A Multipitch Tracking Algorithm Based on Wavelet Packet Analysis

Chen Liting¹, Hu Weiping²*, Mo Jialing³ and Zhang Yixuan⁴

¹College of Electronic Engineering, Guangxi Normal University, Guangxi, 541004, China
²College of Electronic Engineering, Guangxi Normal University, Guangxi, 541004, China
³College of Electronic Engineering, Guangxi Normal University, Guangxi, 541004, China
⁴College of Electronic Engineering, Guangxi Normal University, Guangxi, 541004, China
*Corresponding author’s e-mail: huwp@gxnu.edu.cn;

Abstract. In view of the limitation of many multipitch detection methods in single-channel speech, a robust and accurate multipitch estimation method for multiple voices is processed. In this study, our approach is based on the spectral analysis of the wavelet packet analysis. It utilizes the quasi-periodicity in a short time frame of speech, and gets candidate pitch in every frame by using peak selection and searching algorithm. The multipitch envelope of the mixed signal is obtained by neighborhood analysis and determined from these candidates by virtue of the fact that pitch should change rather smoothly in consecutive frame. Simulation results showed that the proposed method can robustly estimate fundamental frequency for mixed speech from single-channel speech.

1. Introduction

We live in an environment rich in sound from many sources. The presence of multiple sound sources complicates the processing of the target sound that we are interested in, and often causes serious problems for many applications, such as computational auditory scene analysis [1] (CASA), prosody analysis, speech recognition, and speaker identification. However, human beings can concentrate on some interesting voices in noisy environment, the single ear also has the ability to solve the "cocktail party problem"[2]. Perceptually, one of the most potent cues for monaural sound segregation is the fundamental frequency; Especially, the listener can use the difference of the fundamental frequency to separate the harmonic of a sound from the harmonic of the interfering sound. Accordingly, researches on monaural CASA have focussed on the problem of identifying the multiple fundamental frequency present in an acoustic mixture called 'multipitch analysis' and using them to separate the constituent sounds.

Many better performance algorithms for pitch detection have been obtained. However, multipitch detection from a single signal is still a hard task. There are many algorithms have been proposed for multipitch detection of single-channel speech, DeLiang Wang [3] present the autocorrelation algorithm based on CASA model, Michael Wohmayr [4] proposed a probabilistic interaction model...
algorithm based on FHMM, Yuzhou Liu [5] proposed a depth neural network algorithm based on multi-speaker independence, and Rajeev Rajan [6] proposed a double-pitch tracking algorithm based on improved group delay function. The above algorithms have strong robustness, but these algorithms need to do a lot of statistics on their voice database, which requires high statistical characteristics of data and complex algorithm.

In this paper, we propose a robust algorithm for multipitch tracking of mixed speech. Firstly, the vocal cord excitation of mixed speech signal is extracted by wavelet packet transform. Then, through the periodicity analysis of the spectrum to obtain the pitch period of the dominant speaker, and using the relationship between fundamental frequency and frequency multiplication to obtain the pitch period of the concurrent speaker. Finally, utilizing the continuity of speech signals, multipitch envelope of the mixed signal is obtained by neighborhood analysis and determined from these candidates by virtue of the fact that pitch should change rather smoothly in consecutive frame.

This paper is organized as follows. Section 2, we presents some application of wavelet packet analysis in speech signal processing. Detailed explanations of our pitch tracking method are given in Section 3. Section 4 provides evaluation experiments and shows the results. Finally, we discuss related issues and conclude the article.

2. Wavelet Packet Analysis

The wavelet transform[7],[8] can be used for different applications. According to Mallat, the wavelet transform, has shown excellent capacities for the detection of signal singularities [9]. When the wavelet function has specific selected properties, wavelet transform acts as a differential operator. The number of wavelet vanishing moments gives the order of the differentiation. For an appropriately chosen wavelet, the wavelet transform modulus maxima denote the points of sharp variations of the signal [10]. This property of DWT has been proven very useful for detecting pitch periods of speech signals [11].

