AUDIO-VISUAL SOURCE ASSOCIATION FOR STRING ENSEMBLES THROUGH MULTI-MODAL VIBRATO ANALYSIS

Bochen Li
University of Rochester
bli23@ur.rochester.edu

Chenliang Xu
University of Rochester
chenliang.xu@rochester.edu

Zhiyao Duan
University of Rochester
zhiyao.duan@rochester.edu

ABSTRACT

With the proliferation of video content of musical performances, audio-visual analysis becomes an emerging topic in music information retrieval. Associating the audio and visual aspects of the same source, or audio-visual source association, is a fundamental problem for audio-visual analysis of polyphonic musical performances. In this paper, we propose an approach to solve this problem for string ensemble performances by analyzing the vibrato patterns. On the audio side, we extract the pitch trajectories of vibrato notes of each string player in a score-informed fashion. On the video side, we track the left hand of string players and capture their fine-grained left-hand vibration due to vibrato. We find a high correlation between the pitch fluctuation and the hand vibration for vibrato notes, and use this correlation to associate the audio and visual aspects of the same players. This work is a complementary extension to our previous work on source association for string ensembles based on bowing motion analysis. Experiments on 19 pieces of chamber musical performances with at most one non-string instrument show more accurate and robust association performance than our previous method.

1. INTRODUCTION

Music performance, by its nature, is a multi-modal art form. Both the audio and visual aspects of music performance play an important role in engaging the audience in live concerts. Indeed, watching musical performances is often more entertaining than purely listening, since the players' motion visually interprets the melody and attracts the audiences. No wonder, for recorded music, the popularization of video streaming services such as YouTube, has also greatly shifted the dominating recording format from purely audio to video [1]. Despite the rich multi-modality of music, current Music Information Retrieval (MIR) research still mainly focuses on the audio modality of music, ignoring the visual aspects.

Joint analysis of audio and visual aspects of musical performances can significantly boost the state of the art of MIR research. Challenging tasks such as instrument recognition and activity detection in musical performances can be simplified by analyzing the visual aspects [2]. Joint analysis can also augment existing audio-based approaches to chord recognition [3] and onset detection [4] by tracking performers' hands and fingers . These benefits are especially prominent for analyses of polyphonic music, because the visual activity of each player is usually directly observable, whereas the polyphony makes it difficult to unambiguously associate components of the audio to each player [5]. Audio-visual analysis also opens up new frontiers for several existing and new MIR problems such as the analysis of sound articulation and performance expressiveness [6], fingering investigation [7], conductor following [8], source localization [9], etc.

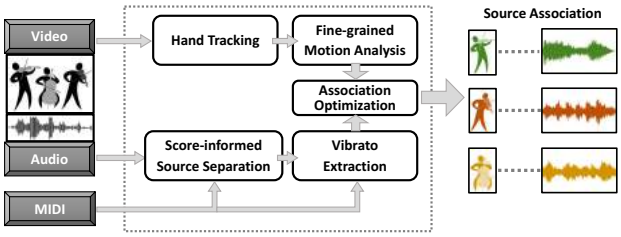

Figure 1. Framework of the proposed system. Separated sources are associated with players in the video, encoded with different colors.

A basic problem in audio-visual analysis of polyphonic music performances is to figure out the association between the audio and visual aspects of each sound source. This task was named *audio-visual source association* in our prior work [10]. It is an important problem because it is required if one wants to incorporate visual analysis in parsing the musical scene. On the application side, such association links audio editing (e.g., remixing) and video editing (e.g., recomposing) of sound sources. It enables cameras to automatically follow and take close-up shots of the instrumental part that is playing the main melody. It can also enable novel multimedia interactions such as music video streaming services that allow users to click on a player in the video and isolate/enhance the corresponding audio components.

In this paper, we propose an approach to solve the audio-visual source association problem for string ensemble performances through multi-modal analysis of vibrato notes. Specifically, we show that the fine-grained motion of the fingering hand has a strong correlation with the pitch fluctuation of vibrato notes, and we use this correlation to solve source association for string ensembles in a score-informed fashion. Fig. 1 shows the framework of the proposed approach. On the audio side, we first estimate the pitch

trajectory of each note of each instrument from the audio mixture in a score-informed fashion. Then, we detect vibrato notes by inspecting the periodicity of the pitch fluctuation. On the visual side, we first track the left hand of each string player, and then capture the fine motion of the hand using optical flow, resulting in a 1-D hand displacement curve along the principal motion direction. For each source-player pair, we define a matching score based on the cross correlation between the concatenated pitch trajectories of all vibrato notes and the concatenated displacement curve segments in the corresponding time ranges. Finally, the bijection between sources and players that maximizes the overall matching score is output as the association result.

This is a complementary extension to our previous work on source association [11] that uses the correlation between bowing strokes and score note onsets. When vibrato is used, the proposed approach offers a more accurate and robust solution, as the correlation between the audio and visual aspects is now performed through the entire process of vibrato notes instead of only the onsets as in our previous work. This advantage is shown by the better association performance on 19 video recordings of chamber musical performances that use at most one non-string instruments. In the following, we first describe related work on audio-visual association in Section 2. We then describe the proposed approach in Section 3. We evaluate the system in Section 4 and conclude the paper in Section 5.

2. RELATED WORK

There has been some previous work on audio-visual source association for speech and music signals. Hershey and Movellan [9] first explored the audio-visual synchrony to locate speakers in the video. In [12], a region tracking algorithm was applied to segment the video into temporal evolution of semantic objects to be correlated with audio energy. Casanovas *et al.* [13] applied non-linear diffusion to extract the pixels whose motion is most consistent with changes of audio energy. In [14], source localization results are incorporated into graph-cut based image segmentation to extract the face region of the speaker. Other methods include time delayed neural networks [15], probabilistic multi-modal generative models [16], and canonical correlation analysis (CCA) [17, 18]. All of the abovementioned systems, however, assume that there is at most one active sound source at a time.

When multiple sources are active simultaneously, cross-modal association can be utilized to isolate sounds that correspond to each of the localized visual features. Barzelay and Schechner [19, 20] detected drastic changes (i.e., onsets of events) in audio and video and then used their coincidence to associate audio-visual components belonging to the same source of harmonic sounds. Sigg *et al.* [21] reformulated CCA by incorporating non-negativity and sparsity constraints on the coefficients of the projection directions, to locate sound sources in movies and separate them by filtering. In [22], audio and video modalities were decomposed into relevant structures using redundant representations for source localization. Segments where only one source is active are used to learn a timbre model for the separation of the source. These methods, however, only deal with mixtures with at most two active sources. The difficulty of source association increases dramatically as the source number increases.

The audio-visual source association problem for chamber music ensembles is more challenging since several sound sources are active almost all the time. In [11], we proposed the first approach to audio-visual source association for string ensembles in a score-informed fashion where there are up to five simultaneously active sources (e.g., string quintets). This approach analyzes the large-scale motion from bowing strokes of string-instrument players and correlates it with note onsets in the score track. The assumptions are that most note onsets correspond to the beginning of bowing strokes, and that different instrumental parts show different rhythmic patterns. When these assumptions are invalid, for example, when multiple notes are played within a single bow stroke (i.e., legato bowing) or when different parts show a similar rhythmic pattern, this approach becomes less robust.

The analysis of the fingering hand motion due to vibrato articulations and their correlation to pitch fluctuations in this paper is a complementary extension of [11]. It extends the large-scale motion analysis of the bowing hand to the fine-grained motion analysis of the fingering hand. It also extends the audio-visual correlation from the discrete note onsets to the entire processes of vibrato notes. As vibrato articulations are common practices for string instrument players (at least for skilled players), and the vibrato parameters (e.g., rate, depth, and phase) of different players are quite different, this vibrato cue provides robust and complementary information for audio-visual source association for string ensembles.

3. METHOD

3.1 Audio Analysis

The main goal of audio analysis is to detect and analyze vibrato notes in a score-informed fashion. There has been much work on vibrato analysis for both singing voices [23] and instruments [24], but most of them only deal with monophonic music. For methods that do deal with polyphonic music, they only analyze vibrato in the melody with the presence of background music [25]. The more challenging problem of vibrato detection and analysis for multiple simultaneous sources is, however, rarely explored. In this paper, we first separate sound sources using a score-informed source separation approach [26] and then perform pitch analysis on each source to extract and analyze vibrato notes.

3.1.1 Score-informed Source Separation

Musical sound sources are often strongly correlated in time and frequency, and without additional knowledge the separation of the sources is very difficult. When the musical score is available, as for many Western classical music pieces, the separation can be simplified by leveraging the score information. In recent years, many different models

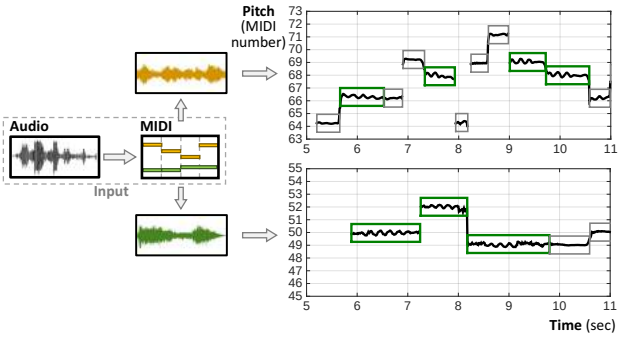

Figure 2. Process of audio analysis. Detected vibrato notes are marked with green rectangles in the pitch trajectories.

have been proposed to explore this idea, including Non-negative Matrix Factorization (NMF) [27–29], Gaussian Mixture Models (GMM) [30], and adaptive instrument models [31]. In this paper, we adopt the pitch-based score-informed source separation approach of [26] and modify it for our task.

This approach consists of three main steps: audio-score alignment, pitch refinement, and source separation. The task of audio-score alignment is to guarantee the temporal synchronization between score events and audio articulations. We apply the Dynamic Time Warping (DTW) framework described in [11] instead of the online particle filtering approach in [26] for more accurate and robust alignment results.

The pitch refinement and source separation steps remain the same as those described in [26]: Pitches that are actually played in the music are first estimated within two semitones around the quantized score-notated pitches. Sound sources are then separated by harmonic masking of the pitches in each frame, where the soft masks take into account the harmonic indexes when distributing the mixture signals' energy to overlapping harmonics.

3.1.2 Vibrato Extraction

For each separated source, we apply the above-mentioned score-informed pitch refinement step again to estimate a robust pitch trajectory. We then segment the pitch trajectory into notes using the onset and offset information from the score. Vibrato notes are detected by checking the periodicity of their pitch trajectories. To do so, we borrow the idea of YIN pitch detection [32] to calculate the difference function of the pitch trajectory for each note i:

$$d_i(\tau) = \sum_{t=t_i^{on}}^{t_i^{off}-\tau} (F(t) - F(t+\tau))^2, \qquad (1)$$

where $F(t)$ is the estimated pitch trajectory of the source, t_i^{on} and t_i^{off} are the onset and offset frame indexes of the i-th note, and $\tau \in \{0, 1, \cdots, t_i^{off}\}$ is the time lag in the calculation. Then we calculate the normalized difference function $d_i'(\tau)$ using its cumulative mean as:

$$d_i'(\tau) = \begin{cases} 1 & \text{if } \tau = 0 \\ d_i(\tau)/[(1/\tau)\sum_{k=1}^{\tau} d(k)] & \text{otherwise} \end{cases}. \qquad (2)$$

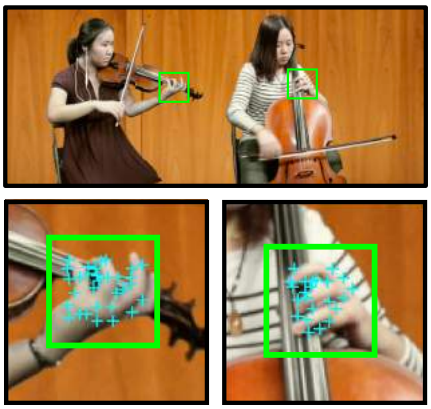

Figure 3. Hand tracking results. Feature points are marked by blue crosses, based on which the fingering hand is localized and bounded by a green bounding box.

As explained in [32], $\min_\tau d_i'(\tau)$ is a good indicator of whether periodicity exists or not, considering the energy of the signal. We use the threshold of 0.5, roughly tuned on one piece. If $\min_\tau d_i'(\tau)$ is smaller than the threshold, the note is considered to have vibrato. Fig. 2 highlights the pitch trajectories of the detected vibrato notes with green boxes.

It is noted that during the score-informed source separation process, the pitch trajectories of all sound sources have already been estimated. We do not, however, directly use this result, because we find that the pitch estimates are less accurate than those estimated from the separated sources. This suggests that the harmonic masking, although not perfect, does significantly reduce the interferences from other sources. It is also noted that the extracted vibrato pitch trajectories are still likely to contain errors, especially when multiple instruments play the same note. However, we will show that they are accurate enough for the audio-visual source association purposes.

3.2 Video Analysis

After retrieving the vibrato pitch trajectories from audio sources, we aim to extract the motion features from the visual aspect so that they can be correlated with the pitch fluctuations. A string instrument player mainly shows three kinds of motions when playing: bowing motion to articulate notes, fingering motion to control pitches, and the whole body motion to express musical intentions. Vibrato articulation is embodied as subtle motions on the left hand that rock forwards and backwards along the fingerboard. It changes the length and tension of the string. This rocking motions usually have a frequency ranging from 4 Hz to 7.5 Hz [33]. To capture this subtle motion, we first identify a region that covers the player's left hand, then estimate the rough position of the hand over time with a feature point-based tracker. Finally, we estimate the fine motion of the hand using optical flow analysis within the hand region.

3.2.1 Hand Tracking

Given the bounding box initialization of the hand position, we apply the Kanade-Lucas-Tomasi (KLT) tracker to track

the rough position of the left hand. This method was originally proposed in [34]. It extracts feature points using the minimum eigenvalue algorithm in a given region specified in the first video frame and then track these points throughout the video based on assumptions such as brightness constancy and spatial coherence. For our implementation (see Fig. 3), we manually specify a rectangle bounding box (70 $\times$ 70 pixels) to cover the left hand in the first frame and initialize 30 feature points within the bounding box using corner-point detector. We then track the points in following frames and relocate the center of the bounding box to the median x- and y- coordinates of the points. In order to prevent the loss of tracking due to abrupt motion or occlusions, we re-initialize the feature points every 20 frames, if some points disappear.

3.2.2 Fine-grained Motion Capture

To capture the fine motion of the left hand, we apply optical flow estimation [35] on the original video frames to calculate pixel-level motion velocities. For each frame, the average motion velocity over all pixels within the bounding box is calculated as $\mathbf{u}(t) = [u_x(t), u_y(t)]$. Notice that the pixel motions in the bounding box contain not only the player's hand vibrato motion but also his/her overall slower body motion while performing the music. This motion not only contains the hand vibration due to vibrato, but it also contains the overall slower body movement that is not associated with vibrato. In order to eliminate the body movement and obtain a clean description of the vibrato motion, we subtract $\mathbf{u}(t)$ by a moving average of itself, as

$$\mathbf{v}(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t), \tag{3}$$

where $\bar{\mathbf{u}}(t)$ is the moving average of $\mathbf{u}(t)$ by averaging over 10 frames. Fig. 4 displays one example with the original hand region in Fig. 4(a) and the optical flow result in Fig. 4(b). The refined motion velocity $\mathbf{v}(t)$ across all the frames is scattered in an x-y plane in Fig. 4(c).

(a) (b) (c)

Figure 4. Fine-grained left hand motion detection from optical flow. (a) Original video frame showing the hand region. (b) Optical flow results encoding motion velocities via colors. (c) Scatter plot of frame-wise average motion velocities of all frames.

We apply Principal Component Analysis (PCA) on $\mathbf{v}(t)$ across all frames to identify the principal direction of motion, which is ideally along the fingerboard. The refined motion velocity vectors are then projected to this principal direction to obtain a 1-D *motion velocity curve* $V(t)$ as

$$V(t) = \frac{\mathbf{v}(t)^T \tilde{\mathbf{v}}}{\|\tilde{\mathbf{v}}\|}, \tag{4}$$

where $\tilde{\mathbf{v}}$ is the eigenvector corresponding to the largest eigenvalue of the PCA of $\mathbf{v}(t)$. We then perform an integration of the motion velocity curve over time to calculate a *motion displacement curve* as

$$X(t) = \int_0^t V(\tau) d\tau. \tag{5}$$

This hand displacement curve corresponds to the fluctuation of the vibrating length of the string and hence the pitch fluctuation of the note. Fig. 5 plots the extracted motion displacement curve $X(t)$ of one player along with the pitch trajectory. We find that when vibrato is detected from audio (labeled in green rectangles), there are always observable fluctuations in the motion.

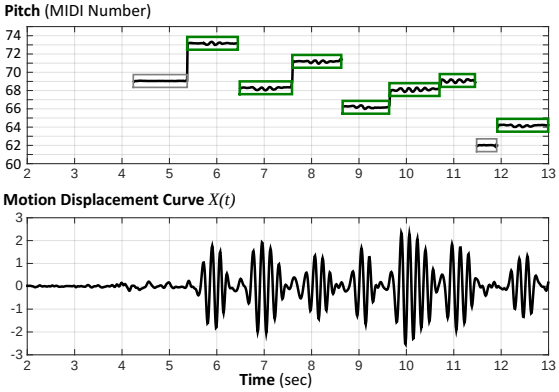

Figure 5. Pitch trajectory with detected vibrato notes marked by green rectangles from audio analysis and the left-hand motion curve of the corresponding player from video analysis.

3.3 Source-Player Association

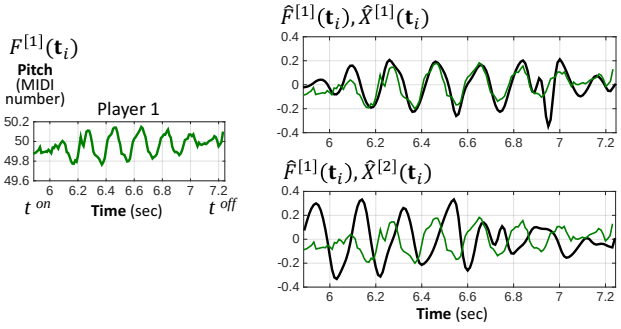

Figure 6. Pitch trajectory of one vibrato note (left), and the normalized left-hand displacement curves of the two players (right, black curves) overlaid with the normalized pitch trajectory (green curves). Note how well the pitch trajectory matches the displacement curve of Player 1 (correct association) but not of Player 2 (incorrect association).

The audio-visual association is a bijection between players and audio sources. This is obtained by matching the pitch trajectories and the visual displacement curves at the note level. Fig. 6 compares the pitch trajectory of one vibrato note with the displacement curves of two players at

the same time range. All of the curves have been normalized to have zero mean and unit Root-Mean-Square (RMS) value. Clearly, Player 1 is a better match. Quantitatively, we measure the similarity between the pitch trajectory of each note of the p-th source and the hand displacement curve of the q-th player in the corresponding time range by

$$m^{[p,q]}(i) = \exp\left\{ \frac{\left\langle \hat{F}^{[p]}(\mathbf{t}_i) \cdot \hat{X}^{[q]}(\mathbf{t}_i) \right\rangle}{\left\| \hat{F}^{[p]}(\mathbf{t}_i) \right\| \left\| \hat{X}^{[q]}(\mathbf{t}_i) \right\|} \right\}, \quad (6)$$

where $\mathbf{t}_i = [t_i^{on}, t_i^{off}]$ represents the time range of the i-th note from its onset t_i^{on} to offset t_i^{off}, $\hat{F}$ and $\hat{X}$ are the normalized pitch trajectories and hand displacement curves, respectively, and $\langle \cdot \rangle$ represents the vector inner product. Then the overall matching score between the p-th source and the q-th player can be defined as the sum of the similarities of all vibrato notes:

$$M^{[p,q]} = \sum_{i=1}^{N_{\text{vib}}^{[p]}} m^{[p,q]}(i), \quad (7)$$

where $N_{\text{vib}}^{[p]}$ is the number of detected vibrato notes of the p-th source.

For an ensemble with K players, there are $K!$ bijective associations (permutations). Let $\sigma(\cdot)$ be a permutation function, then the p-th audio source will be associated with the $\sigma(p)$-th player. For each permutation, we calculate an overall *association score* as the product of all pair-wise matching scores, i.e., $S_\sigma = \prod_{p=1}^{K} M^{[p,\sigma(p)]}$. We then rank all of the association candidates in a descending order and return the first candidate as the final association.

4. EXPERIMENTS

We evaluate the proposed approach on the University of Rochester Musical Performance (URMP) dataset [1] [36]. This dataset contains isolately recorded but synchronized audio-visual recordings of all instrumental parts of 44 classical ensemble pieces ranging from duets to quintets. For our experiments, we use the 11 string ensemble pieces and 8 other pieces that use only one non-string instrument, totaling 19 pieces including 5 duets, 4 trios, 7 quartets, and 3 quintets. Fig. 8 shows screen shots of the performance videos of all the 19 test pieces. Most pieces have a length of 2-3 minutes, providing enough cues for source association. Detailed information about these pieces is listed in Table 1.

Audio is sampled at 48 KHz, and processed with a frame length of 42.7 ms and a hop size of 10 ms for the Short-Time Fourier Transform (STFT). Video resolution is 1080P, and the frame rate is 29.97 frames per second. The audio and video streams have been carefully synchronized in the dataset. The estimated pitch trajectories are interpolated to make the first pitch estimate correspond to time at 0 ms. The estimated hand displacement curves are also interpolated to make them have the same sample rate as the pitch trajectories, i.e., 1 sample every 10 ms. For non-string instruments where the vibrato is not generated by hand motion, we do not track the hand nor capture the motion; Instead, we set the hand displacement curve of these players in Eq. (5) to all zeros.

[1] http://www.ece.rochester.edu/projects/air/projects/datasetproject.html

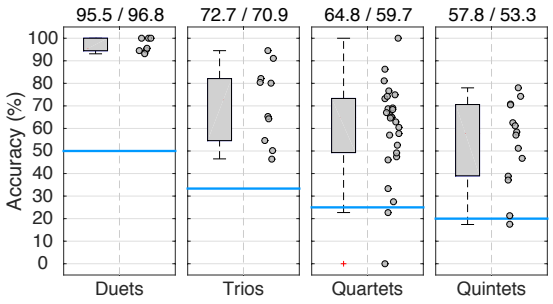

Figure 7. Box plots and scatter plots of note-level matching accuracy of string-instrumental sources of all pieces grouped by polyphony. Each data point represents the accuracy of one source. The blue lines mark the random guess accuracies. Numbers above the figures are median/mean values.

4.1 Note-level Matching Accuracy

As the matching score matrix $M^{[p,q]}$ used to find the best source association is the sum of the matching scores of all vibrato notes $m^{[p,q]}(i)$, we first evaluate how good the note-level matching scores are. To do so, we calculate a *note-level matching accuracy*, which is defined as the percentage of vibrato notes of each source that best matches the correct player according to the note-level matching score. Fig. 7 shows the boxplots and scatter plots of this measure of all sources (excluding the non-string instruments) of all pieces, grouped by their polyphony. Each point represents the accuracy of one source, and the numbers above show the median/mean values. The blue lines mark the accuracies based on random guess, e.g., the chance to randomly match a note to the correct player in a quartet is 25%. We find that this accuracy drops as the polyphony increase because of 1) the more incorrect candidates and 2) the less good quality of the extracted pitch trajectory of vibrato notes from high-polyphonic audio mixtures. Outliers in the figures are sources that only contain a few vibrato notes.

It is noted that a less good note-level matching accuracy does not necessarily lead to a bad piece-level source association, because the latter considers all the vibrato notes in a piece. In fact, as long as the note-level matching is somewhat accurate, the ensemble decision considering all of the vibrato notes at the piece-level is quite reliable, as illustrated in the following experiment.

4.2 Piece-level Source Association Results

We then evaluate the piece-level source association accuracy. The estimated association for each piece is considered correct if and only if all the sources are associated to the correct players. As the number of permutations is the factorial of the polyphony, this task becomes much

Figure 8. Screen shots of the 19 test pieces. Due to limited conditions of the recording facilities, all players faced to the same direction and some parts of certain players were not captured by the camera. These artifacts might have made the visual analysis of left-hand motion easier than in real-world performance recordings.

Metadata				Association Measures	
No.	Dataset Folder Name (with Instrument Types)	Piece Length (mm:ss)	Polyphony - (No. Permutations)	No. Correctly Associated Sources	Rank of Correct Association
1	01_Jupiter_vn_vc	01:03	2 - (2)	2	1
2	02_Sonata_vn_vn	00:46	2 - (2)	2	1
3	08_Spring_fl_vn	00:35	2 - (2)	2	1
4	09_Jesus_tpt_vn	03:19	2 - (2)	2	1
5	11_Maria_ob_vc	01:44	2 - (2)	2	1
6	12_Spring_vn_vn_vc	02:11	3 - (6)	3	1
7	13_Hark_vn_vn_va	00:47	3 - (6)	3	1
8	19_Pavane_cl_vn_vc	02:13	3 - (6)	1	2
9	20_Pavane_tpt_vn_vc	02:13	3 - (6)	3	1
10	24_Pirates_vn_vn_va_vc	00:50	4 - (24)	4	1
11	25_Pirates_vn_vn_va_sax	00:50	4 - (24)	4	1
12	26_King_vn_vn_va_vc	01:25	4 - (24)	4	1
13	27_King_vn_vn_va_sax	01:25	4 - (24)	2	1
14	32_Fugue_vn_vn_va_vc	02:54	4 - (24)	4	1
15	35_Rondeau_vn_vn_va_db	02:08	4 - (24)	4	1
16	36_Rondeau_vn_vn_va_vc	02:08	4 - (24)	4	1
17	38_Jerusalem_vn_vn_va_vc_db	01:59	5 - (120)	5	1
18	39_Jerusalem_vn_vn_va_sax_db	01:59	5 - (120)	5	1
19	44_K515_vn_vn_va_va_vc	03:45	5 - (120)	5	1

