

Multi-Modal Analysis for Music Performances

by

Bochen Li

Submitted in Partial Fulfillment of the
Requirements for the Degree
Doctor of Philosophy

Supervised by Professor Zhiyao Duan

Department of Electrical and Computer Engineering
Arts, Sciences and Engineering
Edmund A. Hajim School of Engineering and Applied Sciences

University of Rochester
Rochester, New York

2020

For my parents who helped me in all things great and small.

Table of Contents

Biographical Sketch	x
Acknowledgments	xiii
Abstract	xv
Contributors and Funding Sources	xvii
List of Tables	xix
List of Figures	xxxiii
1 Introduction	1
1.1 Different Modalities of Music	1
1.2 Motivation	3
1.2.1 Impact on MIR Research	3
1.2.2 Towards Applications	5
1.3 Problem Statement	7
1.3.1 Coordination of Multiple Modalities	8

1.3.2	Multi-Modal MIR	9
1.4	Summary of Contributions	12
2	URMP Dataset Creation	15
2.1	Introduction	16
2.2	Review of Music Performance Datasets	18
2.2.1	Existing Datasets	20
2.2.2	Synchronization Challenges	24
2.3	Approaches to Synchronization	27
2.4	Dataset Creation Procedure	33
2.4.1	Piece Selection	34
2.4.2	Recruiting Musicians	35
2.4.3	Recording Conducting Videos	36
2.4.4	Recording Instrumental Parts	37
2.4.5	Mixing and Assembling Individual Recordings	38
2.4.6	Video Background Replacement	39
2.4.7	Ground-truth Annotation	40
2.5	The Dataset	41
2.5.1	Dataset Content	41
2.5.2	Synchronization Quality	42
2.5.3	Spatial Occlusion & Resolution	44
2.5.4	Limitations of the Dataset	46
2.6	Applications of The Dataset	48

2.6.1	Existing Tasks Using Only Audio Modality	48
2.6.2	New Tasks Using Both Audio and Visual Modalities . .	52
2.7	Conclusions	59
3	Audio-Score Alignment	60
3.1	Introduction	61
3.2	Related Work	65
3.2.1	Early Work	65
3.2.2	Online Alignment	65
3.2.3	Score Following for Piano Music	67
3.3	Sustained Effect in Piano Music	69
3.3.1	Acoustical Properties of Piano Notes	69
3.3.2	Sustained Effect and Its Major Causes	70
3.3.3	Influences of Sustained Effect on Score Following . .	74
3.4	Proposed Approach	75
3.4.1	Online Onset Detection	75
3.4.2	Sustained-sound Reduction	79
3.4.3	What If the Proposed Operations Are Wrongly Applied? .	86
3.4.4	The Score Following Framework	91
3.5	Experiments	92
3.5.1	Experimental Set-up	92
3.5.2	Results in Non-reverberant Cases	100
3.5.3	Parameter Analysis	104

3.5.4	Results in Reverberant Environments	108
3.6	Conclusions	110
4	Source Association	112
4.1	Introduction	113
4.2	Related Work	119
4.2.1	Source Localization	119
4.2.2	Source Association for Separation	120
4.2.3	Source Association for Chamber Ensembles	121
4.3	Method	122
4.3.1	Performance-Score Alignment	123
4.3.2	Onset Correspondence with Body Motion	125
4.3.3	Onset Correspondence with Finger Motion	129
4.3.4	Pitch Correspondence with Vibrato Motion	131
4.3.5	Integrating All Correspondences	137
4.4	Experiments	139
4.4.1	Dataset	139
4.4.2	System Setup	140
4.4.3	Onset Detection Evaluation	141
4.4.4	Source Association Evaluation	143
4.5	Conclusion	149
5	Visually-Informed Audio Analysis	154
5.1	Multi-Pitch Analysis	154

5.1.1	Introduction	155
5.1.2	Proposed Method	158
5.1.3	Multi-Pitch Estimation	162
5.1.4	Multi-Pitch Streaming	164
5.1.5	Experiments	166
5.1.6	Conclusions and Discussions	172
5.2	Vibrato Analysis	173
5.2.1	Introduction	173
5.2.2	Audio-based method	179
5.2.3	Proposed method	182
5.2.4	Experiments	188
5.2.5	Conclusions	195
6	Audiovisual Singing Voice Separation	197
6.1	Introduction	198
6.2	Related Work	202
6.2.1	Singing Voice Separation	202
6.2.2	Audiovisual Source Separation	203
6.3	Method	205
6.3.1	Network Architecture	205
6.3.2	Training	209
6.4	Dataset	211
6.4.1	A Cappella Audition Vocals (AAV)	211

6.4.2	URSing	213
6.5	Experiments	217
6.5.1	Implementation Details	217
6.5.2	Evaluation Metric	217
6.5.3	Baselines	218
6.5.4	Overall Results	221
6.5.5	Different Video Front-End Models	226
6.5.6	Non-Informative or Misleading Visual Input	228
6.5.7	Evaluation on A Cappella Songs	230
6.6	Conclusion	234
7	Visual Performance Generation	235
7.1	Introduction	236
7.2	Related Work	237
7.3	Method	239
7.3.1	Feature Extraction by CNN	239
7.3.2	Skeleton Movement Generation by LSTM	243
7.3.3	Training Condition	244
7.4	Experiments	247
7.4.1	Dataset	247
7.4.2	Objective Evaluations	248
7.4.3	Subjective Evaluation	251
7.5	Conclusions	254

8 Conclusion	258
--------------	-----

Bibliography	261
--------------	-----

Biographical Sketch

The author was born in Luoyang, China. He attended the University of Science and Technology of China, and graduated with a Bachelor of Science degree in Electrical Engineering. He began doctoral studies in Electrical and Computer Engineering at the University of Rochester in 2014. He received the Master of Science degree from the University of Rochester. He pursued his research under the direction of Professor Zhiyao Duan. His research interests lie primarily in the inter-disciplinary area of audio signal processing, machine learning, and computer vision towards multi-modal analysis of music performances, such as video-informed multi-pitch estimation and streaming, source separation and association, and expressive performance modeling and generation. He received a best paper award at the 2017 Sound and Music Computing (SMC) conference and a best paper nomination at the 2017 International Society for Music Information Retrieval (ISMIR) conference.

The following publications were a result of work conducted during doctoral study:

Bochen Li, Yuxuan Wang, and Zhiyao Duan, “Audiovisual Singing Voice

Separation”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, under review.

Bochen Li, Karthik Dinesh, Chenliang Xu, Gaurav Sharma, and Zhiyao Duan, “Online Audio-Visual Source Association for Chamber Music Performances”, *Transactions of the International Society for Music Information Retrieval (TISMIR)*, vol. 2, no. 2, pp. 29-42, 2019.

Bochen Li and Aparna Kumar, “Query by Video: Cross-modal Music Retrieval”, in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2019.

Bochen Li*, Xinzha Liu*, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma, “Creating A Musical Performance Dataset for Multimodal Music Analysis: Challenges, Insights, and Applications”, *IEEE Transactions on Multimedia*, 2018. (* Equal contribution)

Bochen Li, Akira Maezawa, and Zhiyao Duan, “Skeleton Plays Piano: Online Generation of Pianist body Movements from MIDI Performance”, in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2018.

Yapeng Tian, Jing Shi, **Bochen Li**, Zhiyao Duan, and Chenliang Xu, “Audio-Visual Event Localization in Unconstrained Videos”, in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

Xueyang Wang, Ryan Stables, **Bochen Li**, and Zhiyao Duan, “Score-aligned Polyphonic Microtiming Estimation”, in *Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2018.

Bochen Li, Karthik Dinesh, Gaurav Sharma, and Zhiyao Duan, “Video-based Vibrato Detection and Analysis for Polyphonic String Music”, in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2017. (Best Paper Nomination)

Bochen Li, Chenliang Xu, and Zhiyao Duan, “Audio-visual Source Association for String Ensemble Videos through Multi-modal Vibrato Analysis”, in *Proceedings of the Sound and Music Computing Conference*, 2017. (Best Paper Award)

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 *Proceedings of the International Conference on Audio Speech and Signal Processing (ICASSP)*, 2017.

Karthik Dinesh*, **Bochen Li***, Xinzha Liu, Zhiyao Duan, and Gaurav Sharma, “Visually Informed Multi- pitch Analysis of String Ensembles”, in *Proceedings of the International Conference on Audio, Speech and Signal Processing (ICASSP)*, 2017. (* Equal contribution)

Bochen Li and Zhiyao Duan, “An approach to score following for piano performances with sustained effect”, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, 2016, pp. 2425-2438.

Bochen Li and Zhiyao Duan, “Score following for piano performances with sustain-pedal effects”, in *Proceedings of the International Society for Music Information Retrieval (ISMIR)*, 2015.

Acknowledgments

First of all, I would like to express my sincere gratitude to my advisor Professor Zhiyao Duan for his continuous support of my PhD study and research, for his patience, motivation, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. I could not have imagined having a better advisor and mentor for my PhD study.

Besides my advisor, I would like to thank my thesis committee: Professor Gaurav Sharma, Professor Chenliang Xu, Professor Mark Bocko, for their precious time, insightful comments, and encouragement, which helps me to widen my research from various perspectives. I would also like to thank Professor David Temperley (from the Eastman School of Music), for being the Chair of the Committee.

I thank my fellow lab mates and friends at University of Rochester for the precious experience to work together and encourage each other in the past years. I also thank Sarah Rose Smith, Drs. Ming-Lun Lee, Yunn-Shan Ma, and Andrea Cogliati for their contributions in the early-stage creation process of the URMP dataset.

Last but not the least, I would like to thank my family for supporting me spiritually throughout writing this thesis and my life in general.

Abstract

When we talk about music, we tend to view it as an art of sound. However, it is much broader than that. We watch music performances, read musical scores, and memorize lyrics of songs. The visual aspect of musical performances helps express the ideas of performers and engage the audience, while the symbolic aspect of music connects composers with performers across continents and centuries. A successful intelligent system for music analysis should model all these modalities and their relations. This is exactly the objective of my research, multi-modal analysis of music performances. This is at the core of artificial intelligence: It bridges computer audition and computer vision, and connects symbolic processing with signal processing. It will enable novel multimedia information retrieval applications and music interactions for experts and novices alike.

Fundamentally, there are two problems to be addressed: 1) To coordinate the multiple modalities; 2) To leverage this coordination to achieve music analysis tasks that are challenging or impossible by analyzing each modality alone.

The first problem of identifying coordination can be addressed as temporal alignment of audio and music score which are represented in different time units (seconds vs. beats), and spatial source association in videos of ensemble performances, which identifies the affiliation between the players from the visual scene and the audio/score tracks. For temporal alignment I focus on real-time audio-to-score alignment for piano performance with the sustained effect as one of the most challenging cases. For source association I address the problem in chamber ensemble (one player for a track) for common Western instruments including strings, wood-wind, and brass.

For the second problem, I conduct research to prove that the coordination of the multiple modalities of music performance benefits traditional music information retrieval (MIR) tasks and enables new frontiers of emerging research topics. I design multi-modal systems that leverage the visual information to help estimate and stream pitches in polyphonic music from ensemble performances, and to help analyze performance expressiveness (e.g., vibrato characteristics). This concept is implemented in string ensembles and witnesses a great success. I also propose to address the source separation problem on singing performance, and improve the vocal separation quality by incorporating the visual modality, e.g., the mouth movement of the singer. Last but not the least, I propose new topics about expressive visual performance generation, such as generating expressive body movements of pianists given music scores.

Contributors and Funding Sources

This work was supported by a dissertation committee consisting of Professor Zhiyao Duan (advisor), Gaurav Sharma, Mark Bocko of the Department of Electrical and Computer Engineering, and Professor Chenliang Xu of the Department of Computer Science.

The work in Chapter 2 was collaborated with Xinzhaoy Liu and Karthik Dinesh, who took charge of the dataset recording process and video editing respectively, and supervised by Professors Zhiyao Duan and Gaurav Sharma. The work conducted in Chapter 3 was completed by the student independently supervised by Professor Zhiyao Duan. The work in Chapter 4 was collaborated with Karthik Dinesh and supervised by Professors Zhiyao Duan, Chenliang Xu, and Gaurav Sharma. The work in Chapter 5 was collaborated with Karthik Dinesh and supervised by Professors Zhiyao Duan and Gaurav Sharma. The work in Chapter 6 was collaborated with Dr. Yuxuan Wang (from Bytedance Corporation) and supervised by Professor Zhiyao Duan. The work in Chapter 7 was collaborated with Dr. Akira Maezawa (from Yamaha Corporation) and supervised by Professor Zhiyao Duan.

Graduate study was supported by the start-up funds of Professor Zhiyao Duan and National Science Foundation grants No. 1741472. and No. 1846184.

List of Tables

2.1	Summary of several commonly used music performance datasets for music transcription, source separation, audio-score alignment, and multi-modal music analysis. \star): Only 54 seconds are publicly available.	19
2.2	Subjective ranking results of the synchronization quality of the three datasets provided by eight subjects.	44
4.1	The number of pieces for different instrument arrangements from the original and expanded URMP dataset.	140
4.2	The number of evaluation samples with different length and instrumentation for source association.	145
5.1	Results of video-based Play/Non-play detection and MPE accuracy of the 11 test pieces.	169

List of Figures

2.5	Synchronization quality for individual pieces in the URMP, Bach10, and WWQ dataset assessed by onset time deviation for score-notated simultaneous notes. On average, the synchronization quality is ranked as WWQ>URMP>Bach10.	43
2.6	Spatial resolution of ROIs (face, hand, mouth) where most musician-instrument interactions take place.	45
2.7	Comparison between URMP and Bach10 for multi-pitch analysis.	49
2.8	Comparison between URMP and Bach10 for score-informed source separation. Colors encode overlapping categories for easier reference.	51
2.9	Comparison between the proposed visually informed method (white) and the audio-based method (gray) on the multi-pitch analysis task. For each boxplot, the mean and standard deviation values are listed above the plot. Results are reproduced from (Dinesh et al., 2017).	53
2.10	Video-based vibrato note detection and parameter analysis results, reproduced from (Li, Dinesh, Sharma and Duan, 2017). For each boxplot, the mean and standard deviation values are listed above the plot.	55

3.1 Illustration of the sustained effect and the audio-score mismatch problem it causes. The gray notes are extended in the audio waveform longer than their notated length in the score.	71
3.2 Statistics of two causes of the sustained effect. (a) Distribution of the 60 acoustic pieces in the MAPS dataset (Emiya, Badeau and David, 2010) according to the degree of pedal usage. Pedal-down time denotes the percentage of the performing time when the sustain pedal is depressed. (b) Reverberation time of three famous concert halls: Musikvereinsaal in Vienna, Symphony Hall in Boston, and Carnegie Hall in New York.	73
3.3 Chromagrams calculated without (a) and with (b) spectral subtraction. Sustained sounds after onsets (e.g., that marked by the rectangle) are reduced by the sustained-sound reduction operation.	82
3.4 Illustration of the audio-score mismatch reduced by the proposed sustained-sound reduction operation, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation. The blue patch in (b) and (c) indicates the current frame (n -th frame) which lies in the operation region.	83

3.5 Illustration of the spectral peak removal idea. The m -th frame is the reference frame and the n -th frame is a frame under the removal operation. (a) Audio representations in the pianoroll format before and after peak removal. (b) Magnitude spectra with detected spectral peaks in the m -th and n -th frames. Peaks marked by crosses correspond to the first two notes. Peaks marked by circles correspond to the latter two notes.	84
3.6 Illustration of the negligible effect of the proposed sustained-sound reduction operation when there is no sustained sound, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation.	88
3.7 Illustration of the new audio-score mismatch introduced by the proposed sustained-sound reduction operation when the sustained sound is not due to the sustained effect but is notated in the score, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation.	89
3.8 Illustration of the proposed sustained-sound reduction operation when the onset is a false positive, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation. The audio representation in the rectangular region is likely to preserve some energy of the notes after the spectral-subtraction operation but can be totally blank after the peak-removal operation.	90

3.9	The score following framework adopted to implement the proposed sustained-sound reduction approach.	92
3.10	Statistical MIDI information of the testing dataset. (a) Scatter plot of Sustained-effect Frame Rate of the three testing groups, calculated from MIDI performances. Each dot represents one piece. (b) Three measurements of piece complexity of the three testing groups, calculated from MIDI scores. The central circles and vertical bars denote the means and standard deviations.	96
3.11	Boxplot of the Sustained-effect Region Length (SRL) for all unique note onsets (excluding those with a zero SRL.) in pieces in $P1$ (283 onsets), $P2$ (1680 onsets), and $P3$ (2275 onsets). Outliers are displayed in a more dispersed way for better visualization. Numbers above each box show the median/mean value of all the points.	97
3.12	Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (<i>Chr</i>) and the spectral-peak representation (<i>Pk</i>). The number above each box shows the median. .	100
3.13	Onset detection results averaged over pieces within each recording group. Note that the three curves do not necessarily intersect at the same point because they are average values. . .	102

- 3.14 Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*) when the ground-truth onsets are used. The number above each box shows the median. 103
- 3.15 Score following performances (average of AR) of the baseline system and the proposed system built on artificially controlled onset detection results on *P2*. Precision is fixed at 100% in the middle panel while recall is fixed at 100% in the right panel. 104
- 3.16 Boxplot of the Average Onset Deviation of the proposed approach on artificially controlled onset detection results on the group *P2*, where recall is fixed at 100% and precision is varied. Numbers above show the medians. 105
- 3.17 Score following performance (average of AR) of the baseline system and the proposed system with different lengths of the sustained-sound reduction operation region for the three groups. 107
- 3.18 Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*) when reverberation is added to the audio. The number above each box shows the median. 108

3.19 Onset detection results averaged over pieces within each recording group with reverberation added. Note that the three curves do not necessarily intersect at the same point because they are average values.	109
4.1 Outline of the proposed universal source association system for chamber ensemble performances. Three types of motion are modeled and correlated with the audio and score in three modules.	119
4.2 Body motion extraction. Upper body skeletons (second row) are extracted with OpenPose (Cao et al., 2017) in each video frame (first row) followed by temporal smoothing over time. . .	124
4.3 Example correspondence between body motion and note onsets. Top: temporally aligned score track with onsets marked by red circles. Middle: extracted motion salience (primarily bowing motion) from the visual performance of a violin player. Bottom: derived onset likelihood curve from the motion salience.	127
4.4 Optical flow visualization of finger motions in five consecutive frames corresponding to note changes. The color encoding scheme is adopted from (Baker et al., 2011).	132

4.5 Example correspondence between finger motion and note onsets of a flute player. Top: temporally aligned score track with onsets marked by red circles. Bottom: extracted motion flux from finger movements.	133
4.6 Optical flow visualization of the hand motions corresponding to vibrato articulation. Color encoding scheme is adopted from (Baker et al., 2011).	134
4.7 The same segment of normalized pitch contour $f(t)$ (green) overlaid with the motion displacement curve $d(t)$ (black) from the associated track (left) and another random track (right). .	136
4.8 Onset overlap rate for each piece from the original URMP dataset.	141
4.9 Onset detection evaluation results from body motions (a) and finger motions (b) for different instruments.	142
4.10 Source association accuracy only using onset correspondence between score tracks and body motions (the first component \bar{M}_b in Equation (4.5)).	150
4.11 Source association accuracy only using onset correspondence between score tracks and finger motions (the second component \bar{M}_f in Equation (4.5)).	151

4.12 Source association accuracy of string ensembles by (a) only using vibrato correspondence between pitch fluctuation and hand motion (\bar{M}_v in Equation (4.5)), and (b) combining vibrato correspondence with onset correspondence from body motion in (\bar{M}_b and \bar{M}_v in Equation (4.5)).	152
4.13 Source association accuracy of ensembles with different instrumentation using all of the three modules: onset correspondence from body motions, onset correspondence from finger motions, and vibrato correspondence from hand motions (Equation (4.5)).	153
5.1 Proposed framework for enhancing multi-pitch analysis using video-based play/non-play activity detection.	157
5.2 Video analysis for P/NP activity detection.	159
5.3 Sample frame from the performance video (left) and the player detection results (right) with the detected high motion regions in green, detected players in white and the background in black.	161
5.4 Boxplot of MPE accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.	170

5.5	Boxplot of MPS accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.	171
5.6	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 work.	174
5.7	System overview of the proposed video-based vibrato detection and analysis framework.	178

5.8	Audio-based vibrato detection. Detected vibrato notes are marked with green rectangles in the pitch trajectories estimated by score-informed pitch estimation.	181
5.9	Motion capture results from left hand tracking (left), color encoded pixel velocities (middle), and scatter plot of frame-wise refined motion velocities (right).	183
5.10	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.	190
5.11	Vibrato detection performance decreases as polyphony increases for audio-based methods, while it stays the same for video-based methods.	192
5.12	Vibrato detection performance decreases when the fundamental frequency decreases for audio-based methods, while it stays the same for video-based methods.	193
5.13	Distribution of vibrato rate and extent estimation error on all notes of all tracks.	194
5.14	Distributions of vibrato rate and extent for different instruments.	195
5.15	Distributions of vibrato rate and extent of four different violin players.	196

6.1	The proposed model structure. Dashed arrows denote the concatenation operation. Downsample/upsample are applied to both time and frequency dimensions in the outer layers (marked by *), while they are only applied to the frequency dimension in the inner layers.	205
6.2	A sample photo and floor plan of the sound booth for the recording process of the URSing dataset.	213
6.3	Examples of video frames of the URSing dataset and cropped mouth region pictures as the input to the video branch of the proposed method.	215
6.4	The SDR (dB) comparison on separated solo vocals with different methods.	220
6.5	The SI-SDR (dB) comparison on the separated solo vocal from different methods.	220
6.6	One 10-sec example comparing audio-based separation (MM-DenseLSTM) with audiovisual separation (proposed) on a song excerpt with strong backing vocals. The four spectrograms from top to bottom are original mixture, ground-truth vocal, audio-based vocal separation, and audiovisual vocal separation. This sample result has 10-sec long, and one mouth frame of each second is attached.)	223

6.7	The SDR (dB) comparison on the separated solo vocal from the audiovisual method using different video front-end models. “v+” denotes for songs with backing vocals.	227
6.8	The SDR (dB) comparison on the separated solo vocal of MM-DenseLSTM audio-based baseline, the proposed audiovisual method, and the proposed method taking two kinds of irrelevant visual inputs (white noise and mouth region video of another singer). “v+” denotes for songs with backing vocals. .	229
6.9	One sample frame of an a cappella song for subjective evaluation.	230
6.10	Statistics of the 26 subjects’ musical background related to the subjective evaluation.	231
6.11	The subjective ratings of the separation quality in response to the three questions. Each error bar shows mean \pm standard deviation.	231
7.1	The proposed network structure.	240
7.2	The CNN structures and parameters for feature extraction from the (a) MIDI note stream and (b) metric structure information.	242
7.3	The LSTM network structure for body movement generation. .	243
7.4	The two constraints applied during training.	246
7.5	Several generated unnatural skeleton samples without the limb constraint.	247

7.6	One sample frame of the assembled video for subjective evaluation.	251
7.7	Subjective evaluation on expressiveness and naturalness of the generated and human skeleton performance videos. The tracks with significant different ratings are marked with “*”.	256
7.8	The two typical failure cases.	257

Chapter 1

Introduction

Music is universal in all human cultures. People create, play, and enjoy music for fun, for communication, and for healing. When we talk about music, we tend to view it as an art of sound. However, it is much broader than that. We watch music performances; we read musical scores; we memorize lyrics of songs. The visual aspect of musical performances helps express performers' ideas and engage the audience. The symbolic aspect of music connects composers with performers across continents and centuries. A successful intelligent system for music understanding should model all these modalities and their relations. This is exactly the objective of my research: multi-modal analysis of music for better music interactions.

1.1 Different Modalities of Music

The development of recording technologies, starting with Thomas Edison's invention of the phonograph in 1877, has extended music enjoyment beyond

live concerts. For a long time, the majority of music recordings were distributed through various kinds of media that carry only the audio, such as vinyl records, cassettes, CDs, and mp3 files. As such, existing research on music analysis, processing, and retrieval focuses on the audio modality that represents information from the acoustic rendering of music, while other modalities such as visual component was largely forgotten.

About a decade ago, with the rapid expansion of digital storage and internet bandwidth, video streaming services like YouTube gained popularity, which again significantly influenced the way people enjoy music. In addition to listening to the sound, audiences also want to watch the performance. In 2014, music was the most searched topic on YouTube, and 38.4% of YouTube views were from music videos¹.

The visual modality plays an important role in music performances. Guitar players learn new songs by watching how others play online. Concert attendees move their gazes to the soloist in a jazz concert. In fact, researchers have found that the visual component is not just a marginal phenomenon in music perception, but an important factor in the communication of meanings (Platz and Kopiez, 2012). Even for prestigious classical music competitions, researchers have found that visually perceived elements of the performance, such as gesture, motion, and facial expressions of the performer, affect the judge's (experts or novice alike) evaluations, even more significantly than the sound (Tsay, 2013).

¹<http://expandedramblings.com/index.php/youtube-statistics/4/>

Beyond audio and video, symbolic representations are also important components of music. It includes music score, a manuscript or printed form of a musical work that uses modern musical symbols to indicate the pitches (melodies), rhythms or chords of a song, or instrumental musical piece. It is the basic form in which Western classical music is notated so that it can be learned and performed by solo singers or instrumentalists or musical ensembles, across continents and centuries. As the proliferation of computer programs music score can be notated electronically (e.g., MIDI or XML format) and rendered on screen, and in some case, played back using synthesizer or virtual instruments. A large music score database is accessible online at IMSLP².

1.2 Motivation

1.2.1 Impact on MIR Research

Traditional MIR research mainly focuses on audio and symbolic modalities. The visual modality is much more natural, and when available, it can be very helpful for solving many MIR tasks that are challenging using an audio-only approach. Zhang et al. (2007) introduced a method to transcribe solo violin performances by tracking the violin strings and fingers from the visual scene. Similarly, remarkable success has been demonstrated for music tran-

²International Music Score Library Project, accesible at: <https://imslp.org>

scription using visual techniques for piano (Akbari and Cheng, 2015), guitar (Palleari et al., 2008) and drums (Gillet and Richard, 2005). Other MIR tasks include onset detection (Bazzica et al., 2017) and vibrato analysis (Li, Dinesh, Sharma and Duan, 2017). Note that the benefits of incorporating visual information in the analysis of audio are especially pronounced for highly polyphonic, multi-instrument performances, because the visual activity of each player is usually directly observable (barring occlusions), whereas the polyphony makes it difficult to unambiguously associate audio components with each player. Dinesh et al. (2017) proposed to detect play/non-play activity for each player in a string ensemble to achieve improved multi-pitch estimation and streaming results than audio-based methods. A similar idea was applied on different instrument groups among symphony orchestras to achieve performance-score alignment (Bazzica, Liem and Hanjalic, 2014). Gao, Feris and Grauman (2018) proposed to learn the mapping between audio frequency bases and individual visual object to guide audio source separation. Similarly, in (Zhao et al., 2018), a deep cross-modal feature was learned to separate sound for any given region from the video.

Other than traditional MIR research topics, novel research directions are performed when multi-modal music performance data is available, such as fingering analysis for guitarists (Burns and Wanderley, 2006; Kerdvibulvech and Saito, 2007; Radicioni, Anselma and Lombardo, 2004; Scarr and Green, 2010) and pianists (Gorodnichy and Yogeswaran, 2006; Oka and Hashimoto, 2013), conductor’s baton trajectories analysis (Murphy, 2003), audio-visual

source association in chamber ensemble (Li, Dinesh, Xu, Sharma and Duan, 2019), and cross-modal generation (Chen et al., 2017; Li, Maezawa and Duan, 2018; Shlizerman et al., 2018).

1.2.2 Towards Applications

There are plenty of applications that need the underlying techniques to coordinate and joint-model different modalities of music performances. Here I describe them according to different application scenarios.

1.2.2.1 Music Education

Imagine a music tutoring system for music learners. It follows the player's live performance and displays a dynamic progress bar on the music score. It compares the performance with the score notations to rate the tuning, dynamics, and rhythm. Given a music score, it automatically gives annotations of fingering, by either learning from the symbolic context itself, or visually analyzing musicians' hand movements from online video resources. It can also demonstrate the visual performance for a given score, such as hand placement on keyboard, up/down-bow for string instruments, or even expressive whole body movements.

1.2.2.2 Music Entertainment

An AI system could render the expressiveness into a symbolic music, such as a rigid MIDI score downloaded from IMSLP, after learning styles from human

musicians. The expressiveness includes delicate control of sound articulation, dynamics, and tempos in the audio modality, or even body performance in the visual modality. Then the system can create immersive audio-visual music performance by transferring the learned performance style from famous musicians, e.g., Lang Lang, to any rigid symbolic music score downloaded from a public music score database (e.g., IMSLP), to build an immersive home entertainment system.

If you are an instrument player and you are playing with an accompaniment system, the smart system with multi-modal analysis functions built in is able to understand your performance from both audio and vision. One benefit is to predict your tone onsets from your body motions and respond promptly, which is not possible by only capturing audio streams. On the other hand, generating a visual performance of the accompaniment part favors better music interactions, just like the ones between human soloist and human accompaniments in real performances.

1.2.2.3 Concert

Now you are sitting at a chorus concert. An intelligent system would be able to record the audio stream from the live performance and scroll the corresponding lyrics for you at the correct timings in real-time. This greatly enhances the understanding and listening experience of the chorus performance. Or you are sitting at a classical chamber or orchestra performance, an audio-visual system can always take the live video streams and show you

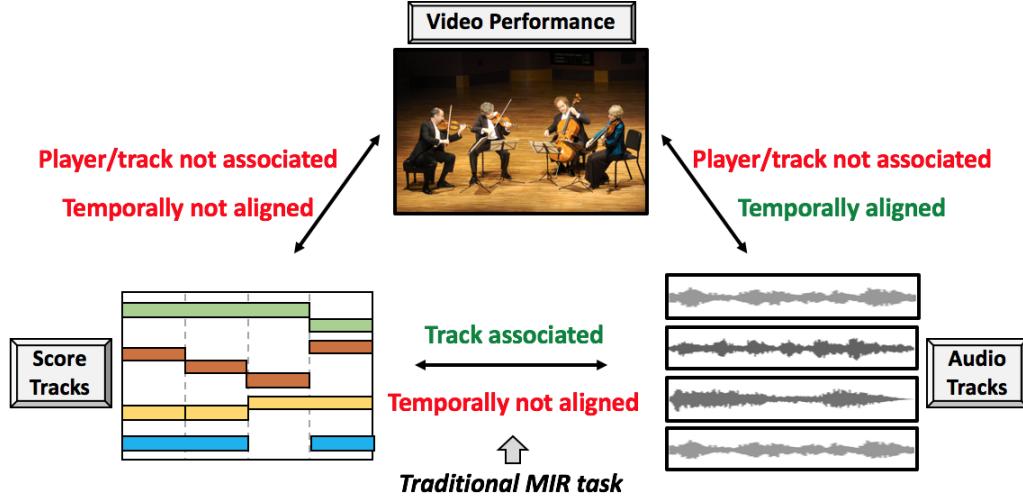


Figure 1.1: The coordinations between different modalities of music.

which player is playing which track. A live broadcast camera with this function built in can automatically focus on a desired track, e.g., the soloist.

1.3 Problem Statement

There are two fundamental problems in multi-modal music analysis: 1) How to coordinate the multiple modalities that are so different at the surface level but are inherently related at much deeper levels? 2) How to leverage this coordination to achieve enhanced music analyses that are impossible by analyzing each modality alone?

1.3.1 Coordination of Multiple Modalities

Considering the three previously discussed modalities in Section 1.1 (e.g., audio, video, and score), there are different kinds of coordinations need to be addressed. Figure 1.1 demonstrates the relationships between the three modalities, where the coordinations that need to addressed are in red text. Between audio and score, it is mainly the temporal alignment as they are represented in different time units (seconds vs. beats). Between audio and video, it is mainly the association between sound sources and players in the visual scene for ensemble performances. Between score and video, both the temporal alignment and the source association are important.

The coordination between audio and score, i.e., *audio-score alignment*, is an active research for decades. The online version of this problem (also known as *score following*) is more challenging but can support more applications. Although score following has been also well explored in the MIR community, there are still some unsolved issues. Most of the challenges come from the mismatch between score notations and the deductive performance, considering trills, staccatos, vibrato, sustain pedal usage, etc. One example is the *sustained effect* in piano music, where the waveform of a note lasts longer than what is notated in the score caused by expressive performing styles such as legato articulation or the usage of the sustain/sostenuto pedals, or the reverberation in the recording environment. This can be addressed by modifying the audio feature representation for sustain sound reduction.

The coordination considering visual modality in ensemble performances,

i.e., *source association*, aims at identifying the affiliation between players and sound sources or score tracks. It is a fundamental problem in many kinds of audio-visual processing tasks, especially in polyphonic setting when multiple sound sources are active at the same time. However, it is rarely addressed in previous literature of MIR due to lack of audio-visual music datasets. Once a music dataset with both auditory and visual modalities is available, the source association can be addressed by analyzing the cross-modal correlations, e.g., joint recognition of the instrument object from vision and the corresponding sound characteristics. A more challenging case can be addressed when considering two or more instances of the same instruments, say violin duets, where modeling the temporal correspondence is required, e.g., a bow motion of the violinist and the corresponding tone onsets.

1.3.2 Multi-Modal MIR

Assuming the multiple modalities of a music performance are all coordinated, we explored what kinds of research problems can be conducted to advance traditional MIR tasks, or even favor the novelty and diversity of emerging MIR tasks. Typical tasks vary from content-level music analysis such as pitch, onsets, to expressiveness analysis such as articulation, tempo, performance styles, or even music synthesis and generation. Here we take some examples to formulate some typical multi-modal MIR problems.

1.3.2.1 Visually-Informed Multi-Pitch Analysis

Multi-pitch analysis of polyphonic music requires estimating concurrent pitches (estimation) and organizing them into temporal streams according to their sound sources (streaming), both are fundamental problems in MIR and computer audition. This is challenging for approaches based on audio alone due to the polyphonic nature of the audio signals. A multi-modal system can leverage visual information of musical instrument players to help solve the multi-pitch analysis problem. Related information includes hand placement on keyboard, finger maneuver on fretboard, play/non-play activity for each individual from ensembles, etc.

1.3.2.2 Visually-Informed Performance Expressiveness Analysis

Visual performance including gestures, motions, and facial expressions of the performer is strongly related to the expressiveness. Some of these directly result in sound articulations, especially for string instrumentalists when the player directly controls the velocity and pressure of the bowing or fingering force. We take the vibrato articulation as an example, which is an important artistic effect contributing to performance expressiveness. It refers to the pitch modulation of a note in a periodic fashion, and can be characterized as vibrato rate and extent. Automatic vibrato detection and analysis is an important topic in MIR, but previous audio-based methods (though some are score informed) are only robust on monophonic audio or one melody

from mixture. For string instruments, vibrato is visible from the left hand motion, and this visual information does not degrade as audio information does when polyphony increases. This brings possibilities to analyze vibrato directly from polyphonic music such as string ensembles by analyzing the fine motion of the left hand for each instrumentalist. Score tracks, after aligned with the performance and associated with the players, helps to segment the polyphony stream into note samples as instances for vibrato detection and analysis.

1.3.2.3 Audiovisual Music Source Separation

Source separation is an interesting research topic of computer audition for decades, and music source separation is to isolate the original music into different audio tracks that correspond to different music components, e.g., instruments. In music performance videos, a multi-modal system can analyze the correspondence between performers' visual performances and audio components and improve the separation quality for the target performer. Visual information is especially beneficial when the different sound components to be separate share similar acoustic features. For example, analyzing a solo singer's mouth movements to separate his/her solo voice from a choir, or analyzing a solo violinist's bowing/fingering motions to separate the solo violin sound from a string orchestra.

1.3.2.4 Visual Performance Generation

In addition to analyzing the visual modalities, another emerging research topic is to generate visual performances from other modalities (e.g., a music score, or a symbolic information of note event sequence in MIDI format). Most previous systems formulate an inverse kinematics problem with physical constraints to generate hand shapes and finger positions, but cannot incorporate musical context information to generate expressive movements of the body joints that further away from the fingers such as head and shoulders. When the visual performance is coordinated with the symbolic data, we can directly train a system to learn this visual expressiveness from real human performance that maps MIDI note stream to a skeleton sequence of the player.

1.4 Summary of Contributions

In this section, I summarize the main contributions of my thesis.

- In Chapter 2, I introduce the University of Rochester Multi-modal Music Performance (URMP) dataset, where I participated in the creation process, post processing, comprehensive analysis, and potential research topics that can be applied on the dataset. We anticipate that the dataset will be further useful for evaluating all the possible audio-visual techniques such as music source separation, transcription, and

performance analysis, etc.

- In Chapter 3, I develop the first system to address the sustained effect for in score following systems for piano performances. It introduces a sustained sound reduction operation on different commonly used audio feature representation and can be built on a state-of-the-art score following framework. Experiments show the proposed approach outperforms previous work.
- In Chapter 4, I develop the first system to address the source association problem for chamber music ensembles. The system provides a universal framework for all common instruments in Western chamber ensembles by automatically and adaptively integrating the different learned correspondence between visual performance and score/sound events, without requiring prior knowledge of the instrument types. Experiments conducted on the URMP dataset demonstrate promising results of association accuracy.
- In Chapter 5, two systems are described to leverage visual information to augment traditional MIR tasks of multi-pitch analysis and vibrato analysis, respectively, for string ensembles from the URMP dataset. The former one detects the play/non-play label on each individual player to guide audio-based methods, which improves the multi-pitch estimation (MPE) and multi-pitch streaming (MPS) accuracy. The latter one analyzes the left hand rolling motion for each player to an-

alyze the vibrato performance instead of analyzing the pitch contours transcribed from audio, which is beneficial especially in high polyphony scenarios.

- In Chapter 6, I address the singing voice separation problem. I develop a system to incorporate the visual analysis of the solo singer’s mouth movement to improve the separation quality of the singer’s voice. I also create the University of Rochester Singing Performance (URSing) dataset to help with the validation of the research. This is the first audiovisual singing performance dataset where singers’ sing with accompaniment tracks while the solo voice is also available (recorded in isolation). Experiments show that the proposed audiovisual method outperforms traditional audio-based method in most evaluation scenarios and this advantage is especially pronounced when the goal is to separate the solo vocal from accompaniments with backing vocal components, e.g., A Cappella songs, which poses a great challenge for audio-only methods.
- In Chapter 7, I develop the first system to take the input of symbolic performance as MIDI note sequence and the corresponding metric structure, and generate the visual performance as upper body key points to simulate a human pianist playing this piece. Subjective evaluations show that the system is capable of learning human’s visual performance in terms of both naturalness and expressiveness.

Chapter 2

URMP Dataset Creation

We introduce a dataset for facilitating audio-visual analysis of music performances. The dataset comprises 44 simple multi-instrument classical music pieces assembled from coordinated but separately recorded performances of individual tracks. For each piece, we provide the musical score in MIDI format, the audio recordings of the individual tracks, the audio and video recording of the assembled mixture, and ground-truth annotation files including frame-level and note-level transcriptions. We describe our methodology for the creation of the dataset, particularly highlighting our approaches for addressing the challenges involved in maintaining synchronization and expressiveness. We demonstrate the high quality of synchronization achieved with our proposed approach by comparing the dataset with existing widely-used music audio datasets.

We anticipate that the dataset will be useful for the development and evaluation of existing music information retrieval (MIR) tasks, as well as for novel multi-modal tasks. We benchmark two existing MIR tasks (multi-

pitch analysis and score-informed source separation) on the dataset and compare with other existing music audio datasets. Additionally, we consider two novel multi-modal MIR tasks (visually informed multi-pitch analysis and polyphonic vibrato analysis) enabled by the dataset and provide evaluation measures and baseline systems for future comparisons (from our recent work). Finally, we propose several emerging research directions that the dataset enables.

2.1 Introduction

Music Information Retrieval (MIR) research, which traditionally focused on audio and symbolic modalities (e.g., musical scores), started to pay attention to other modalities in recent years. For example, players' motion data captured from sensors (MoCap) have been used to analyze players' activities (Caramiaux, Wanderley and Bevilacqua, 2012) and enhance source separation quality (Parekh et al., 2017). However, such data is not easy to obtain from natural music performances because the methodology requires specialized motion capture sensors.

Despite the increased recent interest in multi-modal analysis of music performances, progress in jointly using audio and visual modalities for the analysis of music performances has been rather slow. One of the main reasons, we argue, is the lack of datasets. Although music takes a large share among all kinds of multimedia data, music datasets are scarce. This is because a

music dataset should contain not only music recordings but also ground-truth annotations (e.g., note/beat/chord transcriptions, performance-score alignments) to enable supervised machine learning and the evaluation of proposed methods. Due to the temporal and polyphonic nature of music, the annotation process is very time consuming and often requires significant musical expertise. Furthermore, for some research problems such as source separation, isolated recordings of different sound sources (e.g., musical instruments) are also needed for ground truth verification. When creating such a dataset, if each source is recorded in isolation, it is a challenging task to ensure that different sources are well tuned and properly synchronized.

In this chapter, we present the University of Rochester Multi-modal Music Performance (URMP) dataset. This dataset covers 44 classical chamber music pieces ranging from duets to quintets. For each included piece, the URMP dataset provides the musical score, the audio recordings of the individual tracks, the audio and video recordings of the assembled mixture, and ground-truth annotation files including frame-level and note-level transcriptions. In creating the URMP dataset, a key challenge we encountered and overcame was the synchronization of individually recorded instrumental sources of a piece while maintaining the expressiveness seen in professional chamber music performances. We present our attempts and reflections on addressing this challenge, and hope that this will shed light on similar issues for future dataset creation efforts. We also conduct objective and subjective evaluations on the synchronization quality and compare it with two

widely used datasets. As the first audio-visual multi-instrument multi-track music performance dataset, we anticipate that it will be valuable for MIR research. Therefore, we benchmark URMP with existing widely used music audio datasets on important existing tasks. We also highlight our previous work on URMP to define two novel multi-modal MIR tasks by proposing evaluation measures and providing baseline systems for comparison. We further propose several emerging novel research directions that URMP may support.

In the rest of the chapter, in Section 2.2, we first review existing music performance datasets for MIR tasks and especially highlight the challenges involved in creating multi-track datasets. Then, in Section 2.3, we describe our different attempts aimed at overcoming these challenges while recording the URMP dataset and, in Section 2.4, elaborate on the approach adopted. In Section 2.5, we describe the content of the URMP dataset and analyze the quality of the dataset. In Section 2.6, we compare the URMP dataset with other existing music audio datasets by benchmarking two pre-existing audio-only MIR tasks on the datasets and also mention several novel multi-modal MIR tasks enabled by the multi-modal data in URMP. Finally, we conclude the chapter in Section 2.7.

2.2 Review of Music Performance Datasets

Music performance datasets are not easy to create because recording music performances and annotating them with ground-truth labels (e.g., pitch,

Name	Instrument/Genre	# Pieces	Total Dura- tion	Content
<i>Audio-modality, Single-track</i>				
MAPS	Piano	270	18.6 h	Audio, Note annotation
LabROSA	Piano	130	2.7 h	Audio, Note annotation
Score-informed Transcription	Piano	7	6.4 m	Audio, Note annotation, Performance error annotation
RWC (Subset)	Multi-instrument	100	9.2 h	Audio, Note annotation
Su et al.	Multi-instrument	10	5 m	Audio, Note annotation
<i>Audio-modality, Multi-track</i>				
MedleyDB	Multi-genre	122	7.3 h	Audio, Pitch contour, Instrument activity, Genre label
SSMD	Songs	104	6.8 h	Audio, Structure annotation
MASS	Songs, Multi-genre	6	4.8 m	Audio, Lyrics
Mixploration	Multi-genre	12	4.9 m	Audio
iKala	Songs	252	2.1 h	Audio, Lyrics, Pitch contour
WWQ	Multi-instrument	1	9 m*	Audio, Note annotation
TRIOS	Multi-instrument	5	3.2 m	Audio, Note annotation
Bach10	Multi-instrument	10	5.5 m	Audio, Note annotation, Pitch contour
PHENICX-Anechoic	Multi-instrument	4	10.6 m	Audio, Note annotation
<i>Multi-modality, Single-track</i>				
Multi-modal Guitar	Guitar	10	10 m	Audio, Video
C4S	Clarinet	54	4.5 h	Audio, Video, Visual annotation
<i>Multi-modality, Multi-track</i>				
ENST-Drums	Drum kit	N/A	3.75 h	Audio, Video (multi-camera views)
Abeßer et al.	Guitar, Drum, Bass	N/A	1.2 h	Audio, Video (multi-camera views)
EEP	String quartet	23	N/A	Audio, Note annotation, Bow MoCap data
URMP	Multi-instrument	44	1.3 h	Audio, Video, Note annotation, Pitch contour

Table 2.1: Summary of several commonly used music performance datasets for music transcription, source separation, audio-score alignment, and multi-modal music analysis. *): Only 54 seconds are publicly available.

chord, structure, and mood) require musical expertise and are very time consuming. Commercial recordings can generally not be used due to copyright issues. Recording music performances in research labs is subject to the availability of musicians and recording facilities. Also, when different instruments are recorded in isolation (for evaluating musical source separation), we need to ensure proper methods for synchronization. The annotation process often requires experienced musicians to listen through the musical recording multiple times. It is especially difficult when the annotations are numerical and at a temporal resolution on the order of milliseconds (for evaluating pitch transcription, audio-score alignment, etc.). As a result, music performance datasets are scarce and their sizes are relatively small.

2.2.1 Existing Datasets

In this section, we briefly review several commonly-used music performance datasets that are closely related to the URMP dataset, which can support some MIR tasks like music transcription, source separation, audio-score alignment, etc. A summary of these datasets is provided in Table 2.1. Most of the datasets contain only audio, and only six are multi-modal.

The first group of datasets are single-track polyphonic recordings with MIDI transcriptions for music transcription research. While recording this type of music is straightforward, obtaining the ground-truth transcription is not. A large portion of existing single-track datasets focus on piano music (Emiya, Badeau and David, 2010; Poliner and Ellis, 2007; Benetos, Klapuri

and Dixon, 2012), where the transcription can be obtained automatically: a pianist plays on a MIDI keyboard to generate a MIDI performance with precise note timings and dynamics; the MIDI file is then fed to a reproducing piano (e.g., Yamaha Disklavier¹) to render acoustical recordings. The MIDI file naturally serves as the ground-truth transcription. For other instruments that do not have the MIDI-driven sound reproducing systems, manual annotation of ground-truth transcription is the most accurate approach, which, however, is notoriously labor intensive. To address this issue, the RWC dataset (Goto et al., 2002) (classical and jazz subset) aligns a reference MIDI score to the audio performance in a semi-automatic fashion, and uses the aligned MIDI score as the transcription. The dataset proposed by Su and Yang (2015) uses a different approach, where a professional pianist was employed to follow and play the music on an electric piano to generate well-aligned ground-truth transcriptions.

The second group of datasets are multi-track recordings, where each instrumental source is on one track. A multi-track dataset generally has the merit of better versatility and scalability. First, it can support more MIR tasks (e.g., source separation). Second, it significantly reduces the annotation complexity, from polyphonic to monophonic music. With a robust monophonic pitch analysis tool, fine-grained annotations (e.g., pitch height in musical cents for each time frame) can be acquired with less labor. Third, a large variety of music excerpts can be reproduced by mixing the mono-

¹<http://www.disklavier.com>

phonic tracks of the same piece with different combinations. The difficulty in creating multi-track datasets is during the recording process, which will be discussed in the Section 2.2.2.

The largest multi-track music dataset is MedleyDB (Bittner et al., 2014). It contains multi-track audio recordings of 122 pieces with various styles together with the melody pitch contour and instrument activity annotations. The second largest dataset is the Structural Segmentation Multitrack Dataset (SSMD) (Hargreaves, Klapuri and Sandler, 2012), which contains multi-track audio recordings of 104 rock and pop songs, together with structural segmentation annotations. Most recordings of MedleyDB and SSMD are from third-party musical organizations (e.g., commercial or non-profit websites, recording studios). This lessens the burden of recoding by the researchers themselves. The other multi-track datasets are of a much smaller scale. The MASS dataset (Vinyes, 2008) contains several raw and effects-processed multi-track audio recordings. Mixploration dataset (Cartwright, Pardo and Reiss, 2014) contains 3 raw multi-track audio recordings together with a number of mixing parameters. The iKala dataset (Chan et al., 2015) contains the vocal melody and the accompaniment part of 252 pop songs in separate tracks. The Wood Wind Quintet (WWQ) dataset (Bay, Ehmann and Downie, 2009) contains individual recordings of 1 classical quintet. The original 9-min recording serves as an internal benchmark for the MIREX² Multi-F0 Estimation & Tracking task since 2007; only a 54-second excerpt is

²http://www.music-ir.org/mirex/wiki/MIREX_HOME

publicly available. The TRIOS dataset (Fritsch and Plumbley, 2013) contains 5 multi-track recordings of musical trios together with their MIDI transcriptions. The Bach10 dataset (Duan, Pardo and Zhang, 2010) contains 10 multi-track instrumental recordings of *J.S. Bach* four-part chorales, together with the pitch and note transcriptions and the ground-truth audio-score alignment. The PHENICX-Anechoic (Miron et al., 2016) provides the denoised recordings and note annotations for the Aalto Anechoic Orchestral Database (Pätynen, Pulkki and Lokki, 2008), which contains four symphony pieces, each one has 8-10 instrumental parts and each part was recorded in isolation using multiple microphones.

Existing multi-modal musical datasets include the Multi-modal Guitar dataset (Perez-Carrillo, Arcos and Wanderley, 2015), the Clarinetists for Science (C4S) dataset (Bazzica et al., 2017), the ENST-Drums dataset (Gillet and Richard, 2006), the Abeßer et al. (Abeßer et al., 2011), and the Ensemble Expressive Performance (EEP) dataset (Marchini et al., 2014). The first two are single-track datasets. The Multi-modal Guitar dataset contains 10 audio-visual recordings of guitar performances. The audio was recorded using a contact microphone to capture the vibration of the guitar body and to attenuate the effects of room acoustics and sound radiation. The video was recorded using a high-speed camera with markers attached on joints of the player’s hands and the guitar body to facilitate hand and instrument tracking. The C4S dataset consists of 54 videos from 9 clarinetists, each performing 3 different classical music pieces twice. Visual annotations are also

provided including the coordinates of the face, mouth, left hand, right hand, and the clarinet. The latter three are multi-track datasets. The ENST-Drums contains mixed stereo audio tracks, audio and video recordings of each instrument of a drum kit playing different sequences. All instruments were recorded simultaneously using 8 microphones and 2 cameras. Similarly, instruments in (Abeßer et al., 2011) were also recorded simultaneously. Since different instruments were not recorded in isolation, there is some sound leakage across instrumental tracks. In EEP, each instrument was recorded using a contact microphone; while sound leakage is greatly reduced, the acoustic properties can be very different from using a near-field microphone. There are several other video datasets that contain a subset of music performance such as FCVID (Jiang et al., 2018), YouTube-8M (Abu-El-Haija et al., 2016), Google AudioSet (Gemmeke et al., 2017), etc. We do not include them in Table 2.1 since no content-level annotations are provided, which limits applications in MIR tasks.

2.2.2 Synchronization Challenges

The coordination between simultaneous sound sources differentiates music from general polyphonic acoustic scenes. One important aspect of this coordination is synchronization, which is typically accomplished in real-world music performances by players rehearsing together prior to a performance. During the performance, players also rely on auditory and visual cues to adjust their speed to other players. For large ensembles such as a symphony

orchestra, a conductor sets the synchronization.

In order to have a music performance dataset simulate real-world scenarios, good synchronization between different instrumental parts is desired. However, creating a multi-track dataset without leakage across different tracks is challenging. To eliminate leakage, different instruments need to be recorded separately, which makes it difficult to achieve synchronization because players cannot rely on interactions with other players to adjust their timing. In this subsection, we review existing approaches that researchers have explored for ensuring synchronization when recording multi-track datasets.

For SSMD, MASS, Mixploration, and a large portion of MelodyDB, recordings were obtained from professional musical organizations and recording studios instead of being recorded in a laboratory setting. The pieces are also mostly rock and pop songs, which have a steady tempo, making synchronization easier. In fact, in the music production industry, pop music is almost always produced by first recording each track in isolation and then mixing them and adding effects. This procedure, however, does not apply to classical music, which involves much less processing. Different instrumental parts of a classical music piece are almost always recorded together. This is why these datasets do not contain many classical ensemble pieces. The multi-track recordings in ENST-Drums, Abeßer et al., and EEP were recorded using multiple microphones simultaneously, hence there are no synchronization issues with the recording. However, leakage between microphones is inevitable for the first two of these datasets, and the contact microphone used in EEP

alters the acoustic properties from normal near-field microphone recordings, which makes the dataset less desirable for source separation research.

Existing multi-track datasets that have dealt with the synchronization issue when recording each track in isolation are WWQ, TRIOS, Bach10, and PHENICX-Anechoic. WWQ only has one quintet piece, and the recording process has two stages. In the first stage the performers played together with separate microphones, one for each instrument. Audio leakage inevitably existed in these recordings but they served as a basis for synchronization in the second stage. In the second stage players recorded their parts in isolation while listening to a mix of the other player's recordings in the first stage through earphones. Because these players had rehearsed together and listened to their own performance (the first-stage recordings), the synchronization among individual recordings in the second stage was very accurate. In TRIOS, for each piece, a synthesized audio recording was first created for each instrument from the MIDI score. Each player then recorded his/her part in isolation while listening to the mix of the synthesized recordings of other parts synchronized with a metronome through earphones. Although the players did not rehearse together prior to the recording, the synchronization was easy to address as all the pieces have a steady tempo. In Bach10, instead of using the synthesized recordings and a metronome as the synchronization basis, each player listened to the mix of all previously recorded parts. The first player, however, determined the temporal dynamics and did not listen to anyone else, resulting in a less-than-ideal synchronization in Bach10. Due

to significant variation in the tempo, a listener could easily find many places where notes were not articulated together. In fact, each piece contains several fermata signs, where notes were prolonged beyond their normal duration when the performance was held. For recording the dataset as annotated by PHENICX-Anechoic, a pre-recorded conducting video with a pianist playing was used to set the common timing for the instrumental players. The detailed description of the dataset creation process is presented in (Pätynen, Pulkki and Lokki, 2008). In this work, we arrived at the same synchronization approach as that in (Pätynen, Pulkki and Lokki, 2008) independently. Unlike (Pätynen, Pulkki and Lokki, 2008) where an audio-only dataset was generated, we create a multi-modal audio-visual dataset that we compare comprehensively with existing datasets and also set up performance baselines for several typical tasks on the dataset. Additionally, we also provide a quantitative assessment of alternative synchronization approaches, which has not previously been done.

2.3 Approaches to Synchronization

Similar to the creation of existing multi-track datasets, the creation of URMP dataset faced the synchronization issue. This issue is even more significant because of the following seemingly conflicting goals: 1) *Efficiency*. Our goal is to create a large dataset containing dozens of pieces with different instrument combinations. We also hope that each player could participate in the record-

ing of multiple pieces. Therefore, it would be difficult and time consuming to arrange players to rehearse together before the recording for each piece, which is the approach adopted by the creation of WWQ dataset. 2) *Quality*. We want the players to be as expressive as what they would be in real musical concerts. This requires them to vary the tempo and dynamics significantly throughout a piece. However, without the live interactions between players, this goal makes the synchronization more difficult. We tried different ways to overcome this challenge and eventually arrived at the same approach used in (Pätynen, Pulkki and Lokki, 2008) independently, which achieved both good efficiency and quality. We present our attempts here and hope that this will give some insights into the dataset creation problem. Figure 2.1 (a) summarizes our attempts. We also quantitatively evaluate the quality of several typical attempts by showing the maximal onset time deviation in Figure 2.1 (b), which calculates the maximal absolute time difference among the score-notated simultaneous notes from different tracks. The blue circles represent notes after a rest of at least 2 beats, which are more challenging to synchronize due to fewer temporal hints.

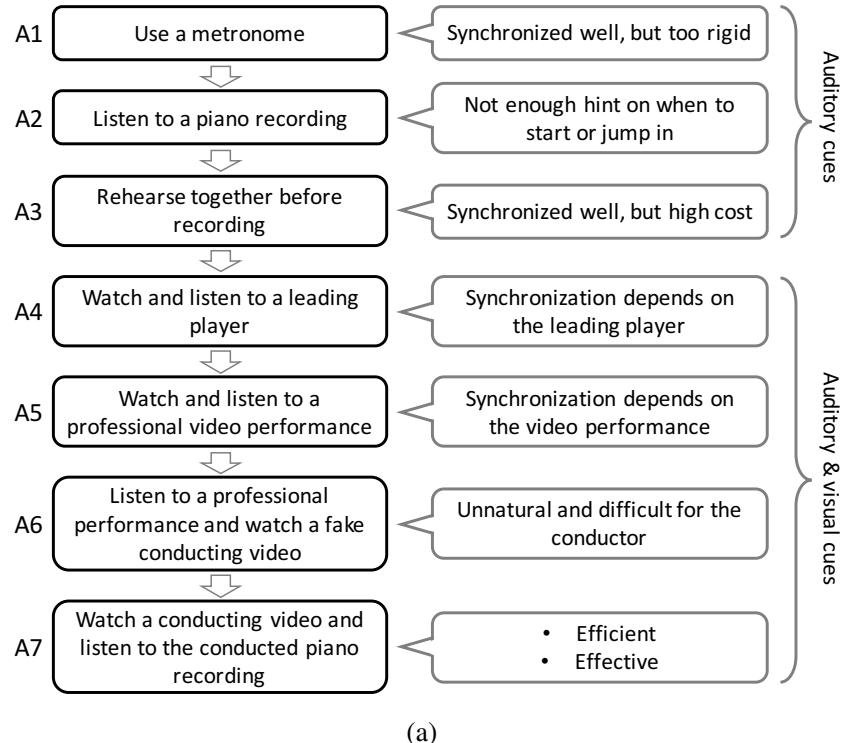
A1) The first approach that we tried was to pre-generate a beat sequence using an electronic metronome, and then have each player listen to the beat sequence through an earphone while recording his/her part. Different instrumental tracks were thus synchronized through the common beat sequence. We tested this approach on a violin-cello duet (*Minuet in G major* by J. S. Bach). Although the synchronization was good, we found that the perfor-

mance was too rigid and did not reach our desired level of expressiveness.

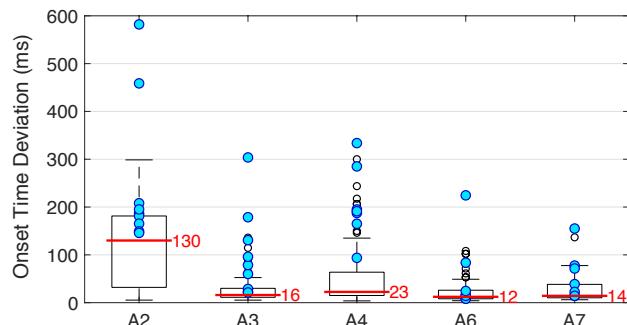
A2) In order to have better expressiveness, we replaced the beat sequence with a pre-recorded piano performance. The pianist played both instrumental parts simultaneously and varied the tempo and dynamics throughout the piece. However, when the players later followed it to record their individual parts, the synchronization was not satisfactory, as shown in Figure 2.1 (b). They did not get enough hints on when to start nor when to jump in after a long break, which resulted in some extreme outliers (shown by the blue circles).

A3) This attempt is inspired by WWQ’s approach: players rehearsed together and used their rehearsal recordings as the basis for synchronization. We further added a conductor to the rehearsal process to improve the expressiveness by varying the tempo and dynamics. The conductor also vocalized several beats before the start of the piece to signal the players. We tested on another violin-cello piece (*Melody by Schumann*). Figure 2.1 (b) shows that the synchronization quality is greatly improved with a median onset microtiming value of 16ms. However, this approach is time-consuming and difficult to scale to many pieces with different instrument combinations.

A4) In the following attempts, we aim at an approach that does not require joint rehearsals while keeping both the synchronization and expressiveness at an acceptable level. Similar to Bach10, we let one leading player record first and let the other(s) follow. Differently, the follower(s) not only listened to but also watched the first player’s recording. We tested on the



(a)



(b)

Figure 2.1: (a): A summary of all attempts for solving the synchronization issue. (b): A quantitative evaluation on the onset time deviation among score-notated simultaneous notes of several key attempts. Red bars and text show the median values. Blue circles are the notes occurring after a rest of at least 2 beats, which are not necessarily outliers.

same piece as in A1 and A2, and found that this approach improved the synchronization from A2, especially at places after long rests. This improvement, reported by the players, was mainly thanks to the visual cues displayed in the first player’s motion. This cue, however, would depend on the leading player and the arrangement of the piece. Overall, its synchronization quality is still much worse than A3.

A5) Building on the previous approach, we asked each player to watch and listen to a professional video performance downloaded from YouTube during the recording. Due to the availability of professional performances, we chose a different piece, *The Art of Fugue No. 1* by *J. S. Bach* string quartet (same for the following attempts). We tested this approach only on the violin and cello parts. Our players reported that the visual cues were not always clear, and it was challenging for them to follow the professional performance even after watching it repeatedly in advance. They were not able to complete the recording using this approach hence we could not quantitatively analyze the synchronization quality in Figure 2.1 (b).

A6) This attempt focused on relieving players’ synchronization burden by applying a professional conductor to “conduct” the YouTube video used in A5. The conducting video was pre-recorded along with the played-back audio. Figure 2.1 (b) shows that this approach achieved synchronization quality similar to the approach with joint rehearsal (A3). However, the conductor needed to practice multiple times to memorize the temporal dynamics of the performance and behave in a timely fashion following the performance.

This was very non-intuitive for conductors. In addition, it is difficult to find YouTube videos that exactly match our arrangements (e.g., instrumentation, key, and notes). Nonetheless, from this attempt, we learned that watching a conductor is more beneficial than watching other players' playing.

A7) In order to strike a balance between the burden placed on the conductor and the players, our final attempt had two key steps. In the first step, we asked the conductor to conduct a pianist performing the piece, and recorded both the conducting video and the piano audio. The conductor varied the tempo and dynamics and the pianist adjusted the performance accordingly. The conductor also gave cues to different instrumental parts in front of the camera to help players jump in after a long rest. As a second step, we asked each player to watch the conducting video as well as listen to the corresponding audio during the recording. The result from this attempt also yielded a satisfying quality. Figure 2.1 (b) shows a median onset deviation value of 14 ms, similar to A3 and A6. Without mandating a joint rehearsal among the players, this method simultaneously meets the requirements of quality, efficiency, and scalability. Furthermore, it is natural for the conductor, the pianist, and the players.

Because the onset times are ambiguous for some soft articulations, the numerical evaluation of onset time deviation in Figure 2.1 (b) is only a limited indicator of synchronization quality. Therefore during the preliminary attempts, we also valued players' subjective evaluation. To collect players' opinions on different pieces, the attempts at synchronization were not al-

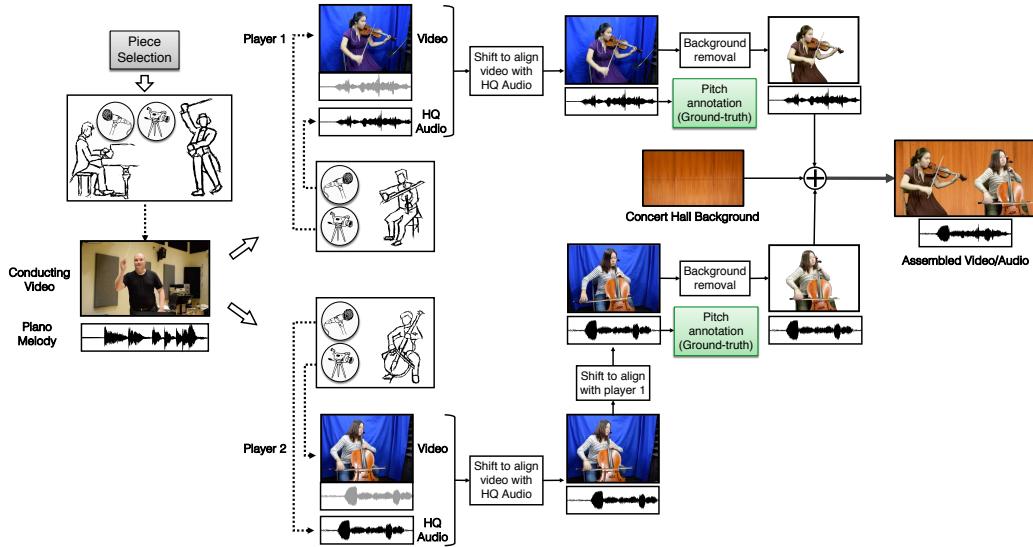


Figure 2.2: The general process of creating one piece (a duet in this example) of the URMP dataset.

ways tested on the same piece. The synchronization difficulty of the pieces is comparable thanks to their similar tempo and expressiveness, and is representative for most of the finally selected pieces in URMP.

2.4 Dataset Creation Procedure

This section explains in detail the execution of the two key steps introduced in A7 in Section 2.3. It covers the entire process of dataset creation, from piece selection and musician recruitment, to recording, post-production, and ground-truth annotation. The whole process is summarized in Figure 2.2 using a duet as an example.

2.4.1 Piece Selection

Our criteria of piece selection were:

- Generality: We want to have a good coverage of polyphony, composers, and instrumentations.
- Complexity: The pieces should be relatively simple so that all players could handle them without much practice. The duration should not be too long (ideally 1 to 2 minutes) to ease the burden of the recording process.
- Expressiveness: To avoid rigidness, the score should allow some self interpretations by the conductor or players, such as tempo rubato, dynamic variations, and ornamentations.

Bearing these guidelines in mind, we select pieces from a sheet music website³, which provides thousands of simplified and rearranged musical scores of different polyphony, styles, composers, and instrumentations. We select a number of classical ensemble pieces, covering duets, trios, quartets, and quintets. Different instrumentations include string groups, woodwind groups, brass groups, and mixed groups. Percussion instruments are not included. The pieces are simple enough so most players could play them by sight-reading or after practicing for one or two times. The durations of these pieces range from 40 seconds to 4.5 minutes, and most are around 2 minutes. In most pieces, *ritardando* (gradual slowing down) appears towards the

³<http://www.8notes.com>

end, and various expressions on notes can be applied such as *trill*, *mordent*, *pizzicato*. This results in 44 piece arrangements, including 11 duets, 12 trios, 14 quartets, and 7 quintets. There are 28 unique pieces from 19 different composers, from which we derive different instrument arrangements and/or keys. After such adaptations, the sheet music was regenerated using *Sibelius* 7.5 software. For the detailed piece list please refer to the documentation included in the dataset.

2.4.2 Recruiting Musicians

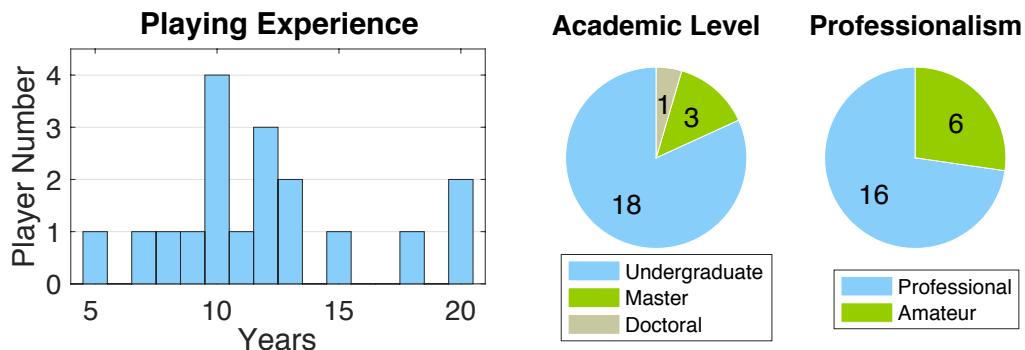


Figure 2.3: Demographic statistics of the total 22 musicians who recorded instrumental parts of the dataset. (The playing experience of 4 players is unknown.)

Creating the URMP dataset requires three kinds of musicians: a conductor, a pianist, and musicians who recorded the instrumental parts of the pieces. All of the musicians were either students from the Eastman School of Music or members of various music ensembles and orchestras from the University of Rochester. The conductor had more than 20 years of conducting

experience. The pianist was a graduate student majoring in piano performance. Background statistics of the instrumental players are summarized in Figure 2.3. In total, 22 players recorded all the instrumental parts, with each player playing only one instrument but maybe multiple tracks. All of the musicians signed a consent form and received a small monetary compensation for their participation.

2.4.3 Recording Conducting Videos

For each piece, a video consisting only of a conductor conducting a pianist playing on a Yamaha grand piano was recorded to serve as the basis for the synchronization of different instrumental parts. These conducting videos were recorded in a $25' \times 18'$ recording studio using a Nikon D5300 camera and its embedded microphone. Before recording each piece, the conductor and the pianist rehearsed several times and the conductor always started with several extra beats for the pianist (and later other players) to follow. The tempo of each piece was set by the conductor and the pianist together after considering the tempo notated in the score. All repeats within a piece were reserved for integrity. All the expression notations in the score were implemented for high expressiveness. Note that although we still need rehearsals between the conductor and the pianist for recording the conducting videos, this is much less effort than arranging joint rehearsals for all instrumental players, especially for larger ensembles and players who played in multiple pieces.

2.4.4 Recording Instrumental Parts

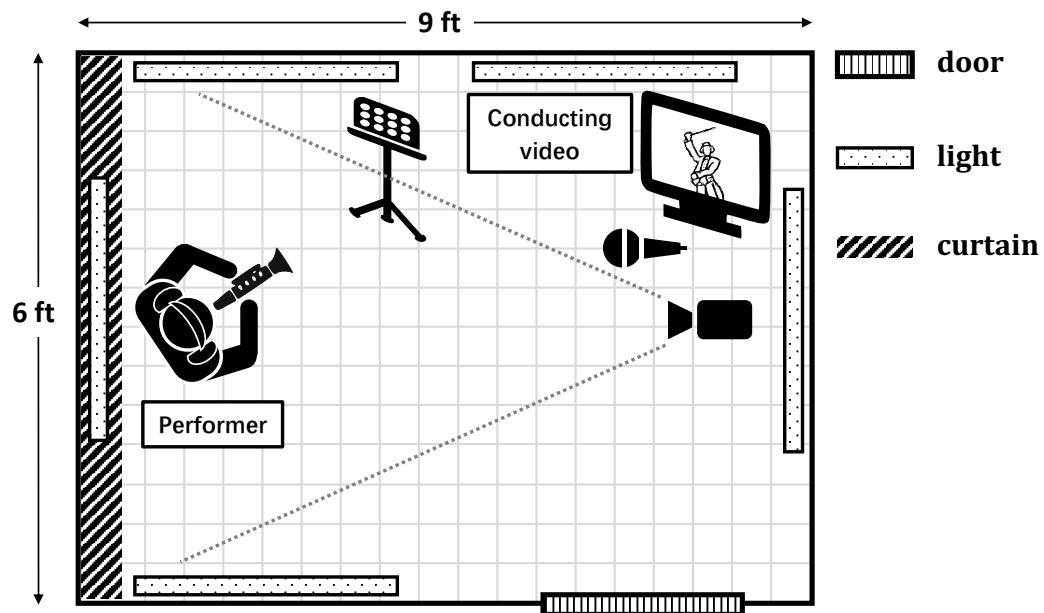


Figure 2.4: The floor plan of the sound booth (top-down view).

The recording of the isolated instrumental parts was conducted in an anechoic sound booth with the floor plan shown in Figure 2.4. The wall behind the player was covered by a blue curtain and the lighting of the sound booth was through fluorescent lights affixed at wall-ceiling intersections around the room. We placed a Nikon D5300 camera on the front-right side of the player to record the video with a 1080P resolution. Since the built-in microphone of the camera did not achieve adequate audio quality, we also used an Audio-Technical AT2020 condenser microphone to record high-quality audio with a sampling rate of 48 KHz and a bit depth of 24. We connected the microphone to a laptop computer running the *Audacity* software for the audio record-

ing thereby making camera and the stand-alone microphone independently controllable.

During the recording, the player watched the conducting video on a laptop with a 13-inch screen placed about 5 feet in front of the player. The player also listened to the audio track of the conducting video through a blue-tooth earphone with no noticeable latency. For simpler pieces, the recording was finished in one shot; while for long and difficult ones, several shots were conducted before we approved the quality of the recordings.

2.4.5 Mixing and Assembling Individual Recordings

For each instrumental part, we first replaced the low-quality audio in the original video recording with the high-quality (HQ) audio recorded using the stand-alone microphone. Because the camera and the stand-alone microphone were controlled independently, the video and the high-quality audio recordings need a relative shift to ensure proper alignment, which was accomplished automatically by using the “synchronize clips” function of the *Final Cut Pro* software.

We then assembled individual instrumental recordings. Although the individual instrumental parts of each piece were all aligned with the conducting video, the starting times of the individual recordings were not aligned with each other. We had to manually time-shift the individual recordings to align them. This was achieved by focusing on the fast sections with clear note onsets. We also manually adjusted the loudness of some tracks to achieve

a better volume balance. This subjective adjustment achieved a more natural balance than objective normalization methods such as root-mean-square normalization. Then the assembled audio is the mixture (addition) of the individual high-quality audio recordings. Finally, we assembled the synchronized individual video recordings into a single ensemble recording. In the video, all players were arranged at the same level from left to right. The order of the players followed the order of score tracks.

2.4.6 Video Background Replacement

In order to make the assembled videos look more natural and similar to live ensemble performances, we used chroma keying (Shimoda, Hayashi and Kanatsugu, 1989) to replace the blue curtain background with a real concert hall image.

We use the Final Cut Pro software for video compositing. The blue background in the videos was unevenly lit and had players' shadow and significant textural variation. To avoid compositing artifacts due to this uneven lighting, we did color correction as a pre-processing step followed by chroma keying. By adjusting the keying and color correction parameters and by setting suitable spatial masks, we were able to get a good separation between the foreground and the background. Once the foreground was extracted, we used a more realistic image as the background for the composite video. The background photo was captured from the Hatch Recital Hall⁴ using a Nikon

⁴<http://www.esm.rochester.edu/concerts/halls/hatch/>

D5300 camera.

2.4.7 Ground-truth Annotation

We also provide ground-truth pitch annotations for each audio track. This annotation was performed on each single audio track using the Tony melody transcription software (Mauch et al., 2015), which implements pYIN (Mauch and Dixon, 2014), a state-of-the-art frame-wise monophonic fundamental frequency (F0) estimation algorithm. For each audio track, we generated two files: a frame-level pitch trajectory and a note sequence. The pitch trajectory was first calculated with a frame hop size of 5.8 ms, and then interpolated to 10 ms according to the standard format of ground-truth pitch trajectories in MIREX. The note sequence was extracted by the Tony software using Viterbi decoding of a hidden Markov model. The pitch of each note takes un-quantized frequencies. To guarantee a good annotation quality, we manually went through all the files introducing necessary corrections. For the frame-level pitch annotation, the annotation from the automatic tool is precise to musical cents, and we only manually corrected insertion, deletion, and octave errors. For the note-level pitch annotation, manual corrections were performed on more than half of the notes, mostly about adjusting the note onset/offset, such as splitting the wrongly merged notes. On average, the correction of each track required about half an hour. We provide the visualizations of all the annotations on the project website⁵.

⁵<http://www.ece.rochester.edu/projects/air/projects/urmp.html>

2.5 The Dataset

2.5.1 Dataset Content

The URMP dataset contains audio-visual recordings and ground-truth annotations for all the 44 pieces, each of which is organized in a folder with the following content:

- **Score:** we provide both the MIDI score and the sheet music in PDF format. The sheet music is directly generated from the MIDI score using Sibelius 7.5 with minor adjustment (clef, key set, note spelling, etc.) for display purposes. The encoded track IDs in MIDI files are ordered following the score track order.
- **Audio:** individual and mixed high-quality audio recordings in WAV format, with a sampling rate of 48 KHz and a bit depth of 24. The naming convention of individual tracks follows the same order as the tracks in the score.
- **Video:** assembled video recordings in MP4 format encoded with an H264 codec. Videos have 1080P resolution (1920×1080), and a frame rate of 29.97 FPS. Players are rendered horizontally, from left to right, following the same order as the tracks in the score. Additional details regarding object-level spatial resolution are provided in Section 2.5.3.
- **Annotation:** ground-truth frame-level pitch trajectories and note-

level transcriptions of individual tracks in ASCII delimited text format.

An overview of the dataset and a sample piece are available online⁶ along with a document that lists all 44 pieces and their instrumentations. The full 12.5 GB dataset is deposited in the Dryad Digital Repository (Li, Liu, Dinesh, Duan and Sharma, 2018).

2.5.2 Synchronization Quality

Because maintaining the synchronization among different instrumental parts is the main challenge in creating the URMP dataset, we compare the synchronization quality of this dataset with that of Bach10 and WWQ. Both datasets have been used in the development and/or testing phases for the MIREX Multi-F0 Estimation & Tracking task in the past. We did not include PHENICX-Anechoic because it used the same approach as URMP and its pieces are symphony pieces with many more parts than the other datasets.

2.5.2.1 Quantitative Evaluation

We first numerically compare the synchronization quality by calculating the onset time deviations as described in Section 2.3. When the polyphony is higher than two, the maximum deviation among the score-notated simultaneous notes is calculated. Figure 2.5 shows a boxplot of the maximum deviation for each piece in URMP, Bach10, and WWQ. The best synchronization

⁶<http://www.ece.rochester.edu/projects/air/projects/urmp.html>

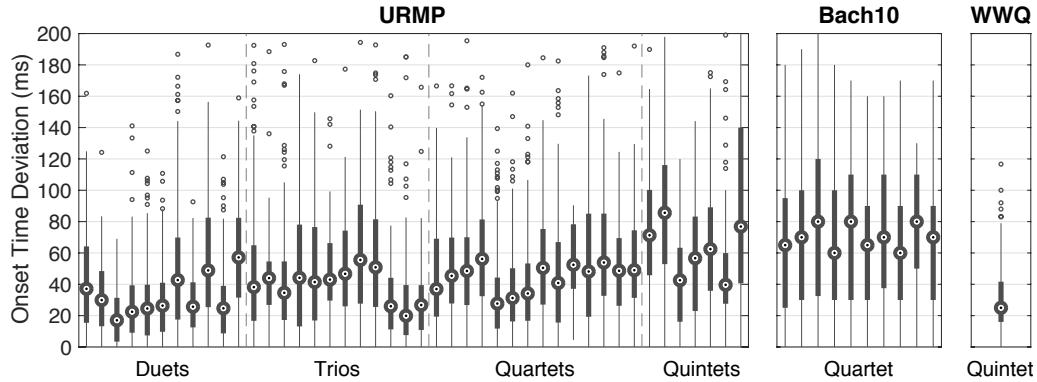


Figure 2.5: Synchronization quality for individual pieces in the URMP, Bach10, and WWQ dataset assessed by onset time deviation for score-notated simultaneous notes. On average, the synchronization quality is ranked as WWQ>URMP>Bach10.

quality is achieved by WWQ, where players rehearsed before recording, a methodology that does not scale to larger datasets. Also note that only a 54-second excerpt out of the 9-minute recording is publicly available and is evaluated here. This excerpt has a strong rhythmic pattern which might help the synchronization. The URMP dataset achieves the second best synchronization quality, with the maximum onset deviation being in the range of 20 to 60 ms. This deviation is larger than our preliminary evaluations in Figure 2.1 (b). This is because we include all of the pieces here and many of them are larger ensembles. Bach10 achieves the worst synchronization quality, showing the maximum onset deviation in the range of 60 to 80 ms.

2.5.2.2 Subjective Evaluation

The numerical evaluation based on onset time deviations has its own limitation, considering the ambiguity of onset instances for some soft articulations. So we also conducted a subjective evaluation. We recruited 8 subjects who were students at the University of Rochester from various fields. Half of them had musical background, and none of them were familiar with these datasets. For each subject, we randomly chose pieces from these datasets to form 4 triplets, one piece from each dataset. We then asked the subjects to listen to the three pieces of each triplet and rank their synchronization quality. Table 2.2 shows the ranking statistics. It can be seen that out of the 32 rankings (4 rankings per subject for 8 subjects), URMP ranks first 9 times, and ranked second 17 times. This is consistent with our quantitative evaluation.

Rank	#1	#2	#3
URMP	9	17	6
WWQ	22	9	1
Bach10	1	6	25

Table 2.2: Subjective ranking results of the synchronization quality of the three datasets provided by eight subjects.

2.5.3 Spatial Occlusion & Resolution

In this section we analyze several aspects of the visual quality of the dataset, i.e., the spatial occlusion and resolution on Regions of Interest (ROI) where the musician-instrument interactions take place. As we mentioned in Section

2.4.4, the videos were captured from the right-side of the players, whose locations and orientations were kept unchanged throughout the piece. So only the right-side faces are in the view without occlusions. From this camera angle, self-occlusions on the players' hands or arms vary for different instrument types:

- Violin/Viola: Both the right-arm bow motion and left arm is visible.
The detailed fingering of the left hand is partially occluded.
- Cello/Bass: The right-arm bow motion and left-hand fingering motion are visible. The left arm is sometimes occluded by the instrument body.
- Woodwind: One arm is in the front and the other arm is occluded. The fingering motion of both hands are visible.
- Brass: Only one hand contributes to the fingering and it is visible.



Figure 2.6: Spatial resolution of ROIs (face, hand, mouth) where most musician-instrument interactions take place.

The spatial resolution on ROIs is a relevant parameter to know which tasks can be tackled using our dataset. Since we used a fixed camera-player distance through the whole recording process, this resolution is roughly the same for all the individual video recordings. After rescaling the players' size in the 1080P assembled video (with some resolution loss), the players' faces, hands, and mouths have the resolutions of about 100×100 , 70×70 , 40×40 pixels, respectively. We sample several typical ROIs from video frames and indicate corresponding spatial resolutions in Figure 2.6. Note that the resolution loss from individual to assembled videos depends on the ensemble size. For example, for the same ROI, the spatial resolution of a quintet is slightly lower than that of a duet, as more players occupy the 1080P video frame.

2.5.4 Limitations of the Dataset

Although we used our resources to create a high-quality audio-visual dataset, there are still limitations we need to point out, which may prevent some potential usage. The limitations mainly exist in the visual part: the limited camera view. Throughout the whole recording process, only one camera was used, so the videos all have a single-camera view. Alternatively, stereo (or even multi-view) datasets are becoming available nowadays and can support more tasks, e.g., depth estimation, 3D reconstruction. Also, our camera view is not always optimal. For example, it is difficult to infer the pitch being played by a violinist from the finger position on the string board, even though

the fingering motion is generally visible. Also, in some scenarios, important objects such as the end point of a violin bow and the head of the bass player, are outside the camera view. This is because of the limited size of the sound booth. The single-camera view limitation also makes the arrangement of players in the assembled videos less natural: players all face to the same direction, which rarely happens in real chamber music performances.

Another limitation of the assembling process is that possible occlusions between players or by the music stand were not considered. This may make the video analysis on our dataset easier than real scenarios. Also, because the instrumental parts were recorded in isolation, natural interactions among players such as eye contacts and body motion interactions do not exist in the assembled videos even though such interactions are commonly observed in real performances. Thus player interactions cannot be visually analyzed using the dataset.

There are several other minor issues that could be avoided in the future work of dataset creation. For example, the bluetooth earphone still has a short wire which resulted in irrelevant movements. The chroma keying operation during the background replacement step sometimes causes slight changes in the color of the foreground.

2.6 Applications of The Dataset

As the first audio-visual multi-track multi-instrument music performance dataset, URMP can support a large variety of MIR tasks, several of which are highlighted in this section. In the first part, we describe two existing MIR tasks that only require the audio modality. We run well-known algorithms on URMP and another widely used multi-track music audio dataset. This also helps benchmark URMP’s audio modality with existing datasets. In the second part, we propose novel tasks that require both the audio and visual modalities of URMP. We also set up evaluation metrics and provide baseline systems. We hope that the baseline results that we provide will invite other researchers to pursue these new research directions and explore other directions with URMP.

2.6.1 Existing Tasks Using Only Audio Modality

There are many existing MIR tasks that URMP can support, and here we only describe two tasks that take the full use of the audio modality and the associated annotations: multi-pitch analysis and score-informed source separation. For these tasks, URMP can be benchmarked with suitable existing multi-track musical audio datasets. Within the multi-track category, only the Bach10, TRIOS, WWQ, and PHENICX-Anechoic have clean individual audio tracks with required annotations. The publicly available audio recording from WWQ is too short for a systematic comparison. For TRIOS, one

instrument is a piano, which makes it difficult to define the polyphony and to perform a fair comparison. Also, PHENICX-Anechoic has orchestra pieces with 8-10 instrumental parts and 10-39 individual tracks, which makes the algorithm performance not comparable for the same reason. Therefore, we just use Bach10 for a comparison with URMP.

2.6.1.1 Multi-pitch Analysis

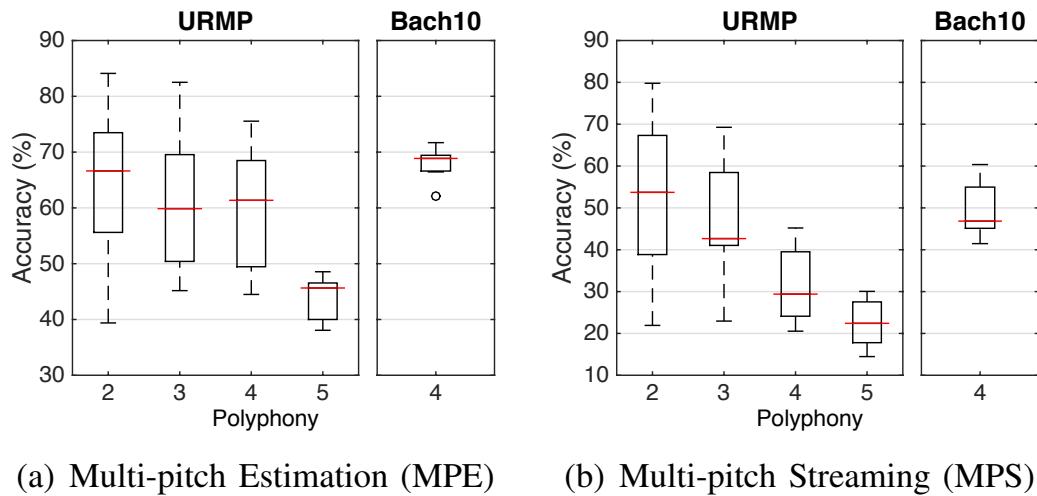


Figure 2.7: Comparison between URMP and Bach10 for multi-pitch analysis.

This task consists of *multi-pitch estimation (MPE)* and *multi-pitch streaming (MPS)*, which are defined as estimating concurrent pitches and organizing them into temporal streams according to their sound sources, respectively. It is a fundamental task towards automatic music transcription and many other MIR applications. For MPE, we run the algorithm described in (Duan, Pardo and Zhang, 2010), which proposes a maximum likelihood method to model

relations between the magnitude spectrum and underlying pitches. For MPS, we run the algorithm proposed in (Duan, Han and Pardo, 2014), which clusters pitches into pitch streams according to their timbre and locality. Both methods are well known and have been tested on Bach10. Performance on both MPE and MPS is often measured by accuracy, which is defined as

$$\text{Accuracy} = \frac{\#TP}{\#TP + \#FP + \#FN}, \quad (2.1)$$

where TP, FP, FN represent true positives, false positives and false negatives, respectively. They are calculated by comparing the estimated and ground-truth pitch with a tolerance of a quarter-tone (Bay, Ehmann and Downie, 2009).

The results on URMP (the first 1-min excerpt of each piece) and Bach10 are shown as boxplots in Figure 2.7, where each piece constitutes one data point, and the red line in each box shows the median value. As expected, both MPE and MPS accuracies decrease when polyphony increases on URMP. When the polyphony is 4, both MPE and MPS accuracies are significantly lower than those on Bach10, suggesting that URMP is a more challenging dataset than Bach10. Indeed, URMP has a larger variety of music pieces, instrumentation, and playing techniques than Bach10, which only contains Bach chorales. Furthermore, different tracks of the same piece of URMP may use the same instrument while Bach10 always uses different instruments. This makes it more difficult to exploit the timbre cues for pitch streaming.

2.6.1.2 Score-informed Source Separation

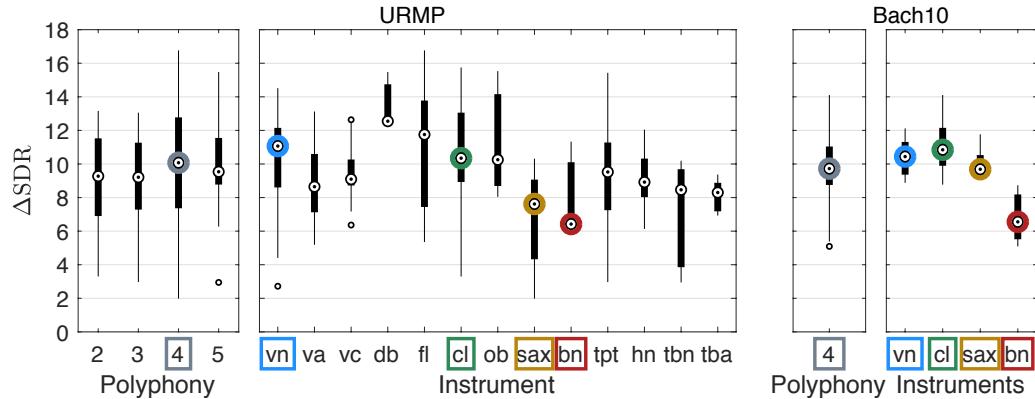


Figure 2.8: Comparison between URMP and Bach10 for score-informed source separation. Colors encode overlapping categories for easier reference.

This task leverages score information to separate musical audio sources. The algorithm we use first aligns the score to the audio mixture using dynamic time warping on chroma feature sequences (Fujishima, 1999). Then audio sources are separated using harmonic masking as described in (Duan and Pardo, 2011a). The quality of the separated audio sources is measured using the Signal-to-Distortion Ratio (SDR) (Vincent, Gribonval and Févotte, 2006). We further calculate the Δ SDR, which measures the improvement of SDR from the audio mixture to the separated source. Figure 2.8 shows box-plots of the results, where each track constitutes one data point, and the circle in each box shows the median value. In contrast with the trends in Figure 2.7, we can see that the performance on URMP and Bach10 is very similar, for pieces with the same polyphony (quartets) and tracks played by the same instrument. This shows that the score information helps sig-

nificantly in overcoming the greater challenges posed by URMP compared with Bach10. Also, the harmonic masking method for source separation does not model timbre information; and thus underexploited the “distinct timbre” advantage of the Bach10 dataset.

2.6.2 New Tasks Using Both Audio and Visual Modalities

With the visual modality available, URMP not only serves for the development and evaluation of audio-based approaches, but also opens up new frontiers for MIR tasks. In this section, we propose two representative tasks that require both the audio and visual modalities. We define the tasks, set up evaluation strategies, and provide baseline results on the URMP dataset to invite the research community to pursue these new research directions.

2.6.2.1 Visually Informed Multi-pitch Analysis

This is the same task as defined in Section ??, but here visual information is available. Visual information about the music performance can significantly help multi-pitch analysis: Observation of the fingering can directly help predict the notes being played; detection of play/non-play activity of instrument players may help estimate the instantaneous polyphony and assign pitches to correct sources. There exist several systems that utilize visual information to estimate pitches for instrument solos such as violin (Zhang et al., 2007), piano (Akbari and Cheng, 2015), and guitar (Paleari et al., 2008), but little

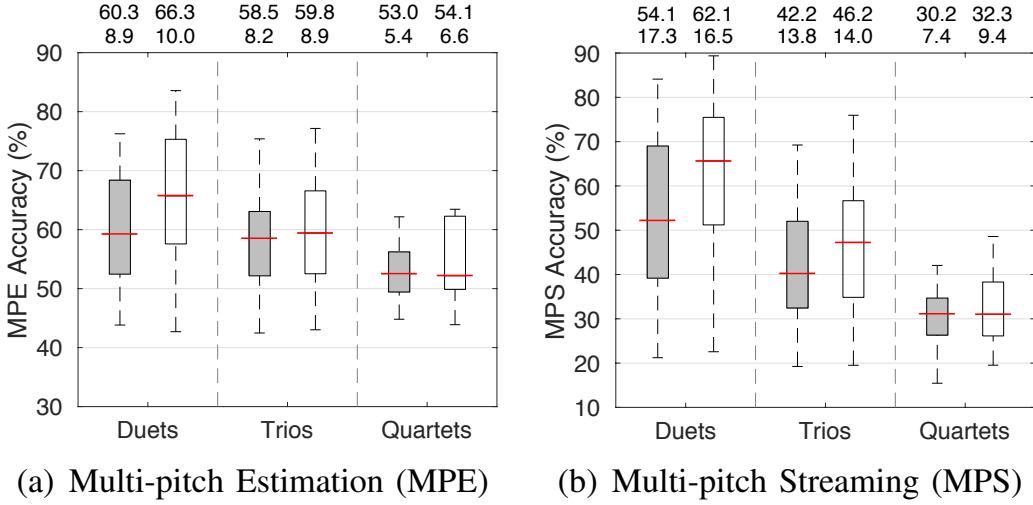


Figure 2.9: Comparison between the proposed visually informed method (white) and the audio-based method (gray) on the multi-pitch analysis task. For each boxplot, the mean and standard deviation values are listed above the plot. Results are reproduced from (Dinesh et al., 2017).

work has been done for other instruments or ensembles, due to the lack of datasets.

We propose to start this task with the 11 string ensembles in the URMP dataset, which provide the most pronounced motion information. Our previous work in (Dinesh et al., 2017) addresses both MPE and MPS for these pieces and can serve as a baseline for future approaches. The basic idea of this work is to model the play/non-play (P/NP) activity of each player from the visual modality and then use it to constrain audio-based pitch analysis. The P/NP activity is classified in each video frame using the bowing motion features that are calculated from optical flow estimation (Sun, Roth and Black, 2010). For MPE, the detected P/NP label provides a more accurate

estimate of the instantaneous polyphony in each frame. For MPS, this label constrains the assignment of pitch estimates to sources: pitch estimates are only assigned to active players. This idea was implemented based on the same audio-based MPE/MPS algorithms as described in Section 2.6.1.1.

Figure 2.9 compares the MPE and MPS accuracies of this method with those of the audio-based method, where each piece constitutes one data point. Note that each polyphony category is the expanded set using all the possible track combinations within each piece. An improvement between 2-12% can be seen across the tasks and pieces.

We want to state that this baseline approach is just a preliminary attempt to address the multi-pitch analysis problem for string instruments. Much visual information such as the fingering is not exploited. In addition, reliable detection of P/NP activity for non-string instruments where motion is more subtle is also an open problem (Bazzica, Liem and Hanjalic, 2014).

2.6.2.2 Polyphonic Vibrato Analysis

In music performances, vibrato is an important artistic effect that adds expressiveness and emotions by slight variations in pitch. Vibrato analysis provides basis for comparing different articulation styles, and thus has broad impact in musicological studies. It also facilitates other tasks such as melody extraction and music synthesis. However, most of the existing automatic vibrato analysis tools are audio-based with a focus on monophonic recordings. In polyphonic cases, even if the score is provided, the task is challenging due

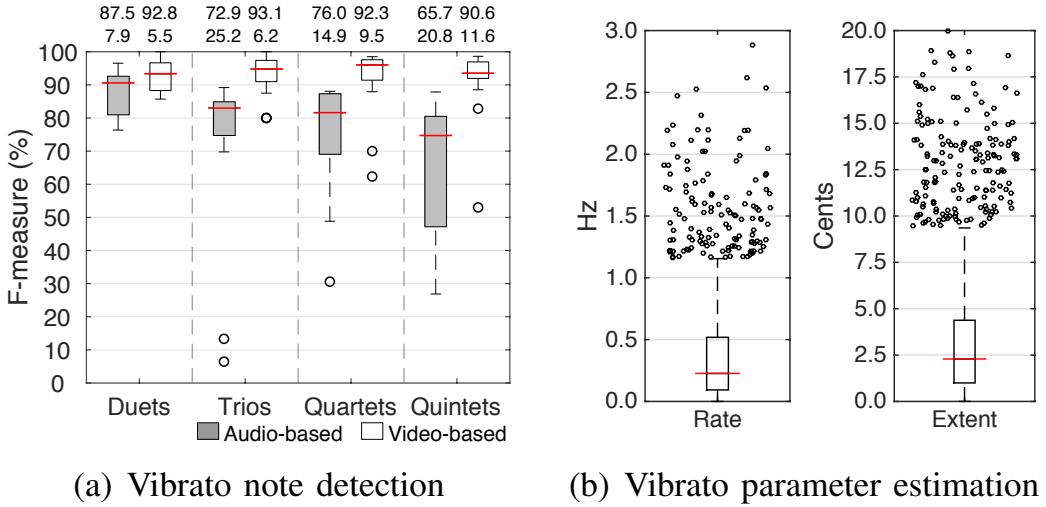


Figure 2.10: Video-based vibrato note detection and parameter analysis results, reproduced from (Li, Dinesh, Sharma and Duan, 2017). For each boxplot, the mean and standard deviation values are listed above the plot.

to the severe interference among sources. Existing audio-based techniques are not yet capable of this task.

The visual modality of a music performance can be very helpful for vibrato analysis. This is especially true for string instruments, where the left-hand fingers' rolling motion along the fingerboard is the direct cause of the fluctuation of pitch. Compared to the audio signals, this motion cue does not degrade as polyphony increases. This makes the polyphonic vibrato analysis task possible.

We define this task on the 19 pieces that use at most one non-string instrument in the URMP dataset. This task contains two subtasks: 1) vibrato note detection and 2) vibrato parameter (rate and extent) estimation. To obtain ground-truth annotations, we first threshold the auto-correlation value

of the ground-truth pitch contour of each note to determine whether the note has vibrato or not, and then calculate the vibrato rate and extent for vibrato notes from the auto-correlation function. To evaluate vibrato note detection performance, we propose to use precision, recall, and F-measure on each track. To evaluate vibrato parameter estimation, we propose to calculate the absolute difference between the estimated value and the ground-truth value.

Our previous work (Li, Dinesh, Sharma and Duan, 2017) serves as the baseline method. It tracks the left hand of each string player using the KLT tracker (Tomasi and Kanade, 1991), and then extracts the hand motion features by optical flow estimation (Sun, Roth and Black, 2010). The aligned score is utilized to temporally segment the raw motion features into temporal-spatial blocks at each note onset/offset time. We then train a support vector machine (SVM) to classify each block as vibrato/non-vibrato. For vibrato parameter estimation, we perform principal component analysis (PCA) on the raw motion features to get a 1D motion curve corresponding to the hand rolling motion along the fingerboard. This amplitude of the motion curve is then normalized by that of the corresponding noisy pitch contour extracted from the audio mixture using a score-informed pitch estimation method. Vibrato rate and extent are finally measured on the motion curve.

We compare this proposed video-based baseline method with an audio-based method that extracts pitch contours in a score-informed fashion on the vibrato note detection subtask. The results are shown in Figure 2.10 (a), where each track constitutes one data point, and the red line in each box

denotes the median value. In all of the polyphony cases, the video-based method always achieves a high F-measure (generally over 90%), while the audio-based method degrades as the polyphony increases. We further evaluate the vibrato parameter estimation performance. Results show that our video-based baseline achieves an average error of 0.38 Hz for rate estimation and 3.47 musical cents for extent estimation. Boxplots of these errors are shown in Figure 2.10 (b), where each vibrato note constitutes a data point, and the red lines denote the median values. 90% of the errors are within 1 Hz and 10 musical cents, respectively.

Although the current task is limited to vibrato analysis, we anticipate that it can be extended to playing technique detection of string instruments in general (Su, Lin and Yang, 2014). These playing techniques may include vibrato and positioning from the left hand, as well as bowing/plucking, up-bow/down-bow, and legato/détaché bowing from the right hand. We hope that this current task will promote the use of multi-modal analysis techniques in musicological studies. Furthermore, we anticipate an extension of music performance analysis to non-string instruments in near future (Bazzica, Liem and Hanjalic, 2016; Bazzica et al., 2017).

2.6.2.3 Other Emerging New Tasks

Besides the two new tasks that we defined above, several other emerging tasks can be developed based on the URMP dataset:

- Visually Informed Source Separation: Audio events (e.g., a violin note)

are often associated with visual movements (e.g., a bowing motion) (Parekh et al., 2017). Designing methods that can leverage visual information for source separation is an interesting task.

- **Audio-visual Source Association:** A related problem to source separation is how to associate sound sources or their components (e.g., a note) to visual objects (e.g., a player). A restricted version of this task has been defined and explored in (Li, Dinesh, Duan and Sharma, 2017) and (Li, Xu and Duan, 2017) for string instruments by modeling their bowing motion and vibrato motion, respectively. Such techniques can be used to design novel music streaming services that allow users to target sound tracks from the visual scene (Zhao et al., 2018).
- **Audio-visual Cross Modality Generation:** By further modeling the audio-visual relations, one may design a system that can generate one modality from the other. Chen et al. (2017) made the first attempt using conditional Generative Adversarial Networks (GAN) to cross-generate static audio spectrograms and instrument-playing images. Extending this task to consider temporal dependencies is an interesting direction (Shlizerman et al., 2018).

2.7 Conclusions

In this chapter, we presented the URMP dataset, a multi-modal music performance dataset that is useful for a broad range of research applications including source separation, music transcription, audio-score alignment, music performance analysis, etc. Synchronization of separately recorded individual instrumental parts while maintaining expressiveness is a key challenge in recording such a dataset and we discussed the approaches for addressing this challenge. The approach successfully adopted for URMP involved having individual instrument players watch and listen to a pre-recorded conducting video when recording their individual parts. Objective and subjective comparisons between URMP and two other widely used multi-track music performance datasets showed that the multi-track synchronization in URMP has a high quality. We highlighted how the URMP dataset supports existing MIR tasks and also defined two novel multi-modal MIR tasks by providing evaluation measures and baseline systems. We further proposed several emerging research directions that URMP can support. We anticipate that the URMP dataset will become a valuable resource for researchers in the field of music information retrieval and multimedia.

Chapter 3

Audio-Score Alignment

The coordination between audio and score is known as audio-score alignment, which temporally warp the symbolic notations from the score to map the corresponding instances from the audio. It is an active research for decades, especially the online version of this problem (also known as score following) that are more challenging but can support more applications.

One challenge in score following for piano music is the *sustained effect*, i.e., the waveform of a note lasts longer than what is notated in the score. This can be caused by expressive performing styles such as the legato articulation and the usage of the sustain and the sostenuto pedals, and can also be caused by the reverberation in the recording environment. This effect creates non-notated overlappings between sustained notes and latter notes in the audio. It decreases the audio-score alignment accuracy and robustness of score following systems, and makes them be prone to *delay errors*, i.e., aligning audio to a score position that is earlier than the correct position. In this work, we propose to modify the feature representation of the audio

to attenuate the sustained effect. We show that this idea can be applied to both the chromagram and the spectral-peak representations, which are commonly used in score following systems. Experiments on the MAPS dataset show that the proposed method significantly improves the alignment accuracy and robustness of score following systems for piano performances, in both anechoic and highly reverberant environments.

3.1 Introduction

The two commonly used digital representations of music are the audio waveform and the musical score. While the audio provides rich information about the musical performance, it is not structured for computers to understand directly. The musical score (e.g., MIDI and music XML), on the other hand, is machine-readable, yet lacks the expressiveness of the musical performance. Aligning musical audio with the score has been an important research topic in music information retrieval since the 1990s (Puckette and Lippe, 1992).

An audio-score alignment algorithm can be classified as offline or online, according to its requirement on the audio input. An *offline* algorithm needs to access the whole sequence of the audio before starting the alignment process. An *online* algorithm (also called *Score Following if runs in realtime*) is able to align each audio frame with the score “without looking into the future” or “by only looking into the near future” (i.e., with some delays) of the audio performance. Since online algorithms require less information than

offline algorithms, they are more challenging to design to achieve the same alignment accuracy and robustness. However, only online systems can support real-time applications if provided with enough computational resources.

Audio-score alignment has many applications. Offline algorithms have been used to construct multi-modal (e.g., video, audio, and score) music digital libraries (Thomas et al., 2009), to build query-by-humming systems (Wang et al., 2013), to design piano tutoring and grading systems (Benetos, Klapuri and Dixon, 2012), to analyze performance techniques (Wang et al., 2018). Online algorithms have been used in automatic music accompaniment for a live soloist (Raphael, 2010), interactive piano pedagogy (e.g., *Tonara* software), real-time enhanced music enjoyment of orchestral performances (Prockup et al., 2013), score-informed source separation (Duan and Pardo, 2011a), automatic coordination of audio-visual equipment (Itohara et al., 2012), and automatic page turning (Arzt, Widmer and Dixon, 2008). Other inspiring potential applications include automatic display of lyrics or opera subtitles and the control of lighting and camera movement on stage.

In this chapter, we focus on score following for piano performances (Li and Duan, 2015, 2016). Piano is one of the most popular instruments in the world (Gordon, 1996). It is played worldwide in a variety of music genres including classical, jazz/blues, rock, and pop. It produces sounds ranging from monophonic (e.g., nursery rhymes) to highly polyphonic (e.g., piano arrangements of symphonies). It is also one of the few instruments that do not require accompaniment, thanks to its wide pitch range and highly

polyphonic nature. However, the flexibility and expressiveness make piano performances difficult to follow by computers. More specifically, in this work, we analyze the *sustained effect* in score following for piano performances (see Section 3.3). This effect can be caused by the legato articulation of notes, the usage of the sustain and the sostenuto pedals, and also the reverberation of the recording environment. It extends the sound of notes longer than the expected length notated in the score, creating overlappings between sustained notes and latter notes, hence causing mismatch between audio and score.

We propose an approach to modify the feature representations of the audio to attenuate the sustained effect. We first perform onset detection and treat all frames within a region immediately after an onset as frames that potentially contain the sustained effect. We then analyze the spectra of the signal in these frames and detect spectral components that are considered as an extension from previous notes. We reduce the energy of these components in the audio representation to attenuate the sustained effect. We implement this basic idea for two commonly used audio representations in audio-score alignment, the chromagram and the spectral-peak representations. We test the modified representations within a hidden Markov model (HMM)-based score following framework (Duan and Pardo, 2011b). Experiments on the MAPS dataset (Emiya, Badeau and David, 2010) show that the proposed approach greatly improves the alignment accuracy and robustness of highly expressive piano performances with different degrees of the sustained effect.

It is noted that our proposed sustained-sound reduction operation does

not always reduce the sustained effect correctly. First, some frames may not have the sustained effect, yet we perform the operation in all frames within a region after an onset. Second, we cannot discriminate whether the extension of a note is due to the sustained effect or is because the note is notated that way. Third, onsets may be wrongly detected. In all these three cases, the sustained-sound reduction operation will cause new audio-score mismatch. However, in Section 3.4.3 we will see that only in the last case the operation is harmful to the alignment, while in the other two cases the operation is still helpful or has no significant effect. We conduct systematic experiments on piano pieces with different degrees of the sustained effect, compare the proposed method with baselines, and analyze the influences of key parameters.

The rest of the chapter is organized as follows. We first introduce related work in Section 3.2. We then illustrate several specific properties of piano music in Section 3.3 to set up the background for the proposed approach. In Section 3.4, we propose a sustained-sound reduction operation with an online onset detection technique to reduce the sustained effect in audio representations, and also discuss the influence when the operation is wrongly applied in three cases. Systematic experiments are illustrated in Section 3.5. Finally, we conclude the work in Section 3.6.

3.2 Related Work

3.2.1 Early Work

Audio-score alignment has been an active research topic for two decades. Early approaches focus on monophonic audio, such as vocals and monophonic instrumental solos (Puckette, 1995; Grubb and Dannenberg, 1997; Orio and Déchelle, 2001). For polyphonic music, Orio and Schwarz (2001) firstly applied Dynamic Time Warping (DTW) for alignment on string and wind ensembles. This method is computationally demanding because it considers alignment between every pair of audio and score segments. Müller, Mattes and Kurth (2006) constrained the alignment within a region found through a multi-scale analysis. They obtained similar alignment results but with 50 times less memory cost. Kaprykowsky and Rodet (2006) proposed a short-time DTW approach to reduce the computational complexity of standard DTW.

3.2.2 Online Alignment

An online audio-score alignment system typically contains two modules: a processor and an observer. The *processor* (or *process model*) is a hypothesis generator. It continuously generates alignment hypotheses for the incoming audio frame. The *observer* (or *observation model*) is a hypothesis evaluator. It evaluates the alignment hypotheses by matching the audio frame

with the hypothesized score positions and chooses the best hypothesis. The contribution of this work lies in the observer module.

One type of commonly used processors is the online DTW algorithm (Dixon, 2005). Alignment hypotheses generated by this method are neighboring cells within an adaptive band centered around the current cell in the alignment matrix. This method, however, has no guarantee of the continuity of the alignment path without retrospective adjustments. This idea is further developed by the “backward-forward strategy” to reconsider the past decisions (Arzt, Widmer and Dixon, 2008), and the incorporation of a tempo model (Arzt and Widmer, 2010) for robustness. Another processor proposed in (Grubb and Dannenberg, 1997) employs stochastic models, where the score position hypotheses are represented by a probability density function. The benefit of this method is that the random factors in the system can better cope with the uncertainties in real performances. Similarly, Duan and Pardo (2011*b*) proposed a hidden Markov process model in a 2-D continuous state space with the score position and tempo being the two dimensions. Pardo and Birmingham (2005) modeled score forms to deal with large-scale structural variations such as skipping or repeating of a section.

The key problem of designing an observer is the choice of representations of audio and score. The most commonly used representation is chromagram (Prockup et al., 2013; Hu, Dannenberg and Tzanetakis, 2003; Ellis and Poliner, 2007; Ewert, Müller and Grosche, 2009; Arzt, Widmer and Dixon, 2012; Wang, Ewert and Dixon, 2014), which well describes the harmonic progres-

sion of music. It can be calculated for both audio and score, and their match can be evaluated using Euclidean or cosine distances. Müller and Ewert (2011) discussed ways of enhancing and implementing chroma features. A similar but richer representation is semigram. It also uses a semitone frequency scale but does not fold different octaves into one as the chromagram does. Semigram was first employed by Dixon (2005) in score following, and was later adopted by other systems (Dixon and Widmer, 2005; Arzt, Widmer and Dixon, 2008; Arzt and Widmer, 2010). Another commonly used audio representation is the spectral-peak representation (Orio and Schwarz, 2001; Duan and Pardo, 2011*b,a*). Spectral peaks are ideally caused by harmonics of notes, hence they convey pitch information (Duan, Pardo and Zhang, 2010), through which the match between audio and score can be defined. Other representations include auditory filter bank responses (Montecchio and Orio, 2009), Nonnegative Matrix Factorization (NMF)-based multi-pitch representation (Cont, 2006; Carabias-Orti et al., 2015), and adaptive-template-based observation models (Joder and Schuller, 2013). Onsets are also modeled in some representations (Ewert, Müller and Grosche, 2009; Miron, Carabias-Orti and Janer, 2014) to achieve more accurate alignment between audio and score.

3.2.3 Score Following for Piano Music

Few methods have been proposed to address score following specifically for piano music. General score following systems may achieve better results on

piano pieces than other instruments in the offline scenario, thanks to the clear onsets and consistent timbre of piano notes (Dixon and Widmer, 2005). In online scenarios, however, this advantage is likely to be dominated by the disadvantages of the high loudness contrast of simultaneous notes and the sustained effect caused by the legato articulation of notes and the usage of the sustain and sostenuto pedals.

The sustained effect for audio-score alignment was firstly observed by Orio and Schwarz (2001). They stated that partials of the previous notes can be still present in the beginning of the next notes, due to the legato articulation or reverberation. To cope with this, they used the first-order positive difference of magnitude spectrum (spectral flux) as the audio representation instead of the original spectrum to emphasize onsets. This was also adopted in (Dixon, 2005; Dixon and Widmer, 2005; Arzt, Widmer and Dixon, 2008). The problem of this representation is that in most inter-onset frames the spectral flux is close to zero, because their spectra are very similar to those of their previous frames. This undermines the inference of the score position in these inter-onset frames, and results in outliers in the alignment path found by the online DTW algorithms, as described in (Dixon and Widmer, 2005).

It is worth to mention Niedermayer *et al.*'s finding on the influence of the sustain-pedal usage on audio-score alignment accuracy (Niedermayer, Böck and Widmer, 2011). They compared alignment results on the same music pieces performed with and without the sustain pedal, and found that

the influence was negligible. However, they explained that this might be due to the rare usage of the sustain pedal in the pieces that they tested, which were all Mozart pieces. In fact, the sustain pedal was rarely used before the Romantic era but has been commonly used since then (see Section 3.3.2). Another reason for Niedermayer *et al.*'s observation, we argue, is that the algorithm used for evaluation was an offline algorithm, which is more robust to the local mismatch between audio and score as a global alignment is employed. For online algorithms, however, they are more sensitive to local audio-score mismatch, which can be greatly introduced by the the usage of the sustain pedal.

