Speech Communication 92 (2017) 77–89

Contents lists available at ScienceDirect

Speech Communication

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

Consonant-vowel unit recognition using dominant aperiodic and transition region detection

Biswajit D. Sarma[a,*], S.R. Mahadeva Prasanna[a], Priyankoo Sarmah[b]

[a] Department of Electronics and Electrical Engineering, Indian Institute of Technology (IIT) Guwahati, Guwahati 781039, India
[b] Department of Humanities and Social sciences, Indian Institute of Technology (IIT) Guwahati, Guwahati 781039, India

ARTICLE INFO

Article history:
Received 4 November 2016
Revised 27 May 2017
Accepted 6 June 2017
Available online 8 June 2017

Keywords:
Consonant-vowel unit
Transition region
Aperiodic region
Vowel onset point
Vocal tract constriction
Sub fundamental frequency filtering
Consonant recognition

ABSTRACT

This work reports a method of Consonant-Vowel (CV) unit recognition by detecting the Dominant Aperiodic component Regions (DARs) and by predicting the Duration of Transition Regions (DTRs) in speech. DAR detection is performed using complementary information from source and vocal tract. While source information is extracted using sub-fundamental frequency filtering of speech, vocal tract information is extracted using a) Dominant Resonant Frequency (DRF) and b) High to Low Frequency component Ratio (HLFR), computed from Hilbert envelope of Numerator Group Delay (HNGD) spectrum of zero-time windowed signal. The DTR is predicted by using vocal tract constriction information. Subsequently, detected DARs and predicted DTRs are compared with manually marked regions and finally used for CV unit recognition of Indian languages. Conventionally, CV unit recognition is performed by anchoring the Vowel Onset Point (VOP) and assuming fixed durations for transition and consonant regions on either side of the VOP. However, in speech, the duration of transition and consonantal regions vary depending on the type of consonants and vowels. In the proposed method, the use of dynamic values for consonant duration and transition regions have resulted in better consonant recognition improving CV unit recognition.

1. Introduction

In conventional speech recognition, speech signal is processed frame-wise and represented as a sequence of feature vectors. In such cases, Mel Frequency Cepstral Coefficients (MFCCs) are used as features, and all speech regions are assumed to be of equal importance. On the other hand, event based speech recognition systems identify certain events or landmarks in the speech signal, where, acoustic-phonetic cues are more salient (Liu, 1996). In such systems, the events are detected first, and subsequently, relevant features are extracted by anchoring to those events. Several previous works have laid emphasis on the event based systems due to its use of explicit acoustic-phonetic knowledge (Gangashetty, 2004; Juneja and Espy-Wilson, 2008; Liu, 1996; Salomon et al., 2004; Sekhar, 1996; Stevens, 2002). In case of Indian languages, the Consonant-Vowel (CV) unit recognition systems proposed in Gangashetty (2004); Sekhar (1996) and Vuppala et al. (2012a) have adopted the event based approach, where, Vowel Onset Points (VOPs) are considered as events and features for consonants and vowels are extracted from the regions around the VOPs. In these works, for recognition of consonants, features were extracted from 40 ms segments of consonant regions before the VOPs and from 40 ms segment of transition regions after the VOPs. However, it has been reported that the voice onset time for certain stop consonants, such as, velars is predominantly more than 40 ms and in case of aspirated stop consonants voice onset time is between 55–154 ms (Cho and Ladefoged, 1999; Prakash, 2012; Prakash et al., 2013). Apart from that, the duration of fricatives and affricates is also longer. Hence, limiting consonant recognition to 40 ms segments may not be optimal, as the limited duration will not be able to capture the transient burst and frication that characterize fricatives and affricates. Perception studies have demonstrated that transient bursts and frication are crucial cues in identifying and characterizing consonants Pisoni and Remez, (Stevens, 1999). Similarly, fixing a 40 ms duration for the analysis of transition region may not be appropriate as transition regions for consonants vary widely depending on the phonetic environment. Hence, in order to capture the consonant characteristics effectively, burst and frication regions, and transition regions have to be estimated dynamically.

CoR, Consonant onset Refinement; DAR, Dominant Aperiodic component Region; DTR, Duration of Transition Region; FA, Forced Alignment; HLFR, High to Low Frequency component Ratio; SFF, Sub Fundamental Frequency; VTC, Vocal Tract Constriction; VTR, Variable duration Transition Region.

[*] Corresponding author.

E-mail addresses: s.biswajit@iitg.ernet.in (B.D. Sarma), prasanna@iitg.ernet.in (S.R.M. Prasanna), priyankoo@iitg.ernet.in (P. Sarmah).

http://dx.doi.org/10.1016/j.specom.2017.06.003
0167-6393/© 2017 Elsevier B.V. All rights reserved.

In the current study, we attempt to estimate the burst and frication regions by detecting the Dominant Aperiodic component Regions (DARs). Transient bursts in stop consonants, and random noise in fricatives are two kinds of aperiodic sources present in unvoiced speech. Due to complete absence of periodic components, all unvoiced sounds can be considered as DARs. Although periodic source is dominant in most of the voiced speech regions, there are some voiced sounds, where, both periodic and aperiodic sources exist. The aperiodic sources in voiced speech are either additive random noise or modulation aperiodicity (d'Alessandro et al., 1998). Additive random noise represents aspiration and frication noises, superimposed on the periodic glottal vibration. Voiced obstruents with strong additive random noise is considered DARs. Modulation aperiodicity is produced due to random variations in the period (jitter) and the amplitude of the signal (shimmer). As modulation aperiodicity is low in normal speech (d'Alessandro et al., 1998), they are not considered DARs for the detection task. However, modulation aperiodicity present in formant transitions are implicitly detected in the second task where the transition regions are detected by predicting their durations.

