VIDEO-BASED VIBRATO DETECTION AND ANALYSIS FOR POLYPHONIC STRING MUSIC

Bochen Li **Karthik Dinesh** **Gaurav Sharma** **Zhiyao Duan**
Dept. of Electrical and Computer Engineering, University of Rochester, Rochester, NY, USA
{bochen.li, kdinesh, gaurav.sharma, zhiyao.duan}@rochester.edu

ABSTRACT

In music performance, vibrato is an important artistic effect, where slight variations in pitch are introduced to add expressiveness and warmth. Automatic vibrato detection and analysis, although well studied for monophonic music, has rarely been explored for polyphonic music, because of the challenge in multi-pitch analysis. We propose a video-based approach for detecting and analyzing vibrato in polyphonic string music. Specifically, we capture the fine motion of the left hand of string players through optical flow analysis of video frames. We explore two methods. The first uses a feature extraction and SVM classification pipeline, and the second is an unsupervised technique based on autocorrelation analysis of the principal motion component. The proposed methods are compared with audio-only methods applied to individual instrument tracks separated from original audio mixture using the score. Experiments show that the proposed video-based methods achieve a significantly higher vibrato detection accuracy than the audio-based methods especially in high polyphony cases. Further experiments also demonstrate the utility of the approach in vibrato rate and extent analysis.

1. INTRODUCTION

Vibrato is an important artistic effect in musical performance. Instrument players use vibrato to color a tone and express emotions. Physically, vibrato is generated by pitch modulation of a note in a periodic fashion [23]. Important characteristics of vibrato include *rate* and *extent* of this periodic modulation [8]. These characteristics vary significantly across instruments, cultures, and personal styles. Compared to woodwind and brass instruments, vibrato is more pronounced in strings.

Automatic vibrato detection and analysis is an important topic in music information retrieval (MIR) with broad impacts. It is useful in musicological studies to compare different articulation styles of different performers and instruments [2]. It is critical in expressive performance pedagogy for singing [19] and violin [28]. It also facilitates other MIR tasks such as singing voice extraction [12], harmonic-percussive decomposition [21], and audio-visual source association [16]. Vibrato analysis also provides the statistical basis for vibrato synthesis of musical instruments [13], singing voices [11], and bird songs [4], through which the synthesized sounds are more realistic and expressive.

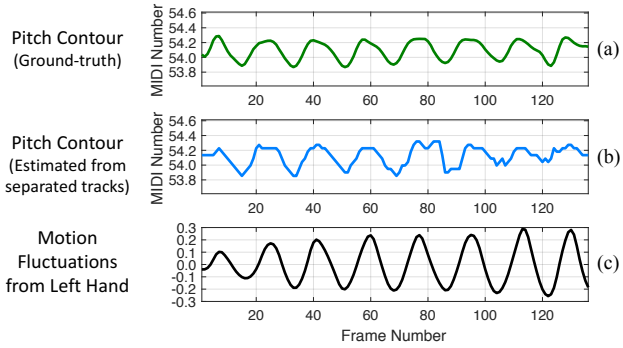

Figure 1. The proposed method tackles the challenging problem of vibrato analysis for polyphonic music by exploiting information from the video to augment audio analysis. (a) The ground-truth pitch contour of a cello vibrato note in a violin-cello duet performance showing a clear vibrato pattern, (b) The estimated pitch contour of this note from the audio mixture using a state-of-the-art score-informed pitch detection method showing corruption due to the interference from the other source, (c) The left hand motion along the fingerboard of the cello player extracted from video analysis is clean and well correlated with the ground-truth pitch contour. The hand motion profile extracted from video is used for vibrato analysis in this paper.

Most of the existing methods for automatic vibrato detection and analysis are audio-based with a focus on monophonic sources, where vibrato can be easily characterized from the pitch trajectory estimated through a monophonic pitch detection algorithm. Methods include thresholding the pitch drift within each note [3], calculating the median distance of the neighboring peaks/troughs of the pitch contour [9], analyzing the spectral peak after a Fourier transform of the pitch contour [25], cross-correlation analysis of frequency/amplitude modulation [26], and a nonlinear sinusoidal decomposition method [27].