Wavelet analysis and wavelet packet analysis[12] are effective methods for analyzing speech signals. Wavelet analysis is a time-frequency localization analysis method with variable time and frequency windows. Although it has good time-frequency analysis ability, but it also has some disadvantages, such as poor time resolution in the low-band and poor frequency resolution in the high-band. In this case, wavelet packet analysis came into being. Wavelet packet analysis overcomes these shortcomings, it can provide a more precise method for signal analysis, And not only decomposes low-band signals, but also decomposes high-band signals.

In voice signal, to perceive the multiple pitches evoked by several simultaneous sounds, the auditory system must estimate their periods. The vocal cord excitation signal is modulated by the channel filter to form different forms of speech. When deconvoluting, the energy of impulse signal is mainly concentrated in low-band and the energy of vocal cord excitation signal is concentrated in high-band. So we need to select an appropriate threshold to exclude the influence of vocal tract factors and separate the vocal cord excitation signal.

3. The Proposed Method

In this section, we first give an overview of the algorithm and stages of processing. As shown in Figure 1, the proposed algorithm consists of three stages. In the first stage, The first step is using the wavelet packet transform to extract the vocal cord excitation signal of the voice signal; The second step is pitch detection, through the periodicity analysis of the spectrum to obtain the pitch period of the dominant speaker, and using the relationship between fundamental frequency and frequency multiplication to obtain the pitch period of the concurrent speaker. The last step is pitch tracking, and get the fundamental frequency envelope of the mixed signal.
3.1. Extraction of Vocal Cord Excitation Signal

As shown in Figure 2, the extraction of vocal cord excitation signal based on wavelet packet analysis is discussed. The algorithm steps are as follows:

- Calculating the Fourier transform of the frame signal \( S(t) \).
  \[
  P(\omega) = \lg |\text{FFT}(s(t))|^2
  \]
  (1)
  Where \( P(\omega) \) is the power spectrum.
- Six-layer wavelet packet decomposition.
  \[
  \left\{d_1, d_2, \ldots, d_i, \ldots, d_k\right\} = \text{DWT} \left( P(\omega) \right)
  \]
  (2)
  Where \( d_i \) is the i-th frequency sub-band, six-layer decomposition is performed on the input speech signal \( P(\omega) \) using Db10 wavelet, and the original signal is decomposed into 64 different frequency sub-bands, and the low-frequency coefficient vector and the high-frequency coefficient vector are respectively obtained for 64 frequency bands according to the Nyquist sampling theorem, the frequency range of the 64 sub-bands obtained by wavelet packet decomposition is 0-62.5Hz, 62.5-125Hz, ..., 3937.5-4000Hz.
- Reconstructing the signal from the lowest sub-band in turn.
  \[
  \begin{bmatrix}
  D_1 \\
  D_2 \\
  \vdots \\
  D_k
  \end{bmatrix}
  = \text{IDWT}
  \begin{bmatrix}
  d_1 & 0 & 0 & \cdots & 0 \\
  d_1 & d_2 & 0 & \cdots & 0 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  d_1 & d_2 & d_3 & \cdots & d_k
  \end{bmatrix}
  \]
  (3)
  Then, reconstruct the frequency coefficient vector to obtain the sub-signal which has the same length with the origin signal.
- Calculating the energy of reconstructed signal.
  \[
  E_i = \sum_n \lg |D_i(n)|^2
  \]
  (4)
- Calculating difference of energy in adjacent bands.
  \[
  b = \left\{ b_1, b_2, \ldots, b_i, \ldots, b_{k-1} \right\}
  \]
  (5)
  Where, \( b_i = |E_{i+1} - E_i| \), is the absolute value of difference between adjacent frequency bands.
- Finding the minimum value of \( b_i \), and i is the separation point.
- Using wavelet packet filter to reconstruct vocal cord excitation signal in high-band, and transform it to spectrum.
Figure 3. Examples of Wavelet Packet Analysis. (a) voice frame. (b) power spectrum. (c) the red dotted line is the spectrum of the impulse signal of the vocal tract, which corresponds to the slowly changing envelope. (d) spectrum of the vocal cord excitation signal, corresponding to the rapidly changing details.

