Multi-Pitch Detection for Co-Channel Speech
Utilizing Frequency Channel Piecewise Integration and Morphological Feedback Verification Tracking

Yong Guan*, Peng Li, Wenju Liu and Bo Xu

National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences,
{yguan,lwj}@nlpr.ia.ac.cn, {pengli,xubo}@hitic.ia.ac.cn

Abstract. We propose an effective algorithm to detect the multi-pitch of co-channel speech with two speakers. A discriminative scheme is exploited to gather harmonic information according to the piecewise continuity of autocorrelation peaks across different frequency channels within a time frame. A group of adaptive morphological filters and feedback verification are employed to coordinate and verify the periodicity information across time frames. Quantitative evaluation shows that the proposed system gains excellent results in veracity and integrality comparing with existing good algorithms of multi-pitch detection for co-channel speech. In addition, the proposed algorithm performs efficiently in time-consuming as well as memory-consuming.

Key words: Multi-Pitch Tracking, CASA, Piecewise Continuity of Harmonic Information, Morphological Filtering

1 Introduction

A reliable algorithm for multi-pitch tracking is critical for many applications, including computational auditory scene analysis (CASA), prosody analysis, speech enhancement, speech recognition, and speaker identification. A typical demand of multi-pitch tracking comes from the processing of co-channel speech, which is composed of speech utterances from two talkers. Sounds from two talkers have large overlap with different harmonic structures. Consequently, it presents considerable difficulty to gain multiple pitches especially for monaural situation. Proposed multi-pitch determination algorithm orients towards the challenge of tracking multiple pitches for monaural co-channel speech.

Pitch determination algorithms (PDA) are generally classified into three categories[2]: time domain, frequency domain, and time-frequency domain. Time-frequency domain algorithms perform time-domain analysis on band-filtered signals obtained via a multi-channel front end. The proposed algorithm is based on

* This work is supported in part by the China National Nature Science Foundation (No. 60172055, No. 60121302), the Beijing Nature Science Foundation (No.4042025) and the National Fundamental Research Program (973) of China (2004CB318105).
time-frequency analysis through a 128-channel gamma-tone filter, which is usu-
ally used as the auditory periphery of CASA [1]. Autocorrelation of the output
from each channel is calculated to gather the periodicity information. A recent
excellent algorithm by Wu et al [3] was designed with the same front end for
multi-pitch detection in noisy conditions, including co-channel speech. In Wu’s
algorithm, channel and peak selection is applied to obtain useful periodicity
information, which is then integrated by a statistical model. Finally an HMM
is used to tracking the multi-pitch in noisy conditions. By channel and peak
selection, peaks in correlation obviously interfered by noise are discarded and
with HMM tracking, the robustness of system is reinforced and the periodicity
information is united in a steady statistical frame. This algorithm is applied
successfully to speech/song segregation [4] and speaker recognition [5]. We will
do a comparative study with Wu’s algorithm in theory, results and efficiency.

In our algorithm, two-phase processing is adopted to tracking multi-pitch in
monaural co-channel speech. First, primary pitch candidates are gathered utilizing
piecewise continuity of autocorrelation peaks in low frequency channels, and
the peaks associated with different harmonics are processed respectively. The
harmonic information in high channels is selectively used since it is not conspic-
uous in many situations. Second, to track the pitch contour across time frames
precisely, a group of adaptive mathematical morphological filters are applied.
In this processing phase, feedback verification utilizing original autocorrelation
information is integrated to get more precise pitch contour. Mathematical mor-
phology [6] is a nonlinear filtering method which is a process of set-theoretical
algebra. It was originally applied in processing binary image, and some attempts
had been carried out including pitch contour smoothing [7].

The outline of this paper is the following. Section 2 gives a detailed descrip-
tion of our system. Specific processing in front end, detailed peaks selection and
integration scheme, and mathematical morphological filtering with feedback pro-
cessing will be introduced here. Then experimental results and Evaluation are
given in section 3. Section 4 concludes our study and gives expectation in farther
direction.

