Speech Communication 89 (2017) 37–46

Contents lists available at ScienceDirect

Speech Communication

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

Two-pitch tracking in co-channel speech using modified group delay functions

Rajeev Rajan*, Hema A. Murthy

Department of Computer Science and Engineering Indian Institute of Technology Madras, Chennai, India

ARTICLE INFO

Article history:
Received 8 October 2016
Revised 29 December 2016
Accepted 22 February 2017
Available online 27 February 2017

Keywords:
Power spectrum
Modified group delay
Comb filter
Spectrum estimation

ABSTRACT

Modified group delay functions are beginning to gain significance in the literature for formant estimation, speaker recognition and speech recognition. In particular, group delay functions have the property that they possess higher resolution compared to that of the magnitude spectrum. In this paper, modified group delay functions are used for the estimation and tracking of two pitches in concurrent speech. The power spectrum of the speech signal is first flattened to annihilate the system characteristics, while retaining the source characteristics. Group delay analysis of the flattened spectrum is performed and the predominant pitch is computed. Next, a comb filter is designed to remove the predominant pitch and its harmonics from the group delay spectrum. The residual spectrum is again subjected to group delay analysis and the next candidate pitch is again estimated using modified group delay processing. The first and second pass pitch trajectories are corrected using post processing. The performance of the proposed algorithm was evaluated on two datasets using two metrics; pitch accuracy and standard deviation of fine pitch error. Our results show that phase based processing holds promise in the context of multipitch estimation.

1. Introduction

In speech and music research, robust pitch detection is a fundamental problem. Pitch is the auditory attribute of a sound that allows its ordering on a frequency related scale. The rising and falling pitch contours help in conveying prosody in speech. In tone languages, pitch helps in determining the meaning of words (Oxenham, 2012). Studies show that in tonal languages, the relative pitch motion of an utterance contributes to the lexical information contained in a word unit (Gerhard, 2003). Pitch detection algorithms are primarily based on processing either the time domain signal or the Fourier transform or both (Hess, 1983; Rabiner et al., 1976; Gerhard, 2003). The most commonly used time domain approaches are autocorrelation function and average magnitude difference function. Frequency domain approaches primarily rely on locating the harmonic peaks (Schroeder, 1968). When a combination of speech utterances from two or more speakers is transmitted through a single channel, pitch cues of the individual sources are weakened due to mutual interference. When the individual sources are weak, pitch estimation is still a challenging problem (Wu and Wang, 2003; Jin and Wang, 2011a).

The multi-pitch estimation problem can be formulated as follows (Christensen et al., 2008):

Consider a signal consisting of several, say K, sets of harmonics with fundamental frequencies ω_k, for $k = 1, \ldots, K$, that is corrupted by an additive white Gaussian noise $\omega[n]$, having a variance σ^2, for $n = 0, \ldots, N-1$, i.e.,

$$x[n] = \sum_{k=1}^{K} \sum_{l=1}^{L} a_{k,l} e^{j\omega_k ln} + \omega[n] \tag{1}$$

where $a_{k,l} = A_{k,l} e^{j\phi_{k,l}}$ is the complex amplitude of the lth harmonic of the source with $A_{k,l} > 0$, $\phi_{k,l}$ being the amplitude and phase of the lth harmonic of the kth source, respectively. The number of sources, K, and the number of harmonics for each source, L, are assumed to be known. The model in Eq. (1) is known as the harmonic sinusoidal model. The objective of any approach based on this model is to estimate the individual pitches ω_k from a finite number of samples of $x[n]$.

The pitch estimation of audio signals has a wide range of applications in Computational Auditory Scene Analysis (CASA), prosody analysis, source separation and speaker identification (de Cheveigne, 1993; Murthy and Yegnanarayana, 2011). Multipitch estimation is inevitable in applications such as extraction of "predominant F_o"(Salamon and Gomez, 2012), content-based indexing of audio databases (Tao Li et al., 2003) and automatic transcription

* Corresponding author.
E-mail addresses: rajeev@cse.iitm.ac.in (R. Rajan), hema@cse.iitm.ac.in (H.A. Murthy).

http://dx.doi.org/10.1016/j.specom.2017.02.004
0167-6393/© 2017 Elsevier B.V. All rights reserved.

(Ryynanen and Klapuri, 2008). In this paper, a group delay based multipitch estimation algorithm is proposed and performance is evaluated on two datasets. As opposed to the literature, the composite frequency spectrum is treated as a sum of sinusoidal signals in noise. This is primarily because the periodicity of pitch results in periodic zeroes in the frequency spectrum. The resultant frequency (after removal of vocal tract information) can be treated as a sinusoidal signal. In the context of multispeaker data, a sinusoid is associated with each speaker.

2. Related work

Numerous methods have been reported for multipitch estimation in speech and music (Li et al., 2008; Nishimoto et al., 2007; Wu and Wang, 2003). The correlogram based algorithm proposed by Wu and Wang (2003) uses a unitary model of pitch perception to estimate the pitch of multiple speakers. The input signal is decomposed into sub-bands using a gammatone filterbank and the framewise normalized autocorrelation function is computed for each channel. The peaks selected from all the channels are used to compute a likelihood of pitch periodicities and these likelihoods are modeled by a Hidden Markov Model (HMM) to generate the pitch trajectories. A subharmonic summation method and a spectral cancellation framework is used in the co-channel speech separation algorithm proposed by Li et al. (2008). Multi-pitch trajectory estimation based on harmonic Gaussian Mixture Model (GMM) and nonlinear Kalman filtering is also proposed for multipitch environments (Kameoka et al., 2004). A constrained GMM based approach based on information theoretic criterion is attempted in Nishimoto et al. (2007).

