Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 80 (2016) 71–83

www.elsevier.com/locate/specom

A pre-processing method for improvement of vowel onset point detection under noisy conditions☆

P. Saha*, R.H. Laskar, A. Laskar

Speech and Image Processing Lab, National Institute of Technology, Silchar, Assam 788010, India

Received 14 September 2015; received in revised form 20 March 2016; accepted 12 April 2016
Available online 29 April 2016

Abstract

Vowel onset point (VOP) is the instant of time at which vowel region starts in a speech signal. VOP plays a vital role in different applications of speech processing, such as syllable detection, speaker verification, duration modification, language identification etc. There are different existing algorithms for the detection of instance of VOP in a speech signal. The algorithm based on the combined evidences extracted from the source excitation, spectral peaks and modulation spectrum has been used as a baseline system for the present work. The baseline system performs well under clean speech data. However, under noisy conditions the performance of the baseline system degrades. The performance of the system degrades in terms of more number of spurious VOPs, which get detected under noisy speech conditions. According to the available literature, this degraded performance is due to the spectral broadening of the speech in the noisy environments. In this paper we have proposed a pre-processing technique on top of the baseline system to reduce this spectral broadening effect of noise. The noisy speech data are passed through the pre-processing algorithm in order to minimize the spectral broadening effect of speech signal. The pre-processed speech is then passed through the baseline system to detect the VOPs in the speech signal. Experiments were carried out under clean and different noisy speech signals. The results of the experiment show an improvement of 16–21% in terms of removal of spurious VOPs, over the existing baseline system under different noisy speech conditions. Further, the performance of the proposed method has been compared with two different best performing techniques for detection of VOP, and found that the proposed method gives a superior level of performance in terms of identification accuracy and identification rate.

Keywords: Vowel onset point; Excitation source; Spectral peak; Modulation spectrum; Perceptual filter.

1. Introduction

Speech can be defined as a pressure of acoustic signal that is articulated in the vocal tract system of human. Speech signal is produced when air is forced from the lungs through the glottis along the vocal tract system. Hence speech is the interaction of the excitation given by the source or glottis with the time varying shape of vocal tract system. VOP is the point in time where the vowel region starts, whereas vocal activity detection (VAD) is a technique for detecting the voiced and unvoiced region of speech signal. Different literature exists for the detection of VOP by the utilizing various features and their combinations. In Prasanna and Yegnanarayana (2005), the detection of VOP is featured by the source excitation technique. The source excitation is evaluated as Hilbert envelope of the Linear Prediction (LP) residual of speech signal. However, the features of the vocal tract are not manifested in this method. The VOP detection method explained in Vuppala et al. (2012b) exploits the fact that the spectral energy at the glottal closure instant (GCI) is higher than the spectral energy at the glottal opening instant. However, the method has not exploited the excitation source information. Another method for the VOP detection in Prasanna et al. (2009) has explored both the source excitation and the vocal tract system features. But, it is seen that the performance of this techniques is not satisfactory under noisy conditions. The performance of this technique degrades resulting in increasing number of spurious

☆ The authors are associated with the Department of Electronics and Communication, National Institute of Technology, Silchar.

* Corresponding author.
E-mail addresses: psaha089@gmail.com (P. Saha), rhlaskar@ece.nits.ac.in (R.H. Laskar).

http://dx.doi.org/10.1016/j.specom.2016.04.004
0167-6393/

VOPs along with the increase in noise level. It is found that the increase in spurious VOPs under noisy speech signal is due to spectral broadening of the speech signal (Prasanna and Pradhan, 2011). Therefore if the spectral broadening effect of noisy speech can be minimized then this method may give a robust VOP detection performance in clean as well as noisy conditions as both the features of source and vocal tract system have been explored in this method. Furthermore the methods proposed in Vuppala et al. (2012a, 2011a) demonstrated a pre-processing technique using speech enhancement and a post-processing technique using the periodicity of GCI respectively. The pre-processing and the post-processing technique used in the two papers are a refinement to reduce the unwanted spurious VOPs over the method explained in Prasanna et al. (2009). However, no substantial work has been done to reduce the spectral broadening effect of noise in speech signals, which might be the core reason for the detection of spurious VOPs.

