A Model For Pitch Estimation Using Wavelet Packet Transform Based Cepstrum Method

Muhaseena T K, Lekshmi M.S

PG Scholar, Ilahia College of Engineering and Technology, Ernakulam, 686673, India

Asst. Professor, Ilahia College of Engineering and Technology, Ernakulam, 686673, India

Abstract

A computationally efficient model for pitch estimation of mixed audio signals is presented. Pitch estimation plays a significant role in music audition like music information retrieval, automatic music transcription, melody extraction etc. The proposed system consists of channel separation and periodicity detection. The input signal is created by mixing two sound signals. First removes the short time correlations of the mixed signal. The model divides the signal into number of channels using wavelet packet transform. Computes the cepstrum of each channel and sums the cepstrum functions. The summary cepstrum function is further processed to extract the pitch frequency of two input signal separately. The model performance is demonstrated to be comparable to those of recent multichannel models. The proposed system can be verified by simulating the system in MATLAB.

1. Introduction

An audio signal is the representation of sound, typically as an electrical voltage. The frequency range of audio signal is between 20 to 20000Hz. The audio signal can be synthesized directly or by using transducers like microphones, phonograph etc. the transducers convert audio signal into corresponding electrical voltages. Loud speakers and headphones converts these electrical voltages into sound. Pitch is an important attribute of the speech. Pitch estimation is very important for the processing of sound signal. Pitch is the basic part of the sound signal. Basically pitch means the highest or lowest tone of a particular sound that is perceived by our ear. It is depend on the number of vibrations that is produced by our vocal chords per second.

Many methods have been already proposed for the estimation of pitch frequency of human sound. For the voice of a single speaker we can easily compute the pitch frequency. But the complexity of the computation of pitch frequency increases by increasing the complexity of the input signal. That is when the input signal is mixed with the voices of more than one speakers then the pitch determination is too difficult. So this paper describes a model that determines the pitch frequency of complex signals.

This paper is organized as follows. Section II presents the literature survey. Section III presents the proposed system which includes the block diagram and expansion of each block. Section IV presents the experimental results, followed by conclusion in section V.

* Corresponding author. Tel.: 9496369684.
E-mail address: lekshmims@gmail.com
2. Review

Recent years many models are proposed for the human pitch perception and periodicity detection. Most of the models use filterbank approach to implement the function of cochlea. The unitary pitch analysis model of Meddis and O’Mard and its predecessors by Meddis and Hewitt are among the best known recent models of time domain pitch analysis. Here the filter bank divide the single channel signal into number of channels.

The envelop of the each channel is computed by using the half wave rectification and lowpass filtering. After computing the envelop of each channel determine the periodicity by the autocorrelation method. In autocorrelation method a signal at a one particular time is compared with its past and future values. The autocorrelation of each channel is summed together to get summary autocorrelation function. This summary autocorrelation functions are the good indicators of the pitch.

Fig. 1 shows the above discussed pitch determination model using filterbank. There are different approaches for the computation of autocorrelation function.

By using the spectral smoothness principle a model is proposed by Anssi P Klapuri[9] for the determination of pitch. In this paper, the authors assumed that the spectral envelopes of real sounds are continues. The smoothing operation corrects approximately half of the pitch errors that occurred in a system. By using spectral smoothness principle the spectral components of a source can be separated from the mixture. Now the multipitch estimation is reduced to single pitch estimation.

Another method of pitch detection is by using the Average Magnitude Difference Function (AMDF)[10]. The fundamental frequency f0 is the main part of the pitch. In AMDF instead of correlating the input speech at various delays a difference signal is formed between the delayed speech and the original. Then the absolute value of the magnitude is taken for each delay. The nonlinear processing is usually used for the pitch tracking to reduce the effect of formant structure.