3.3 Sustained Effect in Piano Music

In this section, we start from basic acoustical properties of piano notes, then we describe in detail the sustained effect, its major causes, and its influences on score following.

3.3.1 Acoustical Properties of Piano Notes

Each piano key is associated with a hammer, one to three strings, and a damper that touches the string(s) by default. When a key is depressed, its hammer strikes the string(s) while the damper is released from the string(s). This yields an impulse-like articulation at the note onset. The loudness of the note is determined by the velocity of the hammer strike, which is affected

by how hard the key is depressed. The string(s) then vibrate freely until the damper returns to the string(s) when the key is released. The free vibration of the string(s) produces an exponential energy decay of the waveform. The pitch of the note, however, does not change. It is determined by the material, length, and tension of the string(s) and is pre-tuned. A performer cannot control the pitch (e.g., playing vibrato) as one could do on other instruments (e.g., strings, winds). To summarize, there are three important acoustical properties of piano notes as follows. In Section 3.4, we will show how these properties can be leveraged to design an approach to reducing the sustained effect.

- *Strong Onset*: Piano note onsets are generally easier to detect thanks to the impulsive articulation.
- *Energy Decay*: The decay time varies for different strings, and can be up to 10 seconds if the strings vibrate freely.
- *Constant Pitch*: The pitch of a note is constant and cannot be controlled during the entire process of a note.

3.3.2 Sustained Effect and Its Major Causes

The *sustained effect* refers to the phenomenon that a note is sustained longer than its notated length in the score. It is very common in piano performances. Figure 3.1 shows the concept. According to the score, the two gray notes (A4 and D4) should stop at Beat 1. Their waveforms in the audio performance,

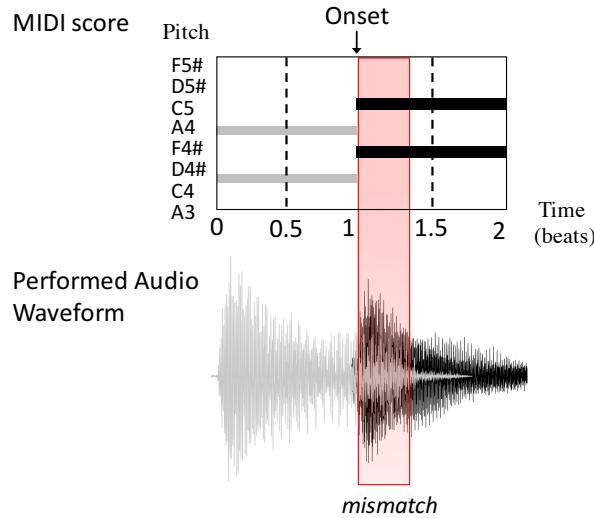


Figure 3.1: Illustration of the sustained effect and the audio-score mismatch problem it causes. The gray notes are extended in the audio waveform longer than their notated length in the score.

however, are sustained longer and are blended with the latter two black notes. This causes mismatch between the audio and the score in the shaded region at the beginning of the two black notes. The length of this region depends on when the gray notes stop sounding. Statistical analysis of this region length in piano performances is presented in Section 3.5.1.1. In the following, we analyze the several major causes of it.

3.3.2.1 Legato Articulation

The musical term *legato* indicates that notes are played smoothly and connected. For piano, the general practice for producing legato notes is to release a key after depressing the key for the following tone with an overlapping (Bresin and Umberto Battel, 2000). To measure the degree of legato artic-

ulation, Repp (1995) introduced the Key Overlap Time (KOT) for adjacent tones, which is defined as “the time interval between the onset of key depression for one tone and the key release for the preceding one” (Repp, 1995). Although the note transition is still not very smooth given the strong onsets of piano notes, this practice does attenuate the percussive sensing.

3.3.2.2 Sustain/Sostenuto Pedal Usage

Modern pianos generally have three foot pedals: sustain, sostenuto, and soft pedals; some models omit the sostenuto pedal. When the sustain pedal is pressed, all dampers of all notes are raised from all strings, no matter whether a key is depressed or released. Therefore, its usage will sustain all notes whose keys are released when the pedal is being pressed. The sostenuto pedal behaves similarly, but only keeps raising dampers that have already been raised without affecting others. Therefore, its usage will sustain notes that are activated before the pedal is pressed *and* that are released while the pedal is being pressed. The soft pedal changes the way that the hammer strikes the string(s), hence it affects the timbre and loudness, but its use is rare compared to the use of the other pedals.

The sustain and sostenuto pedals, especially the sustain pedal, have been commonly used since the Romantic era in Western music history, and in modern piano performances of many different styles. Figure 3.2 (a) shows the proportions of pieces with different degrees of sustain-pedal usage in the MAPS dataset (Emiya, Badeau and David, 2010), which is a commonly used

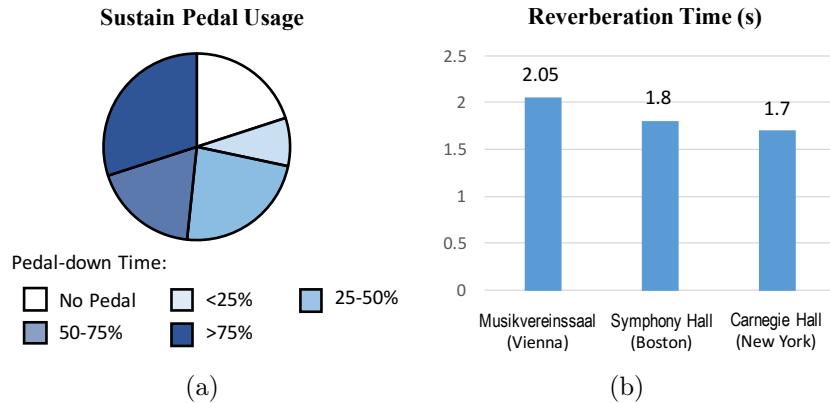


Figure 3.2: Statistics of two causes of the sustained effect. (a) Distribution of the 60 acoustic pieces in the MAPS dataset (Emiya, Badeau and David, 2010) according to the degree of pedal usage. Pedal-down time denotes the percentage of the performing time when the sustain pedal is depressed. (b) Reverberation time of three famous concert halls: Musikvereinsaal in Vienna, Symphony Hall in Boston, and Carnegie Hall in New York.

piano dataset for music transcription and score following with a good coverage of composers in different eras. The usage of the sustain pedal produces round and velvety timbre for lyrical expressions. It also lets pianists sustain notes which would otherwise be out of reach. It can also accomplish legato passages which would otherwise have no possible fingering. While some composers and music arrangers use pedal marks to notate it, appropriate use of the sustain pedal is more often left to the performers' discretion. In addition, the pedals can be partially pressed, causing a slighter sustain effect as the dampers slightly touch the strings. A detailed analysis of the sustain-pedal effects can be found in (Lehtonen et al., 2007).

3.3.2.3 Reverberation

While legato articulation and pedal usage depend on the piece that is being performed, room reverberation is a universal cause of the sustained effect in any real-world piano performances. In this case, all notes are sustained. Because reverberation can be well modeled as a Linear Time-Invariant (LTI) system, it does not change the pitch of the notes. The extent of the sustained effect is determined by the architecture design (e.g., size, shape, material). It can be measured by reverberation time RT_{60} : the time it takes for the room impulse response to decay 60 dB from the original level. It ranges from about 0.5 seconds (e.g., practice room or studio) to 3 seconds (e.g., church or large concert hall) (Meyer, 2009). Figure 3.2 (b) shows the reverberation time of three well-known concert halls¹.

3.3.3 Influences of Sustained Effect on Score Following

The sustained effect causes non-notated overlappings between sustained notes and new coming notes, which influence score following for piano performances. Take Figure 3.1 as an example, when the audio has entered into the shaded region, the score follower may still match to a position before Beat 1. This is because the audio, which contains both the gray and black notes, does not show a much better match with the shaded region of the score than with a location before Beat 1. Therefore, the follower often wrongly aligns

¹Accessed from: <http://hyperphysics.phy-astr.gsu.edu/hbase/acoustic/revtim.html>

to the positions before the correct one (i.e., *delay errors*). This may cause considerable lags in the alignment, and affect the score following accuracy and robustness when such errors are accumulated.

3.4 Proposed Approach

In this section, we propose an approach to address the sustained effect in score following for piano performances. We first locate the regions that contain potential mismatch between audio and score caused by the sustained effect through online onset detection. We then propose an operation in two kinds to reduce the sustained sound in the chromagram and spectral-peak representations of the audio, respectively. Note that not all note onsets are followed by a region containing the sustained effect, so the operation can be wrongly applied. We analyze these cases and the consequences of the wrongly applied operations. Finally, we integrate these operations into an online score following system.

3.4.1 Online Onset Detection

The sustained effect, when it appears, always appears in an audio region right after the onset of new notes, e.g., the shaded region in Figure 3.1. However, not all onsets are followed by the sustained effect. For example, there might be no sustained effect at all between staccato notes. Nevertheless, we propose to locate all the potential audio regions through online onset

detection and apply the sustained-sound reduction operation. The operation may be wrongly applied, but we will see that its consequences are minimal if the onsets are correct in Section 3.4.3.

Onset detection for music signals is a well-studied topic (Bello et al., 2005; Dixon, 2006). A general framework is to first convert audio features (e.g., spectral flux) into a detection function through several signal processing steps or machine learning modules (e.g., neural networks (Eyben et al., 2010)), and then detect onsets through normalization, thresholding, and peak picking. Most methods work in an offline fashion, but there exist online adaptations of these methods (Böck, Krebs and Schedl, 2012). In this work, we propose an online adaptation of a spectral-flux-based onset detection method employing several simple signal processing steps, in favor of simplicity and efficiency.

We assume that the volume of the input piano performance has been normalized. In practice, this normalization can be adjusted during the system setup. Here we simply normalize its Root Mean Square (RMS) value to 1 before the online processing starts. Then let $\mathbf{Y}(n, k)$ be the Short-Time Fourier Transform (STFT) magnitude spectrogram of the music signal calculated with a 46.4 ms Hamming window and a 10 ms hop size, where n and k are time frame and frequency bin indices, respectively. The first step is a logarithmic compression with a ratio $\gamma = 0.2$ of the spectrogram to enhance the high-frequency content:

$$\tilde{\mathbf{Y}}(n, k) = \log(1 + \gamma \cdot \mathbf{Y}(n, k)). \quad (3.1)$$

This compression step yields better results because high frequency components are more indicative of note onsets but are relatively weak in linear-amplitude spectrograms (Rodet and Jaillet, 2001).

The second step is to calculate the spectral flux as an onset salience function $S(n)$, which is the positive first-order difference between consecutive frames of the enhanced spectrogram across all frequency bins:

$$S(n) = \sum_k \left| \tilde{\mathbf{Y}}(n, k) - \tilde{\mathbf{Y}}(n-1, k) \right|_{\geq 0}, \quad (3.2)$$

where $|\cdot|_{\geq 0}$ denotes half-wave rectification, i.e., keeping non-negative values while setting negative values to 0.

The third step is to enhance the salience function to account for loudness variations. This is often done by subtracting the salience function by a smoothed version of itself (Dixon, 2006). For our online setting, the smoothed version is calculated from past frames (Böck, Krebs and Schedl, 2012):

$$\tilde{S}(n) = \frac{1}{\omega + 1} \cdot \sum_{m=-\omega}^0 S(n+m), \quad (3.3)$$

where ω is the sliding window size for the local average calculation. Then the *enhanced onset salience function* $\hat{S}(n)$ can be calculated by taking the positive difference between the original salience function and its smoothed version:

$$\hat{S}(n) = |S(n) - \tilde{S}(n)|_{\geq 0}. \quad (3.4)$$

The last step is to determine onsets from the enhanced salience function. For offline methods, this is often achieved by peak-picking (Bello et al., 2005). For online methods, however, peak-picking requires at least one frame of delay. In our approach, in order to avoid inherent delays, we opt to employ another approach inspired by (Böck, Krebs and Schedl, 2012). We set an amplitude threshold α and a time threshold β . We report an onset in the current audio frame if three conditions are all satisfied: 1) the enhanced salience $\hat{S}(n)$ is greater than α ; 2) the enhanced salience is greater than the salience in all past ω frames, where ω is the window size in Equation (3.3); 3) no other onsets have been reported in the past β frames. This approach ensures real-time onset detection without any delays, but often reports onsets earlier than the true onsets. In our experiments, about 30% of correctly detected onsets (i.e., those deviating less than three frames from ground-truth) are in the same frame as ground-truth. More than 60% are earlier and less than 10% are later. Compared with a delayed detection using peak-picking, this online approach is still preferred in our score following system.

Among the parameters α , β , and ω , the amplitude threshold parameter α has the largest impact on onset detection, specifically on the ratio between false negatives (miss errors) and false positives. A miss error may lead to the miss detection of a sustained-effect region, while a false positive will break an inter-onset interval into two and wrongly apply the sustained-sound reduction operation in the second half. These errors will have different effects on the

score following results. In Section 3.4.3.3 we discuss how false positives affect score following. In Section 3.5 we experimentally investigate the impact of α on onset detection and the final alignment performance.

3.4.2 Sustained-sound Reduction

In this subsection, we propose an approach to reduce the sustained sound in the audio representation of all frames in the audio regions detected in the previous subsection. While the length of the shaded region depends on the degree of the sustained effect, in the experiments we simply use a unified size of $L = 150$ ms or the interval to the next onset if it is within 150 ms. We conduct experiments to show the sensitivity of the parameter L in Section 3.5.3.2.

The basic idea is to inspect the spectrum and reduce the spectral components that are sustained from previous notes. There are two main tasks: 1) Identify components that are sustained, and 2) reduce these components in the audio representation. Answers to these questions depend on the audio representation that is used to match the audio with the score. In the following, we propose methods for two commonly used audio representations in score following: the chromagram and the spectral-peak representations.

3.4.2.1 Spectral Subtraction for Chromagram Representation

The chromagram (Fujishima, 1999) can well represent the harmonic content of the music audio and is less sensitive to timbral variations than the spectro-

gram. It has been commonly used in audio-score alignment approaches (Hu, Dannenberg and Tzanetakis, 2003; Ellis and Poliner, 2007; Ewert, Müller and Grosche, 2009; Arzt, Widmer and Dixon, 2012; Prockup et al., 2013; Wang, Ewert and Dixon, 2014). One way to calculate a 12-d chroma vector \mathbf{Ch}_a for an audio frame is from the STFT magnitude spectrum using weighting functions. Each dimension of the chroma vector is a weighted sum of the energy of all frequency bins in the spectrum. Each weighting function is a mixture of Gaussians that are centered at frequencies of the pitch class at different octaves using the standard $A = 440$ Hz tuning and a standard deviation of a quarter-tone.

To deal with the sustained effect in a detected region after an onset, we calculate the chroma vectors from a modified spectrogram $\mathbf{Y}^*(n, k)$ instead of the original magnitude spectrogram $\mathbf{Y}(n, k)$, by subtracting a reference spectrum $\mathbf{Y}(m, k)$:

$$\mathbf{Y}^*(n, k) = |\mathbf{Y}(n, k) - f(\hat{S}(n)) \cdot \mathbf{Y}(m, k)|_{\geq 0}. \quad (3.5)$$

The reference frame m is set to 5 frames (i.e., 50 ms) before the onset frame in our implementation. The half-wave rectification prevents $\mathbf{Y}^*(n, k)$ from being negative after subtraction. $f(\cdot)$ is a confidence function of onset detection

controlled by the enhanced onset salience $\hat{S}(n)$, and is defined as:

$$f(x) = \begin{cases} 0 & \text{if } x < \alpha \\ \frac{1}{\alpha' - \alpha}(x - \alpha) & \text{if } \alpha \leq x \leq \alpha' \\ 1 & \text{if } x > \alpha' \end{cases} . \quad (3.6)$$

where α is the threshold for onset detection in Section 3.4.1, and α' is another conservative threshold which decides when the full amount of the reference spectrum is subtracted. When the salience $\hat{S}(n) > \alpha'$, the onset detection is confident enough and the full amount of the reference spectrum is subtracted to enhance new notes. When $\alpha \leq \hat{S}(n) \leq \alpha'$, the detected onset might be a false positive. We subtract just a portion of the full amount so that the chromagram does not become totally blank in that case. Note that ideally the reference frame m would serve our purpose the best if it were immediately before the onset, even though the spectral difference among the several frames before the onset is subtle. We pick m as five frames before to leave a safe margin in case the detected onset is a delayed prediction, while it is still not too large to reach the previous onset.

Figure 3.3 compares the chromagrams calculated with and without the spectral subtraction operation in Equation (3.5). With the spectral subtraction operation, spectral components of sustained sound are greatly reduced (instead of totally removed) from frames right after onsets, whereas components of the new notes remained. This greatly reduces the audio-score

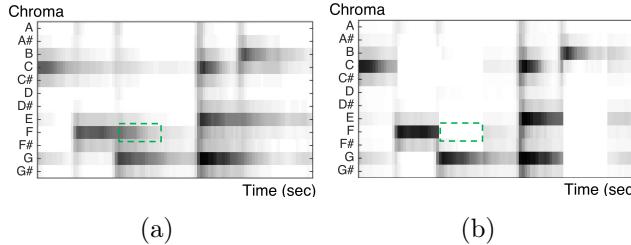


Figure 3.3: Chromagrams calculated without (a) and with (b) spectral subtraction. Sustained sounds after onsets (e.g., that marked by the rectangle) are reduced by the sustained-sound reduction operation.

mismatch as illustrated in Figure 3.4. It shows a MIDI score with two inter-onset segments x_1 and x_2 , where note G is supposed to end when note B starts. For an audio frame right after the onset (e.g., the n -th frame), we can see that it contains both notes, including the extension of note G due to the sustained effect. It is therefore not a precise performance of the correct segment x_2 in the score. Let $M(y_n, x_1)$ denote the match between the audio frame y_n and the score segment x_1 , and let $M(y_n, x_2)$ be defined similarly, then it is difficult to tell which match is better. After the sustained sound reduction operation, however, the note G is greatly reduced in the audio representation of the n -th frame, and it becomes clear that it has a better match to the correct score segment x_2 .

3.4.2.2 Peak Removal for Spectral-peak Representation

The spectral-peak representation is also commonly used in audio-score alignment (Orio and Schwarz, 2001; Duan and Pardo, 2011*b,a*). In this representation, the magnitude spectrum is reduced to a set of frequency-amplitude

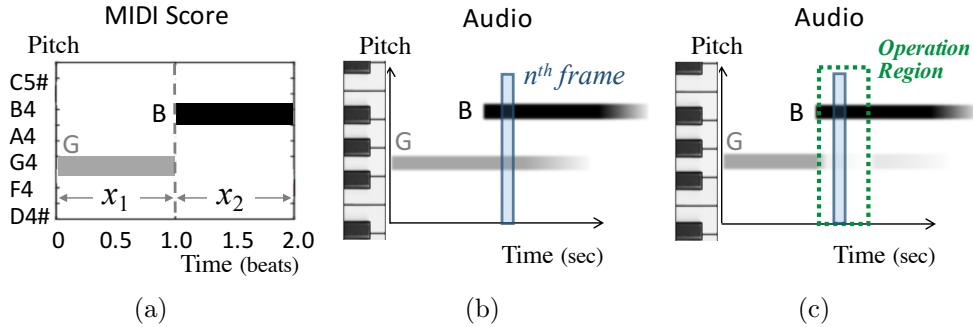


Figure 3.4: Illustration of the audio-score mismatch reduced by the proposed sustained-sound reduction operation, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation. The blue patch in (b) and (c) indicates the current frame (n -th frame) which lies in the operation region.

pairs of significant peaks:

$$\mathcal{P} = \{\langle f_i, a_i \rangle\}_{i=1}^K, \quad (3.7)$$

where K is the total number of peaks detected in the frame. The basis of this representation is that ideally the peaks correspond to harmonics of notes, through which the match between audio and score can be evaluated. For example, a good match would be that the score contains notes whose harmonics appear and only appear at the peaks.

To reduce the sustained effect in a detected region, we propose to remove peaks that correspond to sustained sounds in the spectral-peak representation. Figure 3.5 illustrates the idea. For each frame in a detected region (e.g., the n -th frame), we compare its spectral peaks with those in a reference frame before the onset (e.g., the m -th frame), and remove peaks that

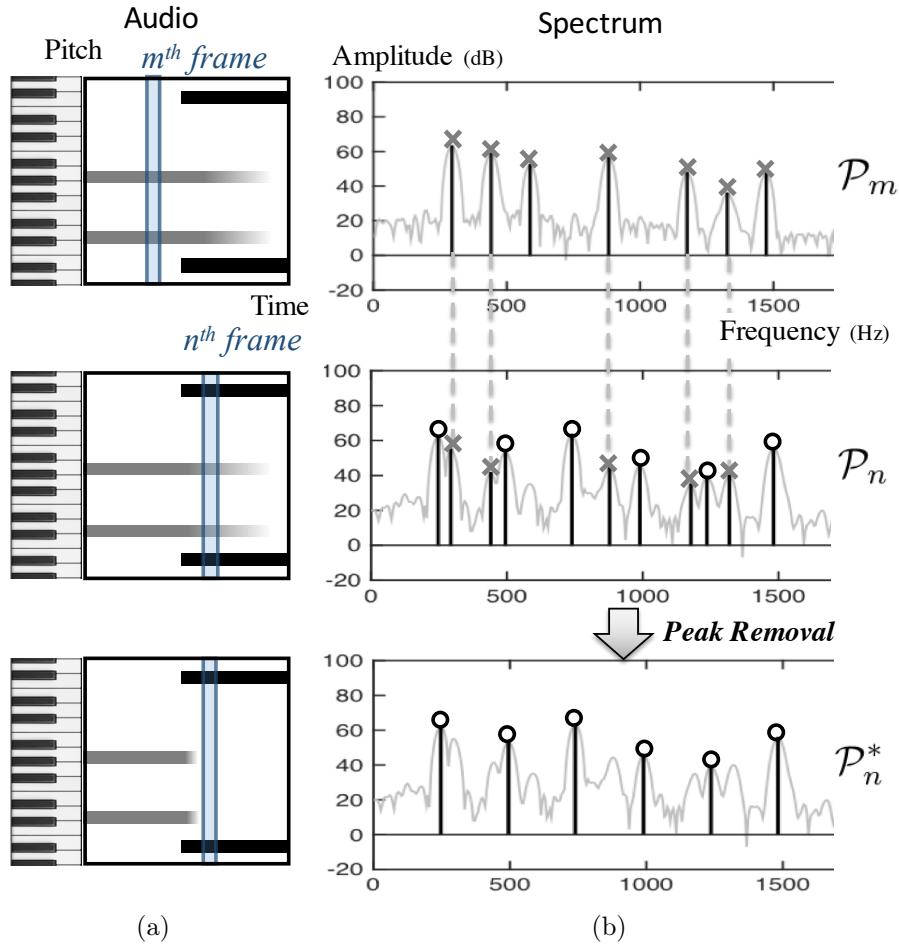


Figure 3.5: Illustration of the spectral peak removal idea. The m -th frame is the reference frame and the n -th frame is a frame under the removal operation. (a) Audio representations in the pianoroll format before and after peak removal. (b) Magnitude spectra with detected spectral peaks in the m -th and n -th frames. Peaks marked by crosses correspond to the first two notes. Peaks marked by circles correspond to the latter two notes.

seem to be extended from the earlier frame. Let $\mathcal{P}_m = \{\langle f_i^m, a_i^m \rangle\}_{i=1}^{K_m}$ be the total K_m peaks detected in the m -th frame, and $\mathcal{P}_n = \{\langle f_j^n, a_j^n \rangle\}_{j=1}^{K_n}$ be the total K_n peaks detected in the n -th frame. A peak in the n -th frame whose frequency is very close to *and* whose amplitude is smaller than those of a peak in the m -th frame is considered as an extension and is removed. Note that repeated notes will not be removed in this way as the amplitude criterion is not met. Thanks to the same reason, partials of the new notes are generally not removed even if they overlap with partials of the extended notes, because a significant energy increase is often observed at the onset of the new notes. After peak removal, a new spectral peak representation of the n -th frame is obtained as

$$\mathcal{P}_n^* = \mathcal{P}_n - \{\langle f_i^n, a_i^n \rangle : \exists j \text{ s.t. } |f_i^n - f_j^m| < d, a_i^n < a_j^m\}, \quad (3.8)$$

where d is the threshold for the allowable frequency deviation, which is set to a quarter-tone in this work. Although this operation is not applied to all frequencies as the spectral subtraction operation does for the chromagram representation, it also reduces the audio-score mismatch and reduces the delay error in score following.

3.4.3 What If the Proposed Operations Are Wrongly Applied?

As described in Section 3.4.2, the proposed sustained-sound reduction operation is applied to the entire 150-ms region after an onset. It reduces audio-score mismatch due to the sustained effect in that region and reduces delay errors in score following. However, not every region after an onset contains the sustained effect. In addition, onset detection has false positive errors. In these cases, the proposed operation will be wrongly applied. Will it be harmful for the audio-score match and score following? There are in total three cases in which the operation will be wrongly applied, and we analyze them in detail in this section.

3.4.3.1 Case 1: Onset Is Correct but There Is No Sustained Sound

In this case, the onset is correctly detected but there is no sustained sound after the onset, i.e., the old notes (e.g., staccato notes) simply cease before the next onset. Figure 3.6 shows an example. The spectral subtraction operation for the chromagram representation in Equation (3.5) will still subtract a portion of the old notes' spectrum and may contaminate the spectrum of the new notes. However, this effect is subtle. The subtraction is half-wave rectified, so it will not affect frequencies that the new notes do not contain. It will not affect the frequencies that the old notes do not contain either, since there is nothing to subtract. The only frequencies that it affects are frequencies shared by the old and new notes. In these frequencies, the new

notes are likely to have a larger amplitude than the old notes as they have just started. Therefore, the subtraction will just reduce the amplitude instead of removing them. This corresponds to a slight timbre change in the modified audio representation, but the same harmonic content is represented, which is the key for audio-score match in the chromagram representation.

For the peak removal operation, no partials of the new notes, whether overlapped with the old notes or not, are likely to be affected. This is because the amplitude criterion in Equation (3.5) is not likely to be satisfied due to the amplitude increase at the onset, as explained in Section 3.4.2.2. Therefore, the proposed operation in both kinds are not harmful to the audio-score match nor score following. However, if the operation region is too long, the amplitude criterion might be satisfied in some latter frames and there can be some slight negative effects. The above analyses are verified by the comparable results achieved with and without the proposed operation on pieces that do not contain apparent sustained effect and the experiment on the operation region length in Section 3.5.

3.4.3.2 Case 2: Onset Is Correct but Sustained Sound Is Notated

In this case, the onset is correctly detected, and there is sustained sound after the onset. However, this sustained sound is not due to the sustained effect but simply because the notes are supposed to sustain according to the score. This case happens very often in piano performances, e.g., the right hand is playing fast melody while the left hand is holding long chords. The left-hand

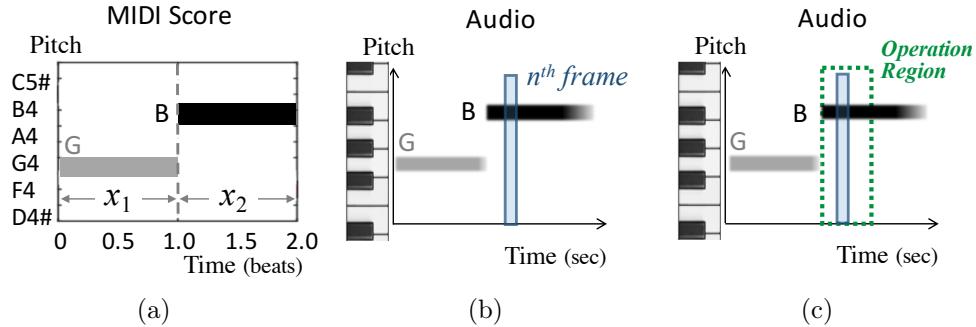


Figure 3.6: Illustration of the negligible effect of the proposed sustained-sound reduction operation when there is no sustained sound, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation.

notes are sustained according to the score. Figure 3.7 shows an example, where note G should be extended after the onset of note B according to the score. In this case, our proposed operation will wrongly remove note G in the n -th frame and decrease the match between the frame with the correct score segment x_2 . However, the match between the n -th frame and the wrong score segment x_1 will be decreased even more. In fact, they will not match at all as they will not share any note after the operation. Therefore, the operation will actually make the score follower favor the correct segment x_2 , even though it introduces new mismatch between audio and score. Good experimental results of the proposed approach on a variety of piano pieces support our claim.

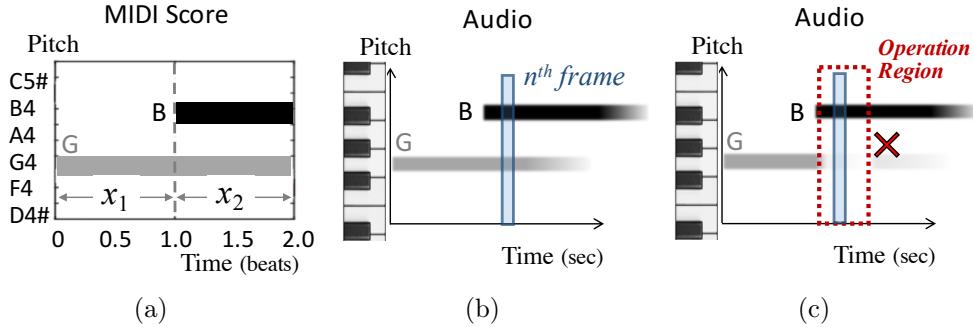


Figure 3.7: Illustration of the new audio-score mismatch introduced by the proposed sustained-sound reduction operation when the sustained sound is not due to the sustained effect but is notated in the score, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation.

3.4.3.3 Case 3: Onset Is A False Positive Error

The last case is that the onset is a false positive, and the sustained sound reduction operation is applied in frames that are in the middle of notes. Figure 3.8 shows an example. In this case, for the spectral subtraction operation designed for the chromagram representation, the n -th frame is not likely to be severely affected. This is because the spectrum of the n -th frame is likely to be just reduced but not totally removed, thanks to the confidence function employed in subtraction in Equation (3.5). For the spectral-peak representation, however, almost all peaks in the n -th frame will be removed, because both the frequency and amplitude criteria in Equation (3.8) for peak removal are likely to be satisfied for these decaying partials. Therefore, the spectral-peak representation can be severely contaminated and mismatch between audio and score can be introduced. The score follower can have difficulty

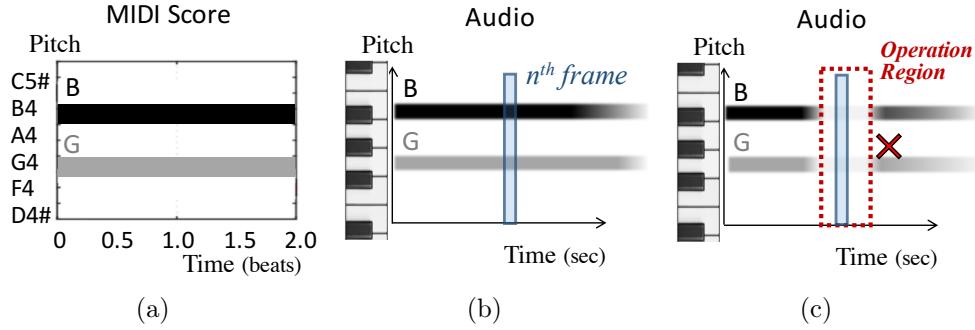


Figure 3.8: Illustration of the proposed sustained-sound reduction operation when the onset is a false positive, with piano-roll representations of: (a) MIDI score, (b) audio before the operation, (c) audio after the operation. The audio representation in the rectangular region is likely to preserve some energy of the notes after the spectral-subtraction operation but can be totally blank after the peak-removal operation.

in matching the audio to any score position. To sum up, false positive errors (which lead to lower precision) of onset detection are not significantly harmful for score following with the proposed sustained-sound reduction operation with the chromagram representation, but it is greatly harmful with the spectral-peak representation. Experiments in Section 3.5.3.1 support our claims.

Note that here we only analyze false positive errors but not miss errors in onset detection. This is because miss errors only prevent the proposed sustained-sound reduction operation from being applied but do not wrongly apply it. In addition, we do not consider offset detection for two reasons: 1) offsets are difficult to detect due to the lack of abrupt changes in energy and spectral content in the audio signal; 2) offsets are related to possible sustained effect in the past, hence their detection is not helpful in an online

system.

3.4.4 The Score Following Framework

The modified chromagram and spectral-peak representations in Section 3.4.2 can be employed by various score following frameworks to cope with the sustained effect in piano performances. In this work, we adopt an effective Markov model-based framework (Duan and Pardo, 2011*b*). This framework uses a 2-d state variable s_n to represent the score position and tempo of the audio in the n -th frame. A process model $p(s_n|s_{n-1})$ is defined to describe how the states transition from one to another: The score position advances from the previous frame according to the tempo and the tempo changes through a random walk. An observation model $p(y_n|s_n)$ is defined to represent the likelihood of the hidden state s_n in explaining the observed spectrum y_n . Provided the process model and the observation model, the hidden state s_n can be inferred by particle filtering from current and previous audio observations y_1, \dots, y_n . The framework is illustrated in Figure 3.9.

The proposed audio representations are integrated into the observation model $p(y_n|s_n)$. For the chromagram representation, $p(y_n|s_n)$ is defined as a Gaussian distribution of the cosine distance between chroma vectors calculated from the audio and the score. For the spectral-peak representation, $p(y_n|s_n)$ is defined through a multi-pitch likelihood function proposed in (Duan, Pardo and Zhang, 2010). Details of these observation models can

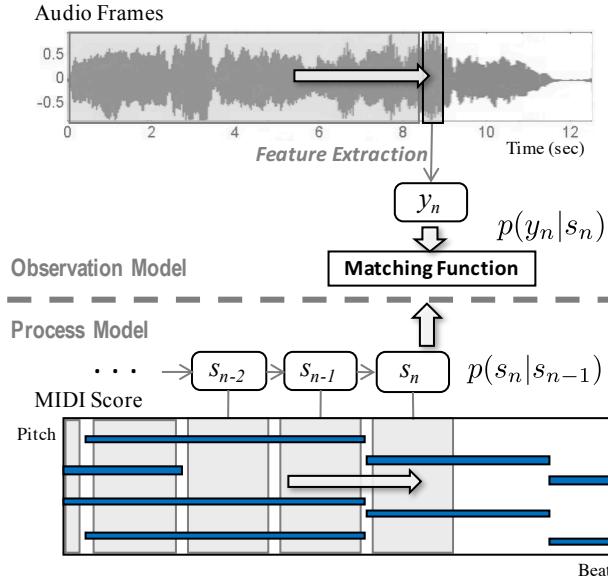


Figure 3.9: The score following framework adopted to implement the proposed sustained-sound reduction approach.

be found in (Duan and Pardo, 2011*b*).

3.5 Experiments

3.5.1 Experimental Set-up

3.5.1.1 Dataset

To evaluate audio-score alignment systems, a set of audio recordings and their musical scores, as well as the ground-truth alignment between the audio and score are needed. For the audio recordings, we used the entire 60 acoustically recorded pieces from the two folders (*close* and *ambient*) of the MAPS dataset

(Emiya, Badeau and David, 2010) to cover a large variety of styles, playing techniques, and the degree of the sustained effect in piano performances. These pieces were from more than 14 different composers including Chopin, Mozart, Liszt, etc., and contained different degrees of the sustain effect. All the pieces were acoustic recordings of a Yamaha Disklavier that took pre-generated MIDI performances as input. Since the degree of the sustained effect varies much even within the same piece, for each piece we choose a short clip with a length between 30 and 60 seconds in our experiments, aiming for a more similar degree of the sustained effect within each clip.

For the musical scores, we downloaded the standard MIDI score for each piece from the Classical Piano MIDI Page². Beside the tempo differences, these MIDI scores also contain other minor differences from the MIDI performances in the MAPS dataset. These differences were due to the occasional missed or added notes, slight desynchronizations between melody and accompaniment (Goebl, 2001), and different renderings of trills in the MIDI performances. Since the MIDI performance of each piece has exactly the same timings as the audio recording, we performed an offline DTW followed by manual corrections to align the MIDI performance with the MIDI score to obtain the ground-truth audio-score alignment.

According to the degree of the sustained effect, we categorized the 60 pieces into three groups. As the sustain pedal usage is a main source of the sustained effect, we first categorized the 12 pieces that were performed

²<http://piano-midi.de/>. Downloaded in October, 2015.

without any sustain pedal usage into the first group, $P1$ (No Pedal). The pedal usage information was extracted from the MIDI performances using java MIDI tools. Note that slight sustained effect may still exist in this group, due to the legato note articulations. For the rest 48 pieces, we then calculated the *Sustained-effect Frame Rate (SFR)*, which is defined as the percentage of mismatch frames due to the sustained effect, i.e., the *active notes* in the MIDI performance are more than those in the aligned MIDI score. Here the offsets of notes in the MIDI performance are extended from a sustain-pedal-down period to the next pedal-release time. The 48 pieces were then divided into two groups, $P2$ (Slight) and $P3$ (Heavy), by a threshold (50%) on the SFR. This threshold was chosen as it is at the low-density region of the distribution and it roughly balances the two groups. Figure 3.10 (a) shows the SFR of all the three groups. One may notice the overlap between $P1$ and $P2$ and ask why we did not use another threshold on SFR to divide $P1$ and $P2$. Our rationale was that SFR is a rather arbitrary measure of the sustained effect while using pedal or not is a clear distinction among pieces. We would like to hold a set of pieces with absolutely no pedal usage.

