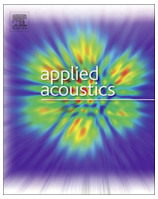
Applied Acoustics 120 (2017) 45–53

Contents lists available at ScienceDirect

Applied Acoustics

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



Technical note

Multiple comb filters and autocorrelation of the multi-scale product for multi-pitch estimation

Jihen Zeremdini [*], Mohamed Anouar Ben Messaoud, Aicha Bouzid

University of Tunis El Manar, National School of Engineers of Tunis, LR11ES17 Signal, Image and Information Technology Laboratory, 1002 Tunis, Tunisia

ARTICLE INFO

Article history:
Received 18 February 2016
Received in revised form 11 November 2016
Accepted 11 January 2017
Available online 19 January 2017

Keywords:
Fundamental frequency estimation
Noisy signal
Two speakers
Multi-scale product
Autocorrelation function
Multiple comb filters
Classification

ABSTRACT

This paper presents a new method that estimates the fundamental frequency in the case of a real noisy environment when many persons speak at the same time and considers the case of two speakers. It essentially gives an accurate estimation of the pitch characterizing the second speaker. The first pitch is determining by detecting the Autocorrelation of the Multi-scale Product (AMP) of the mixture signal. Then a multiple-comb filters is applied to eliminate the dominant signal. After subtracting the resulting signal from the mixture, we obtain the residual signal. Next, we reapply the AMP to the obtained signal to estimate the second pitch. We get a matrix of the second pitch candidates. We classify its elements into three groups. After, we calculate the mean of each column of the appropriate selected group. Finally, the intrusion pitches of each frame are obtained. Experiments are performed using Cooke database. The results show the robustness and effectiveness of the proposed approach.

1. Introduction

Today, the problem of multiple-fundamental frequencies estimation is still being investigated. Indeed, the pitch is an important parameter for speech production, perception, analysis, coding, recognition, identification, synthesis and separation. There are different pitch estimation methods. In the case of a single clean signal, the pitch estimation is done well, but in the case of noisy or composite speech, many current pitch detection techniques still fail to perform well. Multi pitch estimation methods can be classified into three categories: statistical, temporal or spectral filtering and hybrid.

For temporal or spectral filtering methods, we find essentially the approach of Tolonen et al. [1] which is based on the computationally efficient model and periodicity analysis of complex audio signals. This approach uses the summary autocorrelation function (SACF) and enhanced SACF (ESACF) representations for the multi-pitch estimation of an audio signal. Besides, Signol et al. [2] proposed a purely frequency frame to frame algorithm based on the use of plural spectral combs of harmonics suppression (HSP). First, they applied a uniform infinite comb filter in the frequency domain which plays the role of sampler. Then, they used a missing tooth comb filter and a comb filter with negative teeth. Finally, the authors combined the results obtained after weakening the parasitic peaks to select the fundamental frequencies presented. Moreover, Vishnubhotla et al. [3] presented an algorithm based on the AMDF function. It detects the voiced and unvoiced regions in a mixture of two speakers, identifies the number of speakers in voiced regions, and estimates the pitch of each speaker in those regions. They applied a filter bank to the mixture. The fundamental frequency is determined by a two dimensional AMDF representation. In addition, Gilbert et al. [4] used a power spectrum based on a finite impulse response (FIR) filter coming from probability distributions of the speech signal fundamental frequency harmonics. The harmonic windowing function (HWF) greatly improves the difference of arrival time (TDOA) estimation and that of the cross-correlation standard at low signal-noise ratio (SNR). Finally, Huang et al. [5] proposed a Multi-Length Windows (MLW) method. The first fundamental frequency is estimated using the short-time autocorrelation of the mixture. The second estimation is based on the multiple length windows (MLW) which act as an amplifier of the frequency and is used to distinguish all the peaks of harmonic mixing.

For hybrid methods, we note essentially the approach of Gu et al. [6] who use the frequency bin nonlinear adaptive filtering for speech separation. Moreover, a multi pitch contour estimation using HMMs (hidden Markov models) is introduced. This HMM pitch estimator is joined with a pseudo-perceptual pitch estimator

[*] Corresponding author.
E-mail addresses: zeremdini_jihen@hotmail.fr (J. Zeremdini), anouar.benmessaoud@yahoo.fr (M.A. Ben Messaoud), bouzidacha@yahoo.fr (A. Bouzid).

http://dx.doi.org/10.1016/j.apacoust.2017.01.013
0003-682X/© 2017 Elsevier Ltd. All rights reserved.

for the instantaneously estimation of fundamental frequencies of two speakers from summed signals. Besides, Christensen, Højvang, Jakobsson et al. [7] introduced an algorithm based on filtering methods in combination with a statistical model selection criterion. The classical comb filtering method, a maximum likelihood method, and some filtering methods based on optimal filtering are used for fundamental frequencies estimation.

