Computers and Electrical Engineering 000 (2017) 1–18

Contents lists available at ScienceDirect

Computers and Electrical Engineering

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

Determination of instantaneous fundamental frequency of speech signals using variational mode decomposition☆

Abhay Upadhyay[a,*], Manish Sharma[b], Ram Bilas Pachori[a]

[a] Discipline of Electrical Engineering, Indian Institute of Technology Indore, Indore, 453552 India
[b] Department of Electrical Engineering, Institute of Infrastructure Technology Research and Management, Ahmedabad, 380026, India

ARTICLE INFO

Article history:
Received 25 June 2016
Revised 23 April 2017
Accepted 24 April 2017
Available online xxx

Keywords:
Speech signal processing
Hilbert transform
Variational mode decomposition
Voiced speech signal
Low frequency range
Instantaneous fundamental frequency

ABSTRACT

In this paper, a novel approach is proposed to determine instantaneous fundamental frequency (IFF) of speech signals. In the proposed method, the detected voiced speech signals are filtered into low frequency range (LFR) and divided into smaller segments. Further, the fundamental frequency components (FFCs) of each segment of the LFR filtered voiced signals have been obtained using the variational mode decomposition method used in iterative way with suitable convergence criteria. In order to determine IFF, the Hilbert transform and smoothing operation have been applied on the obtained FFC. The FFC is obtained by concatenating all the extracted FFCs corresponding to each LFR filtered voiced speech segment. The performance of the proposed method has been evaluated in different databases and also compared with some other existing methods in the presence of white, babble, and vehicular noises with different signal to noise ratios.

1. Introduction

In speech signal processing, instantaneous fundamental frequency (IFF) is associated with the voiced parts of the speech signals and it has many applications like as, speaker recognition, language identification, and speech recognition [1]. The accurate detection of IFF from speech signals is an essential task due to its dependency on many factors like as emotions, gender, and age of the speaker [2].

The developed methods for determination of IFF from speech signals can be categorized mainly into three classes namely, block-based methods, event-based methods, and instantaneous methods [2]. The block-based methods perform detection of the fundamental frequency from the segments of speech signals [3–6]. The local maxima of the autocorrelation (AC) function has been utilized to determine the IFF of speech signals [3]. However, this method is not found suitable for the noisy speech signals. In multi-band summary correlogram (MBSC) method [4], the four wide-band finite impulse response filters are employed in order to capture the variations of harmonics in each sub-band. Further, the IFF has been estimated using different weighting schemes from the noisy speech signals. However, the method presented in [4] is not able to differentiate the sub-bands of harmonics structure corresponding to speech and noise. In pitch estimation filter with amplitude compression (PEFAC) method [5], the pitch is computed using the log frequency spectrum in combination with pre enhancement normalization method. However, the method presented in [5] affects the coherence of speech harmonics in adverse manner.

☆ Reviews processed and recommended for publication to the Editor-in-Chief by Guest Editor Dr. Zhihong Man.
* Corresponding author.
E-mail addresses: phd1301102002@iiti.ac.in (A. Upadhyay), manishsharma.iitb@gmail.com (M. Sharma), pachori@iiti.ac.in (R.B. Pachori).

http://dx.doi.org/10.1016/j.compeleceng.2017.04.027
0045-7906/© 2017 Elsevier Ltd. All rights reserved.

In YIN method [6], the several modifications like cumulative mean normalized difference function, thresholds, and parabolic interpolation functions are employed in AC functions for estimating the IFF. It should be noted that the estimated IFF in the block-based methods is assumed constant for the selected segment of voiced speech signals. This is the main limitation of block-based methods. On the other hand, the methods which depend on detection of glottal closure instants (GCIs) are categorized as the event-based methods. The intervals between successive GCIs of voiced speech signal have been explored for computing the IFF in event-based techniques [1,2]. In [1], the zero-frequency resonator based filtered signal is used to compute the IFF in efficient way. In [2], the eigenvalue decomposition of Hankel matrix formed from voiced speech signal is used for estimating the IFF. However, the method suggested in [2] requires further study for order selection of Hankel matrix. Furthermore, the main limitation of event based methods is their inability to determine IFF between any two successive GCIs.