Figure 3.10 (b) also shows statistics of three other measurements of the complexity of the pieces calculated from the MIDI score barring the sustain-pedal usage:

- *Note density*: Average number of notes per second, N_{note}/T , where N_{note} is the total number of notes and T is the length of the piece.

- *Polyphony density*: Average number of simultaneous notes, $\frac{1}{N} \sum_n u_n$, where N is the total number of frames and u_n is the number of active notes in frame n .
- *Inter-onset duration*: Average time gap between the note onsets, $\frac{1}{N_{onset}} \sum_i (t_i - t_{i-1})$, where N_{onset} is the total number of unique onsets, and t_i is the time instant of the i -th unique onset.

We performed two-sample t-tests for all group pairs, and found that for all the three measurements, the groups are not significantly different at the significance level of 5%, except for the note density between $P1$ and $P3$, which shows a p value of 0.0485. This helps us to rule out factors other than the degree of the sustained effect that may affect the score following performance, in the situation where controlled experiments are not easy to design.

To measure the length of each region with the sustained effect (the shaded region in Figure 3.1) in our test pieces, we calculated the *Sustained-effect Region Length (SRL)* for the three groups in Figure 3.11. SRL is defined as the length of the mismatch region between the MIDI performance and the aligned MIDI score for each unique onset. Again, the offsets of notes in the MIDI performance are extended from a sustain-pedal-down period to the next pedal-release time. Note that in Figure 3.11 we exclude notes with a zero SRL, e.g., staccato articulations.

We further created another set by adding reverberation to all the 60 pieces

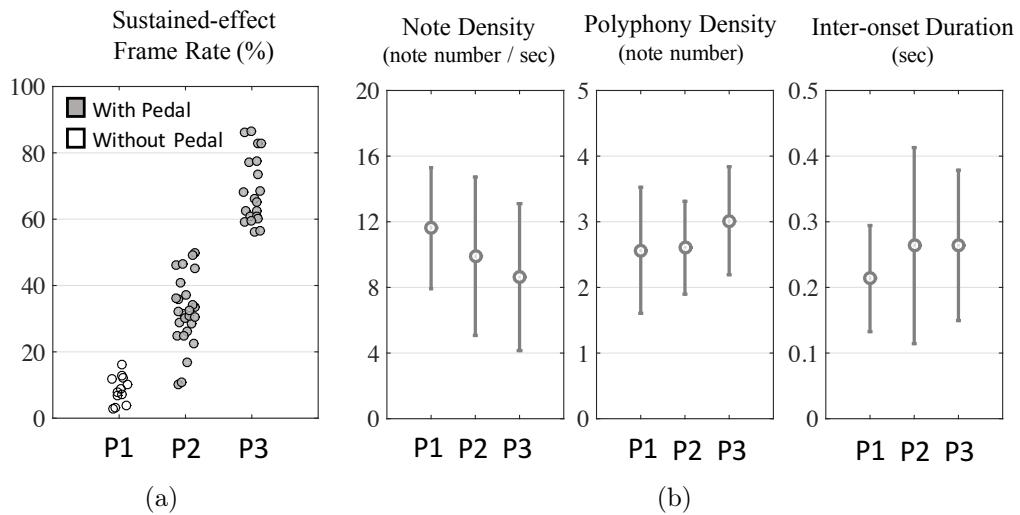


Figure 3.10: Statistical MIDI information of the testing dataset. (a) Scatter plot of Sustained-effect Frame Rate of the three testing groups, calculated from MIDI performances. Each dot represents one piece. (b) Three measurements of piece complexity of the three testing groups, calculated from MIDI scores. The central circles and vertical bars denote the means and standard deviations.

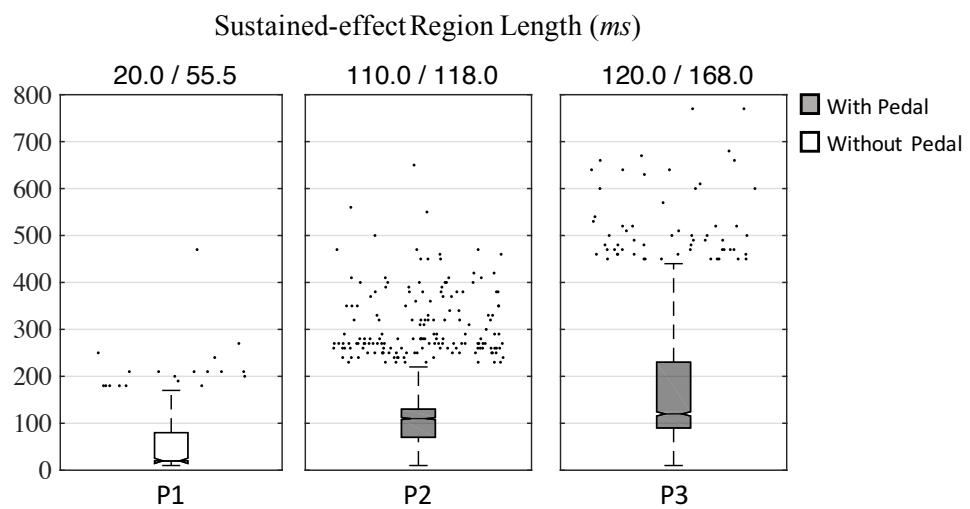


Figure 3.11: Boxplot of the Sustained-effect Region Length (SRL) for all unique note onsets (excluding those with a zero SRL.) in pieces in P_1 (283 onsets), P_2 (1680 onsets), and P_3 (2275 onsets). Outliers are displayed in a more dispersed way for better visualization. Numbers above each box show the median/mean value of all the points.

by convolving each piece with a room impulse response ($RT_{60} = 2.34\text{s}$) sampled from the St. Albans Cathedral³. Note that we did not use the reverberant pieces provided in the MAPS dataset for two reasons: 1) They were synthesized from MIDI using software instead of being acoustically recorded; 2) They use different pieces from the above-mentioned 60 acoustic recordings hence are hard to compare. By adding the external reverberation, we formed a controlled experiment.

3.5.1.2 Evaluation Measures

We use two kinds of evaluation measures for score following:

- *Align Rate (AR)*: It is defined as the percentage of correctly aligned unique onsets among all unique onsets in the score. Simultaneous onsets of different notes in the score are treated as a single onset. An onset i is considered correctly aligned if its aligned audio time \hat{t}_i deviates less than a threshold from the ground-truth reference audio time t_i . Commonly used thresholds range from 50 ms to 1 second depending on the application. For an automatic accompaniment system, a deviation within 50 ms would be required, while for an automatic page turner, 1 second would be still fine. We use 50 ms as the threshold in the experiments.
- *Average Onset Deviation (AOD)*: It calculates the average absolute

³ <http://www.openairlib.net/auralizationdb/content/lady-chapel-st-albans-cathedral>. Downloaded in December, 2015.

time deviation of the alignment of all unique onsets of a piece, i.e.,
 $\frac{1}{N_{onset}} \sum_i |\hat{t}_i - t_i|$.

Since our proposed sustained-sound reduction operation is built upon an online onset detection module, we also evaluate onset detection performance and examine its effect on score following. Again, we only consider unique onsets in the ground-truth. In the audio, onsets of simultaneous notes may deviate from each other slightly, but our algorithm usually only detects one onset thanks to the time threshold β . A detected onset is considered correct if it deviates from a ground-truth onset less than 30 ms. Each ground-truth onset can only be associated with at most one correctly detected onset. We use precision $P = Corr/Est$, recall $R = Corr/Ref$, and F-measure $F = 2PR/(P + R)$ to evaluate onset detection results for each piece, where $Corr$ is the number of correctly detected onsets, Est is the number of all estimated onsets, and Ref is the number of all reference onsets.

3.5.1.3 Parameter Settings

In all the experiments, we set frame size to 46.4 ms and hop size to 10 ms for STFT in audio spectrogram calculation. For onset detection, we set the amplitude threshold α to 40, the time threshold β to 3 frames (30 ms), and the window length threshold ω to 5 frames (50 ms). Section 3.5.3.1 analyzes how onset detection errors affect the score following results. For sustained-sound reduction, the reference frame was set to 5 frames (50 ms) before each onset frame to leave some room to cope with the onset location

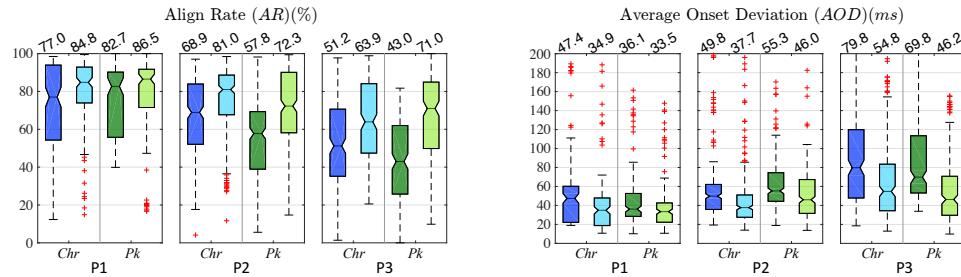


Figure 3.12: Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*). The number above each box shows the median.

estimation inaccuracies. The length of the operation region L where the proposed sustained-sound reduction operation performs is set 15 frames (150 ms) immediately after each onset frame, or until the next onset if it is within the 15 frames. This parameter setting is informed by Figure 3.11, and how its value affects the score following performance is analyzed in Section 3.5.3.2. Parameters of the other parts of the score following system remain the same as those reported in (Duan and Pardo, 2011*b*). Note that all the parameters in this work were set the same for all test pieces and in both non-reverberant and reverberant cases.

3.5.2 Results in Non-reverberant Cases

We first evaluate the system on the three groups in non-reverberant cases, namely $P1$, $P2$, and $P3$. Figure 3.12 shows box plots of the two score following measures (AR and AOD) of the baseline system (dark colors) and the

proposed system (light colors) using both the chromagram representation (blue) and the spectral-peak representation (green). Due to the randomness of particle filtering in the score following framework, we ran each system 10 times per piece. Therefore, each box (along with outliers shown as red crosses) represents $10 \times \# \text{Pieces}$ data points, e.g., 120 points for each box in the group $P1$. The median of each box is shown on the top. To avoid clutter, we cut off some outliers in the figures of AOD.

We conducted paired sign tests on AR between the baseline and proposed system using both audio representations and in the three groups (6 pairs), and all the improvements passed the significance level of 0.005. More interesting observations can be made from Figure 3.12. First, comparing the three recording groups, we see that score following performance using both representations generally degrades when the sustained effect becomes stronger ($P1 \rightarrow P2 \rightarrow P3$) for both measures. This degradation is especially pronounced for the baseline system, while is slighter for the proposed system. Given that the piece complexity of the three groups are similar, this shows that the sustained effect is indeed an issue for piano score following and the proposed approach is able to alleviate this issue. Second, as the sustained effect becomes stronger, the improvement of our proposed system also grows, e.g., 3.8 % in $P1$ and 28% in $P3$, using the spectral-peak representation. Third, comparing the two audio representations, the spectral-peak representation yields more pronounced improvement in the proposed system over the baseline system. This is because the peak removal operation in this

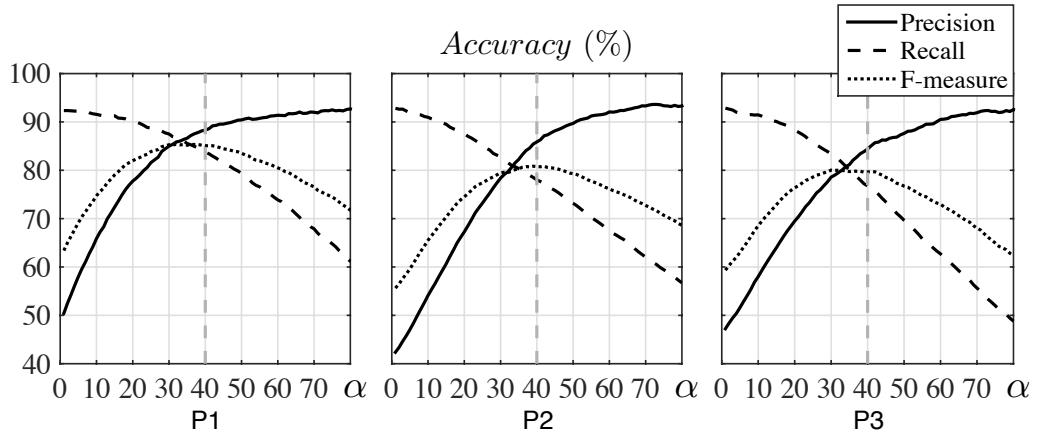


Figure 3.13: Onset detection results averaged over pieces within each recording group. Note that the three curves do not necessarily intersect at the same point because they are average values.

representation removes the sustained peaks entirely, while in the chromagram representation the sustained sound is only reduced (see Equation 3.5). However, the entire removal makes the system sensitive to onset detection errors, as analyzed in Section 3.4.3.3 and shown in the following experiments.

As online onset detection is an important module of the proposed approach, we also evaluate its performance. Figure 3.13 shows the average precision, recall, and F-measure curves of onset detection within each recording group by varying the amplitude threshold α from 1 to 80 with a step of 1. It can be seen that as the sustained effect becomes stronger, onset detection performance becomes worse. The best F-measure is achieved around $\alpha = 35$ for all groups. However, the finally chosen threshold is $\alpha = 40$, as shown by the dashed vertical lines, to prefer a higher precision than recall, as explained in Section 3.4.3.3. Note that we use the same threshold throughout

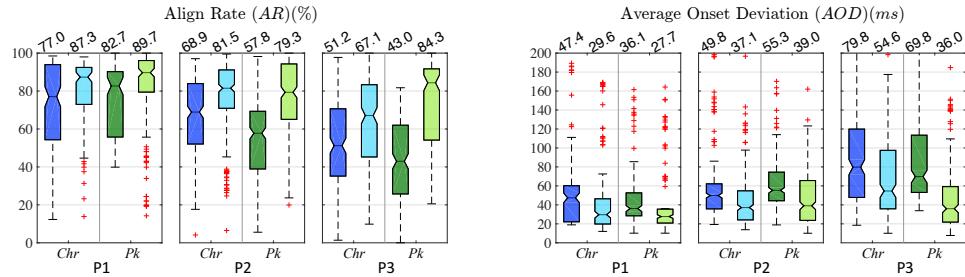


Figure 3.14: Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*) when the ground-truth onsets are used. The number above each box shows the median.

all experiments, although the optimal threshold differs for different pieces and groups. For the second threshold α' used in spectral subtraction for the chromagram representation, we set it to 70, which leads to a high precision for most pieces and an acceptable recall.

To isolate the effects of onset detection on the final score following performance, we also evaluate the proposed system with ground-truth onset information. Figure 3.14 shows results. We can see that the proposed system shows more observable improvement over the baseline system. The improvement is significant in all cases under a sign test at a significance level of 10^{-3} . It reaches to a median AR of 89.7%, 79.3%, and 84.3% for the three recording groups, respectively, using the spectral-peak representation. Also note that *P3* with heavy sustained effect reaches a comparable value with the other groups. These values set the upper bound for the proposed approach when more advanced onset detection technique is employed. Another interesting observation is that the spectral-peak representation yields

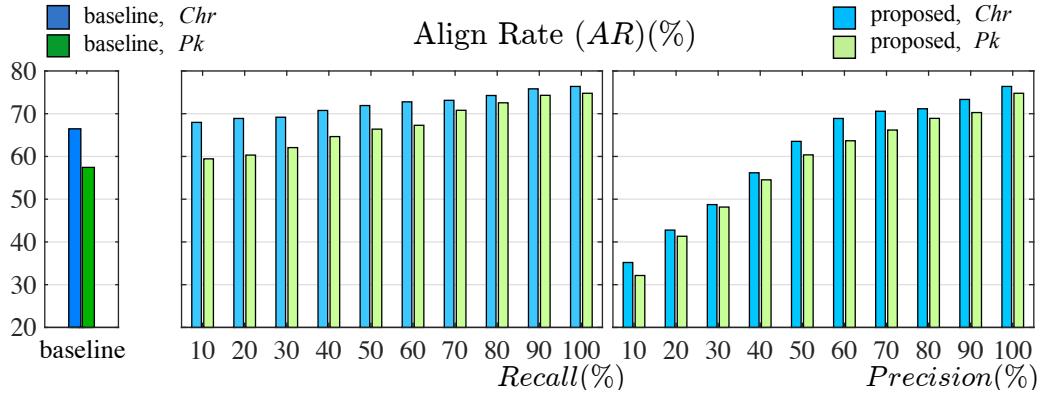


Figure 3.15: Score following performances (average of AR) of the baseline system and the proposed system built on artificially controlled onset detection results on *P2*. Precision is fixed at 100% in the middle panel while recall is fixed at 100% in the right panel.

better performance than the chromagram representation in this case, while their performances are similar in Figure 3.12. This is because the “bolder” peak removal operation in the spectral-peak representation, which is sensitive to onset detection false positives as analyzed in Section 3.4.3.3, removes the sustained effect more effectively than the “more conservative” spectral subtraction operation in the chromagram representation.

3.5.3 Parameter Analysis

3.5.3.1 Effects of Onset Detection Errors

As analyzed in Section 3.4.3.3, miss errors and false positive errors of onset detection have different effects on score following in our proposed approach. To further investigate this, we artificially created onset detection results with

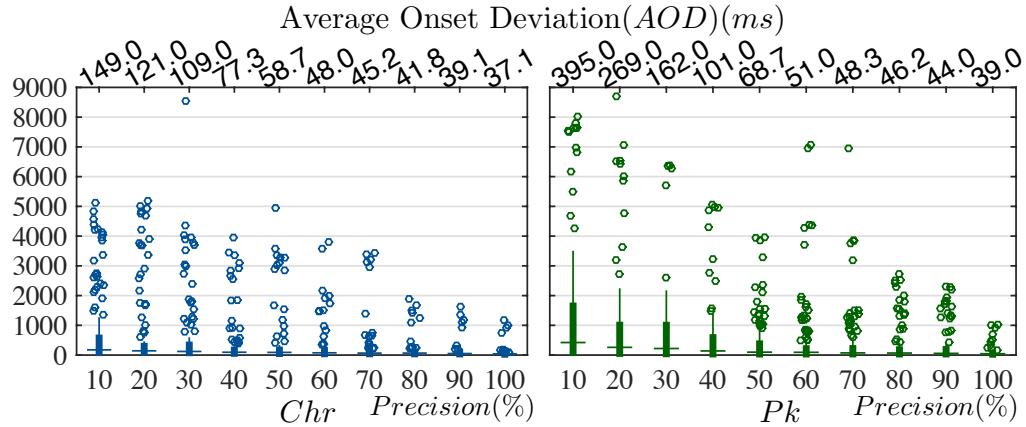


Figure 3.16: Boxplot of the Average Onset Deviation of the proposed approach on artificially controlled onset detection results on the group P_2 , where recall is fixed at 100% and precision is varied. Numbers above show the medians.

separate controls of precision and recall. Starting from the ground-truth onsets, we randomly removed some onsets to control the recall while maintaining the precision at 100%. We also randomly added false positive onsets to control the precision while maintaining the recall at 100%. Taking the group P_2 as an example, score following performance (average of AR values for all the pieces) based on these onset detection results are shown in Figure 3.15. We can see that the result is indeed less sensitive to the recall (miss errors) when the precision is 100% (middle panel), while it is sensitive to precision (false positive errors) even when the recall is 100% (right panel). When the precision is less than 50%, the proposed approach shows inferior performance than the baseline system. Note that 0% recall in the middle panel (i.e., no onset is detected) would make our proposed approach the same as the baseline

system.

In Figure 3.16, we further show the boxplots of AOD with recall fixed as 100% and precision varying for both chromagram and spectral-peak representations. The vertical axis covers a wide range to show outliers while the numbers above the figures show medians. We can see that lowering precision causes significant increase of the median AOD value and the number of outliers with extremely large deviations (e.g., >1000 ms). In fact, those outliers correspond to runs when the system was totally lost during score following. This usually happens on pieces with sparse notes. Comparing the two audio representations, *Pk* shows more outliers than *Chr*, especially when precision is low. This supports our claim in Section 3.4.3.3 that low precision is more harmful for the spectral-peak representation.

3.5.3.2 Effects of the Operation Region Length

To investigate the sensitivity of the proposed system on the sustained-effect operation length L , we conducted another experiment on the three groups in the non-reverberant case with different values of the parameter. Figure 3.17 shows the results. We can see that the average align rate for P1 and P2 stays stable once L reaches 100 ms. The average value of Align Rate for P3 achieves the highest value when L reaches 150 ms and then stays stable with the chromagram representation but slightly decreases when $L > 250$ ms with the spectral-peak representation. This result shows a good correspondence with the statistics of the sustained-effect region length (SRL) in Figure 3.11,

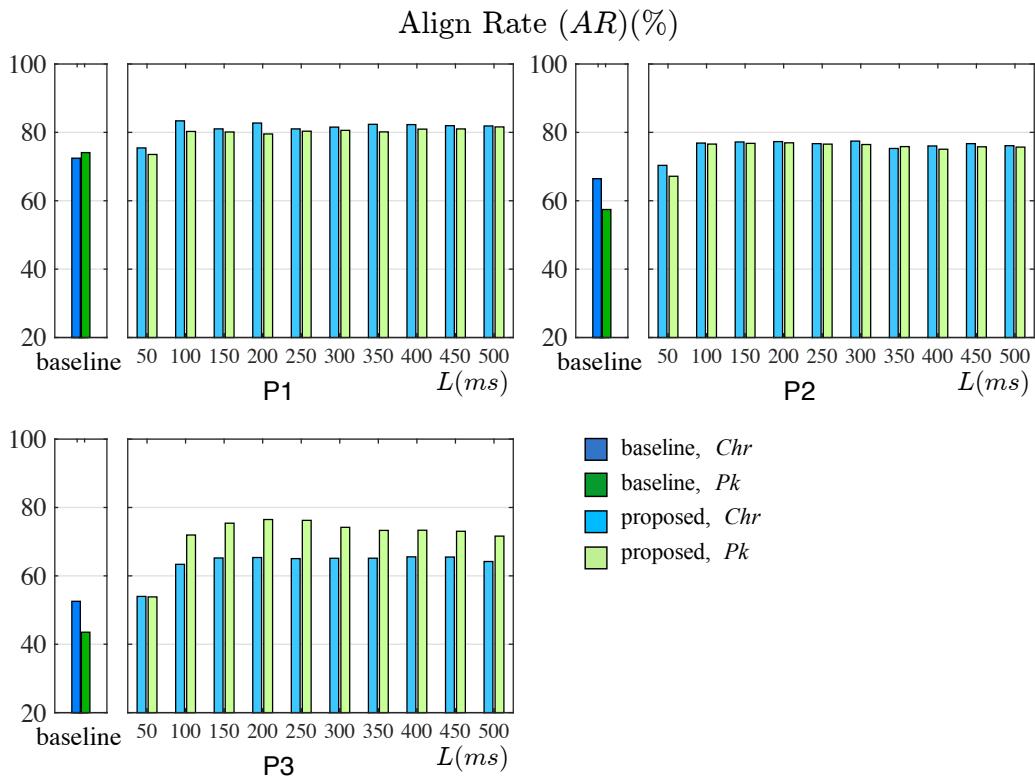


Figure 3.17: Score following performance (average of AR) of the baseline system and the proposed system with different lengths of the sustained-sound reduction operation region for the three groups.

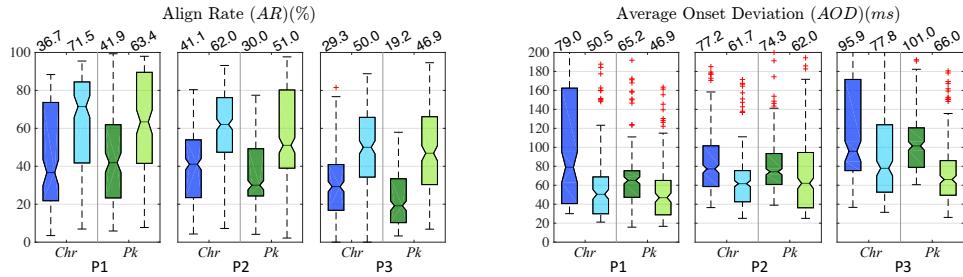


Figure 3.18: Score following results of the baseline system (dark colors) and the proposed system (light colors) using both the chromagram representation (*Chr*) and the spectral-peak representation (*Pk*) when reverberation is added to the audio. The number above each box shows the median.

where the average SRL for the three groups are 55.5 ms, 118 ms and 168 ms, respectively. When a piece has a larger SRL, the operation region should be longer as well. Interestingly, the experiment also shows that there is almost no negative effect if L is set too long. This again validates our analysis in Section 3.4.3.1 and Figure 3.6 that the operation is negligible when it is wrongly applied to regions with little or no sustained sound.

3.5.4 Results in Reverberant Environments

In this section, we further evaluate the proposed approach in reverberant environments. Reverberation imposes the sustained effect on all notes of a piece. It also blurs the onsets and makes onset detection challenging. As described in Section 3.5.1.1, we impose reverberation on all the 60 pieces in our dataset. Figure 3.18 and Figure 3.19 show the score following and onset detection result respectively. Comparing with Figure 3.13, we can see

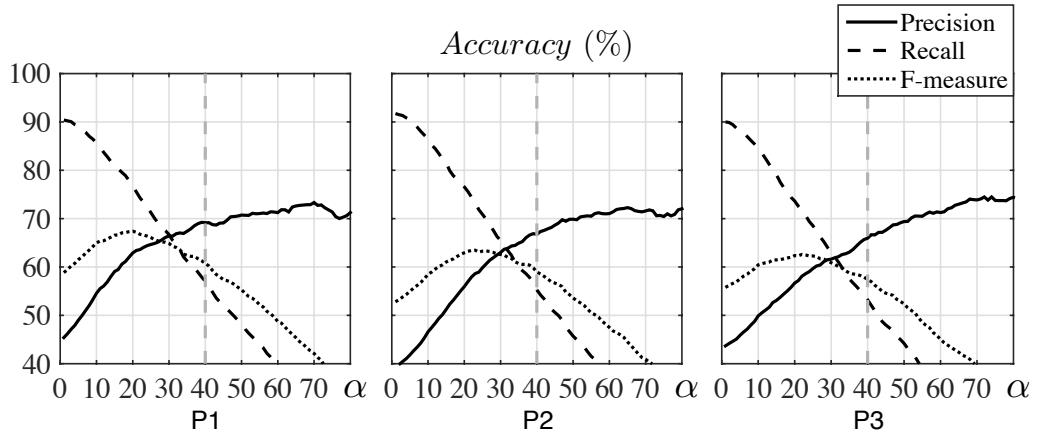


Figure 3.19: Onset detection results averaged over pieces within each recording group with reverberation added. Note that the three curves do not necessarily intersect at the same point because they are average values.

that the performance of our online onset detection degrades from anechoic environments to reverberant environments. For example, the best F-measure drops from 81% to 64% for the group $P2$. The threshold α that achieves the highest F-measure is now around 20. However, we still chose the same threshold 40 in the experiment to prefer higher precision than recall.

Figure 3.18 shows the score following results. Comparing with Figure 3.12, we can see that both the baseline system and the proposed system degrade dramatically. However, the improvement of the proposed system over the baseline system becomes more pronounced, especially in the $P1$ group, as they now have the sustained effect due to reverberation. The improvement on AR is significant in all settings under a sign test at the significance level of 10^{-14} . Another interesting observation is that the baseline system performs badly in some pieces from group $P1$. Further investigation indicates that

some pieces are performed with fast-running notes (see Figure 3.10 (b)). We argue that the reverberation blurred and blended these notes so much that the baseline system was not able to follow the pieces and result in a large number of outliers with extremely large deviations (not shown completely in the figures). The proposed system, however, was able to greatly alleviate this issue through the sustained-sound reduction operation for both chromagram and spectral-peak representations.

Audio examples of the above experiments can be accessed online⁴.

3.6 Conclusions

In this chapter we proposed a score following approach to follow piano audio performances with the sustained effect. The sustained effect refers to the phenomenon that notes are sustained longer than their notated length in the score. It blends the sustained notes with new notes, introduces audio-score mismatch, and causes delay errors of score followers. We analyzed three main causes of the sustained effect, namely the legato playing technique, the usage of the sustain and sostenuto pedals, and the room reverberations. We proposed an approach to reduce the sustained effect by reducing the sustained spectral components in a number of frames immediately after each detected onset. This approach was developed for two commonly used audio representations, the chromagram and the spectral-peak representations. We

⁴<http://www.ece.rochester.edu/~bli23/projects/pianofollowing.html>

integrated the proposed approach with a Markov model-based score following framework to test its effectiveness. We also analyzed effects of the proposed approach when it is wrongly applied. Systematic experiments showed that the proposed approach improved the score following accuracy and robustness significantly on a variety of piano pieces with different degrees of the sustained effect. Detailed analysis of the rationales of parameter settings and their effects on score following were also provided.

For future work, we plan to consider other specific properties of piano music to improve the alignment performance. For example, the sound decay may result in potential mismatch around note offset frames. In this case, a time-varying matching function that considers the exponential energy decay would improve the alignment accuracy.

Chapter 4

Source Association

In audio-visual recordings of music performances, visual cues from instrument players exhibit good temporal correspondence with the audio signals and the music content. These correspondences provide useful information for estimating source associations, i.e., for identifying the affiliation between players and sound sources or score tracks. In this chapter, we propose a computational system that models audio-visual correspondences to achieve source association for Western chamber music ensembles. Through its three modules, the system models three typical types of correspondences between 1) body motions (e.g., bowing for string instruments and sliding for trombone) and note onsets, 2) finger motions (e.g., fingering for most woodwind/brass instruments) and note onsets, and 3) vibrato hand motions (e.g., fingering hand rolling for string instruments) with pitch fluctuations. Although the three modules are designed for estimating associations for different instruments, the overall system provides a universal framework for all common instruments in Western chamber ensembles by automatically and adaptively

integrating the three modules, without requiring prior knowledge of the instrument types. The system operates in an online fashion, i.e., associations are updated as the audio-visual stream progresses. We evaluate the system on ensembles with different instruments and ranging in polyphony from duets to quintets. Results demonstrate that association accuracy increases as the video excerpts become longer. For string quintets, the accuracy is over 90% from just a 5-second video excerpt, while for woodwind, brass, and mixed-instrument quintets, a similar accuracy can be reached after processing 30 seconds of video. This promising result allows the system to enable novel applications such as interactive audio-visual music editing and auto-whirling camera in concerts.

4.1 Introduction

Visual aspects of music performances are as important as the sound in many scenarios. In live concerts, performers use various kinds of body movements to express their emotions and to impress audiences (Parncutt and McPher-
son, 2002; Sörgjerd, 2000). In music ensembles, visual interactions among musicians are important for coordination of timing and dynamics. In pop music, creative visual performances give artists a substantial competitive ad-
vantage. The inclusion of videos in music albums is shown to provide an eight percentage boost, on average, in purchase intent and improved percep-

tion (measured by Nielsen Holdings¹). Even in prestigious classical music performances, research has shown that body movements and facial expressions of performers exert strong influences on the judgment of performance quality, for expert or novice audiences alike (Tsay, 2014).

On the technical side, the rapid expansion of digital storage and Internet bandwidth in the past decades has not only popularized video streaming services like YouTube but also significantly changed the way people enjoy music. With the surge of Virtual Reality (VR) and Augmented Reality (AR) technologies and their adoption in music entertainment, visual aspects of music performances will further gain their importance in innovative music enjoyment experiences.

While Music Information Retrieval (MIR) based on the audio signal and symbolic score/MIDI representations has been widely studied, only limited explorations have been conducted on the interplay of visual and acoustic aspects of music performances. The auditory and visual modalities are intimately related in music performances. The instrument player's movements invariably mediate sounds from acoustic instruments and characteristics of the movements are reflected in the resulting sounds. For example, the amplitude envelope and spectral evolution of a violin note are directly related to the velocity and pressure of a bowing motion (Askenfelt, 1989) and fingering force (Obata et al., 2009); the timing of a clarinet note is often correlated to the fingering fluctuation; the loudness of a drum hit is strongly related to

¹www.nielsen.com

the the drumstick's preparatory height and striking velocity (Dahl, 2004).

Classical chamber music is performed by a small ensemble of instrumentalists, with *one player per score track* (Burkholder and Grout, 2014). In this work, we study the relationship between the instrument players' body movements and sound events in classical chamber ensemble performances, with the aim of solving the *source association* problem, i.e., identifying the bijection between score tracks (MIDI or MusicXML format) and players in the video. This bijection, together with a score-informed audio source separation technique (Duan and Pardo, 2011a), can allow users to separate the audio source for each particular player in the video.

Exploiting information in the video about instrument players' movements for source association is challenging because many body movements (e.g., head movement) are irrelevant to sound articulation (Godøy and Jensenius, 2009) and relevant movements (e.g., maneuver with fingers) can be subtle. In music ensembles, similar body movements can be observed among different musicians when they have similar rhythmic patterns. These challenges are especially pronounced when the video clip is short (e.g., from online streams) and when the ensemble is large. For a quintet, possible associations can be enumerated as 120 permutations, but only one is correct.

Source association enables novel research and applications. It is essential for leveraging the visual information to analyze individual sound sources in music performances. The related techniques include multi-pitch analysis (Dinesh et al., 2017), performance expressiveness analysis (Li, Dinesh, Sharma

and Duan, 2017), source separation (Parekh et al., 2017), etc. By exploiting source associations, one can envision an augmented video streaming service that allows users to click on a player in the video and isolate/enhance the corresponding source of the audio (Zhao et al., 2018). One can also envision an augmented sheet music display interface where on each score track, the visual performance of the specific player is demonstrated. For music production, source association can help enable remixing of audio sources along with automatic video scene recomposition. An online source association system, which does not need to “look into the future”, can be further useful in online video streaming of live concerts. For example, it enables an auto-whirling camcorder to focus on the soloist.

In this work, we build upon our prior work on source association for string instruments using bowing motions (Li, Dinesh, Duan and Sharma, 2017) and vibrato motions (Li, Xu and Duan, 2017), and propose the first universal system to address the problem for all instruments commonly used in Western chamber ensembles, including string, woodwind, and brass instruments. This system does not require prior knowledge of instrumentation of the piece or pre-training of audio-visual correspondence. After temporally aligning the score with the live performance, the system uses three modules to analyze different motion types that may be present in the performance, as shown in Figure 4.1.

As many performance motions are related to note onsets, the first two modules focus on the motion-onset correspondence. The first module extracts

large-scale body motions, which mainly capture bowing motions of string instruments. The second module extracts subtle fingering motions and correlates these with note onsets. This aids with associations for woodwind/brass instruments, as pitch changes are mostly controlled by finger-operated keys. In addition to note onsets, variations of acoustic features throughout tone articulations also show correspondence with certain motions, for instance, for the vibrato articulation in string instruments. Therefore, the third module is designed to detect periodic fingering motions (if any) and to correlate them with the periodic pitch fluctuation estimated from audio. It is expected to only work for string instruments where vibrato articulations can be characterized from the visual modality. Finally, the output of the three modules is integrated through weighted voting according to the motion salience. It is noted that the system does not need to detect the instrument type; it simply extracts the three kinds of motions (if any) for each player and integrates their correspondence with score/audio tracks, jointly.

The proposed system works in an online fashion: The audio-score alignment, the correlation between motion and audio/score, and the association output are all updated in a frame-by-frame fashion without “looking into the future”. Associations in each frame are updated using the Hungarian algorithm (Kuhn, 1955), with a minimum computational cost. Experiments on 17574 audio-visual clips generated from 44 chamber music pieces in the URMP dataset (Li, Liu, Dinesh, Duan and Sharma, 2019) that spans a polyphony range from duets to quintets, show that: 1) Different modules

are helpful for different instruments, and the system is able to integrate them automatically to achieve a high overall accuracy; 2) Accuracy increases as longer video streams are available, reaching an average accuracy of 90% for 5-second video excerpts of string instruments, and for 30-second excerpts of woodwind and brass instruments. In summary, the proposed system for audio-visual source association:

- works universally for all instruments common in Western chamber ensemble performances,
- does not require prior knowledge of instrumentation, and
- relies purely on motion-information for association without modeling instrument characteristics; which allows it to also work for ensembles of the same instrument type, e.g., violin duets.

In the following, we first review existing work on multi-modal modeling in Section 4.2, and highlight challenges involved in source association in music performances. We then describe our proposed method in three modules for the different motion cues for associations in Section 4.3. In Section 4.4, we conduct systematic experiments to evaluate the proposed system. Finally, we conclude the work in Section 4.5.

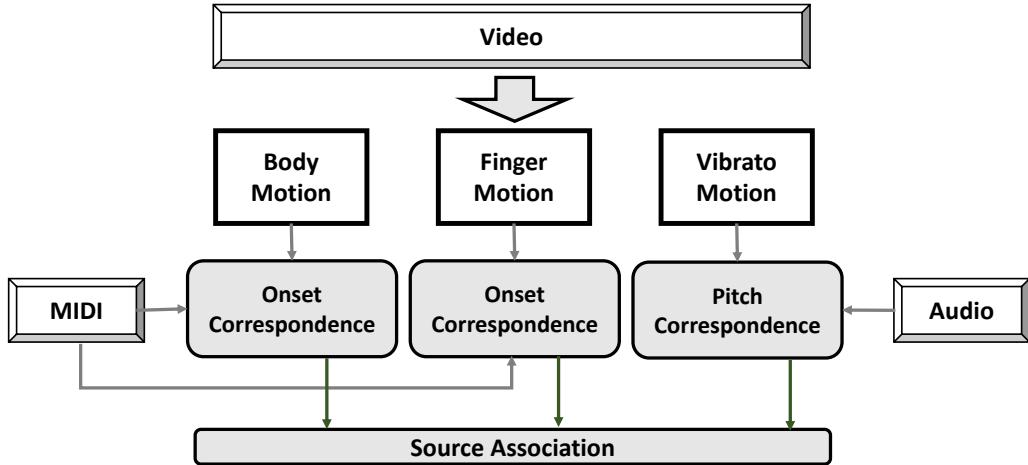


Figure 4.1: Outline of the proposed universal source association system for chamber ensemble performances. Three types of motion are modeled and correlated with the audio and score in three modules.