Finally, for statistical methods, we cite the approach of Wu et al. [8] which integrates an improved channel and peak selection method, a new integration method to extract periodicity information from different frequency channels, and a hidden Markov model (HMM) to form continuous pitch tracks. In this context, Jin et al. [9] designed new techniques to select channel and to estimate pitch in order to improve the system proposed by the authors Wu et al. In addition, Zhang et al. [10] used a subspace analysis technique with time space data model in their approach. They applied an estimator for multiple fundamental frequencies detection and Direction-Of-Arrivals (DOAs) estimation of multiple sources. Moreover, Adalbjornsson et al. [11] used block sparsity for multi pitch estimation. They makes the estimate as a sum of a suitable term and convex sparsity inducing norms, providing then a sparse block solution. In addition, to solve the resulting optimization problem, the author has designed a new ADMM (alternating directions method of multipliers) algorithms. The proposed method is robust to the problem sub harmonics, missing harmonics and closely fundamental frequencies. Also, Nielsen et al. [12] proposed a technique based on a minimum of prior information. They used maximum entropy and invariance arguments to derive the observation model and prior distributions corresponding to the prior information in a consistent fashion. Then, they applied the approximations on the signal-to-noise-ratio (SNR) and the number of observations for that. The posterior distribution is derived from the fundamental frequency. Besides, Jensen et al. [13] estimated jointly (DOA) and the pitch of a periodic source by an uniform linear array (ULA). The methods are based on two estimators: nonlinear least-squares (NLS) and an approximate NLS (aNLS). In addition, to make the estimation as a problem of convex optimization sparse group, Kronvall et al. [14] used the alternating direction of multipliers method (ADMM). This approach allows to estimate both the temporal and spatial signal correlation. In addition, after jointly estimate the two fundamental frequencies and time of arrivals (TOAs) for each sensor and sound source, they formed a nonlinear least squares estimation to obtain the DOAs (directions-of- arrivals). As well, Karimian-Azari et al. [15] proposed a method for multi pitch estimation without knowing informations about the signal sources. Indeed, a dictionary containing a set of possible fundamental frequencies and harmonics groups is applied. After, to formulate sparse signal for individual and grouped sinusoids, authors used ‘1-norm’ penalties. The regularization coefficients of the penalty terms must not be the same for all components of the dictionary. Moreover, these coefficients assigned data-dependent regularization coefficients incorporated with an expectation on individual and grouped sinusoids. Furthermore, Liu et al. [16] suggested speaker-dependent (SD-DNN) and speaker-pair-dependent (SPD-DNN) DNNs to form the probabilistic pitch states of two instantaneous speakers. This work was inspired from Han et al. [17] system which modeled the posterior probability of pitch states for mono-pitch tracking by DNNs. Also, Wohlmayr et al. [18] used speaker-dependent Gaussian mixture models (GMMs) and speaker-dependent FHMM (Factorial Hidden Markov Model) respectively to model the probability of pitch periods and to track fundamental frequencies of two speakers.

On the other hand, in our research group, Ben Messaoud et al. [19] have proposed the Spectral Multi-scale Product (SMP) approach based on the short-time spectral analysis of the multi-scale product of the composite signal, and comb filter to estimate iteratively multiple fundamental frequencies of the voiced frames of the composite speech signal. In addition, Zeremdini et al. have presented in 2013, [20] a multi pitch estimation method based on the autocorrelation analysis of the multi-scale product of the composite signal and its filtered version by modified IIR comb filter. And in 2015, they implemented a new multi pitch estimation method based on the calculation of the autocorrelation function of the Multi-scale product (AMP) of the composite signal, its filtered version by a rectangular improved comb filter and the dynamic programming (DP) of the residual signal spectral density.

In this paper, we present and evaluate an algorithm for estimating fundamental frequencies of composite speech signals. Our algorithm applies the autocorrelation of the multi-scale product (AMP) and multiple-comb filters. The new structure of the applied filter allows a good subtraction of the dominant harmonics and gives a suitable estimation for the second speaker. Indeed, this filter provides several possible values for the intrusion pitch making then a good estimate of the fundamental frequency of this signal. The present paper is organized as follows. In Section 2, we present the different techniques used in the proposed approach. Section 3 describes our approach. The evaluation of its performance and the comparison with other state-of-the-art algorithms are given in Section 4. A discussion and an overview are illustrated in Section 5. Finally, Section 6 concludes this work.

2. Methods

2.1. Multi-Scale Product (MP)

The wavelet (WT) transform shows whether details of a certain scale are introduced in the speech signal and quantifies their respective participation. Generally, the WT is meant to offer good frequency resolution at low frequencies. They have sets of properties, including: uncorrelated coefficients to reduce the temporal correlation, compact support to confirm a local analysis, and null moments to choose the useful information.

In 1970, Rosenfeld proposed forming multi-scale point-wise products [21]. This is designed to detect singularities and enhance signals by eliminating noise.

The multi-scale product consists of multiplication of WT coefficients at some dyadic scales. For a signal x, the MP at scales s_j is given by [22]:

$$p(n) = \prod_{j=1}^{l} Wx(n,s_j) \tag{1}$$

where Wx (n,s_j) is the WT of $x(n)$ at the scale $s_j = 2^j$.

In our paper, we use the following scales ½, 1 and 2 for the first category of the intrusion of Cooke database [23], and we use the three scales 3, 7/2 and 4 for the two other categories. This is can be explained by the fact that at the finest scale, the wavelet coefficients are almost dominated by noise. At the second and third scales, the noise diluted rapidly. It can also be seen that at small scales the positions of the step edges are better localized. But some noise may be falsely considered as edge. At the large scales, the edges can be detected more correctly but with the decreasing of the accuracy of the edge location.

It is well known that WT can be employed for the characterization and the detection of signal singularities [24]. Wavelets with n vanishing moments are described as follows:

$$\psi(t) = (-1)^n \frac{d^n\theta(t)}{dt^n} \tag{2}$$

where θ is the smoothing function. So, the WT of a function f at time u and on the scale s can be written as: