MULTI-PITCH STREAMING OF INTERWOVEN STREAMS

Chih-Yi Kuan¹, Li Su², Yu-Hao Chin², Jia-Ching Wang¹

¹Department of Computer Science and Information Engineering, National Central University, Taiwan
²Institute of Information Science, Academia Sinica, Taiwan

ABSTRACT

In this paper, we discuss the multipitch streaming (MPS) problem for a multi-source audio signal having interweaving pitch contours. We propose two approaches to tackle this challenge, one relates to a feature extracted from the energy levels distributed in multi-channel recordings for better characterization of the source, and the other uses particle swarm optimization (PSO) to enlarge the search space and alleviate the initialization problem in constrained clustering of the features representing different sources. Experiments on music and speech samples having highly interweaving pitch contours are presented to assess its effectiveness.

Index Terms— Multipitch streaming, automatic music transcription, particle swarm optimization, multi-channel

1. INTRODUCTION

In multi-source audio signal processing, recognizing the time-varying behavior of the fundamental frequencies (F0s) of every source is a challenging task.¹ In the literature of automatic music transcription (AMT) [1], this problem was solved at different levels, such as multipitch estimation (MPE) in frame-level, note tracking (NT) in note-level, and multipitch streaming (MPS) in stream-level [2]. MPS is particularly related to a number of music and speech processing problems, including melody tracking, instrument identification [3], source separation [4], speech recognition [5, 6, 7], and prosody analysis [8], to name but a few.

In this paper, we study the problem of MPS, which aims to determine the pitch contour of every source based on a known MPE result, i.e., frame-level pitch estimations. In other words, the main difference between MPE and MPS is that MPE returns activated pitches at every time instance, while MPS returns the pitches and their corresponding source labels. In comparison to MPE, MPS is complicated by the characteristics of sources, such as the timbre of different instruments. Therefore, challenges of MPS include silence, non-pitched sounds, abrupt frequency changes in a stream [2], error propagation from the MPE result, etc. However, extra complexity of this problem is introduced by an essential but under discussed issue, the interwoven streams.

Fig. 1 illustrates what interwoven streams are. The first example shown in the left of Fig. 1 is one of Bach’s chorales, which has four streams, and each of them does not cross one another. This is different from the second example, a three-part country music shown in the right of Fig. 1, where the pitch contours of vocal and guitar cross each other three times in a 15-second recording. This so-called voice crossing phenomenon is avoided in Bach’s composition [9, 10] and often prohibited in pedagogical composition in music theory, but is frequently seen in pop or folk music.² However, in previous studies, discussion on this issue is quite limited.

The lack of discussion motivates us to investigate how interwoven streams affect the performance of an MPS algorithm, and how to improve the algorithm in recognizing interwoven streams. On the basis of the constrained clustering approach [2], we propose two enhanced schemes: first, since most of the recordings are in a dual-channel format, we propose a dual-channel feature to better discriminate different sources by using directional information; second, we introduce particle swarm optimization (PSO) [13, 14] into the constrained clustering method, in order to address the issue of high sensitivity to initialization of cluster centers. Evaluation is then performed on a dataset having highly interwoven streams, and the Bach10 dataset having few interwoven streams, and the results are compared. The proposed method is shown to be useful for interwoven streams, and particularly useful in dual-channel recordings in a real-world environment, where the location of a source affects an energy ratio in both channels.

2. RELATED WORK

An MPS system under our discussion has three inputs: an audio signal, pitch labels (i.e., the MPE result), and the number of sources. The system contains mainly two parts: 1) feature extraction, and 2) streaming. For the feature extraction, most of the previous works considered pitch-informed, spectral-based features, such as the activation matrix from the probabilistic latent component analysis

¹In this paper, the term “multi-source signal” refers to a signal generated by multiple sources including instruments or speakers, with each source assumed to be mono-phonnic; i.e., a source only generates one pitch at a time. With this assumption, a multi-source signal is polyphonnic, while a polyphonnic signal is not necessarily multi-source, such as piano solo music, which is beyond the scope of our discussion.

²In this paper, the terms “voice,” “stream,” “source”, and “pitch contour” are used interchangeably.
(PLCA) [15, 16, 17], correlogram [5], cochleagram [6], and the uniform discrete cepstrum (UDC) [2]. A main challenge in feature extraction is that the harmonic peaks of the spectra of different sources tend to overlap with each other, making it difficult to extract a feature representing a clean source from a multi-source signal.