In a polyphonic context, the overlap between the overtones of different notes and the unknown number of notes occurring simultaneously make the multipitch estimation a difficult and challenging task (Badeau et al., 2007). The algorithms used in polyphonic environment for pitch transcription include auditory scene analysis based methods (Kashino and Tanaka, 1993; Mellinger, 1991), signal model based Bayesian inference methods (Goto, 2004), unsupervised learning methods (Smaragdis and Brown, 2003; Virtanen, 2006) and auditory model based methods (Klapuri, 2008; Tolonen and Karjalainen, 2000; Wu and Wang, 2003). In auditory scene analysis based methods, acoustic features and musical information are used to group the sound sources present in a scene, while signal model based methods employ parametric signal models and statistical methods to transcribe the pitch tracks. Unsupervised learning techniques include independent component analysis, non-negative matrix factorization, usage of source-specific prior knowledge and sparse coding. In auditory model based methods, a peripheral hearing model is used for intermediate data representation of the mixture signal, followed by periodicity analysis and iterative cancellation.

Two different varieties of modified group delay (MODGD) based pitch estimation were proposed (Rajan and Murthy, 2013b; 2013a) : (a) In (Rajan and Murthy, 2013b), a careful choice of cepstral window for smoothing the modified group delay function yields the pitch. (b) Group delay analysis is performed on the flattened power spectrum of the signal (Rajan and Murthy, 2013a). The flattened power spectrum with picket fence harmonics is likened to a sinusoidal signal in noise. Using the property that group delay emphasizes peaks, the modified group delay analysis of the flattened spectrum is computed. This results in peaks at $T_0, 2T_0, \ldots,$ where T_0 corresponds to the pitch period. In this paper, the second formulation of pitch estimation is exploited for multipitch estimation.

The outline of the rest of paper is as follows. Section 3 explains group delay functions and modified group delay function briefly. The theory of pitch detection using modified group delay functions is described in Section 4. In Section 5, the proposed system for multipitch estimation is discussed in detail. Section 6 discusses the datasets and the evaluation strategy. The results are analyzed in Section 7 and conclude in Section 8.

3. Group delay functions and modified group delay functions (MODGD)

Signals can be represented in different domains namely as time domain, frequency domain, z-domain and cepstral domain. In Yegnanarayana et al. (1984), it was shown that signal information can also be represented in the group delay domain.

Consider a discrete time signal $x[n]$. Then

$$X(e^{j\omega}) = |X(e^{j\omega})| e^{j\arg(X(e^{j\omega}))} \tag{2}$$

where $X(e^{j\omega})$ is the Fourier Transform (FT) of the signal $x[n]$, $|X(e^{j\omega}|$ is the magnitude spectrum and $\arg(X(e^{j\omega}))$ is the phase function.

The group delay function $\tau(e^{j\omega})$ is defined as the negative derivative of the unwrapped Fourier transform phase with respect to frequency.

$$\tau(e^{j\omega}) = -\frac{d\{\arg(X(e^{j\omega}))\}}{d\omega} \tag{3}$$

From Eq. (2)

$$\arg(X(e^{j\omega})) = Im[\log\ X(e^{j\omega})] \tag{4}$$

Using Eqs. (3) and (4), the group delay function can be computed directly from the signal as shown below (Oppenheim and Schafer, 1990):

$$\tau(e^{j\omega}) = -Im\frac{d(\log(X(e^{j\omega})))}{d\omega} \tag{5}$$

$$\tau(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|X(e^{j\omega})|^2} \tag{6}$$

where the subscripts R and I denote the real and imaginary parts. $X(e^{j\omega})$ and $Y(e^{j\omega})$ are the Fourier transforms of x[n]and nx[n], respectively.

It is important to note that the denominator term $|X(e^{j\omega})|^2$ in Eq. (6) becomes very small at zeros that are located close to the unit circle. This affects the dynamic range of the group delay function, and results in spikes that mask the spectral structure corresponding to formants. As the spikiness of the group delay function has no role to play in source/system characteristics, the computation of the group delay function is modified such that the source characteristics are deemphasized. The term $|X(e^{j\omega})|$ in the denominator of the group delay function is replaced by its cepstrally smoothed version, $S(e^{j\omega})$. The new function obtained is referred to as the modified group delay function in the literature. The algorithm for computation of the modified group delay function is described in Hegde et al. (2007) and is given as

$$\tau_m(e^{j\omega}) = (\frac{\tau_c(e^{j\omega})}{|\tau_c(e^{j\omega})|})(|\tau_c(e^{j\omega})|)^{\alpha}, \tag{7}$$

where

$$\tau_c(e^{j\omega}) = \frac{X_R(e^{j\omega})Y_R(e^{j\omega}) + Y_I(e^{j\omega})X_I(e^{j\omega})}{|S(e^{j\omega})|^{2\gamma}}. \tag{8}$$

Two new parameters, α and γ are introduced to control the dynamic range of MODGD such that $0 < \alpha \leq 1$ and $0 < \gamma \leq 1$. The algorithm for the computation of modified group delay feature (MODGDF), which is the cepstral representation of MODGD, is given in Murthy and Rao (2003). Modified group delay based algorithms can be used effectively to estimate system and source characteristics in speech processing (Murthy and Yegnanarayana, 2011; Krishnan et al., 2011; Madikeri et al., 2015).