In the literature, many IFF techniques have been developed in the category of instantaneous methods. The B-spline based method which is based on the multistage optimization process, is proposed in [7]. In B-spline multistage optimization process, the prior information of voiced speech signals are required for detecting the IFF from speech signals. The Hilbert–Huang transform (HHT) based method has been suggested to determine the IFF only from clean speech signals [8]. The HHT based method also suffers from mode mixing problem [9]. Hence, the combination of ensemble empirical mode decomposition (EEMD) and discrete energy separation algorithm based method has been explored for IFF determination [10]. Moreover, it has been studied only for vowels. The empirical wavelet transform based method has been suggested to determine the IFF from voiced speech signal [11]. However, the selection for proper segmentation algorithms of frequency-domain spectrum is required in [11] for decomposing the voiced speech signals. In [12], the IFF has been determined from speech signal using variational mode decomposition (VMD) and Hilbert transform (HT) (VMD-HT) based method. However, the method proposed in [12] is not suitable to capture the time-varying nature of the IFF. Moreover, this method also fails in low-frequency noise environments like babble noise and vehicular noise.

The proposed method in this paper overcomes the limitations of IFF determination method presented in [12]. In this work, first, voiced parts of speech signals have been detected using VMD based method which is proposed in [13]. The detected voiced parts of the speech signals are filtered into low frequency range (LFR) from 50 Hz to 500 Hz using the Fourier-Bessel (FB) series expansion based method [2]. The LFR filtered voiced speech signals are divided into segments. Further, the fundamental frequency component (FFC) of each voiced speech segment is separated using VMD method implemented in iterative way with suitable convergence criteria based on estimated center frequencies (denoted by $\overline{\text{MFs}}$) and distance metric values. Finally, the IFF has been computed by performing HT and smoothing operations on separated FFC which is obtained from the concatenation of all the extracted FFCs corresponding to each voiced speech segment.

The proposed methodology of IFF determination from speech signals has following contributions:

1. The proposed method is a new method to determine the IFF from speech signals in the category of instantaneous methods.
2. In this work, the VMD method is applied in iterative way to extract the FFC. In VMD method, the input parameters selection is a challenging task. We have suggested the suitable input parameters of VMD technique and also explored its Wiener filter structure property [14].
3. In [12], the suggested method for determination of IFF from speech signals is not capable to capture the time-varying nature of the IFF. Hence, the proposed method operates segment-wise in LFR filtered voiced speech signal in order to capature the time-varying nature of IFF from speech signals.
4. Our suggested method is also suitable to remove the false components generated due to babble noise and vehicular noise using the distance metric based approach.

The organization of the paper is as follows. The VMD method has been explained in Section 2. The proposed method to estimate IFF from speech signal has been presented in Section 3. Section 4 provides computational complexity of proposed method for determination of IFF from speech signals. The superiority of the developed method for IFF determination in different noise conditions is highlighted in Section 5. The experimental results and discussion have been given in Section 6. Section 7 concludes the paper.

2. VMD method

The VMD method [14] decomposes a real signal $r(t)$ into K narrow-band components $s_k(t)$. It also computes the center frequencies ω_k where $k = 1, 2 \ldots, K$. In order to obtain these narrow-band components and their corresponding center frequencies, this method formulates the constrained optimization problem [14].

Before formulation of optimization problem, the HT is applied to the components $s_k(t)$ in order to compute the single sided frequency spectrum. After that, modulation property is used to shift the frequency spectrum of these components based on the estimated center frequencies. The bandwidth computation of these components is performed using the H^1 Gaussian smoothness of the demodulated signal [14].