For the streaming algorithm, previous studies have adopted a number of supervised or semi-supervised algorithms, such as the hidden Markov model (HMM) [15, 18], deep neural networks [6], discriminate PLCA [19], and hidden-Markov random fields (HMRF) [17]. Besides, there are also unsupervised streaming algorithms being proposed, such as spectral clustering [17] and constrained clustering [2]. In comparison to a supervised algorithm, an unsupervised algorithm does not require a multi-track dataset for training, and becomes more favorable for the MPS of general types of music. An unsupervised algorithm is based on an assumption in which similar features belong to the same stream (i.e., cluster), and resulting streams are musically meaningful (i.e., constraints of monophonic activation of a source, continuity of a pitch contour, etc.).

We base our work on the system proposed in [2], which uses the UDC as feature representation and a constrained clustering scheme in determining the streams. Detailed description of our enhanced schemes is given in the next two sections.

3. FEATURE EXTRACTION

3.1. The UDC feature

Cepstral features have been widely used in timbre classification [20, 21]. A cepstrum is defined as the inverse Fourier transform (IFT) of a log-scale spectrum. To extract the feature of one source in a multi-source signal, the IFT is performed only in the region where energy of a source is distributed in the spectrum of the multi-source signal; this is the basic idea of the UDC. More specifically, a log-amplitude spectrum of the mixture signal \( \mathbf{a} = \{a(i)\}_{i=1}^{N} \) is located in the frequency bins \( \mathbf{f} = \{f(i)\}_{i=1}^{N} \), \( N \in \mathbb{N} \), and \( \hat{\mathbf{a}} = \{\hat{a}(i)\}_{i=1}^{N} \) and \( \hat{\mathbf{f}} = \{\hat{f}(i)\}_{i=1}^{L} \) represent a potential subset of the spectrum and the frequency bins that solely belong to the source of interest. This subset is characterized by the first 50 harmonics of the source from its given pitch obtained from an MPE algorithm. The UDC is obtained by computing the cepstrum of this subset:

\[
\mathbf{c} = \begin{bmatrix}
\sqrt{2}\cos(2\pi f_1) & \cdots & \sqrt{2}\cos(2\pi (p-1)f_1) \\
\sqrt{2}\cos(2\pi f_2) & \cdots & \sqrt{2}\cos(2\pi (p-1)f_2) \\
\cdots & \cdots & \cdots \\
\sqrt{2}\cos(2\pi f_L) & \cdots & \sqrt{2}\cos(2\pi (p-1)f_L)
\end{bmatrix} \begin{bmatrix} a_1^* \\
a_2^* \\
\vdots \\
a_p^* \\
\end{bmatrix}.
\]

(1)

where \( p \) is the order of the cepstral representation. In this paper, we set \( L = 22 \). More implementation details of the UDC can be found in [2] and its source codes provided online (http://www.ece.rochester.edu/~zduan/resource/Resources.html).

3.2. The level-ratio feature

Most of the music recordings are in the multi-channel format. For example, in the scenario of a live concert, sound sources are typically located at different places on the stage, and are commonly recorded by two (or more) microphones. In this case, the energy of each source distributes differently in the two (or more) recording channels, which provides extra information for identifying the source. Therefore, we design features for the MPS problem by using channel information. For a \( C \)-channel recording, \( \mathbf{b}^{(c)} = \{b^{(c)}(i)\}_{i=1}^{N} \) is a magnitude spectrum, where \( 1 \leq c \leq C \). A normalized amplitude distribution for a pitch detection \( g \) as \( I^{(c)} = \frac{p^{(c)}(i)}{\sum_{c=1}^{C} b^{(c)}(i)\delta_{g,m}} \), where \( m \) is a constant, and \( g \) is the frequency bin corresponding to the pitch label. In this work, we set \( m = 4 \). A level-ratio feature derived from the above is represented as

\[
\mathbf{d} = \begin{bmatrix} I^{(1)} \cdots I^{(C)} \end{bmatrix},
\]

(2)

where the dimension is \( C(2m+1) \). Next, a fusion feature concatenating a scaled timbre feature \( \mathbf{c} = \mathbf{c}/\sigma_c \) and a scaled level-ratio feature \( \mathbf{d}/\sigma_d \) is represented as:

\[
\mathbf{u} = [\mathbf{c}, \mathbf{d}],
\]

(3)

where \( \sigma_c \) and \( \sigma_d \) are the standard deviation for feature dimensions of \( \mathbf{c} \) and \( \mathbf{d} \), respectively.

