Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Computer Speech & Language xxx (2017) xxx-xxx

www.elsevier.com/locate/csl

Sequential use of spectral models to reduce deletion and insertion errors in vowel detection☆

Q1 Hamidreza Baradaran Kashani, Abolghasem Sayadiyan*

Department of Electrical Engineering, Amirkabir University of Technology, P.O. Box 15875-4413, 424 Hafez Ave, Tehran, Iran

Received 19 September 2016; received in revised form 28 November 2017; accepted 20 December 2017
Available online xxx

Abstract

From both perspectives of speech production and speech perception, vowels as syllable nuclei can be considered as the most significant speech events. Detection of vowel events from a speech signal is usually performed by a two-step procedure. First, a temporal objective contour (TOC), as a time-varying measure of vowel similarity, is generated from the speech signal. Second, vowel landmarks, as the places of vowel events, are extracted by locating prominent peaks of the TOC.

In this paper, by employing some spectral models in a sequential manner, we propose a new framework that *directly* addresses three possible errors in the vowel detection problem, namely vowel deletion, consonant insertion, and vowel insertion. The proposed framework consists of three main steps as follows. At the first step, two solutions are proposed to essentially reduce the initial vowel deletion error. The first solution is to use the peaks detected by a conventional energy-based TOC, but without utilizing TOC smoothing and peak thresholding processes. The peaks detected by a spectral-based TOC generated on the basis of GMM models are also put forward as the second solution for achieving a smaller vowel deletion error. At the second step, a two-class support vector machine (SVM) classifier is adopted to identify the consonant peaks from the vowel ones. Removing the peaks classified as consonants reduces the consonant insertion error. Finally, a two-class SVM classifier is proposed to classify the consecutive peaks detected within the same vowel from the others. The merging of the peaks classified as "same vowel" considerably reduces the vowel insertion error.

Experiments are separately conducted on three standard speech corpora, namely FARSDAT, TIMIT and TFARSDAT. The effectiveness of the techniques proposed to reduce three types of detection errors is verified. The criteria of total error (as the summation of three detection errors) and F-measure, respectively result in about 9.7% and 95.1% for FARSDAT, 17.5% and 91.3% for TIMIT, and 19.6% and 90.2% for the TFARSDAT corpus. The evaluation results show that the proposed framework outperforms the existing well-known methods in terms of both total error and F-measure on both read and spontaneous speech corpora.

Keywords: Vowel landmark detection; Temporal objective contour (TOC); Vowel deletion error; Consonant insertion error; Vowel insertion error

☆ This paper has been recommended for acceptance by Roger K. Moore.

Q2 * Corresponding author.
E-mail address: hr_baradaran@aut.ac.ir (H.B. Kashani), sayadiyan@outlook.com (A. Sayadiyan).

http://dx.doi.org/10.1016/j.csl.2017.12.008
0885-2308/ 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In speech signal, syllable units enjoy significant characteristics. They are considered as linguistically-motivated structural units for phonological representation that capture co-articulation between sounds and contain important information about prosodic and rhythmic aspects of speech. Accordingly, syllable (or generally syllable-based) units have long been employed as the basic units in different speech processing applications (e.g., Choueiter et al., 2008; Bartels and Bilmes, 2010; Tachbelie et al., 2014; Cernak et al., 2015; Yadav and Rao, 2016).

Syllable-based units are typically detected or extracted by locating the vowels as the nuclei of syllables. Vowels constitute the most significant speech events in terms of both speech production and speech perception. The open configuration and no constriction along the vocal tract during vowel production cause the vowels to be sounds with high resonance and energy as well as having clear pitch and formant structures. In contrast, consonants are produced with a constriction or a closure at some points along the vocal tract.

Estimating the speech rate based on the detection of syllable nuclei and, consequently, developing automatic speech recognition (ASR) systems resistant to speech rate variabilities are among the major applications of vowel detection (Chu and Povey, 2010; Zeng et al., 2015; Jiao et al., 2015). In different researches (e.g., Thomas et al., 2006; Narendra and Rao, 2012; Reddy and Rao, 2013), concatenative speech synthesis systems based on syllable or syllable-like units have been developed. Other basic applications related to syllable nucleus detection can be introduced as speaker recognition (Prasanna and Pradhan, 2011; Li et al., 2013), language recognition (Ng et al., 2013), and voice conversion (Rao; 2010; Veaux and Rodet, 2011) tasks that widely employ prosodic features extracted at the syllable-based levels.

1.1. Previous work

Various methods have been presented for vowel detection in the literatures. The majority of methods adopt a two-step procedure. First, a *temporal objective contour* (TOC), as a time-varying measure of vowel similarity, is generated from the speech signal (TOC generation). Second, the places of vowel events, called vowel landmarks (VLs), are extracted by determining prominent peaks of the TOC (TOC-VL extraction). All methods can be classified into two general groups: *heuristic-based* or *empirical* versus *model-based* methods.

In empirical methods (e.g., Nagarajan and Murthy, 2004; Xie and Niyogi, 2006; Wang and Narayanan, 2007; Obin et al., 2013), no training and statistical modeling processes are used for both detection stages, i.e. TOC generation and TOC-VL extraction. In fact, empirical methods directly employ the definition of *syllable* for detecting vowel landmarks. According to International Phonetic Association (1999), a syllable is normally considered to be a string of phonetic units that has an energy peak in the vowel existing within the nucleus and energy minimums in the potential consonants existing at the two sides of the nucleus. So, empirical methods adopt speech energy (either all-pass or band-pass) as the main acoustic feature for TOC generation. For example, Xie and Niyogi (2006) considered the TOC as the all-pass energy of speech signal. In Wang and Narayanan (2007), speech energy was firstly extracted from 19 frequency subbands. Then, the TOC was generated by calculating the temporal correlation per subband followed by the spectral correlation on several high-energy subbands. Moreover, Dekens et al., (2014) generated the TOC by multiplying the square of a low frequency energy contour by the summation of three high frequency energy contours.

In contrast, model-based methods (e.g., Yuan and Liberman, 2010; Khonglah et al., (2014); Yarra et al., 2016), train the acoustic models with different spectral features for either TOC generation or TOC-VL extraction. For example, Yuan and Liberman (2010) proposed the broad phonetic HMM-based recognizers trained with PLP features for detecting VLs. Landsiedel et al., (2011) suggested bidirectional long-short-term memory neural networks trained with PLP and modulation spectrum features for this goal. Deep belief networks trained with MFCC features were used by Khonglah et al., (2014) for detecting vowel and semivowel frames.

In almost all methods (either empirical or model-based), once the TOC is generated, the locations of vowels (or VLs) usually appear as prominent peaks on the contour. Next, VLs are extracted on the basis of some rules and also by comparing peak prominences with a number of predefined thresholds. Note that the terms “empirical TOC” and “model-based TOC” are respectively used for the TOCs generated by empirical and model-based approaches throughout the rest of this paper.