4.2 Related Work

4.2.1 Source Localization

When there is at most one active sound source at a time, the problem of audio-visual source association is also known as *source localization*, i.e., visualizing the sound source in the video. For audio-visual speech, this is helpful for speaker face segmentation (Liu and Sato, 2008). Early work on speaker localization correlates audio energy changes with pixel motions via non-linear diffusion (Casanovas and Vandergheynst, 2010) or with semantic regions via video segmentation and tracking (Li, Ye and Hua, 2014). Other methods include time-delayed neural networks (Cutler and Davis, 2000), probabilistic multi-modal generative models (Fisher and Darrell, 2004), and Canonical

Correlation Analysis (CCA) (Kidron, Schechner and Elad, 2007; Izadinia, Saleemi and Shah, 2013).

Later work proposes to localize semantic objects in unconstrained videos by learning deep multi-modal representations. In (Owens and Efros, 2018), a fused multi-sensory network is proposed to learn an audio-visual representation, which further localizes the sound objects on the video frames. The similar two-stream network structure is employed in (Senocak et al., 2018), where an attention mechanism is developed for sound source localization. Similar idea is adopted in (Arandjelović and Zisserman, 2018) for cross-modal retrieval and source localization, and in (Tian et al., 2018) for both spatial and temporal localization.

4.2.2 Source Association for Separation

Other work deals with mixtures of active sources, where cross-modal association can be utilized to isolate sounds that correspond to each visual object. Barzelay and Schechner (2007, 2010) 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. (2007) reformulated CCA by incorporating non-negativity and sparsity constraints on the coefficients of the projection directions to locate and separate sound sources in movies. In (Casanovas et al., 2010), auditory and visual modalities are 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. In (Ephrat et al., 2018), a deep network-based model is proposed to isolate a single speech signal from a mixture of sounds given the target speaker from the video. In (Gao, Feris and Grauman, 2018), audio frequency bases are mapped to individual visual objects via an audio-visual object model, which further guides audio source separation. These methods, however, either deal with mixtures with at most two active sources or only focus on isolating one source from multiple active sources (e.g., background noises). The association problem for each individual source is not addressed.

4.2.3 Source Association for Chamber Ensembles

The source association problem for music ensembles is more challenging since all the available sound sources (the players) are active almost all the time, and the difficulty increases dramatically as the number of sources increases. Although each track is performed by one player in chamber music, the same kind of instruments are often used for different parts (e.g., a violin duet). Therefore, we cannot learn a deep representation that maps audio features with visual appearances to localize each source as in (Owens and Efros, 2018; Senocak et al., 2018; Arandjelović and Zisserman, 2018). Instead, one needs to recognize the distinct motions of different players and correlate them with the music content to achieve association.

In (Li, Duan and Sharma, 2016; Li, Dinesh, Duan and Sharma, 2017), we first proposed an approach to solving the association problem for string

ensembles with up to five simultaneously active sources in a score-informed fashion. This approach analyzes the bowing motion and correlates it with note onsets in score tracks. The assumptions are that many note onsets correspond to the beginning of bowing strokes and that different instrumental parts often have 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. Later we proposed a complementary approach in (Li, Xu and Duan, 2017) to correlate the fingering hand rolling motion with pitch fluctuations of vibrato notes for the association of string instruments. It, however, only works when vibrato notes are played. To our best knowledge, there is no existing work on integrating the bowing motion and vibrato motion for source association for string instruments nor extending the concept to deal with non-string instruments as what the proposed universal system does in this work.

4.3 Method

The proposed system takes data in three modalities as the input, namely, the audio and video recordings of the performance and the music score. As illustrated in Figure 4.1, the system uses three parallel modules to model three types of temporal correspondence between motions detected in the video and note events captured in other modalities for different instrumentalists. In this

section, we present the system in detail.

4.3.1 Performance-Score Alignment

As the proposed approach is score informed, a preliminary step for the system is to temporally align the music score with the dynamic timing of the audiovisual ensemble performance (assuming audio and video are pre-synchronized). This is achieved through audio-score alignment on their harmonic content. To do so, the audio is first converted to short-time Fourier spectral magnitudes with a 42.7 ms frame length (2048 samples in 48 kHz sampling rate), 10 ms hop size, hamming window, and 4-times zero padding, and then mapped to 12-D chroma vectors, where each element represents a pitch class. Each chroma vector is normalized by its root mean square (RMS) value. A similar operation is applied to the score, which is segmented into non-overlapping frames of the same size using the default tempo. A 12-D binary chroma vector is calculated for each frame to indicate the presence (taking more than 50% of the frame) and absence of a pitch class. The chroma vector is then normalized by its RMS.

In offline scenarios where the entire performance is available beforehand, the alignment can be obtained by the dynamic time warping (DTW) algorithm, where the alignment path is retrieved backward using Viterbi decoding. In online scenarios where the performance data arrives as a live stream, one commonly used framework is the online DTW algorithm (Dixon, 2005), optionally with a “backward-forward strategy” to reconsider the past deci-

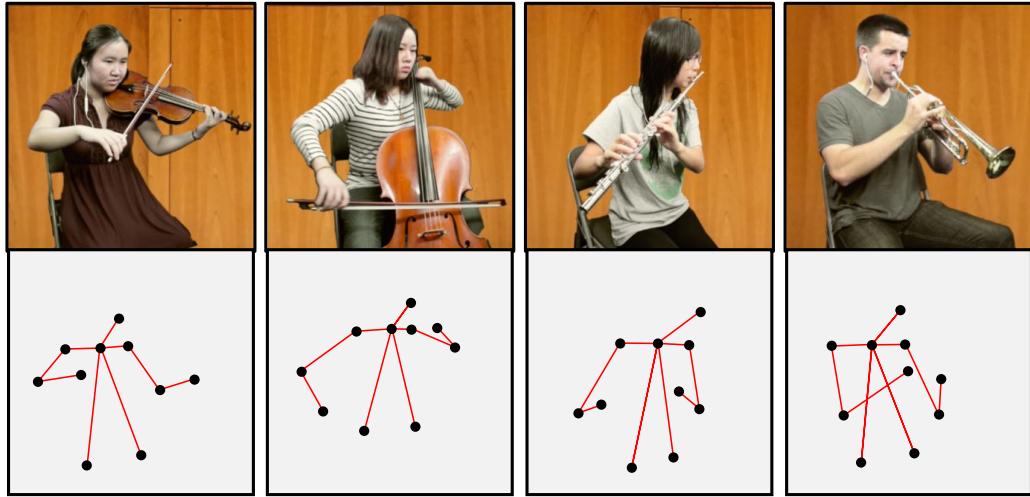


Figure 4.2: Body motion extraction. Upper body skeletons (second row) are extracted with OpenPose (Cao et al., 2017) in each video frame (first row) followed by temporal smoothing over time.

sions (Arzt, Widmer and Dixon, 2008), and a step to incorporate a tempo model (Arzt and Widmer, 2010) for robustness. Another framework is to employ a stochastic model (Grubb and Dannenberg, 1997; Duan and Pardo, 2011*b*), where the score position hypotheses are represented by a probability density function. We apply a hidden Markov process model proposed in (Duan and Pardo, 2011*b*) that uses a 2D continuous state space to represent the score position and tempo. This framework was previously evaluated on the Bach10 dataset (Duan, Pardo and Zhang, 2010) showing decent results.

4.3.2 Onset Correspondence with Body Motion

4.3.2.1 Body Motion Extraction

In music performances, body motion of performers conveys important musical expressions and ideas, e.g., the head nodding at leading notes. For some instruments, body motion directly articulates notes (e.g., strings, drums) or controls the pitch (e.g., trombones). To capture body motion from video recordings, one approach is optical flow estimation. In our prior work (Li, Dinesh, Duan and Sharma, 2017), we applied optical flow estimation to extract bowing motion of string players. However, we argue that this pixel-level analysis may not be ideal for semantic-level understanding of body gestures and movements, and can be less robust to occlusions and camera viewpoint changes.

In this work, we propose to extract body skeletons of all instrumentalists in each video frame using OpenPose (Cao et al., 2017), a real-time multi-person pose estimation approach. A skeleton in each frame is represented as a 20-D vector $\mathbf{y}(t)$ corresponding to the horizontal and vertical coordinates in the video frame of the 10 upper body joints, including nose, neck, shoulders, elbows, wrists, and hips. We do not include lower body joints as they are often less relevant to note events. Figure 4.2 shows video frames of several instrumentalists with the extracted body skeletons. To form a continuous skeleton sequence across time, we eliminate joint coordinates if the confidence score from OpenPose is smaller than 0.2 *and* the L_2 distance be-

tween consecutive frames is larger than 10% of the head-hip distance, which is considered the maximally possible regular movement in a 29.97-FPS video. We also temporally smooth their coordinates using a moving average with a 5-frame window size. We then take the two hips as reference coordinates to align the body position across frames. Finally, we calculate motion velocities $\mathbf{z}(t)$ as the derivative of $\mathbf{y}(t)$ w.r.t. time. Compared to optical flow estimation, this gesture-based motion analysis approach is semantically more meaningful, less computationally expensive, and more robust to occlusions and camera viewpoint changes.

To extract motions related to note onsets in each video frame, for each player we denote the motion velocities of n frames in the past as $\mathbf{Z} \in R^{n \times 20}$ and apply principal component analysis (PCA) as $\mathbf{Z}^T \mathbf{Z} = \mathbf{V}^T \Sigma \mathbf{V}$. We then project the motion velocity $\mathbf{z}(t)$ onto the first principal component and take its absolute value as the motion salience $s(t)$. This discards the direction information of the motions (e.g., up/down-bow for violinists), which is less relevant to timings than the amplitude information is. We set n to 150 frames, i.e., 5 seconds in time. To reduce the computational cost, we update \mathbf{V} every 1 second (assuming consistent motion patterns within a short period).

4.3.2.2 Onset Likelihood

From the motion salience $s(t)$, we infer the timings of the motion strokes that are potentially related to score note onsets. As a note onset often corresponds to the beginning or ending of a sound articulation motion (e.g., a bowing

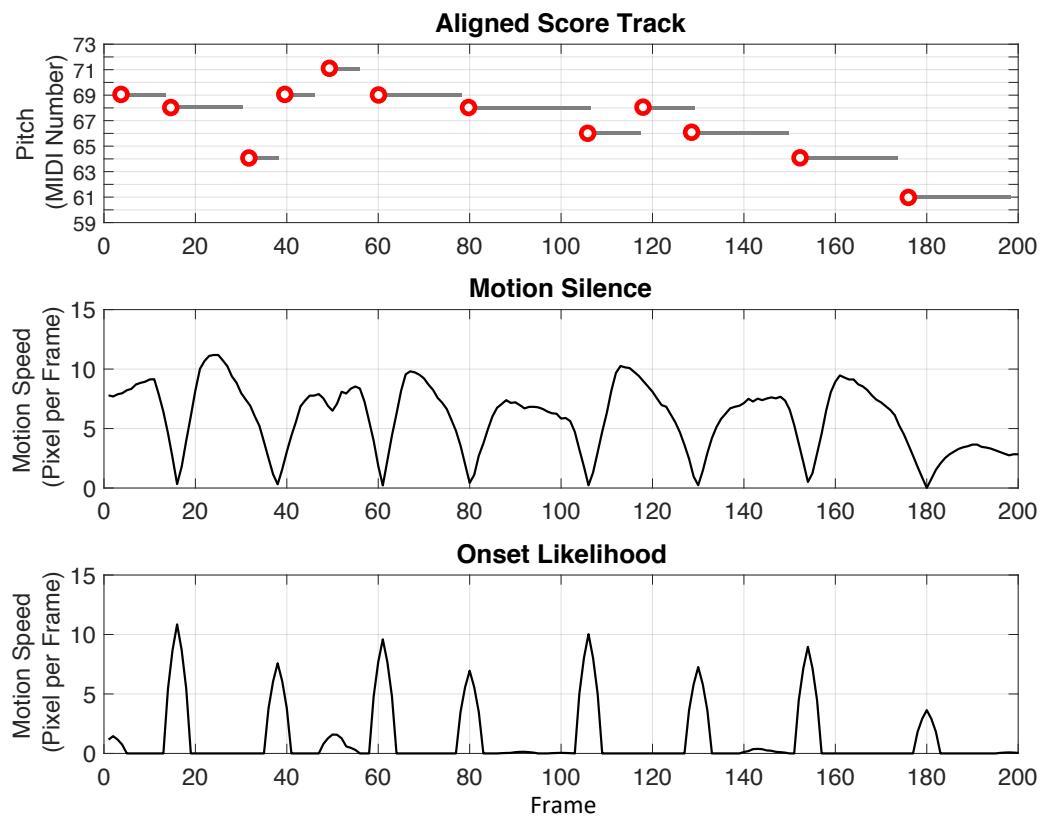


Figure 4.3: Example correspondence between body motion and note onsets. Top: temporally aligned score track with onsets marked by red circles. Middle: extracted motion salience (primarily bowing motion) from the visual performance of a violin player. Bottom: derived onset likelihood curve from the motion salience.

stroke for string instruments), the motion speed at the onset is often small. Therefore, local minima of the motion salience $s(t)$ are often indicative of note onsets. Let Ω be the set of all these local minima throughout a piece. For each local minimum $\tau \in \Omega$, we represent the likelihood of a note onset as $a(\tau) = \max_{\gamma \in [\tau, \tau+30]} s(\gamma) - s(\tau)$ that is determined by the maximum speed of the motion stroke considering the following 30 frames: the larger the more likely that a note onset is activated by this motion stroke. Therefore, we can define an *onset likelihood* curve $\phi_b(t)$ derived from body motion analysis as

$$\phi_b(t) = \left(\sum_{\tau \in \Omega} a(\tau) \cdot \delta(t - \tau) \right) * \mathcal{N}(t), \quad (4.1)$$

where $\delta(t)$ is the Kronecker delta function, and $\mathcal{N}(t)$ is a Gaussian function to give each predicted onset time a tolerance (width) with a standard deviation of 3 frames (30 ms). It is noted that $\phi_b(t)$ can be calculated in an online fashion, with a delay of up to 1 second due to the search for the local maximum after each local minimum. Figure 4.3 plots the onset likelihood curve $\phi_b(t)$ along with the associated and temporally aligned score track, where the note onset timings are marked as red circles. We find that many of the note onsets can be associated with peaks of $\phi_b(t)$. This correspondence sets the basis for the association between score and motion, as described below.

4.3.2.3 Pair-wise Correspondence

We extract the motion-based onset likelihood curve for each player from the video performance as $\phi_b^{[p]}(t)$, where p is the player index. From each track of the temporally aligned score, we use a binary impulse train $\psi^{[q]}(t)$ to represent the note onsets, where q is the track index, $\psi^{[q]}(t) = 1$ if there is a note onset in the t -th frame of the q -th track and $\psi^{[q]}(t) = 0$ otherwise. Then the pair-wise *matching score* between the p -th player and the q -th score track, up to the t -th frame, can be calculated through inner product:

$$M_b^{[p,q]}(t) = \sum_{\tau=0}^t \phi_b^{[p]}(\tau) \cdot \psi^{[q]}(\tau). \quad (4.2)$$

This can be updated in an online fashion as new time frames arrive.

4.3.3 Onset Correspondence with Finger Motion

4.3.3.1 Finger Motion Extraction

While note articulation is visible on body movements for string instrumentalists, this is generally not the case for woodwind/brass instrumentalists, where notes are articulated by blowing to the reed/mouthpiece, showing a less visible motion around the mouth. However, pitch changes of these instruments are mostly controlled by finger-operated keys, which often results in synchronized events between finger movements and note onsets (Palmer et al., 2007). Compared to body motions, finger motions are more subtle

and more prone to occlusion. In this section, we propose to extract finger motions and to correlate them with note onsets.

We apply OpenPose again to extract the positions of all the finger joints from each player. Due to the limited video resolution and occlusion, this result is not robust enough to estimate the motion. Inspired by (Li, Xu and Duan, 2017), we use optical flow estimation (Sun, Roth and Black, 2010) to capture this subtle motion at the pixel level. To reduce the computational cost, we set a region of interest (ROI) around the detected finger joints for optical flow estimation. The ROI centers at the median of all the finger joints for each hand, and spans to cover all the joints. Similar to body skeletons, we smooth the joint coordinates using moving average with a window size of 5 frames. Then we compute the optical flow estimation inside the ROI. Again, to eliminate the rigid and affine motion, each optical flow vector is subtracted by the average motion vector of the ROI, resulting in a motion vector $\mathbf{u}^{(ij)}(t)$ at the pixel (i, j) and t -th frame. Figure 4.4 takes one flute player and one clarinet player as examples to visualize the optical flow estimation of one-hand finger motion in five consecutive frames, where the estimated finger joint positions are overlaid on the first video frames.

4.3.3.2 Onset Likelihood

On each frame we take the maximum value of pixel-wise motion magnitude $|\mathbf{u}^{(ij)}(t)|$ across all the pixels in the ROI as the motion flux, which captures the finger movements corresponding to pitch changes and is directly consid-

ered as *onset likelihood* $\phi_f(t)$ from finger motions. Figure 4.5 plots the onset likelihood curve $\phi_f(t)$ along with the associated and temporally aligned score track piano-roll. We can observe salient motion flux around most note onset frames. Compared to Figure 4.3, the correspondence to note onsets to fingering motion for woodwind/brass instruments is not as robust as that to body motion for string instruments. This is because this fine-grained motion is more sensitive to irrelevant motions. In addition, repeated notes for most woodwind/brass instruments are not reflected by fingering motion on the keys.

Analogue to Equation (4.2), the pair-wise matching score from finger motions can be calculated as:

$$M_f^{[p,q]}(t) = \sum_{\tau=0}^t \phi_f^{[p]}(\tau) \cdot \psi^{[q]}(\tau). \quad (4.3)$$

4.3.4 Pitch Correspondence with Vibrato Motion

In addition to the onset time, variations of acoustic features throughout the entire process of some note articulations show correspondence with certain motions. Vibrato is one of them. Vibrato is a commonly used artistic note articulation method to color a tone and express emotions in music performances. Physically, vibrato is generated by pitch modulation of a note in a periodic fashion. For some instruments such as strings, vibrato is often

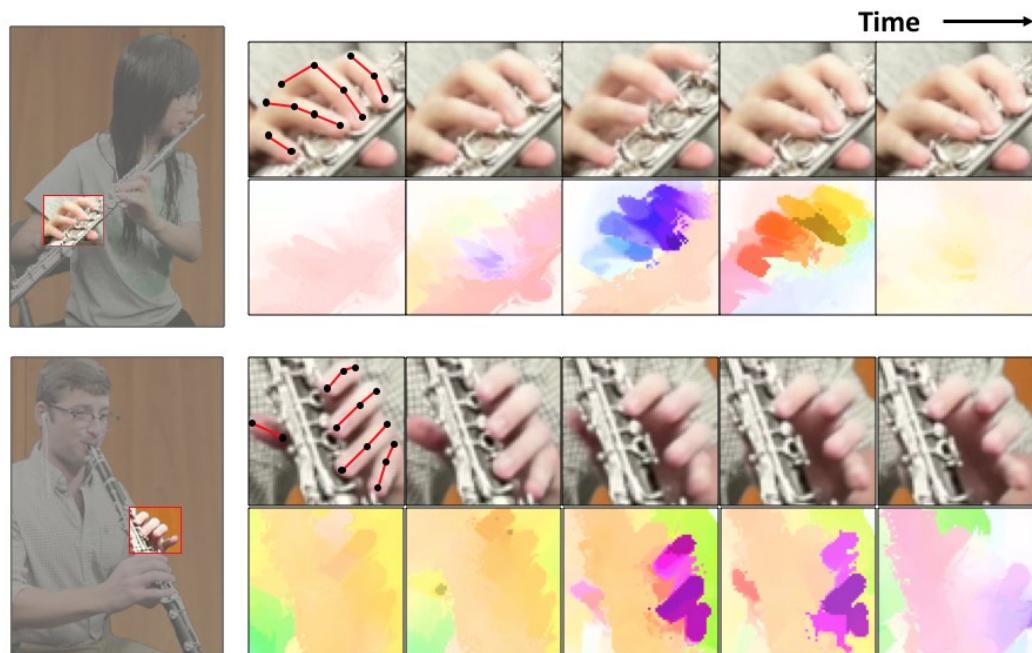


Figure 4.4: Optical flow visualization of finger motions in five consecutive frames corresponding to note changes. The color encoding scheme is adopted from (Baker et al., 2011).

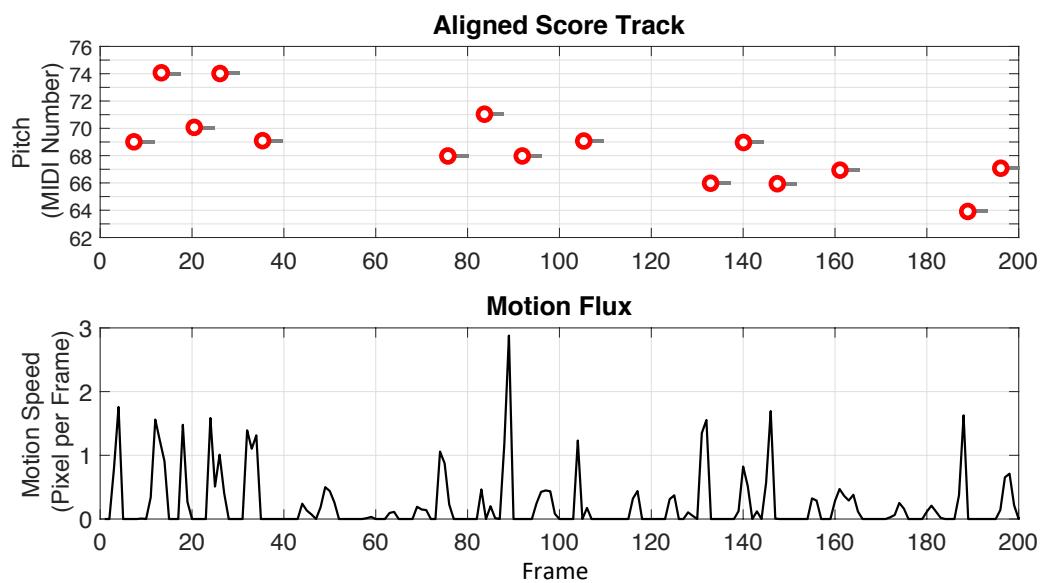


Figure 4.5: Example correspondence between finger motion and note onsets of a flute player. Top: temporally aligned score track with onsets marked by red circles. Bottom: extracted motion flux from finger movements.

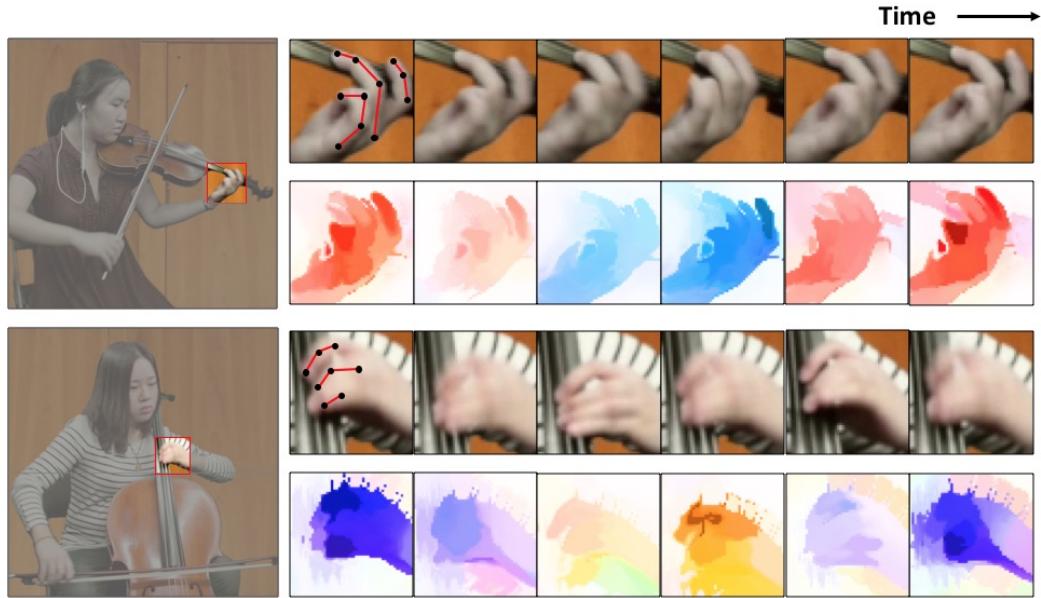


Figure 4.6: Optical flow visualization of the hand motions corresponding to vibrato articulation. Color encoding scheme is adopted from (Baker et al., 2011).

visible as the left hand rolling motion on the fingerboard. This motivates us to extract this fine motion and find the correspondence with pitch contours extracted from the audio modality.

4.3.4.1 Vibrato Motion Extraction

We retrieve the fingering motion $\mathbf{u}^{(ij)}(t)$ as computed from the previous section. Although the vibrato motion is mostly a rigid motion (fingers move together with little relative movements), it is periodic and very fast (usually about 4-7.5 Hz (Geringer, MacLeod and Allen, 2010)), hence it is not removed as other slow rigid/affine motions. For each frame t , we take the av-

verage motion vector across all pixels within the ROI as $\mathbf{u}(t) = [u_x(t), u_y(t)]^T$, where the motion direction is preserved for vibrato detection.

The vibrato detection module works as a binary classifier as proposed and trained in (Li, Dinesh, Sharma and Duan, 2017). The classifier is implemented as a support vector machine (SVM) that takes the input of a 8-dimensional feature extracted from each sample, including the zero crossing rate of the x - and y - motion velocities and their auto-correlations, the energy of 3-9 Hz frequency range, and the auto-correlation peaks. According to (Li, Dinesh, Sharma and Duan, 2017), this method achieves a vibrato detection accuracy of over 90% from visual motions regardless of the polyphony number and instrument type within the string instrument family. Here each input sample is a 1-second segment of $\mathbf{u}(t)$ (again introducing an average 0.5-second delay of the association system).

For each detected vibrato segment, we perform PCA on $\mathbf{u}(t)$ within this 1-second segment to obtain the 1-D principal motion velocity curve $v(t)$. We then integrate $v(t)$ over time to calculate a *motion displacement curve*, $d(t)$, which corresponds to the length fluctuation of the vibrating string hence the pitch fluctuation of the note. We normalize each vibrato segment of $d(t)$ with zero mean and unit variance. We set the non-vibrato segments of $d(t)$ to zero.

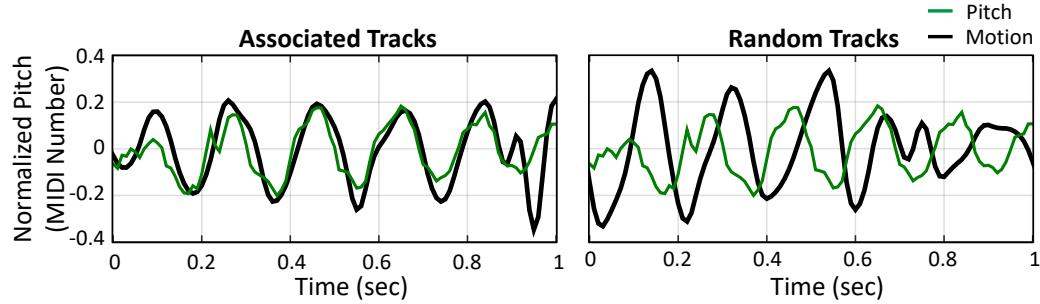


Figure 4.7: The same segment of normalized pitch contour $f(t)$ (green) overlaid with the motion displacement curve $d(t)$ (black) from the associated track (left) and another random track (right).

4.3.4.2 Pitch Contour Extraction

Utilizing the score information, we apply Soundprism (Duan and Pardo, 2011a), an online score-informed source separation system, to separate the polyphonic audio mixture into each individual sources. To extract the pitch contour, we perform a score-informed pitch estimation step on each separated audio source, as described in (Li, Xu and Duan, 2017). The pitch contour of each note segment is normalized to have zero mean and unit variance for each note, and is denoted as $f(t)$. This operation discards the original pitch height information, and only preserves the pitch drift from the central frequency within each note. Figure 4.7 plots a 1-second segment of the normalized pitch contour overlaid with a motion displacement curve from the associated track (left) and a random track (right). Similar to Equations (4.2)

and (4.3), we calculate the vibrato correspondence as:

$$M_v^{[p,q]}(t) = \sum_{\tau=0}^t d^{[p]}(\tau) \cdot f^{[q]}(\tau). \quad (4.4)$$

4.3.5 Integrating All Correspondences

We integrate the three modules to calculate the pair-wise correspondence between visual motion and score/audio events considering both onset timing and the entire note articulation process. This is calculated as

$$\begin{aligned} M^{[p,q]}(t) &= w_b(t) \cdot \bar{M}_b^{[p,q]}(t) \\ &+ w_f(t) \cdot \bar{M}_f^{[p,q]}(t) \\ &+ w_v(t) \cdot \bar{M}_v^{[p,q]}(t). \end{aligned} \quad (4.5)$$

Here $\bar{\cdot}$ means the normalization across all of the pair-wise combinations between N players and N tracks, i.e.,

$$\bar{M}_b^{[p,q]}(t) = M_b^{[p,q]}(t) / \sum_{p,q=1}^N M_b^{[p,q]}(t), \quad (4.6)$$

and we use the weighting parameters w_b , w_f , w_v to re-scale the different modules. Weight w_v is set as $2w_f$, emphasizing more on finger motions with vibrato patterns. Weights w_b and w_f are linearly related to their motion

salience in the past frames as

$$w_b(t)/w_f(t) = \sum_{\tau=0}^t s(\tau) / \sum_{\tau=0}^t \phi_f(\tau), \quad (4.7)$$

It recovers to the original scale of M_b and M_f , which represents the motion extracted from body and finger respectively. This allows the system to focus on the part with stronger motion cues, such as body motion for string instrumentalists, and finger motion for wind/brass instrumentalists. In Section 4.4, we test the components in isolation as well as some combinations of them.

For an ensemble with N players, the number of all of the possible associations is the factorial of N . Let $\sigma(\cdot)$ be a permutation function from $p \in [1, N]$ to $q \in [1, N]$ that represents one association candidate, where the p -th player is associated with the $\sigma(p)$ -th track. For each association candidate σ , we calculate an overall association score as the product of the N pair-wise correspondence values. The final association solution $\hat{\sigma}$ is returned to maximize the association score as:

$$\hat{\sigma} = \arg \max_{\sigma} \prod_{p=1}^N M^{[p, \sigma(p)]} = \arg \min_{\sigma} \sum_{p=1}^N -\log M^{[p, \sigma(p)]}. \quad (4.8)$$

The right side replaces the product with the sum of negative logarithms, to make the efficient Hungarian algorithm (Kuhn, 1955) directly applicable for finding the best association.

4.4 Experiments

4.4.1 Dataset

The proposed source association system is evaluated on the URMP dataset (Li, Liu, Dinesh, Duan and Sharma, 2019). The dataset contains 44 classical chamber ensemble pieces ranging from duets to quintets, with the frame-level annotations for each track. The video data was recorded at a frame rate of 29.97 frames per second (FPS) and a resolution of 1080P with a fixed camera view. The audio performance was recorded with a sample rate of 48 kHz and bit depth of 24 bit.

We further expand the dataset by creating all possible track combinations within each piece. For the example of a quartet, we further generate 6 duets and 4 trios from the 4 original tracks. This in total expands the dataset into 171 duets, 126 trios, 47 quartets, and 7 quintets. Note that we do not combine tracks across pieces, to ensure the naturalness of the expanded set. The number of pieces for different instrument arrangements are listed in Table 4.1.

To further understand the dataset, we calculate the *onset overlap rate* for each original piece. This statistic is defined as the percentage of onset positions that are shared by two or more tracks for each piece. This statistic is relevant to the performance of the proposed source association approach, as two out of the three motion analysis modules rely on onset patterns to

		String	Wind/Brass	Mix	Total
Original	Duet	2	6	3	11
	Trio	2	6	4	12
	Quartet	5	6	3	14
	Quintet	2	4	1	7
Expanded	Duet	57	91	23	171
	Trio	41	65	20	126
	Quartet	15	25	7	47
	Quintet	2	4	1	7

Table 4.1: The number of pieces for different instrument arrangements from the original and expanded URMP dataset.

associate players with tracks. Figure 4.8 plots this statistic for all of the original 44 pieces. While the rate varies much from one piece to another, we see a general increasing trend as the polyphony increases.

4.4.2 System Setup

For implementation, the audio is processed with a frame length of 42.7 ms and a hop size of 10 ms for score following and pitch contour extraction. When calculating the vibrato correspondence, the motion curve extracted from the 29.97-FPS video is up-sampled to 100 FPS, enabling a synchronized time resolution between the audio and video. As vibrato detection is performed on 1-second segments and the onset likelihood curve from body motion is derived from a local maximum within future 30 frames (1 second), the system has a 1-second inherent delay when it runs for real-time applications. The past 5-second of body and finger motion velocities are stored in memory to apply PCA in Section 4.3.2.1 and to calculate the weighting parameters in

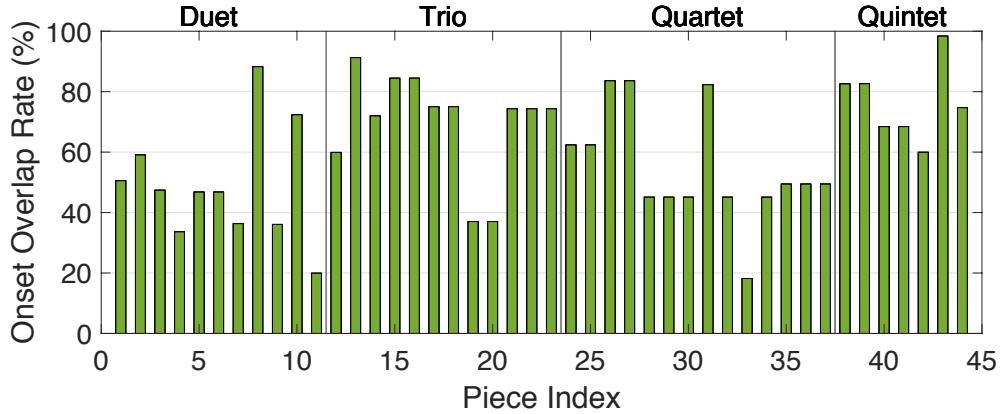


Figure 4.8: Onset overlap rate for each piece from the original URMP dataset.

Equation (4.7).

For evaluation, we first address each track independently to investigate the quality of the extracted onset likelihood features, using the traditional onset detection measures. Then we evaluate the association performance on the expanded set of ensemble pieces. The result is grouped by different ensemble types, from duets to quintets, which directly correlate with the difficulty levels. Note that whatever number of tracks presented in the performance, only one association is correct. We do not include a quantitative evaluation of the score following and vibrato detection modules in this work, since they have been fully evaluated in our previous work.

4.4.3 Onset Detection Evaluation

As two modules of the proposed system rely on the synchronization cues of onset timing between different modalities, we evaluate the quality of our

proposed onset likelihood curves that are extracted from body motions and finger motions. To do so, we set up an onset detection task. We take the onset likelihood curve as the onset detection function (Bello et al., 2005), and perform peak-picking to retrieve the onsets. A true positive detection is counted when a detected onset is within a tolerance window of 3 video frames (100ms). This is wider than the standard 50ms in the literature, since the precise timing is not the main focus of the source association system.

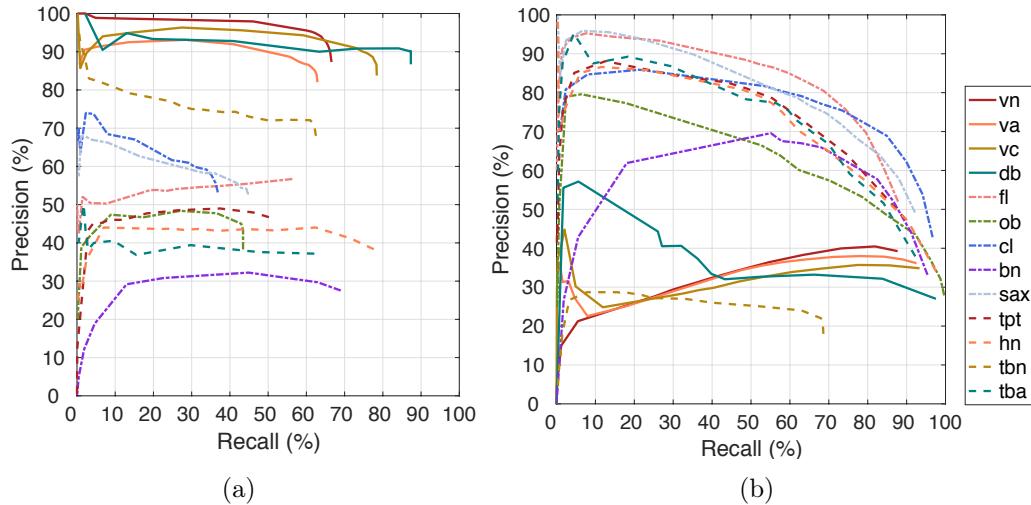


Figure 4.9: Onset detection evaluation results from body motions (a) and finger motions (b) for different instruments.

