Available online at www.sciencedirect.com

SciVerse ScienceDirect

Journal of the Franklin Institute 350 (2013) 698–716

Journal of The Franklin Institute

www.elsevier.com/locate/jfranklin

Marginal energy density over the low frequency range as a feature for voiced/non-voiced detection in noisy speech signals

Pooja Jain, Ram Bilas Pachori*

School of Engineering, Indian Institute of Technology Indore, Indore 452017, India

Received 4 May 2012; received in revised form 2 November 2012; accepted 6 January 2013
Available online 21 January 2013

Abstract

In this paper, we present a pseudo Wigner–Ville distribution (PWVD) based novel method for the voiced/non-voiced (V/NV) detection in noisy speech signals. The energy distribution of the speech signal on the time–frequency plane is obtained by computing the PWVD coefficients of the analytic speech signal over the low frequency range (LFR). The marginal energy density with respect to time (MEDT) over the low frequency range (LFR) derived from the energy distribution of the speech signal on the time–frequency plane is used as a feature to provide the instantaneous V/NV detection. The experimental results on speech signals from the CMU-Arctic database under white, babble and vehicular noise environments taken from the NOISEX-92 database at various signal to noise ratio (SNR) are obtained to assess the performance of the proposed method. A significant performance improvement in the V/NV detection accuracy is obtained by the proposed method over the existing methods for the V/NV detection under the white noise and babble/vehicular noise environments, respectively.

1. Introduction

Voiced/non-voiced (V/NV) detection refers to identification of regions in the speech signal with strong vocal fold activity. During the production of voiced speech, the vocal tract system is excited by the vibration of vocal folds, resulting in a quasi-periodic speech signal. The unvoiced speech is produced when the air is passed through a narrow constriction in the

*Corresponding author. Tel.: +91 7324240716.
E-mail addresses: poojaj@iiti.ac.in (P. Jain), pachori@iiti.ac.in (R.B. Pachori).

0016-0032/$32.00
http://dx.doi.org/10.1016/j.jfranklin.2013.01.002

wind pipe, generating a noise like random output signal. Silence occurs in the absence of any excitation to the vocal tract system and contains only background noise. Non-voiced speech includes unvoiced speech and silence. While speech signal processing applications like language identification [1], multi-rate speech coders [2,3], speech signal modeling [4] require classification of the speech signal into voiced, unvoiced and silence (V-UV-S) regions, there are some prominent speech signal analysis applications like identification of the glottal closure instants (GCIs) [5], pitch frequency estimation [6,7], which require knowledge of only the voiced regions of the speech signal. The prerequisite of boundaries of voiced regions of these applications can be catered by a V/NV detection method requiring much less computational complexity than V–UV–S classification methods. Detection of voiced regions from the speech signal in the presence of noise finds use in automatic speech recognition (ASR) [8]. Applications like speech enhancement [9], diagnosis of pathological voice disorders [10,11], emotion recognition [12,13] rely on the estimation of pitch frequency and detection of GCIs from noisy speech signals. A noise resilient V/NV detection method can provide reliable detection of voiced regions for pitch frequency determination and extraction of GCIs from speech signals distorted by noise.

Several methods have been proposed in the literature to distinguish V/NV regions in the speech signal. Various time domain parameters like zero crossing rate (ZCR), short-term energy estimates have been used to separate voiced/unvoiced (V/UV) regions of the speech signal [14]. However, the method is susceptible to noise. Features extracted from the linear prediction (LP) analysis of the speech signal such as the first predictor coefficient, LP residual energy have been considered to perform V–UV–S classification in [15]. The normalized low frequency energy ratio and merit of periodicity evaluated from the LP residual, harmonicity measure computed from the LP residual have been employed to decide V/UV regions in the noisy speech signal [16,17]. The reliable estimation of the parameters of the assumed statistical distributions of multiple features used in [15] to achieve the V–UV–S classification requires large amount of training data. In order to enable the adaptive modification of the classifier, multi-layer feedforward network was employed in [18] and the feature vector comprising waveform features and cepstral coefficients derived from the LP coefficients and LP residual energy was used to accomplish the V–UV–S classification. The LP based analysis assumes the speech signal to be stationary for about 20–25 ms which is not true for quickly varying phonemes such as plosives [19]. Methods based on frequency domain parameters exploit the periodic structure of the spectrum of voiced regions of the speech signal, such as in [20], the harmonic measure computed from the instantaneous frequency amplitude spectrum (IFAS) was used to perform the V/UV detection and in [8], the similarity between the shape of the signal's short-term magnitude spectrum and the spectrum of the frame analysis window was employed for voicing detection.

The Gabor atomic decomposition was proposed in [21] and the generalized likelihood ratio test which measures the ratio of the energy of the harmonic part of the signal to the energy of the complementary orthogonal non-harmonic part of the signal was proposed in [22] to distinguish V/UV regions in the speech signal degraded with noise. The method in [21] requires training of the radial basis function neural network for different types of background noises. The energy of the zero frequency resonator (ZFR) filtered signal was shown to provide efficient characterization of the glottal activity in the presence of noise [23]. However, an estimate of the pitch period is a prerequisite for this method. The property of noise robustness of GCIs present during voiced regions of the speech