Fig. 2 shows another method to find out the multipitch of the complex signals [1]. In this paper first the short time correlation of the input signals are removed by using the pre-whitening filter.
Then the single signal channel is divided into two channels that is below and above 1000Hz. The highpass channel is again halfwave rectified and lowpass filtered inorder to separate the low channel. Compute the periodicity of the each channel using autocorrelation. Autocorrelation uses the discrete fourier transform and inverse fourier transform. Autocorrelation is computed by using the following equation

\[ x_2 = IDFT\left( iDFT(x_{low}) \right)^K + IDFT\left( iDFT(x_{high}) \right)^K \] (1)

Where, \( x_{low} \) and \( x_{high} \) are the low and high channel signals. \( K \) determines the frequency domain compression and normally choose \( K \) as 0.67. The Fast Fourier Transform and Inverse Fast Fourier Transform is used here to increase the computational speed.

Output of summary autocorrelation function is good indicators of the pitch period. But from this we cannot estimate the true peaks. Because summary autocorrelation function generates peaks at every fundamental periods. So to extract the correct pitch frequency from SACF signal an enhancement is required. The SACF is first clipped to positive values only. Thus removes the negative peaks from the summary autocorrelation function. Then expand the time period by time scaling by a factor of two. Then it is substracted from the original positive clipped SACF Signal. This difference also contains positive and negative values. So to remove negative peaks its again clipped to positive values. This is the enhanced summary autocorrelation function and the repetative peaks are eliminated in Enhanced Summary Autocorrelation Function (ESACF).

The output of the summary autocorrelation may contain basic peaks as well as duplicate peaks. The enhancement process doubles the values of basic peaks than duplicate peaks. It also removes the near zero time lag part of the SACF curve.

3. Proposed method

Figure 3 shows the block diagram of proposed model consists of four sections they are pre-whitening, channel separation, periodicity detection and peak detection.
3.1 Pre-whitening

Inputs signal to this system will be combinations of two or more voices. So it may contain short time correlations. In order to remove these short time correlations, pre-whitening is required. The pre-whitening is implemented using warped linear prediction (WLP). Warped linear prediction was first proposed by Hans Werner Strube. It is the modification of linear predictive coding and it is a signal modeling technique.

By using warped linear prediction we can reduce the filter order. In this particular model 12th order filter with sampling rate of 22kHz, hamming windowing, frame size of 23.3ms are used.

3.2 Channel separation

The linear prediction coefficients obtained for the input signal is a single channel signal. In this step we divide the single channel into four channels using wavelet packet transform. The half of the signal channel is half wave rectified and lowpass filtered. We used Daubechies 3 wavelet packets in the design as it gave better results compared to other wavelets available in Matlab wavelet toolbox.

3.2 Periodicity detection

Periodicity detection is the important part of the pitch estimation model. In this proposed model Cepstrum method is used for pitch detection. Periodicity of each channel is computed separately and its sum is computed. Each channel output is framed using hamming window with a frame size of 775.

Hamming window is the sum of rectangular window and Hanning window. Speech signal is non-stationary and its properties change over time. And it may be remain invariant for a short period of time. So the speech processing is done for this short period of time and this short period is extracted using window function. Window function of finite length is multiplied with long speech signal giving finite length version of the original signal. Compared to other
windows. Hamming window is a smoother one. This is zero at the edges and rises gradually to be one in the middle. Hamming window reduced the edge effects.

Then the window signal is added with each of the signal channels. Take the Fast Fourier Transform of the signal. Perform the Inverse Fast Fourier Transform of logarithmic amplitude of the frequency components of the signal and it is the cepstrum coefficients. These are computed for each channel separately and finally summed together to get the combined cepstrum coefficients of the input signal.