Figure 4.9 plots the precision versus recall by varying the peak-picking threshold on the onset likelihood curves extracted from body motions and finger motions respectively. They are calculated for each instrument across all pieces in the original dataset. It reveals that the onset likelihood curve extracted from body motion shows better correlation with the ground-truth

onset timings for string instruments, while that from finger motion shows better correlation for woodwind/brass instruments. An exception is trombone, where the onset likelihood curve extracted from body motion shows better correlation than that from finger motions. This is reasonable as its pitch change (hence note transition) is mainly performed by moving the slide using the right arm (body motion).

Another interesting observation is that although the onset likelihood curve $\phi_b(t)$ in Figure 4.3 shows less noisy than $\phi_f(t)$ in Figure 4.5, the recall calculated from $\phi_b(t)$ for string instruments cannot reach an as high value as that of woodwind/brass instruments calculated from $\phi_f(t)$. We analyze that this is because legato bowing (i.e., articulating a sequence of notes from one sustained bowing action) is widely used in string instrument performances, where onset detection from bow motions misses some true positives. This explains the upper bound of recall rates (around 80% as in Figure 4.9 (a)) for string instruments. For woodwind/brass instruments, there are also onsets not visible such as repeated notes, but the amount is much smaller. This explains why the recall rates can reach closely to 100% in Figure 4.9 (b).

4.4.4 Source Association Evaluation

In this section we evaluate the source association performance, first for each module (corresponding to each component in Equation (4.5) independently, then for the finally integrated approach. We use *association accuracy* as the evaluation measure, which is defined as the percentage of correctly associated

pieces among all testing pieces. A piece is considered correctly associated if the exactly correct bijection between players and score/audio tracks is retrieved. Note that the difficulty of source association increases dramatically from small to large ensembles. In a quintet ensemble, there are in total $5! = 120$ bijection candidates, and only one is considered correct. Therefore, we divide our evaluation based on the size of ensembles.

Besides the ensemble size, the length of the performance also affects the difficulty of the association problem, assuming longer pieces provide richer cues. In an online setting, we hope that the proposed system can retrieve the correct association as quickly as possible. Therefore, in the experiments, we segment the testing pieces into non-overlapping excerpts for each of the following lengths: 5, 10, 15, 20, 25, and 30 seconds. When doing so, we first remove the beginning and the last 5 seconds of each piece as the performance may not cover the entire length of those segments. This further expands the testing pieces to a large number of evaluation samples, totaling 17574 samples, as presented in Table 4.2.

4.4.4.1 Body Motion

We first evaluate the source association performance using the onset correspondence \bar{M}_b between score tracks and body motions (the first component of Equation (4.5)). Figure 4.10 shows the association accuracy for ensembles consisting of string, woodwind/brass, and mixed instruments with different polyphony. Note that the mixed ensemble contains all the available pieces

String	excerpt duration (sec)					
	5	10	15	20	25	30
Duet	1323	642	420	303	236	200
Trio	1044	506	333	240	189	158
Quartet	355	172	114	82	65	54
Quintet	64	31	21	15	12	10
Wind/Brass	excerpt duration (sec)					
	5	10	15	20	25	30
Duet	1809	887	557	435	323	266
Trio	1275	626	391	309	229	187
Quartet	474	232	145	115	86	68
Quintet	66	32	20	16	12	9
Mix	excerpt duration (sec)					
	5	10	15	20	25	30
Duet	441	203	141	96	82	60
Trio	380	174	121	82	70	51
Quartet	199	92	64	44	37	28
Quintet	22	10	7	5	4	3

Table 4.2: The number of evaluation samples with different length and instrumentation for source association.

(Corresponding to the last column of Table 4.1 and all the elements in Table 4.2). For each piece, we plot how the association accuracy varies as the duration of the input stream increases from 5 to 30 seconds. Each marker in the figure is the association accuracy calculated from the number of excerpts shown in Table 4.2.

Comparing different ensemble sizes, the association accuracy decreases as the number of players/tracks increases. From Figure 4.10 (a), we find that correlating onsets with body motions is beneficial for string instruments. The accuracy increases as the duration of video stream increases, which provides more cues to solve the association. It reaches around 90% for all ensemble sizes when the video stream duration reaches 30 seconds. This strategy, however, is not effective for woodwind/brass instruments, where the association accuracy remains around random guess accuracy as shown in Figure 4.10 (b), e.g., 1/6 for trios. This is consistent with our expectations and the onset matching evaluations in Figure 4.9.

4.4.4.2 Finger Motion

We then evaluate the source association performance using the onset correspondence \bar{M}_f between score tracks and finger motions (the second component of Equation (4.5)). The association accuracy is plotted in Figure 4.11, with the same set of pieces used in Figure 4.10. It can be seen that different from body motion, finger motion is a more prominent cue for matching with note onsets for woodwind/brass instruments (except for trombone). When a

30-second video excerpt is available, the association accuracy reaches about 90% for all sizes of woodwind/brass ensembles. This is also consistent with our onset detection evaluations in Figure 4.9. For string instruments, however, the extracted finger motions are mostly vibrato motions, which are not relevant to note onsets.

Figures 4.10 and 4.11 also reveal some limitations of the source association solution based on onset-motion correspondence. First, there are many note onsets not revealed from body or finger motions, such as notes played with legato bowing for string instruments and repeated notes from woodwind/brass instruments, as analyzed in Section 4.4.3. Second, as note synchronization between players is at the foundation of music performances, note onsets between tracks have high chances to overlap with each other, as shown in Figure 4.8. This limits the association performance of approaches that only rely on onset-motion correspondence, especially from short video excerpts.

4.4.4.3 Vibrato Motion

The correspondence between pitch contour and vibrato motion (denoted as \bar{M}_v , the third component of Equation (4.5)) helps to retrieve the source association on a finer level for string instrumentalists. The evaluation result is plotted in Figure 4.12 (a) for the same set of pieces performed by string ensembles as in Figure 4.10 (a). We do not include the woodwind/brass instrument group here since no vibrato pattern can be detected from finger

motions. We can find that the source association can reach a high accuracy from shorter video clips, i.e., 90% after 10 seconds. The limitation of this approach is that vibrato articulation is not guaranteed to always present in the performance. We thus combine this module with the onset correspondence from body motions, the two dominant cues to solve association for string instruments, to evaluate the association accuracy in Figure 4.12 (b). These two components from M_b and M_v work together to reach a high association accuracy from a short video stream.

4.4.4.4 The Integrated System

At last, we evaluate the proposed complete source association system after integrating all the modules together, as presented in Equation (4.5). The pieces are the same as plotted in Figures 4.10-4.11. This presents a universal source association system regardless of instrument types. Comparing Figure 4.13 (a) with 4.12 (b), or Figure 4.13 (b) with 4.11 (b), adding components with irrelevant association cues does not harm the system, thanks to the weighting strategy in Equation (4.5) over different modules. Comparing Figure 4.13 (c) with Figure 4.10(c)/4.11(c), the integrated system greatly improves the association accuracy for pieces with mixed types of instruments. The association accuracy for mixed ensembles is between that for string and woodwind/brass ensembles.

4.5 Conclusion

In this chapter, we proposed an online source association system for Western chamber ensembles, which aims to retrieve the association between players in the video and the audio/score tracks through the analysis of the cross-modal correspondences. We designed three modules to model different correspondences between 1) body motions and note onsets, 2) finger motions and note onsets, and 3) vibrato motions and pitch fluctuations. Although these correspondences apply to different kinds of instruments, the proposed system automatically integrates them in an adaptive fashion, without the need for knowing the instrument types. This makes the system a universal framework for all instruments common in Western chamber ensembles. In addition, the system runs in an online fashion to update association results as the video stream progresses. Experiments with audio-visual recordings of performances with different polyphony and instrumentation demonstrate that the accuracy of the proposed system increases with the length of video streams, and high accuracy is achieved within a relatively short interval. The accuracy for string ensembles is generally better than that for woodwind, brass, and mixed-instrument ensembles because more correspondences are modeled for these instruments.

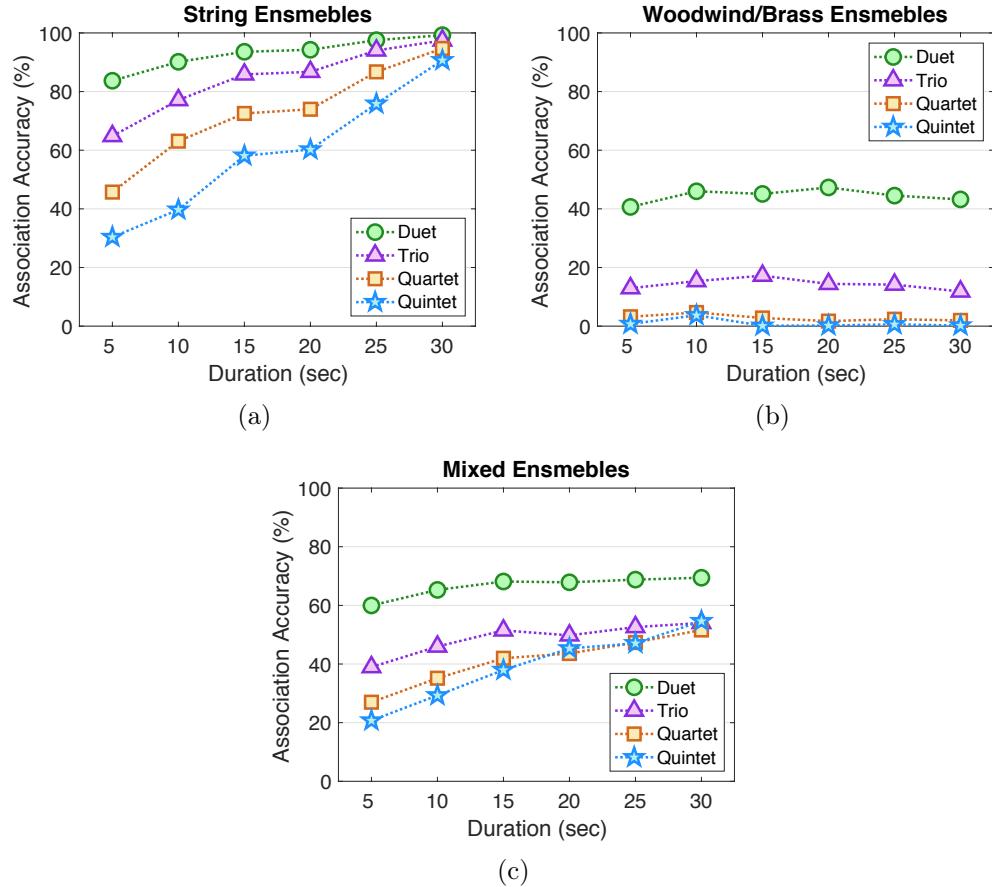


Figure 4.10: Source association accuracy only using onset correspondence between score tracks and body motions (the first component \bar{M}_b in Equation (4.5)).

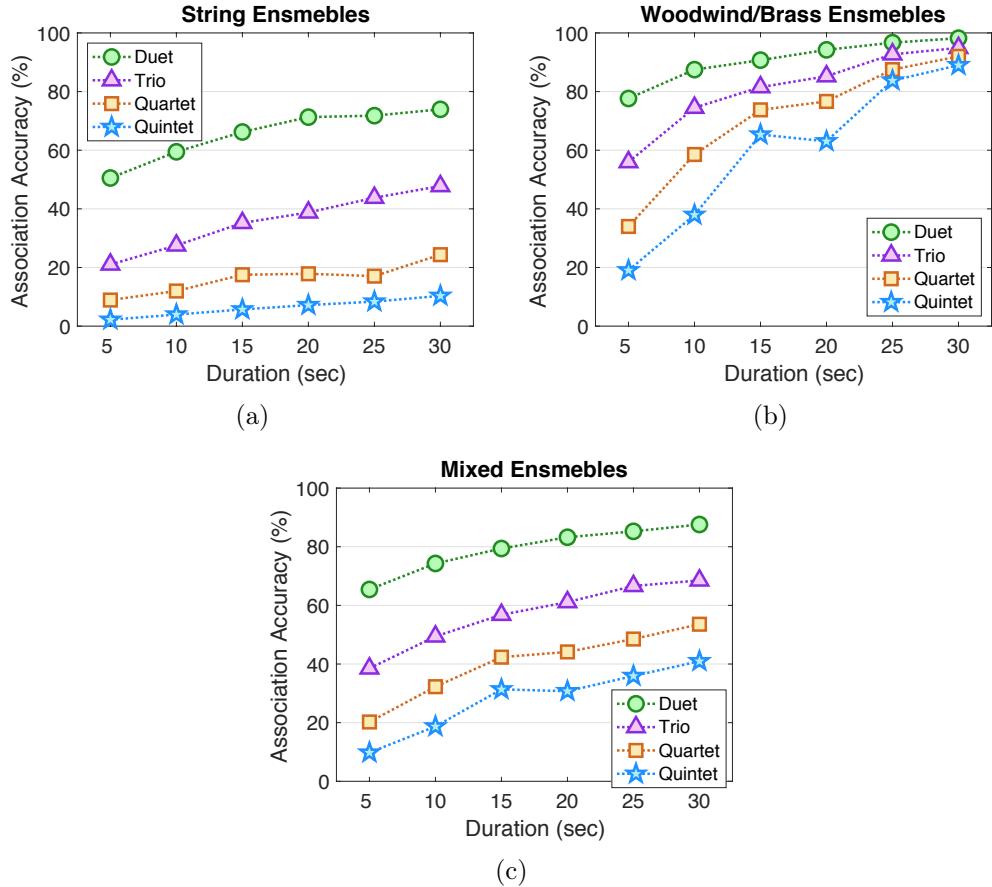


Figure 4.11: Source association accuracy only using onset correspondence between score tracks and finger motions (the second component \bar{M}_f in Equation (4.5)).

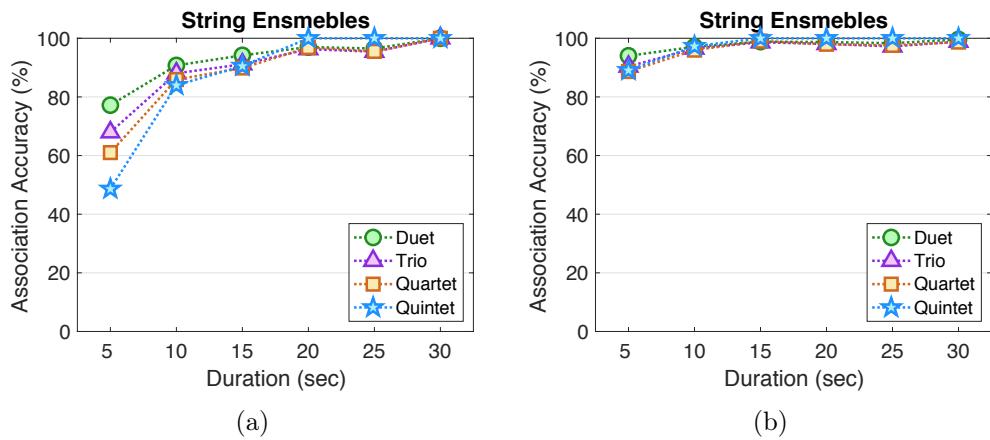


Figure 4.12: Source association accuracy of string ensembles by (a) only using vibrato correspondence between pitch fluctuation and hand motion (\bar{M}_v in Equation (4.5)), and (b) combining vibrato correspondence with onset correspondence from body motion in (\bar{M}_b and \bar{M}_v in Equation (4.5)).

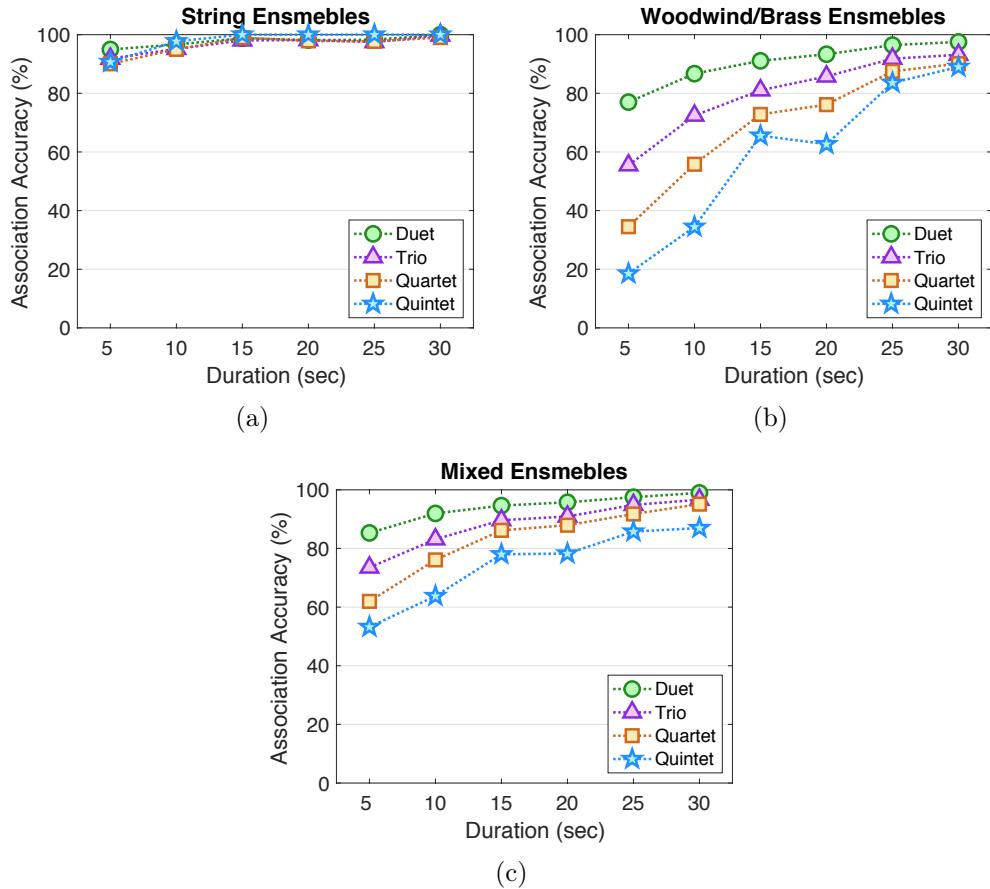


Figure 4.13: Source association accuracy of ensembles with different instrumentation using all of the three modules: onset correspondence from body motions, onset correspondence from finger motions, and vibrato correspondence from hand motions (Equation (4.5)).

Chapter 5

Visually-Informed Audio Analysis

In Chapters 2-4, we introduced an audio-visual music performance dataset, and discussed the work to coordinate the auditory, visual, and symbolic modalities of music performance. In this chapter, we take two typical MIR tasks to demonstrate how multi-modal analysis advance the traditional MIR tasks when multi-modal data is available and coordinated.

5.1 Multi-Pitch Analysis

Multi-pitch analysis of polyphonic music requires estimating concurrent pitches (estimation) and organizing them into temporal streams according to their sound sources (streaming). This is challenging for approaches based on audio alone due to the polyphonic nature of the audio signals. Video of the performance, when available, can be useful to alleviate some of the difficulties. In this work, we propose to detect the play/non-play (P/NP) activities from musical performance videos using optical flow analysis to help with audio-

based multi-pitch analysis. Specifically, the detected P/NP activity provides a more accurate estimate of the instantaneous polyphony (i.e., the number of pitches at a time instant), and also helps with assigning pitch estimates to only active sound sources. As the first attempt towards audio-visual multi-pitch analysis of multi-instrument musical performances, we demonstrate the concept on 11 string ensembles. Experiments show a high overall P/NP detection accuracy of 85.3%, and a statistically significant improvement on both the multi-pitch estimation and streaming accuracy, under paired t-tests at a significance level of 0.01 in most cases.

5.1.1 Introduction

Multi-pitch analysis of polyphonic music is important in many music information retrieval (MIR) tasks including automatic music transcription, music source separation, and audio-score alignment. It can be performed at different levels: *Multi-pitch Estimation (MPE)* is to estimate concurrent pitches and the number of pitches (polyphony) in each time frame; *Multi-pitch Streaming (MPS)* goes one step further to also assign the pitch estimates to different sound sources.

There exist various audio-based methods for multi-pitch analysis. For MPE, methods include auto-correlation (Tolonen and Karjalainen, 2000) and Bayesian inference (Davy, Godsill and Idier, 2006) in the time-domain, and harmonic amplitude summation (Klapuri, 2006) and peak/non-peak modeling (Duan, Pardo and Zhang, 2010) in the frequency domain. For MPS,

methods often rely on modeling the timbre of sound sources to organize pitch estimates. Supervised methods, which learn timbre models from isolated training excerpts of sources, employ Bayesian models (Vincent, 2006), hidden Markov models (Wohlmayr, Stark and Pernkopf, 2011), and probabilistic latent component analysis (PLCA) (Bay et al., 2012). Unsupervised methods that infer timbre models of sound sources directly from the mixture audio are also proposed (Hu and Wang, 2013; Duan, Han and Pardo, 2014; Arora and Behera, 2015). The common idea is to cluster pitch estimates that have similar timbre features into the same stream, while the clustering process is often aided by constraints that model the locality relations between pitches.

These state-of-the-art audio-based methods, however, cannot achieve satisfactory performance for many applications as yet. This is due to the core challenge that polyphonic audio signals have: signals of different sound sources mix together and interfere with each other. More specifically, multi-pitch estimation needs to estimate the number of mixed sound sources at each time instant (instantaneous polyphony). This is difficult for audio-based approaches due to large variety of harmonic relations and timbre combinations of concurrent pitches. Furthermore, even if the instantaneous polyphony were correctly estimated in each frame, identifying which sources are active in these frames for the estimated pitches to assign to is also challenging purely from audio. These issues, however, could be alleviated when videos are available. Specifically, availability of video can help identify Play/Non-

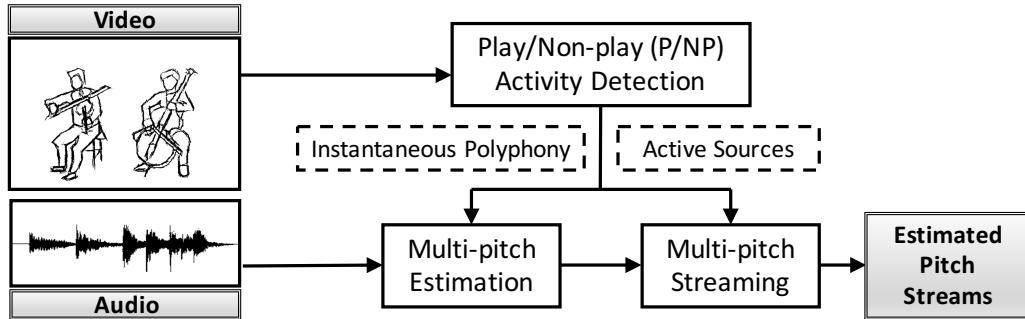


Figure 5.1: Proposed framework for enhancing multi-pitch analysis using video-based play/non-play activity detection.

play (P/NP) activities of instrument players, helping with the estimation of the instantaneous polyphony and the detection of active sound sources for pitches to be assigned.

Advances in the field of multimodal signal analysis have propelled the use of visual features along with audio features to solve a variety of problems like information retrieval (Kuo, Shan and Lee, 2013), multimedia content authoring (Gillet, Essid and Richard, 2007), sentiment analysis (Poria et al., 2016), shot change detection (Essid and Richard, 2012), and audio-visual feature extraction (Acar, Hopfgartner and Albayrak, 2014). In the field of music performance analysis, visual information has been exploited to detect instrument playing activities in an orchestra for audio-score alignment (Bazzica, Liem and Hanjalic, 2014). Video analysis has also been employed to track the fret-board and movement of hands to transcribe guitar performances (Paleari et al., 2008). However, to date, there is remarkably little visually informed work on the fundamental problem of multi-pitch analysis.

In this chapter, we build upon our prior work on audio-based MPE (Duan, Pardo and Zhang, 2010) and MPS (Duan, Han and Pardo, 2014) to propose the first method that leverages visual information for multi-pitch analysis of string ensembles. Figure 5.1 shows the system overview. The video analysis module detects players as well as their P/NP activity through an optical-flow-based motion analysis in each video frame. The instantaneous polyphony of each audio frame is then derived from the P/NP activity and is used to inform the audio-based MPE module. The P/NP information is also passed to the MPS module so that estimated pitches are only allowed to be assigned to active players in each frame. Experiments on 11 string ensembles show that the proposed video-based P/NP detection achieves a high overall accuracy of 85.3%. The incorporation of these detected P/NP results to our audio-based baselines results in a statistically significant improvement on both the MPE and the MPS accuracy at a significance level of 0.01 in most cases.

5.1.2 Proposed Method

5.1.2.1 Play/Non-Play Activity Detection

We employ optical flow estimation and supervised classification to detect P/NP activities of players in each video frame. Figure 5.2 summarizes the analysis workflow.

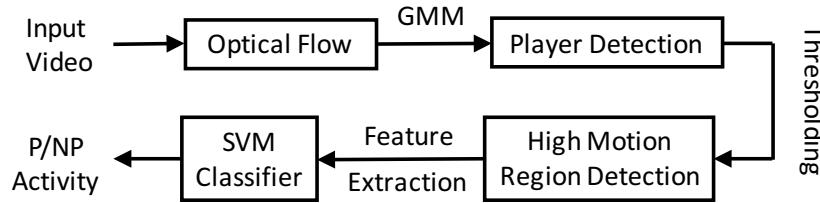


Figure 5.2: Video analysis for P/NP activity detection.

5.1.2.1.1 Optical Flow Estimation

Optical flow (Wang, Ostermann and Zhang, 2002), which estimates the motion field using the observed pattern of brightness displacements from frame to frame, forms the basis of our motion analysis. The assumption that the brightness is preserved as the pixels get displaced due to motion in the scene, yields the classic optical flow equation (Horn and Schunck, 1981)

$$\frac{\partial I}{\partial x}u + \frac{\partial I}{\partial y}v + \frac{\partial I}{\partial t} = 0, \quad (5.1)$$

where (x, y) represent the spatial (pixel) coordinates, t denotes time, and $I = I(x, y, t)$ denotes the observed spatio-temporal pattern of image intensity and $(u, v) = [\frac{dx}{dt}, \frac{dy}{dt}]$ represents the flow vector in horizontal (x) and vertical (y) directions. The collection of flow vectors over spatial extent of the frame forms the motion field $u_t(x, y), v_t(x, y)$, where t indexes the frames. Optical flow techniques estimate the flow field for each frame by minimizing an energy function that combines a data term based on Equation 5.1 with regularization terms that ensure smoothness of the flow field. In this work, we adopt the

approach of (Sun, Roth and Black, 2010) which improves upon the classical objective function of (Horn and Schunck, 1981; Black and Anandan, 1993) by incorporating flow and image boundary information in the regularization function and provides highly accurate motion field estimations.

5.1.2.1.2 Player Detection

Instrument play movements are typically the dominant motion in the video. For a given camera viewpoint, these movements localize in spatial regions corresponding to individual players across the span of temporal frames. We therefore identify distinct regions of significant motion in the video to estimate the locations of the players. Specifically, we compute a temporally aggregated motion magnitude function for each spatial location as the sum of the optical flow motion field magnitudes over the frames. The motion magnitude function is modeled as a mixture of Gaussians, and a rough estimate of the locations of the players is obtained by identifying the spatial locations that associate with individual components (with high probability).

5.1.2.1.3 High Motion Region Detection

Within the spatial regions corresponding to a string player's movements, pixels and time intervals with high motion (e.g., the bowing hand) are indicative of the P/NP activity. Therefore, we detect high motion regions (pixels) from the initial estimate of individual player locations obtained in the previous step. Given that different players (instruments) exhibit different degrees of

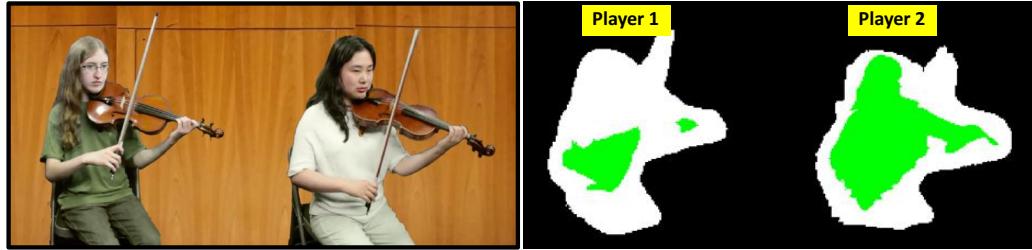


Figure 5.3: Sample frame from the performance video (left) and the player detection results (right) with the detected high motion regions in green, detected players in white and the background in black.

motion in the video, we use an adaptive threshold on the motion magnitude for each player, using a threshold equal to the temporal mean + twice standard deviation of the histogram of the flow vector magnitude. A sample frame is shown in Figure 5.3 where we can see clusters of high motion regions (green) within the detected players (white) who are detected from the background (black).

5.1.2.2 Feature Extraction and Classification

We train a support vector machine classifier (SVM) to classify P/NP activities of each player in each video frame. A 20-dimensional feature vector is extracted from the motion vectors of the detected high motion pixels for each player. These features include: (a) Mean, variance, and standard deviation of the flow vectors in x and y directions separately. (b) Mean, variance, and standard deviation of the flow vector magnitude. (c) Sum of the motion vector magnitude in each region of the frame characterizing the total amount of motion. (d) The major directions of the motion present in each region, char-

acterized by the eigenvectors and eigenvalues of the Principal Component Analysis. (e) Statistics from the Gray Level Co-occurrence Matrix (GLCM), which is obtained from the flow vector magnitude in the high motion region detected in the previous step. The statistics include (1) Energy, measuring the orderliness or regularity of flow vector magnitude, (2) Correlation, measuring the joint probability of the occurrence of flow vector magnitudes, (3) Contrast, measuring local variation of flow vector magnitude, and (4) Homogeneity, measuring similarity of flow vector magnitudes.

To train the SVM, we collect solo string performances that are distinct from the test set. The ground-truth P/NP labels for the training pieces are obtained from audio-based single-pitch detection results followed by manual corrections: If a pitch is detected in the audio of a video frame, then the frame is annotated as Play; otherwise, it is annotated as Non-play. To parameterize the SVM training algorithm we, (a) set the kernel function parameter to radial basis function kernel (RBF), (b) set the kernel scale parameter to automatic scaling. The relative amount of play/non-play classes in the training data varied from 76%-80% for play labels and 20%-24% for non-play labels.

5.1.3 Multi-Pitch Estimation

5.1.3.1 Audio-based MPE

The proposed method is built upon an audio-based method proposed in (Duan, Pardo and Zhang, 2010). It is a maximum-likelihood approach mod-

eling both spectral peaks and non-peak regions of the audio frame to be analyzed. Assume that the audio frame has N monophonic sound sources and let θ be a set of N fundamental frequencies. Fundamental frequency of each source is estimated by maximum likelihood estimation,

$$\hat{\theta} = \arg \max_{\theta \in \Theta} (O|\theta), \quad (5.2)$$

where Θ denotes the space of all possible sets of fundamental frequencies, and O is the observation, i.e., magnitude spectrum. This method estimates pitches in an iterative way from more prominent pitches to less prominent ones using a greedy strategy. After pitches and polyphony are estimated in each frame, a post-processing step is employed to smooth the estimates within several neighboring frames to remove inconsistent estimation errors.

5.1.3.2 Visually Informed MPE

An important disadvantage of this audio-based MPE method (and other audio-based methods as well) is that the instantaneous polyphony estimation is not that accurate. For low-polyphony pieces (e.g., duets and trios) it tends to overestimate, while for high-polyphony pieces (e.g., quartets and above) it tends to underestimate. This is due to the polyphonic nature of string ensembles and the harmonic relations among the sources. The P/NP labels detected from the visual scene, on the other hand, can provide a more accurate estimation for the instantaneous polyphony. We therefore count the

number of active players in each video frame and use it to replace the audio-based polyphony estimates in the corresponding audio frames. To account for the possible errors in the P/NP detection in individual frames, we still adopt the post-processing module to refine the pitch and polyphony estimates within neighboring frames.

5.1.4 Multi-Pitch Streaming

5.1.4.1 Audio-based MPS

We also build the visually informed MPS algorithm upon our prior audio-based MPS framework (Duan, Han and Pardo, 2014). It formulates the MPS problem as a constrained clustering problem. It takes pitch estimates in individual frames (MPE results) as input and clusters them into different pitch streams. Two kinds of constraints are considered: must-links and cannot-links. Must-link constraints are added to pitches that are close in both time and frequency. Cannot-link constraints are used to prevent assigning pitches in the same time frame to the same source. Pitches from the same source have similar timbre features, so the objective function is designed to minimize the timbre inconsistency of pitches within the same stream as

$$f(\Pi) = \sum_{m=1}^M \sum_{t_i \in S_m} \|t_i - c_m\|^2, \quad (5.3)$$

where Π is the clustering of pitches, M represents the number of monophonic sound sources, t_i denotes the timbre feature vector of the i -th pitch, and c_m is the centroid of timbre features in stream S_m .

An iterative algorithm was proposed to solve this constrained clustering problem in (Duan, Han and Pardo, 2014). After an initialization, in each iteration the clustering is updated to decrease the objective function while satisfying all the constraints that have already been satisfied. The new clustering is found through a *swap operation*: swapping cluster labels of two streams within a *swap set*. The swap set is defined as a connected subgraph of the pitches between two clusters using the already satisfied constraints as edges. If the swap operation is accepted (i.e., it decreases the objective function), then the set of already satisfied constraints is updated to all constraints satisfied by the new clustering. The algorithm was proven to converge.

5.1.4.2 Visually Informed MPS

To design the visually informed MPS system, we inject the P/NP information obtained from the video into the audio-based framework to prevent assigning pitches to non-playing players. The algorithm is described in Algorithm 1, where ‘*’ indicates the changes from the audio-based algorithm (Duan, Han and Pardo, 2014) for incorporation of the P/NP information.

As shown in the algorithm, the P/NP information is incorporated at two places. First, during the clustering initialization, estimated pitches are sorted in a descending order and are assigned to only the active performers

from high-pitched instruments to low-pitched instruments. Second, when updating the clustering through the swap operations, only swaps that satisfy the P/NP constraints are accepted. The satisfaction criterion is that for each source in the swap set, among the frames that the source has a pitch after the swap operation, at least 50% of the frames are labeled as Playing according to the P/NP information. This criterion prevents the algorithm from assigning too many pitches to an inactive source during the clustering update processing. We chose this threshold to account for possible errors in the P/NP detection results. As a preliminary study, we did not investigate the effect of this parameter on the MPS results.

5.1.5 Experiments

5.1.5.1 Dataset

We evaluate the proposed system on the URMP dataset¹ (Li, Liu, Dinesh, Duan and Sharma, 2019). Each piece was assembled (mixed for audio and composed for video) from isolated recorded but well coordinated performances of individual instrumental tracks. We selected all the string-instrument (violin, viola, cello, and bass) pieces which include 3 duets, 2 trios, 4 quartets and 2 quintets. Video files are downsampled to 240P for optical flow estimation. Note that as an initial demonstration here we only evaluate on a few pieces due to the lack of large audio-visual datasets.

¹<http://www.ece.rochester.edu/projects/air/projects/urmp.html>

Algorithm 1: Visually informed MPS algorithm. '*' indicates places of the incorporation of the P/NP information.

M - the number of monophonic sound sources

PNP - the binary P/NP matrix indicating which player is playing at which frame (1-playing, 0-not playing)

begin

```

* Initialization: Assign pitches to only active players in the
pitch-descending order;  $t \leftarrow 0$ ; repeat
   $t \leftarrow t + 1$ ;  $f_{max} \leftarrow f(\Pi_{t-1})$ ;  $\Pi_t \leftarrow \Pi_{t-1}$  while  $f_{max} == f(\Pi_{t-1})$ 
  & not all pitches  $p_1, \dots, p_N$  are traversed do
    Randomly pick  $p_n$  which is in stream  $S_m$  and not be
    replaced; for  $j = 1 : M$  do
      Find the swap set of  $p_n$  between  $S_m$  and  $S_j$ ; * if  $PNP$  is
      satisfied in the swap set then
        Do the swap to get a new clustering  $\Pi_s$ ; if
         $f(\Pi_s) < f_{max}$  then
           $| f_{max} \leftarrow f(\Pi_s)$ ;  $\Pi_t \leftarrow \Pi_s$ ;
        end
      end
    end
  end
   $C_T$  = constraints satisfied by  $\Pi_t$ ;
until  $\Pi_t = \Pi_{t-1}$ ;
  Return  $\Pi_t$  and  $C_t$ ;
end

```

5.1.5.2 Evaluation of Play/Non-play Activity Detection

Since we have only 11 videos, we adopt a training strategy where the piece to be evaluated is considered a test piece and the remaining videos are considered as the training set from which the features are extracted to form training and test feature matrix whereas for the training and test labels we use the ground truth P/NP information from the annotated audio file. The training feature matrix with the training label is fed into the SVM training algorithm to develop a model which is used on the test feature matrix to get the predicted labels and the predicted labels are compared with test labels to find a match which is used as a measure of accuracy. The left half of Table 5.1 shows the video-based P/NP detection accuracy for all 11 videos. We can see good match between predicted labels and the ground truth test labels which has resulted in an average accuracy level 85.3% for the pieces. For piece #7 and #11, as we can observe, the accuracy has decreased because of the limited bowing motion due to the nature of the composition of the two pieces. Higher the accuracy, higher is the probability of an improvement on multi-pitch estimation and streaming accuracy.