2 System Description

Proposed algorithm includes three main phases, as showed in Fig.1.: the first
phase is front end, where input co-channel speech is decomposed with cochlear
filter in time-frequency domain and the normalized autocorrelation is computed;
the second phase is primary pitch decision using low frequency information and
selectively using high frequency information; the last phase tracks and verifies
the primary pitch using morphological processing and feeding back to get more
convinced pitches from original normalized autocorrelation information. In the
following subsections, we will present these three phases mentioned above sepa-
rately.
2.1 Multi-Channel Front End

The input signal (16kHz sampled) is passed through a bank of fourth-order gamma-tone filters, which is a standard model for cochlear filtering [1]. The bandwidth of each filter is set according to its equivalent rectangular bandwidth (ERB), and a bank of 128 gamma-tone filters with center frequencies equally distributed on the ERB scale between 80 Hz and 5 kHz are used. This is a typical auditory periphery [1], [3], [4], [5].

The channels are classified into low frequency channels (center frequency below 800 Hz) and high frequency channels (center frequency above 800 Hz) [3]. In each high frequency channel, the envelope of the output is extracted. Then for each channel, at a given time step $j$, which indicates the center step of a 16 ms long time frame, the normalized correlation $A(c, j, \tau)$ for channel $c$ with a time lag $\tau$ is computed by running the following normalized autocorrelation function in every 10-ms interval [3]:

$$
A(c, j, \tau) = \frac{\sum_{n=-N/2}^{N/2} r(c, j+n)r(c, j+n+r)}{\sqrt{\sum_{n=-N/2}^{N/2} r^2(c, j+n)} \sqrt{\sum_{n=-N/2}^{N/2} r^2(c, j+n+r)}}.
$$

(1)

Where $r$ is the filter output, $N = 256$ corresponds to the 16 ms window size (one frame) and normalized autocorrelations are computed for $\tau = 0, 1, \ldots, 200$. The autocorrelation is computed directly on the output of a filter in low frequency channels while on the output envelope in high frequency channels. The peaks in autocorrelations contain the periodicity information to be used in pitch hypotheses evaluation. This phase is the same as the process of Wu [3].
2.2 Periodicity Information Selection and Integration

In this sub phase, we gain the pitch candidates from fundamental periodicity information provided by autocorrelation function of the output from each frequency channel after auditory periphery.

Monaural co-channel speech is a combination of speech utterances from two talkers, and it is usually produced when two speech signals are transmitted over a single communication channel. Consequently, speeches from two talkers have large overlap and the autocorrelation of the output from each frequency channel after auditory periphery is affected by different harmonics from two talkers. Still, since sound has an energy masking effect in one time-frequency unit, the autocorrelation is mainly associated with an individual harmonic in a certain unit. Also, for multi-channel filtering, the energy is continuous in conjoined frequency channels. Therefore, the autocorrelation is piecewise continuously associated with a certain harmonic in frequency direction, especially in low frequency channels.

As an example, Fig. 2. shows this phenomenon. We see that, the peak positions of autocorrelation function from co-channel speech are consistent with the ones in target speech from channel 1 to channel 40, but they are consistent with the ones in interferer speech from channel 42 to channel 50.

![Fig. 2. Schematic Picture of ACF Peaks Distribution in Low Frequency Channels.](image-url)

In Fig. 2, Upright line represents the true pitch of co-channel speech, + marks the peak position of autocorrelation function of original target speech, and x marks the peak position of autocorrelation function of original interferer speech, and o marks the peak position of autocorrelation function of co-channel speech.

A low frequency channel will be selected if the maximum value of the autocorrelation in the plausible pitch range exceeds threshold $\theta = 0.9$. Across
different channels, we compute the mean normal correlation where the normal autocorrelation has the same peak numbers in low frequency channels, in which the peak positions of autocorrelation function from co-channel speech are consistent with the ones coming from only one harmonics just like Fig.2 showed. Such peak numbers are piecewise continuous in frequency direction according to the piecewise continuity of autocorrelation in co-channel speech.

\[
A_k(j, \tau) = \frac{1}{N_e} \sum_{c \in E_k} A(c, j, \tau).
\]

(2)

Where \(E_k\) is the set of channels with the same \(k\) autocorrelation peak numbers, \(N_e\) is the total element numbers in \(E_k\). Then the maximal peak in each mean normal correlation \(A_k(j, \tau)\) is retained as the pitch candidate. Such candidates are integrated to not more than two pitch candidates according to the relation of candidate values and their channels coming from.

In the high frequency range, we retain all the channels so that more information is available. Since the peak values fluctuant with large range and a unitary threshold is hard to choose, we compute the statistic numbers of the delay where the first peak occurs instead of mean autocorrelation. Then we gain the first local maximum of such numbers and regard its delay value as a supplementary pitch candidate in each time frame. These candidates can give additional information for the frames where the pitches are not sure in low frequency channels.