Few approaches have focused on polyphonic music, and when they do, they only characterize vibrato of a single source (usually the solo instrument) in the mixture. This

is mainly due to the difficulty of reliably estimating simultaneous pitches in polyphonic music [5]. Abeßer et al. [1] proposed a score-informed approach to first estimate the pitch contour of the solo instrument from the audio mixture and then perform vibrato detection and analysis on the pitch contour through autocorrelation. The performance of this approach, however, depends heavily on the pitch estimation performance. Spectrogram-based approaches such as harmonic partial tracking [12] and template convolution [6] reduce the dependency on pitch estimation. However, these operations are still error-prone when harmonics of different sources overlap. To our best knowledge, there is no existing approach for vibrato detection and analysis of multiple simultaneous sources of a polyphonic music mixture, such as a string ensemble. Existing polyphonic audio analysis techniques are not yet sufficient.

Figure 1 shows the limitation of audio-based analysis and motivates the video-based analysis proposed in this paper. In Figure 1 (a), the ground-truth pitch contour of a cello vibrato note in a violin-cello duet performance is shown. This pitch contour is estimated using a monophonic pitch detection algorithm [17] on the isolated (ground truth) signal of the cello note prior to mixing. Vibrato characteristics are clearly observable in this pitch contour. Figure 1 (b) shows the estimated pitch contour of this cello note obtained from a state-of-the-art score-informed source separation and pitch estimation algorithm [7]. Due to the interference from the violin, the estimated pitch contour is corrupted and the vibrato patterns are obscured, especially toward the later time instants represented on the right side of the plot. Note that this example is just a duet of instruments with distinct pitch ranges. For music with higher polyphony using instruments with similar pitch ranges, the estimated pitch contours are further corrupted, making audio-based vibrato detection and analysis unsatisfactory.

For some instruments such as strings, vibrato is often visible from the left hand motion, and this visual information does not degrade as audio information does when polyphony increases. This motivates our proposed approach of vibrato detection and analysis through video-based analysis of the fine motion of the left hand. Figure 1 (c) shows the left hand rolling motion along the principal motion direction (i.e., the fingerboard) of the cello player playing the note. We can see that this motion curve is smooth and it aligns with the ground-truth pitch contour in Figure 1 (a) very well.

The overview of our proposed approach is illustrated in Figure 2. This approach integrates audio, visual and score information, and assumes that the players in the video are well associated with score tracks. Our previous work has addressed the association problem accurately [14]. For each string player, we track the left hand, and then estimate optical flow motion vectors at the pixel level around the left hand. We use audio-score alignment to identify the onset and offset of each note, and perform vibrato detection and analysis on each note from the motion vectors. We develop two approaches for vibrato detection. One uses

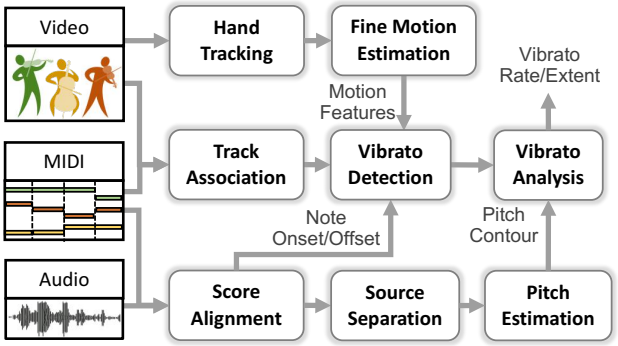

Figure 2. System overview of the proposed video-based vibrato detection and analysis framework.

a Support Vector Machine (SVM) to classify motion features extracted from the pixel-level motion vectors, and the other is based on autocorrelation analysis of the left hand motion along the principal direction (i.e., fingerboard). We further propose a framework to analyze vibrato characteristics: rate and extent. The vibrato rate is estimated from the period of the hand motion curve, and the vibrato extent is estimated from the amplitude of the motion curve after it is scaled to match the estimated noisy pitch contour from score-informed audio analysis.

Experiments are carried on 19 pieces of polyphonic string music from an audio-visual music performance dataset, and the proposed video-based approach is compared with two audio-based baseline methods for vibrato detection. Results show a significant improvement for video-based vibrato detection over the audio-based methods. Further analysis reveals that video-based vibrato detection is robust irrespective of polyphony and instrument types. We further show that the video-based approach is able to estimate the vibrato rate and extent with a deviation from the ground-truth smaller than 1 Hz and 10 musical cents for 90% of the notes, respectively.

2. AUDIO-BASED METHOD

In this section, we introduce an audio-based framework to detect vibrato in polyphonic music to serve as a baseline method. Vibrato can be detected from the pitch contour of each source using either autocorrelation or Fourier transform. However, estimating the pitch contour of each source from the audio mixture is challenging. Inspired by [1], score information can be utilized to alleviate the difficulty of pitch estimation and its assignment to sources.

2.1 Score-informed Pitch Estimation

To utilize the score information for pitch estimation of each source, robust audio-score alignment is required to guarantee the temporal synchronization between the score events and audio articulations. We apply the Dynamic Time Warping (DTW) framework with chroma feature to represent audio and score, as described in [14]. Then the audio mixture is separated using harmonic masking as described in [7]: Pitches of each source are first estimated within two semitones around the quantized score-notated

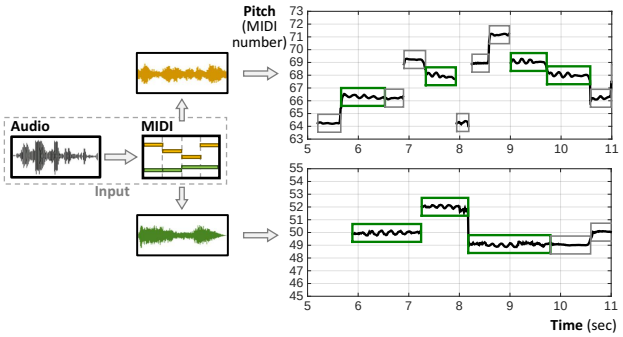

Figure 3. Audio-based vibrato detection. Detected vibrato notes are marked with green rectangles in the pitch trajectories estimated by score-informed pitch estimation.

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 signal's energy to overlapping harmonics.

We then re-estimate the pitch contour of each source from its separated signal for vibrato analysis. We again apply the above-mentioned score-informed pitch refinement step to further reduce interference from other sources. The output pitch contour is segmented into notes using the onset/offset information provided by the aligned score. Note that although we can refine the pitches directly from the audio mixture without source separation, it is reported in [16] that the result is more robust on the separated sources. Besides, the availability of separated audio sources is advantageous for other vibrato detection methods that do not rely on pitch contours.

2.2 Vibrato Detection from the Pitch Contour

After obtaining the pitch contour, vibrato detection can be achieved by analyzing the periodic pattern for each note. The pitch contour is analyzed in the MIDI scale, and its DC component is removed by subtracting the average value over the contour. Then we implement two methods to detect the fluctuation rate of the pitch contour: autocorrelation [1] and spectral analysis [25]. For the autocorrelation method, prominent peaks are detected from the autocorrelation function, and the median value of all the neighboring peak distance is used to calculate the fluctuation rate. If the rate is within the range of 3-9 Hz (considering a typical vibrato rate range of [4, 7.5] Hz for strings [10]), the note is detected as vibrato. For the spectral analysis method, we first calculate the magnitude spectrum of the pitch contour of a note through Fourier transform. We then check if the frequency of the maximum peak lies in the rang of 3-9 Hz. Quadratic interpolation is applied in both methods to get a more precise peak location estimation.

The audio-based methods are simple, yet sufficient to detect vibrato in the score-informed fashion. Figure 3 reviews this process and illustrated the detected vibrato notes in green boxes. This approach achieves high detection accuracy in low polyphony settings, but the performance degrades rapidly with increasing polyphony.

3. PROPOSED METHOD

Motivated by the fact that the motion features from the video are correlated with the pitch fluctuations, we propose a video-based vibrato detection and analysis framework. A string instrument player exhibits three kinds of motions: bowing motion to articulate notes, fingering motion to control pitches, and the whole body motion to express musical intentions. Fine periodic fingering motion on the left hand along the fingerboard which changes the length and tension of the string results in vibrato articulations. In this section, we will present the method to extract this fine motion for vibrato detection and analysis.

3.1 Motion Capture

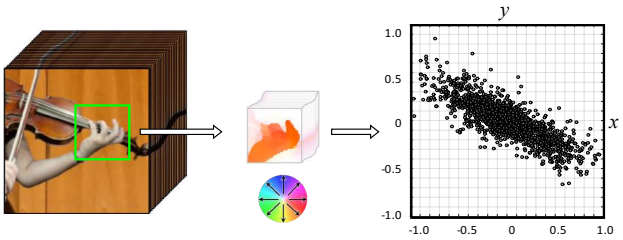

Figure 4. Motion capture results from left hand tracking (left), color encoded pixel velocities (middle), and scatter plot of frame-wise refined motion velocities (right).