4. CLUSTERING

4.1. Constrained clustering

For an input signal having \( K \) sources, a tuple \( (t, f, k) \) denotes a detected pitch activation of a \( k \)-th source at a time \( t \) and a pitch \( f \), where \( t \) and \( f \) are both positive real numbers, and \( k = 1, 2, \ldots, K \). \( D \) represents the number of all pitch activations in one music piece. To define an one-to-one mapping \( \mathcal{S} : (t, f)_{\text{activated}} \rightarrow \{1, 2, \ldots, D\} \), the set of all pitch activations are indexed as \( \mathbf{p} = \{p(d)\}_{d=1}^{D} \), where \( p(d) \) (or \( p(t, f) \)) represents the stream label of the \( d \)-th pitch activation at time \( t \) and pitch \( f \) (i.e., \( p(d) = k \)). In this paper, \( p \) is referred to as a stream partition, \( S_k = \{i \in \mathcal{S} | p(i) = k\} \) is defined as the set of all indices of the \( k \)-th stream, \( \mathbf{u}_d \) represents the fusion feature (i.e., Equation (3)) of the \( d \)-th pitch activation, and \( \mathbf{c}_k \) represents the center of the \( k \)-th source. Then, the multipitch streaming problem is solved by minimizing the following objective function:

\[
f(p) = \sum_{k=1}^{K} \sum_{d \in S_k} \|\mathbf{u}_d - \mathbf{c}_k\|^2.
\]

(4)

The minimization can be solved by using the well-known \( K \)-means algorithm. Moreover, every \( (t, f) \) in \( S_k \) should satisfy two constraints derived from the domain knowledge of music [2]:

1. **Must-link**: different sources rarely have the same pitch at the same time; therefore, a pitch activation different from another pitch activation in a neighboring frame by less than one semitone should be assigned to the same source as the other pitch activation.

2. **Cannot-link**: one source does not perform two pitches at the same time. Namely, \( p(t_1, f_1) \) and \( p(t_2, f_2) \) should belong to different clusters if \( t_1 \) is close to \( t_2 \), while \( p_1 \) and \( p_2 \) are far apart. Two pitches occurring at the same time must satisfy this relation. The set of all partitions satisfying this cannot-link condition is defined as \( \Gamma \).

After checking the above two constraints against all pitches, a refined list of constrained pitch candidates is obtained.

4.2. PSO-based constrained clustering

The \( K \)-means algorithm suffers a possible convergence to local minima due to its sensitivity to initialization of cluster centers. Previous studies suggested pitch order initialization [2], but this method is not applicable to the case of interwoven streams. Therefore, a
Algorithm 1 PSO-based constrained clustering.

**[INPUT]** Feature set of all pitch activations \((u_i)_{i=1}^D\), number of particles \(\gamma\), maximum number of iteration \(\eta\)

**[OUTPUT]** Optimal partition \(p \in \mathbb{R}^D\)

for \(j = 1, \ldots, \gamma\) do

    Initialize \(q^{(1,j)}(d), v^{(1,j)}(d), p_{gb}^{(j)}(d) \sim U(0.5, K + 0.5), d = 1, \ldots, D\)

end for

Initialize \(p_{gb}(d) \sim U(0/5, K + 0.5), d = 1, \ldots, D\)

while \(i < \eta\) do

    for \(j = 1, \ldots, \gamma\) do

        \(p^{(i,j)}(d) \leftarrow \arg\min_{p : p \in \Gamma} \|q^{(i,j)} - p\|^2\)

        \(p^{(i,j)}(d) \leftarrow \text{FindNewPartition}(p^{(i,j)})\) \([2]\]

        if \(g(p^{(i,j)}) < g(p_{gb})\) and \(f(p^{(i,j)}) < f(p_{gb})\) then

            \(p_{gb} \leftarrow p^{(i,j)}\)

        end if

    end for

    \(v^{(i,j)}(d) \leftarrow v^{(i,j)}(d) + b_1\phi_1(p_{gb} - p^{(i,j)}) + b_2\phi_2(p_{gb} - p^{(i,j)})\)

    \(q^{(i+1,j)} \leftarrow p^{(i,j)} + v^{(i,j)}\)

end while

\(p \leftarrow p_{gb}\)

\[\text{Algorithm 1 PSO-based constrained clustering.}\]

1. \(g(p)\) outputs a total number of must-link violations. Here, we apply an additional process used in [2], called swap operation, to trace all pitches and change their stream labels to the respective clusters, and to check if there is a stream label better satisfies the constraints and minimizes the total within-source feature distance at the same time. If the swap operation finds a better set of stream labels, we update this result to that particle. Its purpose is to find better must-links.

2. \(f(p)\) outputs the sum of the intra-source feature distance represented as (4).

A locally optimal candidate, \(p_{gb}\), has the best fitness function value in all iteration steps of that candidate, and a globally optimal candidate, \(p_{gb}\), has the one with the best fitness function value among all candidates. The direction of optimization (i.e., the velocity of the swarm) of each particle is therefore defined as follows for the \(i\)-th iteration and the \(j\)-th particle:

\[v^{(i,j)} \leftarrow v^{(i,j)} + b_1\phi_1(p_{gb} - p^{(i,j)}) + b_2\phi_2(p_{gb} - p^{(i,j)})\],

where \(b_1\) and \(b_2\) are two constants, and \(\phi_1\) and \(\phi_2\) are two Gaussian random variables with a zero mean and unit standard deviation. In this paper, \(b_1 = b_2 = 0.2\), and we choose \(\gamma = 7\) particles, five of which are initialized randomly, one by a simple \(K\)-means result, and one by pitch ordering. The maximal number of iteration \(\eta\) is set to 5. The proposed algorithm is detailed in Algorithm 1.

4.3. Post-processing

We apply two post-processing steps. First, we examine the pitches against the must-link constraints, and group every 10 frame-level pitches (i.e., every 0.3s) satisfying the must-link relation into a non-overlapping segment. We make a vote over the stream labels, and assign the 10 pitches to the voting result. This is a smoothing process which is usually required in refining the frame-level estimation [22]. Then, if there are still two pitches violating the cannot-link constraint, we take the pitch farther from the cluster center as a false positive and remove it.

5. EXPERIMENT AND RESULT

5.1. Data

We collect a dataset consisting of five clips by selecting 2-3 tracks having highly interwoven streams from several music and speech datasets (see Table 1). To quantify how much the streams are interwoven in a clip, we consider the number of frames whose stream labels in pitch order is not equal to the average pitch order of the full clip; a higher number indicates more interwoven streams. Then, we define a term named interwoven rate (IR), which is the ratio between this number and the total frame number in a clip. The IR of each clip in the “Interwoven” dataset is listed in Table 1.

We also consider the Bach10 dataset, which has ten of Bach’s chorales played by four different instruments (violin, clarinet, saxophone, bassoon) [11, 2]. The four streams in the dataset interweave rarely. The average IR of the Bach10 dataset is only 4.761%.

All recordings are sampled at 44.1 kHz. We use a Hamming window of size 46.4 ms and a hop size of 30 ms for feature extraction. To simulate a room performance recorded by two microphones, all songs are remixed using the open-source room acoustics simulator Roomsim [23]. For PSO-based clustering, each setting is repeated for 3 rounds in order to get stable results. We evaluate the accuracy of MPE using the average frame-level F-score, which counts the average F-score over all clips and rounds.
Table 1. The dataset with interwoven streams. The first three clips are selected from the MedleyDB dataset [12], the fourth is from the MIREX dataset, and the final from the CHAIN-Corpus [24].

<table>
<thead>
<tr>
<th>Filename</th>
<th>Parts (selected)</th>
<th>IR (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MusicDelta_ChineseJiangNan</td>
<td>erhu, guzheng, ruan</td>
<td>12.92</td>
</tr>
<tr>
<td>MusicDelta_ChineseHenan</td>
<td>erhu, guzheng, liuqin</td>
<td>19.18</td>
</tr>
<tr>
<td>MusicDelta_Country1</td>
<td>vocal, guitar, bass</td>
<td>56.22</td>
</tr>
<tr>
<td>MIREX Multi-F0 training data</td>
<td>clarinet, horn</td>
<td>30.96</td>
</tr>
<tr>
<td>CHAIN-Corpus</td>
<td>two male talkers</td>
<td>10.45</td>
</tr>
</tbody>
</table>

Table 2. Comparison of average F scores (in %) over all clips and rounds for various features and clustering algorithms between the Bach10 and the Interwoven datasets using ground-truth MPE result.

<table>
<thead>
<tr>
<th></th>
<th>K-means</th>
<th>CC</th>
<th>PSOCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bach10</td>
<td>UDC</td>
<td>70.63</td>
<td>94.24</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>71.82</td>
<td>85.96</td>
</tr>
<tr>
<td></td>
<td>UDC+LR</td>
<td>75.21</td>
<td>92.27</td>
</tr>
<tr>
<td>Interwoven</td>
<td>UDC</td>
<td>59.32</td>
<td>67.63</td>
</tr>
<tr>
<td></td>
<td>LR</td>
<td>72.77</td>
<td>73.55</td>
</tr>
<tr>
<td></td>
<td>UDC+LR</td>
<td>67.06</td>
<td>71.48</td>
</tr>
</tbody>
</table>

5.2. Result

We perform two MPS experiments. The first one takes the ground truth pitch annotation as the input, i.e., the result in the MPE stage is assumed to be perfect. Results are shown in Table 2. We compare three different kinds of features, including the UDC, the proposed level-ratio (LR) and the fusion of both (UDC+LR). We also compare three clustering methods, namely the K-means, constrained clustering (CC) [2] and the proposed PSO-based constrained clustering (PSOCC) method. Table 2 shows that although LR does not improve the F-scores with respect to the UDC in the Bach10 dataset (except for the case of K-means), it outperforms the UDC in the Interwoven dataset for all cases. In general, the performance of UDC+LR is better than the UDC while worse than LR. For the clustering method, we found that both CC and PSOCC outperform K-means, and the performance of PSOCC and CC are similar.

The results can be explained by two reasons; first is that the UDC is pitch-dependent while LR is not, and second is that imposing the constraints actually prefers not to interweave streams having less distinguishable features when interweaving of streams actually occurs. This is why the pitch-dependent UDC+LR outperforms LR in the pitch-ordered Bach10 dataset, while LR outperformed the UDC in the Interwoven dataset. The result of UDC+LR having an intermediate performance between the UDC and LR again explains that when the notes are highly interwoven, the MPS task is solved with a trade-off between feature similarity and the pitch-linking constraints. Comparing the clustering method, we found that randomly-initialized PSOCC performs slightly better than pitch-initialized CC in both datasets. Although the improvement is marginal, the effectiveness of this clustering scheme is demonstrated under the same objective function and constraints as CC. In addition, this method can be further improved in the future by using different number of particles and velocity parameters in the optimization.

Fig. 2 illustrates the MPS results of a selected clip using baseline (UDC+CC) [2] and the proposed method ({UDC+LR}+PSOCC). The major difference between the two methods can be illustrated with two interwoven streams, vocal and guitar. The baseline method prefers to keep the pitch order of the streams, namely, keeping the guitar stream higher than the vocal stream. Conversely, the proposed method has more flexibility in assigning stream labels by employing the information of LR, which explains why the proposed method works effectively in those “voice-crossing” events in the clip.

The second experiment uses the results of real-world MPE algorithms as the input. We consider two MPE algorithms; one is proposed by Duan et al. [11], and the other is the combined frequency and periodicity (CFP) method [22]. The average F-scores of MPE on Bach10 dataset are 72.25% for [11] and 76.78% for CFP. Since both methods have an average F-scores lower than 40% on the Interwoven dataset, the error of MPE dominates the result. Therefore, we consider the comparison of MPS on the Interwoven dataset unreliable, and report only the results on Bach10 dataset, see Table 3. One can observe that, for an MPE result having a lower F-score, the proposed method performs better. This is mainly because when frame-level pitch activations are noisy, missing or false alarm pitches can virtually create a scenario similar to interweaving streams.

Table 3. Comparison of average F scores (in %) between the baseline and proposed experiment schemes using real-world MPE.

<table>
<thead>
<tr>
<th></th>
<th>Duan</th>
<th>CFP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (UDC+CC)</td>
<td>55.99</td>
<td>68.89</td>
</tr>
<tr>
<td>Proposed ({UDC+LR}+PSOCC)</td>
<td>66.21</td>
<td>68.45</td>
</tr>
</tbody>
</table>

Fig. 2. Illustration of the MPS result of MusicDelta_Country1. Top: Ground truth. Middle: baseline. Bottom: proposed.

6. CONCLUSION

To solve the MPS problem of a multi-source signal having interwoven streams, we have presented the level-ratio feature using directional information, and also the PSO-based constrained clustering method providing extra freedom in selecting better solution. Evaluation on recordings having either a high or low interwoven rate gives a promising result, while at the same time reveals a trade-off between feature similarity and the pitch-linking constraint. The performance gaps between interwoven streams and non-interwoven streams, and between ground-truth MPE input and real-world MPE input also suggest more room for improvement in handling this problem. Possible future work can target at how to better exploit timbre and directional information, how to incorporate musical knowledge into a clustering scheme, and how to control the computational complexity.
7. REFERENCES