3.2. Fundamental Frequency Detection

In sick voice, the spectrum of the vocal cord excitation signal will appear periodic [13] between 0~1700Hz, and extend this conclusion to normal voice. In order to ensure the robustness of the experiment, we intercepted the frequency spectrum between 0~1600Hz. In this case, the spectrum is the superposition of two spectrum in mixed speech, so we assume that the fundamental frequency of two speakers do not overlap.

As shown in Figure 4, in this stage, We intercept the spectrum from 0 to 1600Hz, and then Fourier transform is applied to it. Considering the relationship between the two speakers' amplitudes in mixed speech, the experiment discussed two cases. In this paper, We use the peak detection function to detect its peak value. When the peak is one, judge it as the first type, shown in Figure 5(b); and the peak is two, judge it as the second type, shown in Figure 5(d).

Figure 4. Diagram of Amplitude Type Judgment

```plaintext
Fourier transform  Peak detection  Amplitude Type Judgment
```

Figure 4. Diagram of Amplitude Type Judgment
Figure 5. Examples of type judgment. The first column (a) and (c) are the spectrum of vocal cord excitation signals with frequency ranging from 0~1600Hz. The second column(b) and (d) are the Fourier transform of spectrum, and discussed two amplitude types separately.

For the first type, the difference between the two amplitudes is large. Figure 6(a) and 6(c) show the spectrum of the single voice, respectively. What we have observed is that the obvious periodicity is pitch period, and we can obtain the accurate value from Figure 6(b) and 6(d). However, when they mix, there are differences in the amplitude of the original voice. The concurrent speech will annihilate in the dominant speech. So we can only detect the maximum peak position is the fundamental frequency of the dominant speaker of mixed speech.

Figure 6. Examples of the first type. (a) spectral of the dominant speech. (b) Fourier transform of the dominant spectrum. (c) spectral of the concurrent speech. (d) Fourier transform of the dominant spectrum. (e) spectral of the mixture. (f) Fourier transform of the mixture.
And in order to obtain the fundamental frequency of the concurrent speaker, the strategy is to remove the impact of the dominant speaker’s fundamental frequency and their multiplier information. The algorithm is as follows:

- Center clipping the spectrum. The initial clipping value is 1% of the maximum amplitude of the spectrum, increasing by 0.01 steps, and the iteration stops when the clipping value increases to 90% of the maximum amplitude of the signal. Simultaneously detect the peak of the clipping signal and count the number of peaks.
- Draw distribution curve. (the x-axis is the clipping value, the y-axis is the peak number).
- Find the clipping threshold. When the value of y-axis is one, the minimum value of the x-axis is the optimal threshold.
- Clipping the spectrum. Using the optimal threshold in step 3, signals larger than clipping threshold are retained and lower than clipping level are set to zero.
- Spectral subtraction. subtraction of the original spectrum and the clipping spectrum. At this point, we think that the signal contains more the concurrent speaker’s signals.
- Perform Fourier transform and find the peak, the value of this frequency is the fundamental frequency of the concurrent speaker in the mixed voice.

Figure 7. Diagram of the first type

- Center clipping the spectrum. The initial clipping value is 1% of the maximum amplitude of the spectrum, increasing by 0.01 steps, and the iteration stops when the clipping value increases to 90% of the maximum amplitude of the signal. Simultaneously detect the peak of the clipping signal and count the number of peaks.
- Draw distribution curve. (the x-axis is the clipping value, the y-axis is the peak number).
- Find the clipping threshold. When the value of y-axis is one, the minimum value of the x-axis is the optimal threshold.
- Clipping the spectrum. Using the optimal threshold in step 3, signals larger than clipping threshold are retained and lower than clipping level are set to zero.
- Spectral subtraction. subtraction of the original spectrum and the clipping spectrum. At this point, we think that the signal contains more the concurrent speaker’s signals.
- Perform Fourier transform and find the peak, the value of this frequency is the fundamental frequency of the concurrent speaker in the mixed voice.