The first step is to detect and track the left hand for each player, where the vibrato motions come from. The hand tracking is based on the Kanade-Lucas-Tomasi (KLT) tracker [24] and implemented using the same parameters as presented in [16]. The KLT tracker results in a dynamic region of tracked hand location where we apply the optical flow estimation [22] to obtain the raw motion velocities for each pixel in x and y directions within that region. The motion velocities are spatially averaged as $\mathbf{u}(t) = [u_x(t), u_y(t)]$, where u_x and u_y represents the mean motion velocities in x and y directions respectively, and t is the time index. Notice that the motion velocities in the hand region contain not only the player's fine motion corresponding to vibrato playing, but also his/her large-scale body motions during the performance. In order to eliminate the body movement and obtain a refined motion velocities for vibrato observation, we subtract a moving average of the signal $\mathbf{u}(t)$ from itself, to obtain

$$\mathbf{v}(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t), \tag{1}$$

where $\bar{\mathbf{u}}(t)$ is the moving average of $\mathbf{u}(t)$ over a 10 frame window. Figure 4 illustrates the original video frame with the tracked hand position, the raw motion velocities from optical flow estimation, and the refined motion velocities $\mathbf{v}(t)$ across all the frames.

3.2 Vibrato Detection from Motion Features

The proposed vibrato detection methods are score informed, where the note onset/offset information from the score is utilized to temporally segment the refined mean motion velocities into $\mathbf{v}^i(t)$, where i is the note index.

To achieve this, each score track needs to be temporally aligned with the video frames, and spatially associated with the players. The first issue is resolved using audio-score alignment, assuming video and audio frames are naturally synchronized. The second issue is addressed as in [14], where player locations are segmented and associated with the score tracks by correlating the bow motions with note events. By utilizing the mean motion velocities and the extracted features, we introduce two methods for vibrato detection. The first method is based on a SVM framework, where each $\mathbf{v}^i(t)$ is classified as vibrato or non-vibrato. The second method is analogous to the audio-based technique, where we perform auto-correlation on the extracted 1-D motion curve along the fingerboard after principal component analysis.

3.2.1 SVM

We train a Support Vector Machine (SVM) as a classification framework for vibrato/non-vibrato detection. We utilize the refined motion velocity segments $\mathbf{v}^i(t) = [v_x^i(t), v_y^i(t)]$ obtained from the procedure explained in Section 3.1. From each $\mathbf{v}^i(t)$, we have velocity components in x and y directions from which 8 dimensional features are extracted. The features are

(a) **Zero crossing rate** (4-D): Vibrato has inherent periodicity when compared to non-vibrato regions. Hence we utilize the zero crossing rate, which is the ratio of total zero crossings to total frame length for $v_x^i(t)$, $v_y^i(t)$ and their auto-correlations, respectively.

(b) **Frequency** (2-D): Vibrato has a typical frequency in the range of 3-9 Hz. Hence we calculate the sum of the absolute value of Fourier coefficients in the 3-9 Hz frequency range for $v_x^i(t)$ and $v_y^i(t)$.

(c) **Auto-correlation peaks** (2-D): Auto-correlation of $v_x^i(t)$ and $v_y^i(t)$ is calculated within a fixed lag of 10 video frames, where total number of local maximum values is utilized as one of the features.

The SVM is trained on tracks which are distinct from the test set using the leave-one-out training strategy. The ground truth vibrato/non-vibrato labels are obtained from ground-truth audio tracks and associated with the corresponding player. For the SVM training algorithm we set the kernel function and scale parameters as radial basis function and automatic scaling, respectively.

3.2.2 PCA

We also propose an unsupervised framework for vibrato detection. From Figure 4, we find that the distribution of the refined motion velocities for vibrato motions are along the fingerboard. So we perform Principal Component Analysis (PCA) on $\mathbf{v}(t)$ across all frames to identify this principal motion direction, and project the motion velocity vectors to this principal direction to obtain a 1-D *motion velocity curve* $V(t)$ as

$$V(t) = \frac{\mathbf{v}(t)^T \tilde{\mathbf{v}}}{\|\tilde{\mathbf{v}}\|}, \tag{2}$$

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

$$X(t) = \int_0^t V(\tau) d\tau. \tag{3}$$

This displacement curve corresponds to the fluctuation of the vibrating length of the string and hence the pitch fluctuation of the note. Figure 1 (c) shows the motion displacement curve for one vibrato note, which is matched with the ground-truth pitch contour. Similar to the audio-based approach, vibrato is detected through peak picking on the autocorrelation function of the motion displacement curve. Note that different thresholds on the peak picking will affect the sensitivity of the vibrato detection, and we use the uniform threshold for all the notes which yields the best overall results.