In Figure 3, we give the primary pitch candidates through piecewise integration of low frequency information and supplement with high frequency information. Where solid line represents the true pitch of the target and interfere speech, diamond square represents the candidate pitch point and so does \(\times\) marks.

**Fig. 3. Pitch contours after Primary Processing.**
2.3 Pitch Tracking and Verifying by Mathematical Morphology

Mathematical morphology [6] was proposed as a signal processing method. It is easy to process parallel and realize by hardware. Structuring element is very important and essential, and it is a set in Euclidean space with a compact region of support. Mathematical morphology consists of two fundamental operators, dilation and erosion from which other morphological operators are derived. Dilation and erosion are Minkowski addition and subtraction with structuring element. The definitions of these two operations are expressed as follows:

\[
(f \oplus g)(x) = \sup_{y \in G} \{f(y) + g(x-y)\}. \tag{3}
\]

\[
(f \ominus g)(x) = \inf_{y \in G} \{f(y) - g(x-y)\}. \tag{4}
\]

Where \(f(x)\) is gray-level signal, \(g(x)\) is structuring elements. In digital signal processing, supremum and infimum of function can be replaced by maximum and minimum. The combinations of dilation and erosion can form further morphology operations, open, close and so on. The definitions of open and close operations are:

\[
(f \blacklozenge g)(x) = [f \ominus g] \oplus g. \tag{5}
\]

\[
(f \oslash g)(x) = [f \oplus g] \ominus g. \tag{6}
\]

In our algorithm, gray-level pitch curve is processed by a group of mathematical morphological filters and the structure elements showed in Figure 4. These structure elements are selected according to pitch continuity and trend.

![Fig. 4. Schematic diagrams of structure elements.](image-url)

In Fig.4., hollow circle with a black point in it represent the datum mark of morphological, solid circles represent other points, and hollow circle without
a black point denotes the datum mark in pitch contour is zero. First, structure elements in Figure 4 are used as erosion morphological filter. The dilation operators with same structure elements are performed followed by feedback verification by using of the original autocorrelation information. Therefore, these series of morphological processing make up of open operators despite of some revision to utilize feedback verification.

Through series of morphological filtering, the pitch contour obtained from original peak candidates is adjusted to get rid of single pitch point, and to infill visible valley at the assistance of feedback verification from autocorrelation information. The result can be seen in the following Fig.5., where solid line represents the true pitch of the target and interfere speech, diamond square represents the proposed pitch point and so does × marks. Comparing with the pitch contour in Fig.3, we can find these effects by morphological filtering.

![Pitch Tracking](image)

**Fig. 5. Pitch contours after Morphology Processing.**

3 Experimental Results and Evaluation

A corpus of 100 mixtures of speech and interference collected by Cooke [1] is commonly used for CASA research. Part of the corpus, with the last three kinds of noise, n7, n8 and n9, which are speech interference, is selected as the experimental data to evaluate our system [3]. These data are typical co-channel speech.

In evaluation, we take Wu’s algorithm to calculate reference pitches with pure speech before mixing. To revise some obvious mismatching, manual correcting is selected to improve the veracity of reference pitch contour. Just like Wu’s description [3], we evaluate a group of detection errors: $E_{i \rightarrow j}$, $E_{Gross}$, $E_{Fine}$, $E_{Dom}$, and the results are shown in Table 1 and Table 2, where $E_{i \rightarrow j}$
$i, j \in \{0, 1, 2\}$ represents the error rate of time frames with $i$ pitch points misclassified as $j$ pitch points, $E_{\text{Jump}}$, the jump error, is the summation of $E_{i \rightarrow j}$ for all $i$ and $j$. $E_{\text{Gross}}$, the gross error, is the error rate of time frames with more than 20 percent pitch values error without jump errors, $E_{\text{Fine}}$, the fine error, is the average pitch values error without jump and gross errors, and $E_{\text{Dom}}$, $E_{\text{Dom}}$, are the gross and fine errors for dominant pitches.