Figure 8. (a) distribution curve. (b) spectrum of central clipping. (c) spectral subtraction. (d) Fourier transform of spectrum.

For the second type, the difference between two amplitudes is small. So we can detect two peaks in detection and get two pitch period of two speakers easily.
3.3. Pitch Tracking

The two pitch periods of mixed speech is obtained, and the envelopes of fundamental frequency need to be further determined [14]. By the virtue of the fact that pitch should change rather smoothly in consecutive frame, the fundamental frequency variation within two adjacent frames will not exceed 15 Hz. In addition, we assume that the pitch period of the fundamental envelope will not change abruptly, that is, the angle difference between the straight line formed by two adjacent points and the front line is not more than 90 degrees.

Using the above characteristics, we analyze the fundamental frequency points of each frame in Section 3, then achieve the goal of multipitch tracking.

4. Results Analysis

In this section, we produce the results obtained after applying the proposed algorithm to mixture of male-female speakers. Experiments were conducted on synthetic data from THCHS-30[15], which is an open Chinese speech database published by Center for Speech and Language Technology (CSLT) at Tsinghua University.

Our algorithm is evaluated on a corpus of 100 mixture sounds composed of ten male utterances pronounced and ten female voices utterances pronounced for one hundred combinatorial experiments. And then we count the error detection of our proposed algorithm.

Figure 10 illustrates the estimated pitch periods from the speech signal of the sentence pronounced by the two simultaneous utterances of a male and a female speaker. We note that the pitch periods obtained from the speech signal match exactly with the estimated ones using the single signal. Figure 11 depicts a more difficult case where the speech signal mixe with tone changes, and there’s a pause in the statement pronounced. In this case, although the detection effect at the beginning and end of the statement is not ideal, our algorithm allows a good multi-pitch tracking in continuous frames. The pitch periods of both dominant and concurrent sounds are estimated by our proposed approach.
Statistical the results of one hundred sentences, our algorithm achieves a total gross errors of 7.56% for the pitch tracking. To put the above performance in perspective, we compare with two recent multipitch detection algorithms proposed by Tolonen and Karjalainen [16] and Mingyang Wu and DeLiang Wang [3]. The total gross error for mixtures is 7.56% for ours, and for others it ranges from 7.17% to 10.04%. Generally, on the basis of simplifying the computation, our algorithm achieves the average detection effect.

Figure 10. Examples of pitch tracking (a) Time-frequency energy plot for a mixture of two simultaneous utterances of a male and a female speaker. The utterance are vowels in Pinyin “a”. (b) Result of tracking the mixture. The solid lines indicate the true pitch tracks. The “*” and “o” tracks represent the pitch tracks estimated by our algorithm.

Figure 11. Examples of pitch tracking. (a) Time-frequency energy plot for a mixture of two simultaneous utterances of a male and a female speaker. The utterance are Chinese Pinyin “lantian baiyun” and “shangma mama maren.” (b) Result of tracking the mixture. The solid lines indicate the true pitch tracks. The “*” and “o” tracks represent the pitch tracks estimated by our algorithm.

5. Conclusion
The results show that our algorithm performs reliably for tracking multipitch tracks. However, there are still some shortcomings. First, when there is a pause or burst occurs during a period of speech, the head and tail of speech are not as periodic as the vocal cord excitation signal, there will be "wild point" deviating from the fundamental frequency at the time of detection. Second, the detection effect is not ideal when the speaker's fundamental frequency is closer to each other, which requires further study.