3.3 Vibrato Analysis

The video-based method also enables new techniques for analyzing the vibrato features, i.e., vibrato rate and vibrato extent, which describe the speed and the amount by which the pitch is varied. Here extent is defined as the dynamic range of the pitch contour, i.e., the peak-trough difference. Vibrato rate can be directly extracted from video by observing how fast the left hand is rolling along the fingerboard. Again this is solved by analyzing the autocorrelation on the motion displacement curve $X(t)$. Quadratic interpolation is required for peak picking due to the low frame rate of videos. Vibrato extent, however, cannot be estimated by capturing the motion extent, which varies upon different camera distance and angles. Besides, to generate the same vibrato extent, the extent of motion also depends on the vibrato articulation style, the hand position on the fingerboard, and the instrument type. Therefore, we combine the audio analysis together with the extracted motion displacement curve for vibrato extent estimation.

We first estimate the vibration extent of the motion displacement curve as $\hat{w}_e$ by calculating the median of the distance between all the peaks and troughs within each note. We then scale the displacement curve to fit the pitch contour, and the vibrato extent can be calculated from the scaling factor. Specifically, assuming $F(t)$ is the estimated pitch contour (in MIDI number) of the detected vibrato note from audio analysis after subtracting the DC component of itself, the vibrato extent v_e (in musical cents) is estimated as $\hat{v}_e$ as:

$$\hat{v}_e = \arg\min_{v_e} \sum_{t=t^{on}}^{t^{off}} \left| 100 \cdot F(t) - v_e \frac{X(t)}{\hat{w}_e} \right|^2. \tag{4}$$

where $100 \cdot F(t)$ is the pitch contour measured in musical cents; $\frac{X(t)}{\hat{w}_e}$ is the normalized hand displacement curve. Since $X(t)$ is calculated from the video modality, temporal interpolation is applied beforehand to guarantee the same frame rate as the audio, i.e., the hop size for Short-Time Fourier Transform. Note that temporal shift may be applied to $X(t)$ to maximize the cross correlation between $X(t)$ and $F(t)$ to compensate the slight asynchrony between the two modalities (usually within 20ms).

4. EXPERIMENTS

4.1 Dataset and Evaluation Measures

The vibrato detection and analysis system is tested on the URMP dataset [15]. The dataset contains individually recorded audio-visual tracks of various instruments, which are synchronized and assembled to form 44 classical ensemble pieces ranging from duets to quintets. Ground-truth audio tracks and pitch/note annotations are provided in the dataset. The ground-truth annotation of the vibrato rate/extent is acquired by the autocorrelation method as described in Section 2.2 on ground-truth individual audio tracks, and the presence of vibrato is manually examined. For our experiments, we use the 19 ensemble pieces that contains at most one non-string instrument, including 5 duets, 4 trios, 7 quartets, and 3 quintets. 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 STFT. Video resolution is 1080P, and the frame rate is 29.97 frames per second.

In the experiments, we evaluate the two proposed video-based methods, i.e., the classification method using SVM framework (Vid-SVM) and autocorrelation analysis on the principal motion component (Vid-PCA). Two audio-based methods described in Section 2.2 are also compared as baseline methods, i.e., peak-picking of the autocorrelation (Aud-AC), and Fourier transform of the pitch contour (Aud-FT). Since the vibrato detection can be viewed as a retrieval task, we compute the note-level precision (P), recall (R), and F-measure (F) using the number of true positives, false positives and false negatives on each track. For the two audio-based methods and the Vid-PCA method, we adjust the peak-picking threshold for a balanced value of precision and recall and fix it for all the tracks. For vibrato rate and extent estimation, we calculate the error between the estimated and ground-truth values on the true positive detections from the Vid-PCA method.

4.2 Results

4.2.1 Overall Evaluation on Vibrato Detection

We first evaluate the vibrato detection results using precision, recall and F-measure for the four methods on all of the 57 tracks from the 19 pieces excluding non-string instrument ones, as plotted in Figure 5. Each bar is the average of the 57 tracks. We find that in polyphonic music, both audio-based methods achieve limited performance; lower than 75% for the F-measure. Video-based methods can get a pronounced improvement on the F-measure, which is as high as 90%. The supervised classification method based on SVM further outperforms the unsupervised method, because of the richer features.