The cepstrum of the signal is computed using the formula,

$$x_2 = IDFT \left( \log(DFT(x_2 + \text{window})) \right) + IDFT \left( \log(DFT(x_3 + \text{window})) \right)$$

(2)

3.4 Peak detector

Output of the periodicity detection is good indicators of the pitch period. But from this we cannot estimate the true peaks. Because cepstrum coefficients generates peaks at every fundamental periods. So to extract the correct pitch frequency from combined cepstrum coefficients, it is smoothened and coefficients of desired pitch frequencies are selected [80Hz to 250Hz]. The peaks in the desired frequency range are computed and dominant two peaks are considered as the pitch of the speaker’s in the mixture.

The model is analyzed for various input signal combinations and its Root Mean Square Error is computed with the correct pitch values obtained from the clean speeches of individual speaker. RMSE is calculated as,

$$\text{RMSE} = \sqrt{(F_1-F_2)^2}$$

(3)

$F_1$: Pitch frequency obtained using proposed cepstrum model/ existing autocorrelation model

$F_2$: Pitch frequency of individual speakers obtained using cepstrum/autocorrelation model

4. Experiment results

The proposed model is verified for different input mixed signals like male-male voice, female-female voice, male and female with male dominant, male and female with female dominant etc. The TIMIT data base speech is used for the speech processing.

The results are compared with the two channel method which uses autocorrelation function for periodicity detection [1].

<table>
<thead>
<tr>
<th>Type of mixture</th>
<th>Root Mean Square Error</th>
<th>Dominant</th>
<th>Interference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Autocorrelation</td>
<td>Cepstrum</td>
<td>Autocorrelation</td>
</tr>
<tr>
<td>A</td>
<td>72</td>
<td>73</td>
<td>69</td>
</tr>
<tr>
<td>B</td>
<td>95</td>
<td>95</td>
<td>88</td>
</tr>
<tr>
<td>C</td>
<td>67</td>
<td>69</td>
<td>84</td>
</tr>
<tr>
<td>D</td>
<td>89</td>
<td>91</td>
<td>69</td>
</tr>
</tbody>
</table>

Table 1. RMSE of proposed method

<table>
<thead>
<tr>
<th>Type of mixture</th>
<th>Root Mean Square Error</th>
<th>Dominant</th>
<th>Interference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2. RMSE of existing method
<table>
<thead>
<tr>
<th>Type of mixture</th>
<th>Root Mean Square Error</th>
<th>Dominant</th>
<th>Interference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Autocorrelation</td>
<td>Cepstrum</td>
</tr>
<tr>
<td>A</td>
<td>122</td>
<td>111</td>
<td>140</td>
</tr>
<tr>
<td>B</td>
<td>134</td>
<td>137</td>
<td>99</td>
</tr>
<tr>
<td>C</td>
<td>123</td>
<td>113</td>
<td>99</td>
</tr>
<tr>
<td>D</td>
<td>132</td>
<td>131</td>
<td>140</td>
</tr>
</tbody>
</table>

A-Male-male signal  
B-Female-female signal  
C-Male female with male dominant  
D-Male female with female dominant

The experimental result reveals that the proposed model gives better results in terms of RMSE.

5. Conclusion
In this paper we proposed a new approach for estimating the pitch frequency of a sound signal based on wavelet packet transform and cepstrum method. The model has been developed as a compromise for computational efficiency. Pre-whitening based on warped linear prediction removes the short time correlation between the signals. The proposed model provides the better root mean square error and reduced computational complexity.

Acknowledgment
The authors gratefully acknowledge the contributions of TeroTolonen, for his work on the original version of this document.

References
[1] TeroTolonen, “A Computationally Efficient Multipitch Analysis Model” Student Member, IEEE, and MattiKarjalainen, Member, IEEE


[9] Anssi P. Klapuri “Multipitch estimation and sound separation By the spectral smoothness principle” Tampere University of Technology, P.O.Box 553, FIN-33101 Tampere, Finland.

[10] Li Tan and Montri Karnjanadecha “Pitch detection algorithm: autocorrelation method and amdf” Department of Computer Engineering Faculty of Engineering Prince of Songkhla University Hat Yai, Songkhla, Thailand, 90112