICSI Switchboard (Godfrey et al., 1992), TIMIT (Zue et al., 1990) and CTIMIT (Brown and George, 1995) corpora for all experiments in this work. Switchboard is a spontaneous speech corpus consisting of sentences spoken by 370 speakers with a wide range of speech rate, ranging from 1.26 to 9.2 syllables per second. The audio in the Switchboard corpus was collected through the telephone channel. A subset of 7300 audio segments, each of duration greater than 200 ms, is used for our experiments. TIMIT is a read speech database, which has phonetically balanced 6300 sentences spoken by 630 speakers with a speech rate ranging