4.2.2 Vibrato Detection Evaluation on Different Cases

We further investigate how the vibrato detection performance changes along with polyphony and instrument types. Figure 6 illustrates the scatter plot of the vibrato detection F-measure for the four methods (with different colors) in four different polyphony levels corresponding to duets, trios, quartets, and quintets. Each sample point represents the evaluation on one track, and the average

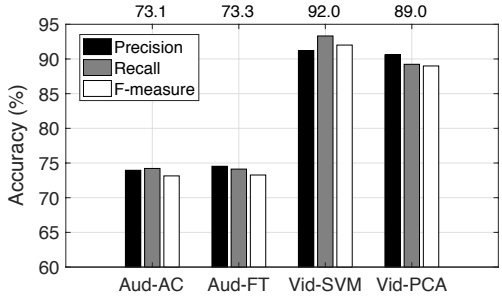

Figure 5. Overall vibrato detection results showing the precision, recall, and F-measure (shown on top) accuracies for 2 audio-based methods and 2 video-based methods.

value in each subset is marked as the red line. We see that the two audio-based methods can reach performance comparable with the video-based methods in low-polyphony pieces, but their performance drops when polyphony increases. This is because of the decreased quality of the pitch contour that is extracted from high-polyphony audio. However, polyphony does not affect the vibrato detection performance for the two video-based methods, since the left hands are always directly observable from visual scene in this dataset. Note that there are several extremely low F-measure values for video-based methods. These come from tracks with plucking-vibrato articulations, where the vibrato is captured from hand motion but is not annotated in the ground truth as its duration and extent are different from the bowing-vibrato articulations.

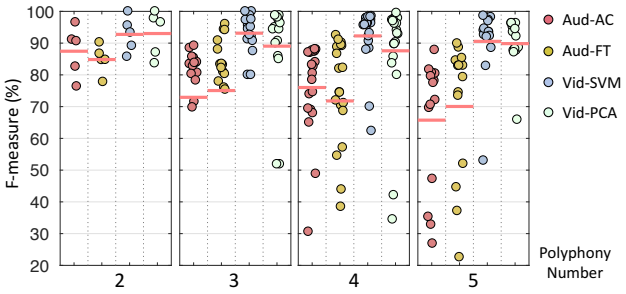

Figure 6. Vibrato detection performance decreases as polyphony increases for audio-based methods, while it stays the same for video-based methods.

Figure 7 further reveals how the vibrato detection results vary for different instruments: violin, viola, cello, and double bass. Again, the audio-based methods are sensitive to instrument types while video-based methods are not. The reason is that the separated track of the low-pitch instrument (such as double bass) is likely to get contaminated by other higher-pitch voices using the harmonic mask method for source separation. In contrast, the vibrato motions for the four different instruments have similar patterns, thus easy to capture by our proposed methods.

4.2.3 Evaluation of Vibrato Characteristics

Due to the unsatisfactory performance of audio-based vibrato detection, we evaluate the accuracy of vibrato rate

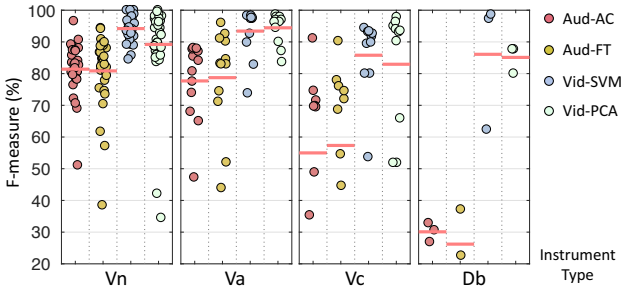

Figure 7. Vibrato detection performance decreases when the fundamental frequency decreases for audio-based methods, while it stays the same for video-based methods.

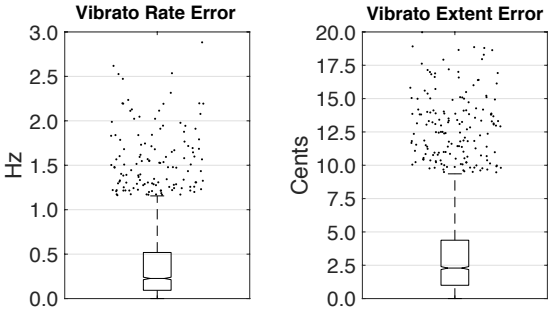

Figure 8. Distribution of vibrato rate and extent estimation error on all notes of all tracks.

and extent estimation only based on the video modality. We conduct this analysis on the true positive detections from the Vid-PCA method, totaling 2290 vibrato notes from the 57 tracks. We calculate the absolute deviation of the estimated value from the ground-truth value for all the notes, and get an average vibrato rate estimation error of 0.38 Hz and median of 0.23 Hz. For vibrato extent, we have an average estimation error of 3.47 cents and a median of 2.29 cents. Figure 8 plots the vibrato rate and extent error distribution for all the notes. We find that for 90% of the vibrato notes, the proposed approach estimates the vibrato rate and extent within an error of 1 Hz and 10 cents, respectively.