Table 1. Detailed information of the 19 test pieces together with the source association performance of the proposed approach. Abbreviations of instruments are: violin (Vn.), viola (Va.), cello (Vc.), double bass (D.B.), flute (Fl.), oboe (Ob.), clarinet (Cl.), saxphone (Sax.), and trumpet (Tp.).

more challenging for pieces with a high polyphony. In our experiments, the proposed approach successfully outputs the correct association for 18 of the 19 pieces, yielding a success rate of **94.7%**. This outperforms our previous method [11] whose accuracy was 89.5% on the same pieces. Among the 19 pieces, there are in total 65 individual sources, and 63 of them (96.9%) are associated with the correct player, while in our previous system only 58 sources are correct.

Table 1 lists the association results on all of the testing pieces. For each piece, we show the number of correctly associated sources and the rank of correct association among all permutations. The proposed approach only fails gently on one piece (No. 8), where the correct association is ranked the second. This failure case is a trio containing a clarinet while the two string instruments are always playing plucking notes, for which vibrato is not generally used.

The main reason that the proposed approach outperforms our previous system [11] on this dataset, as we stated in the introduction, is that the vibrato analysis provides more reliable cues for association than the note onset analysis in our previous system. Vibrato, if present, often occurs throughout a note; Vibrato patterns of different players are also distinct even if they are playing notes with the same rhythm. We have to admit that for performances that vibrato is not used, the proposed approach would fail. Nonetheless, it provides a complementary approach to our previous system and combining these two approaches is a future direction of research.

5. CONCLUSIONS

We proposed a methodology for source association by synergistic analyses of vibrato articulations from both audio

and video modalities for string ensembles. Specifically, we found that the fine-grained motion of the fingering hand shows a strong correlation with the pitch fluctuation of vibrato notes, which was utilized to solve source association in a score-informed fashion. Experiments showed a high success rate on pieces of different polyphony, and proved that vibrato features provide more robust cues for source association than bowing motion features that were utilized in our previous work. This technique enables novel and richer music enjoyment experiences that allow users to isolate/enhance sound sources by selecting the players in the video. For future work, we would like to combine the vibrato cues and the bowing cues to achieve more robust association results. We also would like to explore scenarios where the score is not available.

6. REFERENCES

[1] C. Vernallis, *Unruly media: YouTube, music video, and the new digital cinema*. Oxford University Press, 2013.

[2] A. Bazzica, C. C. Liem, and A. Hanjalic, "Exploiting instrument-wise playing/non-playing labels for score synchronization of symphonic music." in *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, 2014.

[3] A. Hrybyk, "Combined audio and video analysis for guitar chord identification," Master's thesis, Drexel University, 2010.

[4] A. Bazzica, J. C. van Gemert, C. C. Liem, and A. Hanjalic, "Vision-based detection of acoustic timed events: a case study on clarinet note onsets," in *International Workshop on Deep Learning for Music (DLM) in conjunction with the International Joint Conference on Neural Networks (IJCNN)*, Anchorage, Alaska (USA), 2017.

[5] K. Dinesh, B. Li, X. Liu, Z. Duan, and G. Sharma, "Visually informed multi-pitch analysis of string ensembles," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2017.

[6] S. Dahl and A. Friberg, "Visual perception of expressiveness in musicians' body movements," *Music Perception: An Interdisciplinary Journal*, vol. 24, no. 5, pp. 433–454, 2007.

[7] J. Scarr and R. Green, "Retrieval of guitarist fingering information using computer vision," in *Proc. Intl. Conf. Image and Vision Computing New Zealand (IVCNZ)*, 2010.

[8] D. Murphy, "Tracking a conductor's baton," in *Proc. Danish Conf. Pattern Recognition and Image Analysis*, 2003.