Most of the works on periodic and aperiodic component rely either on decomposing speech into periodic and aperiodic components or on estimating proportion of periodic and aperiodic energy (d'Alessandro et al., 1998; Deshmukh et al., 2005; Yegnanarayana et al., 1998). In Yegnanarayana et al. (1998), an iterative algorithm was proposed for decomposition of speech signal into periodic and aperiodic components. A study regarding effectiveness of periodic and aperiodic component decomposition method for analysis of voice sources was made in d'Alessandro et al. (1998). Another study used temporal information for obtaining proportion of periodic and aperiodic energy in speech in addition to estimating pitch period in the periodic component (Deshmukh et al., 2005). Unlike these methods, in this work, we propose a novel approach to detect the DARs using source and vocal tract information. Motivation behind the use of source information is that basic source characteristics, such as, nature of discontinuities due to impulse-like excitations are different in periodic and aperiodic sources. Proper exploitation of these characteristics may help to separate the aperiodic sources from the periodic ones. We propose a method using Sub Fundamental Frequency (SFF) filtering of speech to capture the discontinuities due to aperiodic sources. The SFF filtering is a modification of the Zero Frequency Filtering (ZFF) method proposed in Murthy and Yegnanarayana (2008) to detect the discontinuities due to significant impulse-like excitations in voiced speech region. Vocal tract information, such as, Dominant Resonant Frequency (DRF) and High to Low Frequency component Ratio (HLFR) are different in DARs and other regions, and hence, these parameters can also be used for detection of DARs. Unlike the source information, vocal tract information is extracted by performing block processing. In order to get better time and frequency resolution, DRF and HLFR are computed from Hilbert envelope of Numerator Group Delay (HNGD) spectrum of zero time windowed speech. DARs detected using combined source and vocal tract information are evaluated by comparing them to the manually marked DARs existing in the database.

Several methods have been reported for detecting the acoustic landmarks associated with stop consonants (Jayan and Pandey, 2009; Juneja and Espy-Wilson, 2008; Lin and Wang, 2011; Liu, 1996), but none of them automatically marked the transition regions. A method for detecting the VC and CV transitions in VCV utterances was presented using a measure of the rate of change of vocal tract area function (Jagbandhu et al., 2012). But the method rely on the estimation of the vocal tract shape which itself is a difficult task. Alternatively, in this work, the Duration of Transition Region (DTR) is predicted using vocal tract constriction information in the vowel region, as vowel height differences have direct effects on the duration of transition. For example, consonant-vowel transitions are expected to be more in case of open (low) vowels as a direct consequence of the need for longer articulation duration in their productions (Diehl et al., 1987; Fant, 1960; Johnsen, 2013; Stevens, 1999). In vowels, the Vocal Tract Constriction (VTC) represents the vowel height and hence, it should be possible to predict the DTR using the knowledge of amount of constriction. A vocal tract constriction evidence was proposed in Sarma and Prasanna (2014) to approximately measure the amount of constriction in the vocal tract. In the current work, VTC evidence is used to predict transition regions by mapping the values to time durations. Later, predicted transition regions are compared with the ones that are manually marked by four trained acoustic phoneticians.

The method for detection of DARs is detailed in Section 2 of this paper. Section 3 demonstrates the procedure for predicting the duration of transition regions. The performance of the proposed methods are evaluated in Section 4. Section 5 describes the use of detected DARs and duration of transition regions for CV unit recognition. In Section 6, the proposed consonant recognition method is evaluated and compared with the baseline system. Finally, Section 7 summarizes and concludes the work.

2. Detection of dominant aperiodic component regions (DARs) in speech

In order to detect the DARs, source and vocal tract information are required to be explored. While source information is obtained using SFF filtering of speech, vocal tract information is obtained by computing DRF and HLFR. DARs detected using individual evidences, derived from two complementary information, are combined to get the final output. Following subsections describe the DAR detection methodology.

2.1. Sub-Fundamental frequency (SFF) filtering

Both periodic and aperiodic sources contain impulse excitations. In periodic sources, the impulse excitations are due to glottal closure and opening instants and they occur at regular time intervals. In aperiodic sources, the impulse excitations are due to transient bursts and frication noises, and they occur at every instant of time with arbitrary amplitude. These time instants of occurrence of the excitation impulses, are reflected as discontinuities in the signal. Discontinuities are also observed in the transition points between obstruents and sonorants, and sometimes also at the onsets and offsets of sonorants and obstruents, primarily due to sudden change in the signal energy. Discontinuities due to epochs in voiced regions were detected using ZFF method proposed by Murthy and Yegnanarayana (2008). Here, we attempt to detect some of the discontinuities arising due to aperiodic sources, using SFF filtering.

The SFF filtering method is motivated from the ZFF method where the signal is passed through a cascade of two 0 Hz resonators. The output of the resonators grows as a polynomial function of time. The effect of discontinuities arising due to impulse sequences are overridden by the large values of the filtered output. To extract the characteristics of the discontinuities arising due to impulse excitation, the deviation of the filtered output was computed from the local mean as shown in 1.

$$z(n) = y(n) - \frac{1}{2M+1}\sum_{m=-M}^{M} y(n+m) \qquad (1)$$

In 1, $y(n)$ is the output of the 0 Hz resonator and $2M+1$ is the length of the window, over which the local mean was computed. The trend removed signal $z(n)$ is the ZFF signal. Later, an FIR implementation of these sequence of operations was also proposed (Srinivas and Prahallad, 2012). The output of the ZFF filter