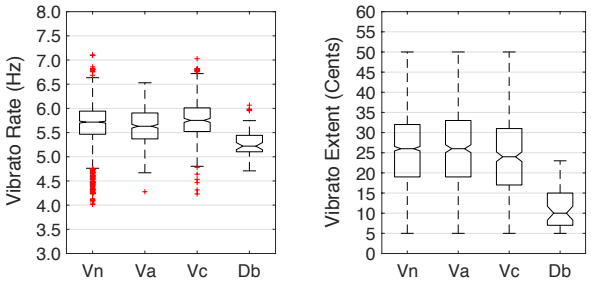

Figure 9. Distributions of vibrato rate and extent for different instruments.

In order to further demonstrate the potential applications of our approach in musicology studies, we analyze how the vibrato rate and extent vary on different instruments and players in this dataset. Figure 9 plots the distributions of rate and extent for the four string instruments, where each sample point represents one track. Similar vibrato rate and extent can be observed for violin and viola whereas, in contrast, we observe a significant drop for the double bass, where a slower rate and subtler extent is inferred. This was explained in [18]; to produce audible pitch fluctuations on the thicker and longer strings on double bass requires more effort to overcome the strength, flexibility, and coordination than other string instruments. Thus vibrato rates of double bass players (4-5 Hz [20]) are typically slower than other string instrumentalists.

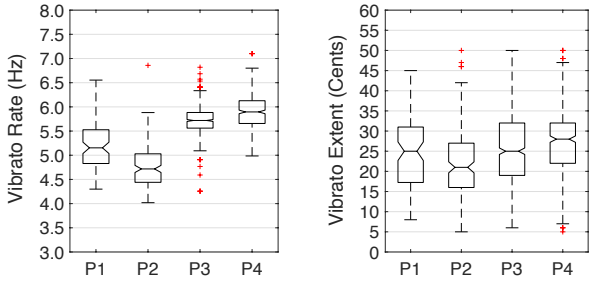

Figure 10. Distributions of vibrato rate and extent of four different violin players.

We also analyze the vibrato patterns of the four different violinists among the 31 violin tracks, as plotted in Figure 10. Vibrato rate is more dispersed among players than vibrato extent, and both rate and extent show a similar trend among the players. For example, the second player exhibits a slower vibrato rate with a subtler vibrato extent, while the forth player exhibits a faster vibrato rate with a pronounced vibrato extent. This may be because of different players' articulation styles, or different characteristics of the pieces. Detailed discussion is not included in this paper, but our proposed system can provide a powerful tool for further analyses on the musicology side.

5. CONCLUSION

We proposed a video-based vibrato detection and analysis framework for polyphonic string music. Specifically, we developed two methods that utilize the motion features from the video for vibrato detection based on the observed correlation between the motion vibrations and the vibrato pitch fluctuations. We also extended the framework to estimate the vibrato rate and extent. Experiments show that the proposed method is successful and offers much better performance than audio-based methods, particularly on pieces with high polyphony, where the strong interference between sources severely degrades the performance of audio-based methods. In future work, it would be helpful to develop a non-score-informed framework for vibrato detection and analysis.

6. ACKNOWLEDGMENT

We thank the Center for Integrated Research Computing (CIRC), University of Rochester for providing computational resources for the project.

7. REFERENCES

[1] Jakob Abeßer, Estefanía Cano, Klaus Frieler, Martin Pfleiderer, and Wolf-Georg Zaddach. Score-informed analysis of intonation and pitch modulation in jazz solos. In *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, pages 823–829, 2015.

[2] Jakob Abeßer, Klaus Frieler, Estefanía Cano, Martin Pfleiderer, and Wolf-Georg Zaddach. Score-informed analysis of tuning, intonation, pitch modulation, and dynamics in jazz solos. *IEEE/ACM Trans. Audio, Speech, and Language Process.*, 25(1):168–177, 2017.

[3] Isabel Barbancho, Cristina de la Bandera, Ana M Barbancho, and Lorenzo J Tardon. Transcription and expressiveness detection system for violin music. In *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pages 189–192. IEEE, 2009.

[4] Jordi Bonada, Robert Lachlan, and Merlijn Blaauw. Bird song synthesis based on hidden markov models. In *Proc. InterSpeech*, volume 2016, 2016.