[9] J. R. Hershey and J. R. Movellan, "Audio vision: Using audio-visual synchrony to locate sounds." in *Proc. Neural Inform. Process. Syst.*, 1999, pp. 813–819.

[10] B. Li, Z. Duan, and G. Sharma, "Associating players to sound sources in musical performance videos," in *Late Breaking Demo, Intl. Society for Music Information Retrieval (ISMIR)*, 2016.

[11] B. Li, K. Dinesh, Z. Duan, and G. Sharma, "See and listen: Score-informed association of sound tracks to players in chamber music performance videos," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2017.

[12] K. Li, J. Ye, and K. A. Hua, "What's making that sound?" in *Proc. ACM Intl. Conf. Multimedia*, 2014, pp. 147–156.

[13] A. L. Casanovas and P. Vandergheynst, "Audio-based nonlinear video diffusion," in *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*. IEEE, 2010, pp. 2486–2489.

[14] Y. Liu and Y. Sato, "Finding speaker face region by audiovisual correlation," in *Proc. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications*, 2008.

[15] R. Cutler and L. Davis, "Look who's talking: Speaker detection using video and audio correlation," in *Proc. IEEE Intl. Conf. Multimedia and Expo,*, vol. 3, 2000, pp. 1589–1592.

[16] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 406–413, 2004.

[17] E. Kidron, Y. Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, no. 4, pp. 1390–1404, 2007.

[18] H. Izadinia, I. Saleemi, and M. Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Trans. Multimedia*, vol. 15, no. 2, pp. 378–390, 2013.

[19] Z. Barzelay and Y. Y. Schechner, "Harmony in motion," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2007, pp. 1–8.

[20] Z. Barzelay and Y. Y. Schechner, "Onsets coincidence for cross-modal analysis," *IEEE Trans. Multimedia*, vol. 12, no. 2, pp. 108–120, 2010.

[21] C. Sigg, B. Fischer, B. Ommer, V. Roth, and J. Buhmann, "Nonnegative CCA for audiovisual source separation," in *Proc. IEEE Workshop on Machine Learning for Signal Processing*, 2007, pp. 253–258.

[22] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Trans. Multimedia*, vol. 12, no. 5, pp. 358–371, 2010.

[23] L. Regnier and G. Peeters, "Singing voice detection in music tracks using direct voice vibrato detection," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2009, pp. 1685–1688.

[24] S. Rossignol, P. Depalle, J. Soumagne, X. Rodet, and J.-L. Collette, "Vibrato: detection, estimation, extraction, modification," in *Proc. Digital Audio Effects Workshop (DAFx)*, 1999, pp. 1–4.

[25] J. Driedger, S. Balke, S. Ewert, and M. Müller, "Template-based vibrato analysis of music signals," in *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, 2016.

[26] Z. Duan and B. Pardo, "Soundprism: An online system for score-informed source separation of music audio," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1205–1215, 2011.

[27] R. Hennequin, B. David, and R. Badeau, "Score informed audio source separation using a parametric model of non-negative spectrogram," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2011.

[28] J. Fritsch and M. D. Plumbley, "Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2013, pp. 888–891.

[29] S. Ewert and M. Müller, "Using score-informed constraints for NMF-based source separation," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2012, pp. 129–132.

[30] P. Sprechmann, P. Cancela, and G. Sapiro, "Gaussian mixture models for score-informed instrument separation," in *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2012, pp. 49–52.

[31] F. J. Rodriguez-Serrano, Z. Duan, P. Vera-Candeas, B. Pardo, and J. J. Carabias-Orti, "Online score-informed source separation with adaptive instrument models," *Journal of New Music Research*, vol. 44, no. 2, pp. 83–96, 2015.

[32] A. De Cheveigné and H. Kawahara, "Yin, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002.

[33] J. M. Geringer, R. B. MacLeod, and M. L. Allen, "Perceived pitch of violin and cello vibrato tones among music majors," *Journal of Research in Music Education*, vol. 57, no. 4, pp. 351–363, 2010.

[34] C. Tomasi and T. Kanade, *Detection and tracking of point features*. School of Computer Science, Carnegie Mellon Univ. Pittsburgh, 1991.

[35] D. Sun, S. Roth, and M. J. Black, "Secrets of optical flow estimation and their principles," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.

[36] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, submitted. Available: https://arxiv.org/abs/1612.08727.