In summary, we propose a novel multipitch tracking algorithm. This algorithm is based on the spectral analysis of the wavelet packet analysis. We test this approach and prove its efficiency in estimating the fundamental frequency of both the dominant and the concurrent sound when it exists.
In the future, we will test our algorithm with ten intrusions by Cooke [17].

Acknowledgments
This work was supported by the National Natural Science Foundation of China (No. 61861005).

References
[1] D.L. Wang, and G.J. Brown, "Separation of speech from interfering sounds based on oscillatory correlation," IEEE Trans. Neural Net, vol. 10, pp. 684-697, 1999.
[2] P. G. Parande and T. G. Thomas, "A study of the cocktail party problem," 2017 International Conference on Electrical and Computing Technologies and Applications (ICECTA), Ras Al Khaimah, 2017, pp. 1-5.
[3] Mingyang Wu, DeLiang Wang and G. J. Brown, "A multipitch tracking algorithm for noisy speech," in IEEE Transactions on Audio, Speech and Language Processing, vol. 11, no. 3, pp. 229-241, May 2003.
[4] M. Wohlmayr, M. Stark and F. Pernkopf, "A Probabilistic Interaction Model for Multipitch Tracking With Factorial Hidden Markov Models," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 4, pp. 799-810, May 2011.
[5] Y. Liu and D. Wang, "Robust pitch tracking in noisy speech using speaker-dependent deep neural networks," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, 2016, pp. 5255-5259.
[6] Rajan, Rajeev & Murthy, Hema. (2017). Two-pitch tracking in co-channel speech using modified group delay functions. Speech Communication. 89.10. 10.1016/j.specom.2017.02.004.
[7] S. Mallat, A Wavelet Tour of Signal Processing, Academic Press, second edition, 1999.
[8] C.S. Burrus, R.A. Gopinath and H. Guo, "Introduction to Wavelets and Wavelet Transform : A Primer," Prentice Hall, 1998.
[9] M. A. B. Messaoud, A. Bouzid and N. Ellouze, "Spectral Multi-Scale Analysis for Multi-Pitch Tracking," 2009 IEEE 13th Digital Signal Processing Workshop and 5th IEEE Signal Processing Education Workshop, Marco Island, FL, 2009, pp. 26-31.
[10] Z. Berman, and J.S. Baras, "Properties of the multiscale maxima and zero-crossings representations," IEEE Trans.on Signal Processing, pp. 3216-3231, 1993.
[11] S. Kadambe, and G. Faye Boudreaux-Bartels, "Application of the Wavelet Transform for Pitch Detection of Speech Signals," IEEE Trans, on Info. Theory, vol. 38, pp. 917-924, March 1992.
[12] W. Nogueira, A. Giese, B. Edler and A. Buchner, "Wavelet Packet Filterbank for Speech Processing Strategies in Cochlear Implants," 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, Toulouse, 2006.
[13] Hu Weiping, Wang Xiuxin, Yao Tiejun, Xu Aihua, Zeng Sien, Wang Qiang, Liang Dongdong. The application of frequeney-domain harmonics-to-noise ratio in the acoustic analysis of the normal voice and the abnormal voice [J]. Journal of Clinical Otolaryngology. 1998(06):263-265.
[14] Zhou Chao, Hong Hong Hong. Detection of double pitches for Chinese Putonghua [J]. Journal of Acoustics. 2011(2):239-243.
[15] Wang D, Zhang X. THCHS-30: A Free Chinese Speech Corpus[J]. arXiv preprint. arXiv: 1512.01882, 2015.
[16] T. Tolonen and M. Karjalainen, “A computationally efficient multipitch analysis model,” IEEE Trans. Speech Audio Processing, vol. 8, pp.708–716, Nov. 2000.
[17] M.P. Cooke, Modeling auditory processing and organisation, Press, U. K: Cambridge Univ. 1993.