[5] Karthik Dinesh, Bochen Li, Xinzhao Liu, Zhiyao Duan, and Gaurav Sharma. Visually informed multi-pitch analysis of string ensembles. In *Proc. IEEE Intl. Conf. Acoust., Speech, and Signal Process. (ICASSP)*, 2017.

[6] Jonathan Driedger, Stefan Balke, Sebastian Ewert, and Meinard Müller. Template-based vibrato analysis of music signals. In *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, 2016.

[7] Zhiyao Duan and Bryan Pardo. Soundprism: An online system for score-informed source separation of music audio. *IEEE J. Sel. Topics Signal Process.*, 5(6):1205–1215, 2011.

[8] Harvey Fletcher and Larry C Sanders. Quality of violin vibrato tones. *The Journal of the Acoustical Society of America*, 41(6):1534–1544, 1967.

[9] Anders Friberg, Erwin Schoonderwaldt, and Patrik N Juslin. Cuex: An algorithm for automatic extraction of expressive tone parameters in music performance from acoustic signals. *Acta acustica united with acustica*, 93(3):411–420, 2007.

[10] John M Geringer, Rebecca B MacLeod, and Michael L Allen. Perceived pitch of violin and cello vibrato tones among music majors. *Journal of Research in Music Education*, 57(4):351–363, 2010.

[11] Hung-Yan Gu and Zheng-Fu Lin. Singing-voice synthesis using ann vibrato-parameter models. *J. Inf. Sci. Eng.*, 30(2):425–442, 2014.

[12] Chao-Ling Hsu and Jyh-Shing Roger Jang. Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion. In *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, pages 525–530, 2010.

[13] Hanna Järveläinen. Perception-based control of vibrato parameters in string instrument synthesis. In *Proc. International Computer Music Conference (ICMC)*, 2002.

[14] Bochen Li, Karthik Dinesh, Zhiyao Duan, and Gaurav 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.

[15] Bochen Li, Xinzhao Liu, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. Creating a classical musical performance dataset for multimodal music analysis: Challenges, insights, and applications. *IEEE Trans. Multimedia*. submitted. Available: `https://arxiv.org/abs/1612.08727`.

[16] Bochen Li, Chenliang Xu, and Zhiyao Duan. Audio-visual source association for string ensembles through multi-modal vibrato analysis. In *Proc. Sound and Music Computing (SMC)*, 2017.

[17] Matthias Mauch and Simon Dixon. PYIN: a fundamental frequency estimator using probabilistic threshold distributions. In *Proc. IEEE Intl. Conf. Acoustics, Speech, and Signal Process. (ICASSP)*, pages 659–663. IEEE, 2014.

[18] James Paul Mick. *An analysis of double bass vibrato: Rates, widths, and pitches as influenced by pitch height, fingers used, and tempo*. PhD thesis, The Florida State University, 2012.

[19] Tomoyasu Nakano, Masataka Goto, and Yuzuru Hiraga. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Proc. InterSpeech*, 2006.

[20] George Papich and Edward Rainbow. A pilot study of performance practices of twentieth-century musicians. *Journal of Research in Music Education*, 22(1):24–34, 1974.

[21] Jeongsoo Park and Kyogu Lee. Harmonic-percussive source separation using harmonicity and sparsity constraints. In *Proc. Intl. Society for Music Information Retrieval (ISMIR)*, pages 148–154, 2015.

[22] Deqing Sun, Stefan Roth, and Michael J Black. Secrets of optical flow estimation and their principles. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2010.

[23] Johan Sundberg. Acoustic and psychoacoustic aspects of vocal vibrato. *STL-QPSR*, pages 35–62, 1995.

[24] Carlo Tomasi and Takeo Kanade. Detection and tracking of point features. Technical Report CMU-CS-91-132, School of Computer Science, Carnegie Mellon University, Apr. 1991.

[25] José Ventura, Ricardo Sousa, and Anibal Ferreira. Accurate analysis and visual feedback of vibrato in singing. In *Proc. Intl. Symposium on Communications Control and Signal Process. (ISCCSP)*, pages 1–6. IEEE, 2012.

[26] Henrik Von Coler and Axel Roebel. Vibrato detection using cross correlation between temporal energy and fundamental frequency. In *Proc. 131st Audio Engineering Society Convention*, 2011.

[27] Luwei Yang, Khalid Z Rajab, and Elaine Chew. The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation. *Journal of Mathematics and Music*, pages 1–19, 2017.

[28] Jun Yin, Ye Wang, and David Hsu. Digital violin tutor: an integrated system for beginning violin learners. In *Proc. ACM Intl. Conf. Multimedia*, pages 976–985. ACM, 2005.