This paper focuses in reducing the spectral broadening effect of noise in speech signals, and to the knowledge of the author no such work to reduce the spectral broadening effect of noise in the detection of VOP has been reported in literature. The combined evidence plot of source excitation, spectral peak and modulation spectrum in Prasanna et al. (2009) has been used as the baseline method for this paper. On top of the baseline system a pre-processing technique has been used to counter the effect of spectral broadening. Perceptual filter has been used as the pre-processing technique in this paper. Speech signals are passed through the pre-processing block. The pre-processed signal is then subjected to the baseline system for the detection of VOPs. The proposed system is tested in clean and different noisy conditions and is found to give a satisfactory level of performance in all the conditions. The contributions of the paper may be highlighted as follows:

- The spectral broadening effect of noise on different vowel regions has been studied for different noise levels for various types of noises.
- A novel pre-processing filter has been introduced for the improvement of the detection of VOP by addressing the spectral broadening effect of noisy speech signals. The pre-processing filter has been developed using the concept of perceptual filter. The pre-processing filter has been followed by the stability check of the signal by utilizing the concept of pole-zero analysis of the speech signal.
- An experimental study has been carried out to investigate the effect of VOP detection for varying perceptual weighing factors γ_1 and γ_2.
- Experimental study has been carried out to investigate the effect of the proposed pre-processing filter with three different VOP detection techniques and their combinations. The three different techniques are: source excitation, spectral peak and modulation spectrum.
- Comparison of the proposed method with two different existing methods has been evaluated.

The rest of the paper is organized as follows: The study of spectral broadening effect of noisy speech signal is presented in Section 2. The proposed pre-processing technique has been explained in Section 3. Section 4 discusses the effect of the proposed pre-processing technique in speech signal. The baseline method of combination of three techniques for VOP detection has been explained in Section 5. In Section 6 the improvement of detection of VOP using proposed technique has been elaborated. The choice of the parameters for the pre-processing filter is given in Section 7. In Section 8, results are discussed. Section 9 presents the summary of the paper.

2. Effect of noise in spectral domain of speech signal

The speech signal gets affected by various types of noise in the environment. In this section we have presented a study on the spectral behavior of noisy speech signal. The noise when incorporated in speech signal has a characteristic that it tries to flatten the peaks of the formant frequency. In other words, the noisy speech has a tendency to increase the 3 dB bandwidth of the formant frequencies. To establish the fact we have made a study of the effect of noise in the vowels taken from continuous utterance of speech signal. The continuous vowels are taken from speech database of TIMIT. The reason to take the vowels from the continuous speech rather than synthetic vowel or single vowel utterance is because this will help us in the correct detection of vowel in continuous speech signals. The study has been carried out for different vowel utterances in different speech signals. A total of 50 vowels have been chosen randomly to study the effect under noisy conditions. The effect has been studied in the spectral domain of the signal. A study of one of the vowels has been shown in Fig. 1.

Fig. 1 also shows the spectral broadening effect of noise. Fig. 1(a) is a plot of clean speech signal of the vowel /i/ taken from the word /in/ of TIMIT database (Garofolo, 1993). The corresponding figures from (b)–(f) are the spectral envelopes of a frame of the vowel at different noise levels (20, 10, 5, 0 dB babble noise). It can be clearly seen from the figures that as the noise level is increasing from 20 dB to 0 dB babble noise, there is a significant change in the spectral behavior of the speech signal. With the increase in noise level there is an increase in spectral broadening effect which broadens the envelope around the formant frequency of the speech signal. Fig. 1(e) shows the highest amount of spectral broadening effect on speech as it has been corrupted with 0 dB babble noise whereas Fig. 1(b) shows the least affected speech by spectral broadening as it is corrupted with 20 dB babble noise.

From Fig. 1 it is clear that there occurs spectral broadening of the speech signals as the noise level is increased.

The spectral envelope of a frame of speech signal can be determined by the following steps:

1. A frame of 20 ms duration has to be selected from a speech signal by multiplying the speech signal with a hanning window of same size as that of the frame i.e. 20 ms.