### Table 1. Error Rates (in Percentage) I of Pitch Detection: Jump Errors

<table>
<thead>
<tr>
<th></th>
<th>$E_{0 \rightarrow 1}$</th>
<th>$E_{0 \rightarrow 2}$</th>
<th>$E_{1 \rightarrow 0}$</th>
<th>$E_{1 \rightarrow 2}$</th>
<th>$E_{2 \rightarrow 0}$</th>
<th>$E_{2 \rightarrow 1}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF</td>
<td>0.47</td>
<td>0</td>
<td>2.37</td>
<td>0.13</td>
<td>0.96</td>
<td>27.93</td>
</tr>
<tr>
<td>WW</td>
<td>0.55</td>
<td>0</td>
<td>0.65</td>
<td>0.12</td>
<td>0</td>
<td>31.63</td>
</tr>
<tr>
<td>WWR</td>
<td>0.55</td>
<td>0</td>
<td>0.65</td>
<td>0.12</td>
<td>0</td>
<td>31.63</td>
</tr>
<tr>
<td>BF</td>
<td>1.62</td>
<td>0</td>
<td>2.25</td>
<td>2.95</td>
<td>1.39</td>
<td>28.44</td>
</tr>
</tbody>
</table>

### Table 2. Error Rates (in Percentage) II of Pitch Detection: Gross and Fine Errors

<table>
<thead>
<tr>
<th></th>
<th>$E_{\text{Jump}}$</th>
<th>$E_{\text{Gross}}$</th>
<th>$E_{\text{Dom}}$</th>
<th>$E_{\text{Dom}}$</th>
<th>$E_{\text{Fine}}$</th>
<th>$E_{\text{Dom}}$</th>
<th>$E_{\text{Dom}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF</td>
<td>31.94</td>
<td>2.21</td>
<td>1.27</td>
<td>4.48</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WW</td>
<td>32.95</td>
<td>5.24</td>
<td>0.52</td>
<td>7.28</td>
<td>0.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td>WWR</td>
<td>32.95</td>
<td>1.86</td>
<td>0.52</td>
<td>3.91</td>
<td>0.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BF</td>
<td>36.65</td>
<td>5.26</td>
<td>1.44</td>
<td>4.87</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Table 1 and Table 2, we compare results from four algorithms: with proposed final results as PF, proposed before morphological processing as BM, in which we can see the validity of piecewise continuity processing, Wu’s results directly obtained from Wu’s program as WW and Wu’s revised results, in which we manually adapt the obvious mismatched points in the end of pitch contour in Wu’s results, as WWR, to get one better results than Wu’s. By the way, there is no similar midterm results like BM form Wu’s method to compare.

In Table 1, we can see that the main portion of jump errors $E_{2 \rightarrow 1}$ the result we proposed is better than Wu’s algorithm.

In Table 2, take note of that the $E_{\text{Jump}}$ target, the most important criterion, is better than Wu’s method, $E_{\text{Gross}}$, $E_{\text{Dom}}$, are better than Wu’s and slightly worse than Wu’s revised results, but comparable, and $E_{\text{Fine}}$, $E_{\text{Dom}}$, the fine errors are about double of the results Wu proposed, but acceptable for not more than two delay points about 0.13ms.

Therefore, we can summarize that the results we proposed are better than Wu’s and slightly worse than Wu’s Revised, but it runs faster than Wu’s HMM tracking algorithm and needs smaller memory. For 1.4 seconds speech, it takes 7 seconds and 57MB memory spaces in our algorithm, whereas respectively 30 seconds and 380MB memory spaces in Wu’s algorithm. The experiment was
performed under Inter(R) Xeon(TM) CPU 2.80 G and 1.00GB of Memory. Particularly, Wu’s algorithm consumes CPU time and Memory space mainly on pitch tracking phase, whereas time-frequency front end phase in our algorithm.

In Table 1 and Table 2, we also list the results of our algorithm before morphological processing, and we can recognize the effect of it.

4 Conclusion and Expectation

In this paper, we have proposed a multi-pitch detection algorithm for co-channel speech. A new peak information piecewise integrating method is proposed to gain more sufficient pitch candidates. The morphological processing with feedback verification eliminates the errors from individual points and discontinuous pitch contours and gains supplement from trustful pitch contours. The results are comparable with Wu’s revised, and the algorithm executes more efficiently.

In the future work, we will integrate our multi-pitch detection algorithm into CASA system. Our new peak information piecewise integrating method has an easy connection with the fragment in CASA, and also adaptive morphological method can be used to the continuity detection in fragment both in time and frequency domain. Another important aspect, some pitch detection algorithms gather the period information from average magnitude difference function [8], which has faster realization than correlation function in time consuming. We consider adopting the function and therefore our algorithm will has fast realization in gathering period information and tracking pitch. In addition, these are all parallel methods and easy to implement in hardware.

References