5.1.5.3 Evaluation of Multi-Pitch Estimation

For audio analysis, we first evaluate the multi-pitch estimation results using the MPE accuracy measure proposed in (Dixon, 2000) with the error tolerance of one quarter tone. The right half of Table 5.1 lists all of the 11 testing

Piece No.	P/NP Detection Accuracy (%)					MPE Accuracy (%)		
	P1	P2	P3	P4	P5	Audio	Video PNP	GT PNP
# 1	97.4	91.5	-	-	-	70.2	83.6	85.1
# 2	93.6	93.3	-	-	-	68.7	72.2	74.2
# 3	81.1	71.3	-	-	-	58.5	62.7	70.0
# 4	92.5	91.4	78.4	-	-	59.8	65.9	68.6
# 5	93.9	92.9	89.4	-	-	75.0	76.7	79.0
# 6	83.4	88.4	78.6	73.4	-	49.5	52.3	56.3
# 7	69.3	73.6	75.1	70.1	-	52.1	52.0	59.0
# 8	90.0	90.9	84.6	86.4	-	62.2	62.3	66.6
# 9	93.1	95.5	82.4	91.5	-	62.2	63.3	65.7
# 10	91.9	92.3	88.5	94.1	91.2	47.4	52.3	53.3
# 11	74.2	75.1	70.0	75.3	62.5	46.4	44.0	48.8

Table 5.1: Results of video-based Play/Non-play detection and MPE accuracy of the 11 test pieces.

pieces and compares the audio-based baseline method with the proposed visually informed method. For almost all of the testing pieces, the proposed method achieves prominent improvements (13% on the first piece) based on the audio-based method. The MPE accuracy drops when the polyphony number increases, and the improvements become less pronounced when the P/NP detection accuracy decreases. For 7th and 11th pieces the proposed approach even drops from baseline method due to the relatively low accuracy of P/NP detection, which supports our analysis in Section 5.1.5.2. We further add another testing group where the ground-truth P/NP labels are incorporated into the MPE process. This gives a upper bound of the MPE accuracy improvement by using a perfect visual activity detection module.

To further prove the effectiveness of the proposed approach, we also create more subsets using the 11 pieces for a statistical evaluation. We arrange all

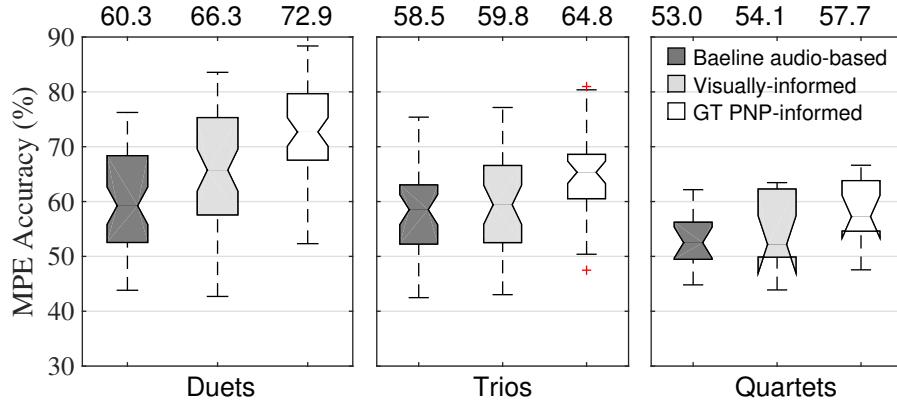


Figure 5.4: Boxplot of MPE accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.

possible track combinations within each piece. For the example of a quartet, we can further arrange 6 duets and 4 trios using the 4 original tracks. This operation on all of the 11 pieces totally results in 53 duets, 38 trios, 14 quartets and 2 quintets, on which the average increase of the MPE accuracy from the baseline method to the proposed visual-based method is 3.71%. We group all these subsets by polyphony (excluding the 2 quintets) and show the boxplot of MPE accuracy in Figure 5.4. The improvements of the first two polyphony groups are statistically significant under a paired t-test with $p < 10^{-19}$ and $p < 10^{-3}$, respectively. The improvement for quartets is not statistically significant under the same test at the significance level of 0.05.

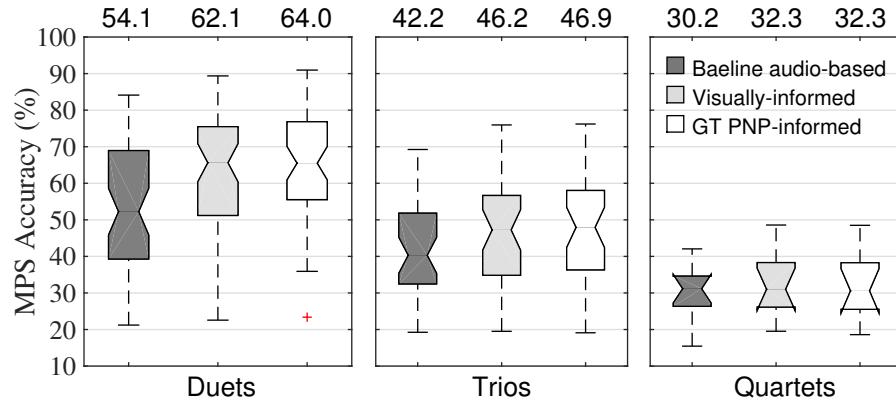


Figure 5.5: Boxplot of MPS accuracy grouped by polyphony on all subsets derived from the 11 pieces, comparing the baseline audio-based method (dark gray), proposed visually informed method (light gray), and the incorporation of ground-truth PNP labels (white). The number above each box shows the mean value of the box.

5.1.5.4 Evaluation of Multi-Pitch Streaming

We evaluate the proposed visually informed MPS system on the same derived track combinations from the 11 pieces. The criterion for a pitch to be considered correctly streamed needs to satisfy both the frequency deviation condition and the stream assignment condition: it should deviate less than a quarter tone from the ground-truth pitch in the stream that it is assigned to (Duan, Han and Pardo, 2014). Figure 5.5 shows the boxplot of the MPS accuracy of the three comparison methods on three polyphony groups. It can be seen that the visually informed approach improves over the audio-based baseline consistently over all three groups, reaching close to the ground-truth P/NP-informed upper bound. A paired t-test shows that the improvement is statistically significant for all groups at a significance level of

0.01. Further analyses show that the improvement is more pronounced when the pieces have a layered structure (up to 30% improvement), i.e., different tracks come and go at different times. This is intuitive as this is when the video-based P/NP detection is most informative for streaming.

5.1.6 Conclusions and Discussions

In this section, we propose and demonstrate a framework for visually-informed multi-pitch analysis of string ensembles. The play/non-play activity of different players is detected from analysis of the video and incorporated into the techniques used for multi-pitch estimation (MPE) and multi-pitch streaming (MPS). Our results demonstrate that, in most cases, the proposed incorporation of visual information offers statistically significant improvements (under a paired t-test) in pitch analysis accuracy over purely audio-based approaches.

5.2 Vibrato Analysis

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.

5.2.1 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 (Sundberg,

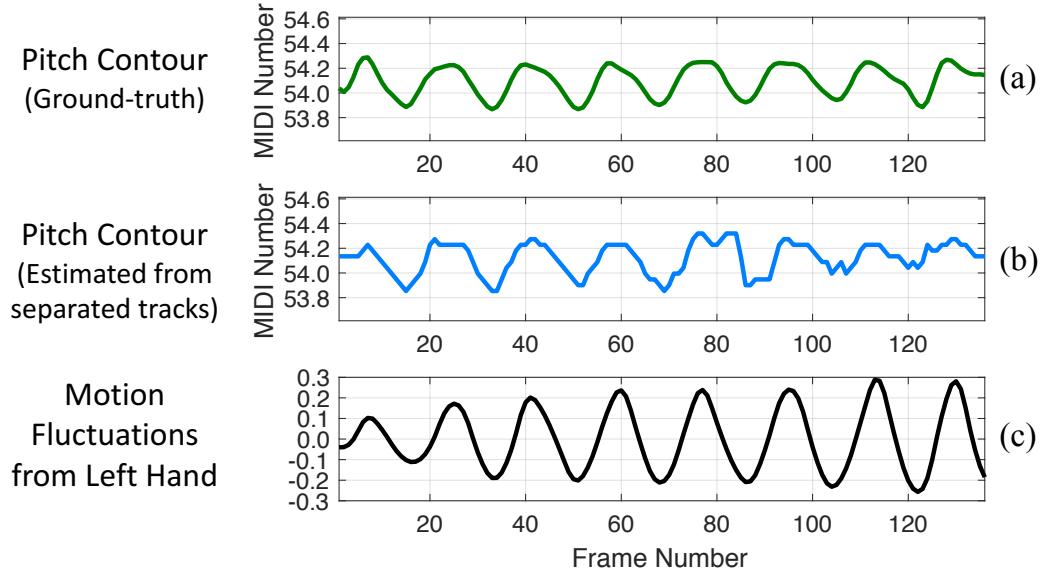


Figure 5.6: The proposed method tackles the challenging problem of vibrato analysis for polyphonic music by exploiting information from the video to augment audio analysis. (a) The ground-truth pitch contour of a cello vibrato note in a violin-cello duet performance showing a clear vibrato pattern, (b) 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 work.

1994). Important characteristics of vibrato include *rate* and *extent* of this periodic modulation (Fletcher and Sanders, 1967). 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 (Abeßer et al., 2017). It is critical in expressive performance pedagogy for singing (Nakano, Goto and Hiraga, 2006) and violin (Yin, Wang and Hsu, 2005). It also facilitates other MIR tasks such as singing voice extraction (Hsu and Jang, 2010), harmonic-percussive decomposition (Park and Lee, 2015), and audio-visual source association (Li, Xu and Duan, 2017). Vibrato analysis also provides the statistical basis for vibrato synthesis of musical instruments (Järveläinen, 2002), singing voices (Gu and Lin, 2014), and bird songs (Bonada, Lachlan and Blaauw, 2016), through which the synthesized sounds are more realistic and expressive.

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 (Barbancho et al., 2009), calculating the median distance of the neighboring peaks/troughs of the pitch contour (Friberg, Schoonderwaldt and Juslin, 2007), analyzing the spectral peak after a Fourier

transform of the pitch contour (Ventura, Sousa and Ferreira, 2012), cross-correlation analysis of frequency/amplitude modulation (Von Coler and Roebel, 2011), and a nonlinear sinusoidal decomposition method (Yang, Rajab and Chew, 2017).

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 (Dinesh et al., 2017). Abeßer et al. (2015) 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 (Hsu and Jang, 2010) and template convolution (Driedger et al., 2016) 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 5.6 shows the limitation of audio-based analysis and motivates the video-based analysis proposed in this work. In Figure 5.6 (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 (Mauch and Dixon, 2014) on the isolated (ground truth) signal of the cello note prior to mixing. Vibrato characteristics are clearly observable in this pitch contour. Figure 5.6 (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 (Duan and Pardo, 2011a). 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 5.6 (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 5.6 (a) very well.

The overview of our proposed approach is illustrated in Figure 5.7. This approach integrates audio, visual and score information, and assumes that

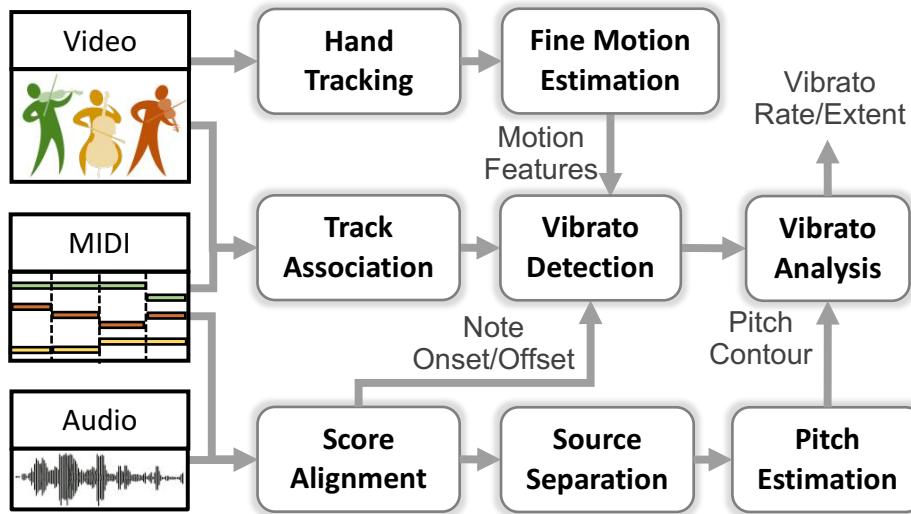


Figure 5.7: System overview of the proposed video-based vibrato detection and analysis framework.

the players in the video are well associated with score tracks. 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 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.

5.2.2 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 (Abeßer et al., 2015), score information can be utilized to alleviate the difficulty of pitch estimation and its assignment to sources.

5.2.2.1 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 (Li, Dinesh, Duan and Sharma, 2017). Then the audio mixture is separated using harmonic masking as described in (Duan and Pardo, 2011*a*): Pitches of each source 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 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 (Li, Xu and Duan, 2017) 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.

5.2.2.2 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

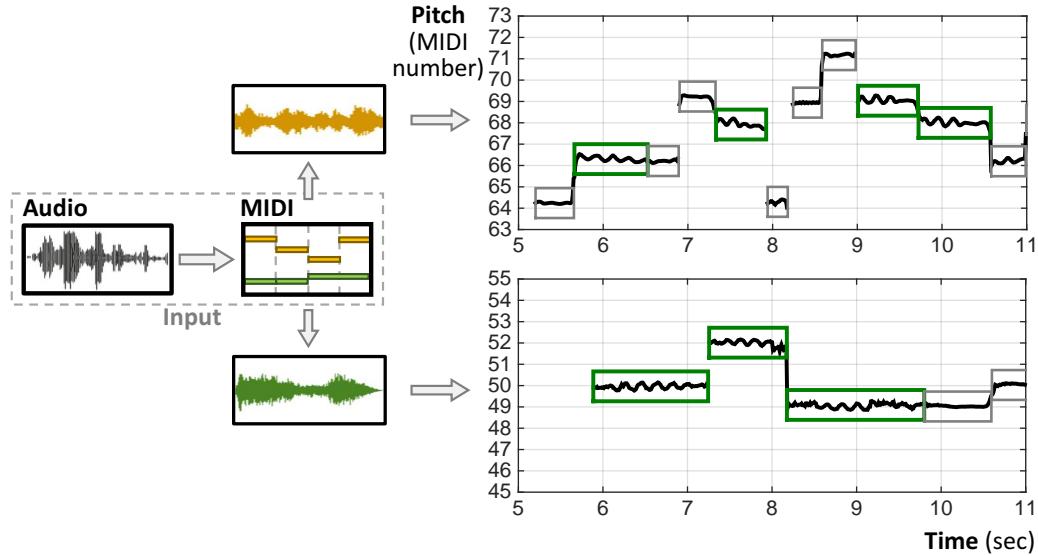


Figure 5.8: Audio-based vibrato detection. Detected vibrato notes are marked with green rectangles in the pitch trajectories estimated by score-informed pitch estimation.

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 (Abeßer et al., 2015) and spectral analysis (Ventura, Sousa and Ferreira, 2012). 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 (Geringer, MacLeod and Allen, 2010)), 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 range 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 5.8 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.

5.2.3 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.

5.2.3.1 Motion Capture

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 (Tomasi and Kanade, 1991) and implemented

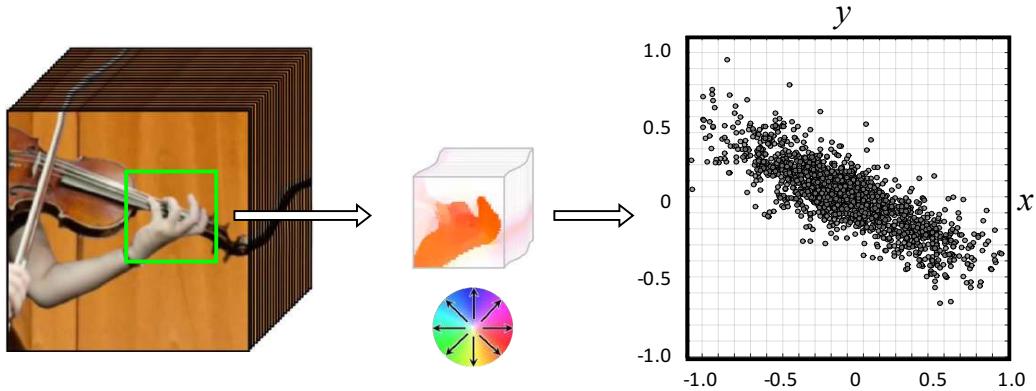


Figure 5.9: Motion capture results from left hand tracking (left), color encoded pixel velocities (middle), and scatter plot of frame-wise refined motion velocities (right).

using the same parameters as presented in (Li, Xu and Duan, 2017). The KLT tracker results in a dynamic region of tracked hand location where we apply the optical flow estimation (Sun, Roth and Black, 2010) 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

$$\mathbf{v}(t) = \mathbf{u}(t) - \bar{\mathbf{u}}(t), \quad (5.4)$$

where $\bar{\mathbf{u}}(t)$ is the moving average of $\mathbf{u}(t)$ over a 10 frame window. Figure 5.9 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.

5.2.3.2 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 (Li, Dinesh, Duan and Sharma, 2017), 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.

5.2.3.2.1 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 5.2.3.1. 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.

5.2.3.2.2 PCA

We also propose an unsupervised framework for vibrato detection. From Figure 5.9, 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

$$V(t) = \frac{\mathbf{v}(t)^T \tilde{\mathbf{v}}}{\|\tilde{\mathbf{v}}\|}, \quad (5.5)$$

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

$$X(t) = \int_0^t V(\tau) d\tau. \quad (5.6)$$

This displacement curve corresponds to the fluctuation of the vibrating length of the string and hence the pitch fluctuation of the note. Figure 5.6 (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.

5.2.3.3 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 auto-correlation 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:

$$\hat{v}_e = \arg \min_{v_e} \sum_{t=t^{on}}^{t^{off}} \left| 100 \cdot F(t) - v_e \frac{X(t)}{\hat{w}_e} \right|^2. \quad (5.7)$$

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).

5.2.4 Experiments

5.2.4.1 Dataset and Evaluation Measures

The vibrato detection and analysis system is tested on the URMP dataset (Li, Liu, Dinesh, Duan and Sharma, 2019). 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 5.2.2.2 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 5.2.2.2 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.

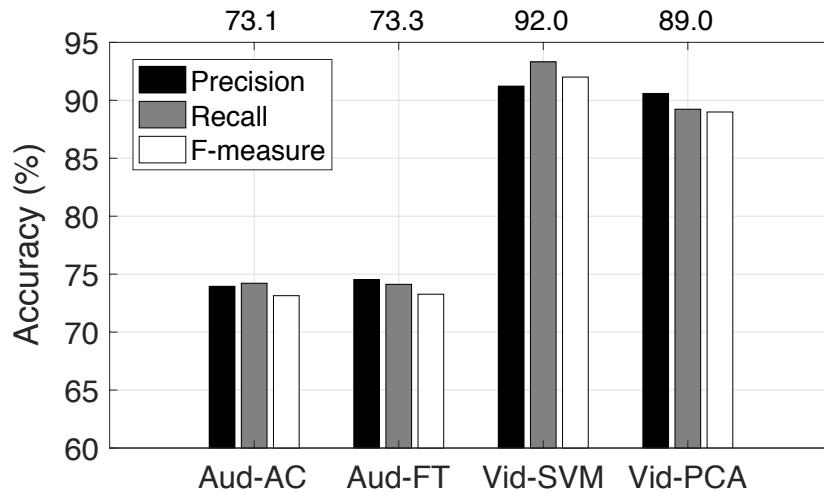


Figure 5.10: Overall vibrato detection results showing the precision, recall, and F-measure (shown on top) accuracies for 2 audio-based methods and 2 video-based methods.

5.2.4.2 Results

5.2.4.2.1 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 5.10. 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.

5.2.4.2.2 Vibrato Detection Evaluation on Different Cases

We further investigate how the vibrato detection performance changes along with polyphony and instrument types. Figure 5.11 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 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 5.12 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

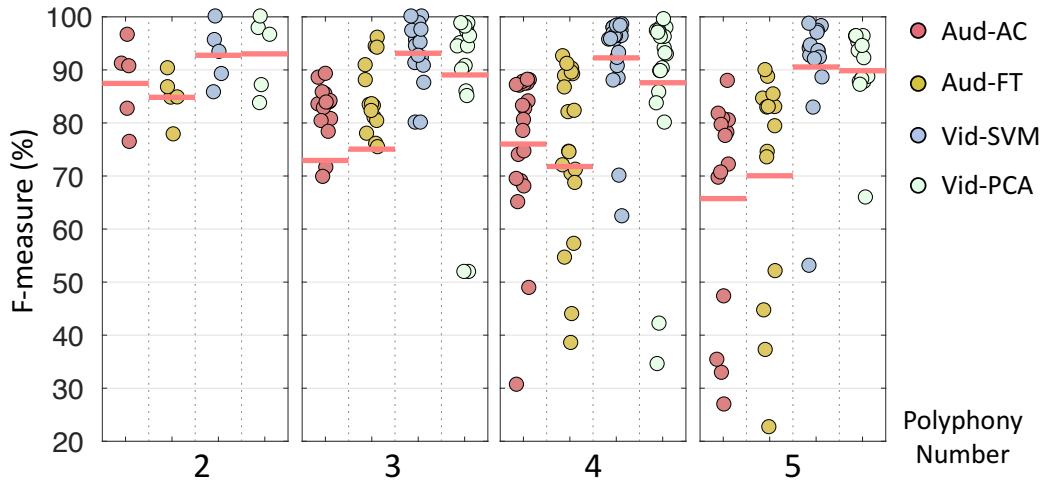


Figure 5.11: Vibrato detection performance decreases as polyphony increases for audio-based methods, while it stays the same for video-based methods.

(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.

5.2.4.2.3 Evaluation of Vibrato Characteristics

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

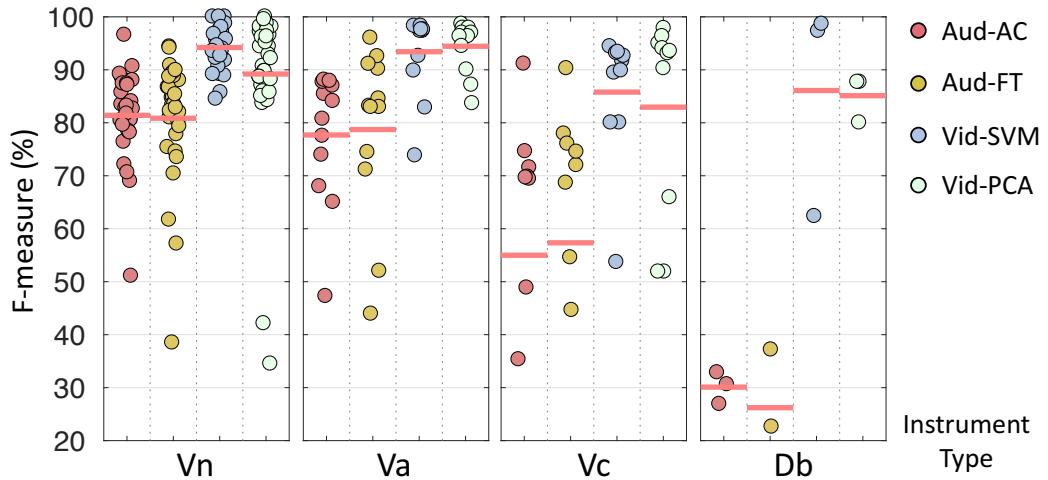


Figure 5.12: Vibrato detection performance decreases when the fundamental frequency decreases for audio-based methods, while it stays the same for video-based methods.

estimation error of 3.47 cents and a median of 2.29 cents. Figure 5.13 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.

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 5.14 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 (Mick, 2012); to produce audible pitch fluctuations on the

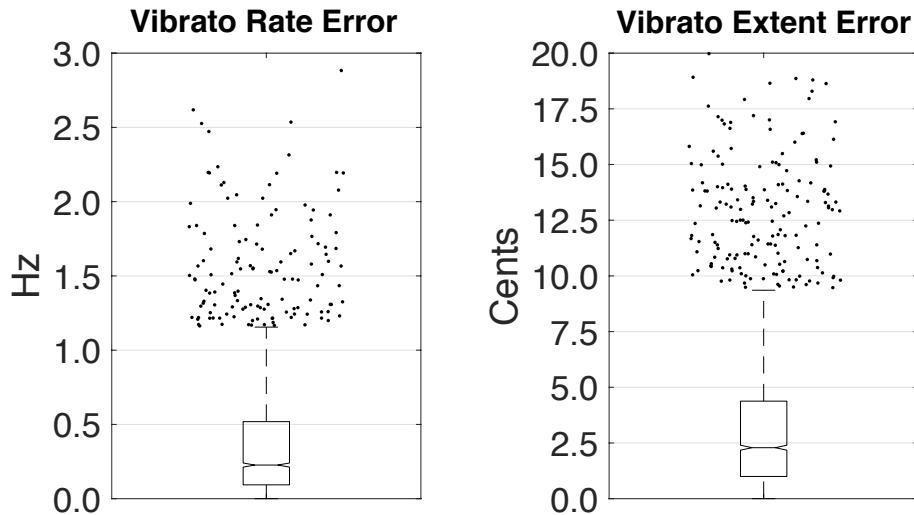


Figure 5.13: Distribution of vibrato rate and extent estimation error on all notes of all tracks.

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 (Papich and Rainbow, 1974)) are typically slower than other string instrumentalists.

We also analyze the vibrato patterns of the four different violinists among the 31 violin tracks, as plotted in Figure 5.15. 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 work, but our proposed system

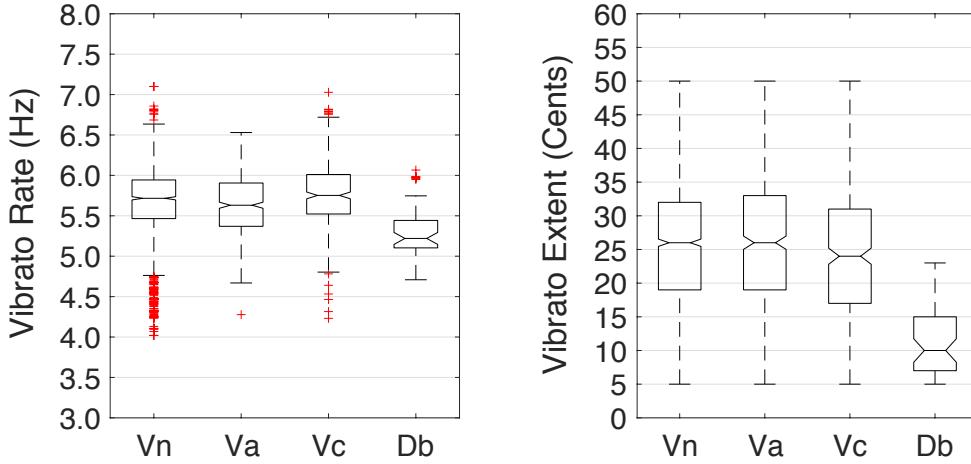


Figure 5.14: Distributions of vibrato rate and extent for different instruments.

can provide a powerful tool for further analyses on the musicology side.

5.2.5 Conclusions

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

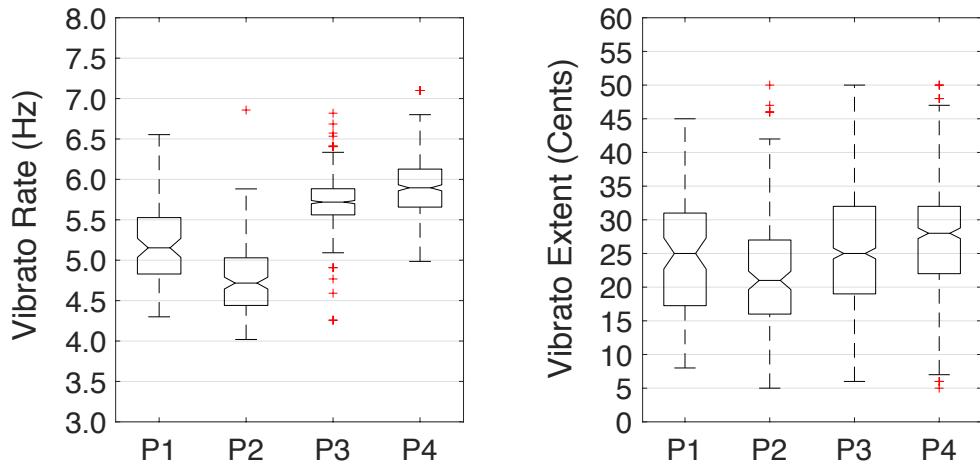


Figure 5.15: Distributions of vibrato rate and extent of four different violin players.

detection and analysis.

Chapter 6

Audiovisual Singing Voice Separation

Separating a song into vocal and accompaniment components is an active research topic, and recent years witnessed an increased performance from supervised training using deep learning techniques. We propose to apply the visual information corresponding to the singers' vocal activities to further improve the quality of the separated vocal signals. The video frontend model takes the input of mouth movement and fuses it into the feature embeddings of an audio-based separation framework. To facilitate the network to learn audiovisual correlation of singing activities, we add extra vocal signals irrelevant to the mouth movement to the audio mixture during training. We create two audiovisual singing performance datasets for training and evaluation, respectively, one curated from audition recordings on the Internet, and the other recorded in house. The proposed method outperforms audio-based methods in terms of separation quality on most test recordings. This advantage is especially pronounced when there are backing vocals in the accompaniment, which poses a great challenge for audio-only methods.

6.1 Introduction

Vocal performance is an important art form of music. The task of singing voice separation is to isolate vocals from the audio mixture, which contains other instrumental sounds that help to define the harmony, rhythm, and genre. Singing voice separation is often the first step towards many application-oriented vocal processing tasks including pitch correction, voice beautification, and style transfer, as implemented in some mobile Apps such as WeSing and Smule. It is also often a pre-processing step for other research tasks such as singer identification (Berenzweig, Ellis and Lawrence, 2002), lyrics alignment (Fujihara et al., 2006), and tone analysis (Fujihara and Goto, 2007).

There are various scenarios when video recordings are available for singing performances, such as operas, music videos (MV), self-recorded singing activities. In pop music, creative visual performances give artists a substantial competitive advantage. Moreover, due to the rapid growth of Internet bandwidth and smartphone users, videos of singing activities are becoming popular in a number of video sharing platforms such as TikTok and Instagram.

Visual information, e.g., lip movement, has been incorporated and shown its benefits in speech signal processing, such as audiovisual speech separation (Lu, Duan and Zhang, 2019), enhancement (Afouras, Chung and Zisserman, 2018), and recognition (Petridis et al., 2018). Visual information has also been incorporated in music instrumental performance analysis (Duan et al.,

2019), such as source association (Li, Dinesh, Xu, Sharma and Duan, 2019), source separation (Zhao et al., 2019), multi-pitch analysis (Dinesh et al., 2017), and playing technique analysis (Li, Dinesh, Sharma and Duan, 2017), and cross-modal generation (Chen et al., 2017; Li, Maezawa and Duan, 2018). For singing performances, however, little work has been done. It is reasonable to think that visual information would also help to analyze singing activities, and in particular, separate singing voices from background music. This is based on the fact that mouth movements and facial expressions of the singer are often correlated with the singing voice signal fluctuations. The advantages of audiovisual analysis over audio-only analysis can be best demonstrated in songs with backing vocals in the accompaniment and songs with multiple singers singing simultaneously. However, to what extent does the incorporation of visual information help singing voice separation is still a question. Different from speech signals, singing voices (except for rap) generally change slower, showing less frequent matching with mouth movements. Furthermore, some musically important fluctuations of the singing voice such as pitch modulations show little, if any, correlation with mouth movements.

Therefore, it is our intention to answer the following research question in this chapter: *Can visual information about the singer improve singing voice separation, and if yes, how much?* It is noted that in this work we define the singing voice separation tasks as separating the solo singing voice from the background music, which itself may contain backing vocals. This is different

from the definition in SiSEC2018¹ or MIREX where all vocal components in a song are treated as the singing voice. We argue that our definition is widely useful in many application scenarios.

To answer the above-mentioned research question, we design an audiovisual neural network model to separate the solo singing voice from the background music that may contain backing vocals. This network model takes both the audio mixture signal and the mouth region of the singing video as input. The audio processing sub-network is designed based on the MM-DenseLSTM (Takahashi, Goswami and Mitsufuji, 2018), the champion of SiSEC2018, a biennial campaign of music source separation. The visual processing sub-network uses several convolutional and LSTM layers to encode mouth movements of the singer. The audio and visual encodings are fused before they are used to reconstruct the solo singing magnitude spectrogram. The training target of the proposed audiovisual network is to minimize the Mean-Square-Error (MSE) loss of the magnitude spectrogram reconstruction of the solo singing voice. To facilitate the network to learn audiovisual correlation of singing activities, we add extra vocal signals irrelevant to the solo singer to the audio mixture during training. To investigate the benefits of visual information, we compare the proposed audiovisual model with several state-of-the-art audio-based singing separation methods and an audiovisual speech enhancement method. We further vary the architecture and input of

¹A community-based Signal Separation Evaluation Campaign, accessible at: <https://sisec18.unmix.app>

the visual processing sub-network to compare their performances.

One challenge we encounter in this work is the lack of audiovisual datasets of singing. For training, this can be addressed by randomly mixing solo singing videos downloaded from the Internet with irrelevant accompaniment music. We download A cappella audition vocal performance videos and randomly mix their audio with accompaniment audio tracks from the MUSDB18 dataset (officially provided by SiSEC2018) to generate mixtures. We name this the *Audition-RandMix* dataset, and partition it into training, validation and test subsets. For evaluation, however, we need audiovisual recordings of singing with its relevant accompaniment music in separate tracks. To our best knowledge, no such dataset exists. Therefore, we record a new audiovisual dataset named *URSing*, where singers are recruited to sing along with prepared accompaniment tracks. Some of these accompaniment tracks also contain backing vocals.

We conduct experiments on both the Audition-RandMix test set and the URSing dataset. Results on both sets show that the proposed audiovisual method outperforms baseline methods in most test conditions, no matter if the accompaniment tracks contain the backing vocals or not. We further conduct subjective evaluations on a cappella video performances in the wild to prove the advantages of our proposed method.

The contributions of this chapter include:

- The first work to incorporate visual information to the state-of-the-art music source separation framework to address the singing voice

separation problem,

- A proposal of solo voice separation where backing vocal components are regarded as accompaniment tracks, which better fits many application scenarios, and
- The first audiovisual singing performance dataset, URSing, free for download².

6.2 Related Work

6.2.1 Singing Voice Separation

Early methods for singing voice separation include non-negative matrix factorization (Vembu and Baumann, 2005), adaptive Bayesian modeling (Ozerov et al., 2007), robust principal component analysis (Huang et al., 2012), and auto-correlation (Rafii and Pardo, 2011). Recently, deep learning based methods are proposed to model convolutional (Chandna et al., 2017) or recurrent structures (Uhlich et al., 2017) of magnitude spectral representations of music signals. Some works also learn to reconstruct spectral phases in addition to magnitudes (Takahashi et al., 2018), while others directly work on time-domain waveforms with an end-to-end training strategy (Lluis, Pons and Serra, 2019; Stoller, Ewert and Dixon, 2018). A direct comparison of recently proposed methods is available at the SiSEC2018 post. The best per-

²<http://www.ece.rochester.edu/projects/air/projects/URSing.html>

forming methods in SiSEC2018 use a DenseNet structure with a recurrent structure to process magnitude spectrograms (Takahashi and Mitsufuji, 2017; Takahashi, Goswami and Mitsufuji, 2018), where the feature reuse strategy inside each dense block greatly reduces the model size. Later some open-sourced methods/tools have been proposed with comparable results, such as Open-Unmix (Stöter et al., 2019) and Spleeter (Hennequin et al., 2019). In this chapter, we build upon the DenseNet to propose an audiovisual model.

6.2.2 Audiovisual Source Separation

Most audiovisual separation works are proposed for speech signals. Hou et al. (2018) address the speech enhancement problem, where speech is separated from background noise. The model has a two-stream structure that takes both noisy speech and frames of the cropped mouth regions as inputs to compute their features. These features are then concatenated by a fusion network which also outputs corresponding clean speech and reconstructed mouth regions. Another audiovisual speech enhancement work proposed in (Afouras, Chung and Zisserman, 2018) uses 1D convolutional layers to reconstruct the magnitude spectrogram of the clean speech and uses it to further estimate its phase spectrogram. The input of the visual branch is the feature embeddings on the lip region that are pre-trained on lip reading tasks. For speech separation, one challenge is the permutation problem where the separated components need to be assigned to the correct talkers. Lu, Duan and Zhang (2018) specifically address the problem by applying the visual

information as a post-processing step to adjust the separation mask. Later the same group proposes to fuse the visual information to an audio-based deep clustering framework to propose an audiovisual deep clustering model for speech separation in (Lu, Duan and Zhang, 2019). Another work is described in (Ephrat et al., 2018), where the input is the mixture spectrogram and the face embeddings of all the appeared speakers in the audio sample. The training target is the complex mask that can be applied to the original spectrogram to recover the complex spectrogram of each speaker.

Less work has been proposed for audiovisual music separation. Parekh et al. (2017) apply non-negative matrix factorization (NMF) to separate string ensembles, where the bowing motions are used to derive additional constraints on the activation of audio dictionary elements. This method, however, is only evaluated on randomly assembled video scenes of string instruments where distinct bowing motions of each player are clearly captured. Zhao et al. (2018) propose to learn static audiovisual correspondence with cross-modal source localization; The correlation between each pixel in a given video frame and the sound component can be constructed. Later the same group proposes to learn dynamic audiovisual correspondence (Zhao et al., 2019) which captures correlations between motions and sound fluctuations. Another followup work is to recognize visual gestures for sound source separation (Gan et al., 2020). This line of research achieves promising results in audiovisual music separation, but have not addressed singing voice separation.

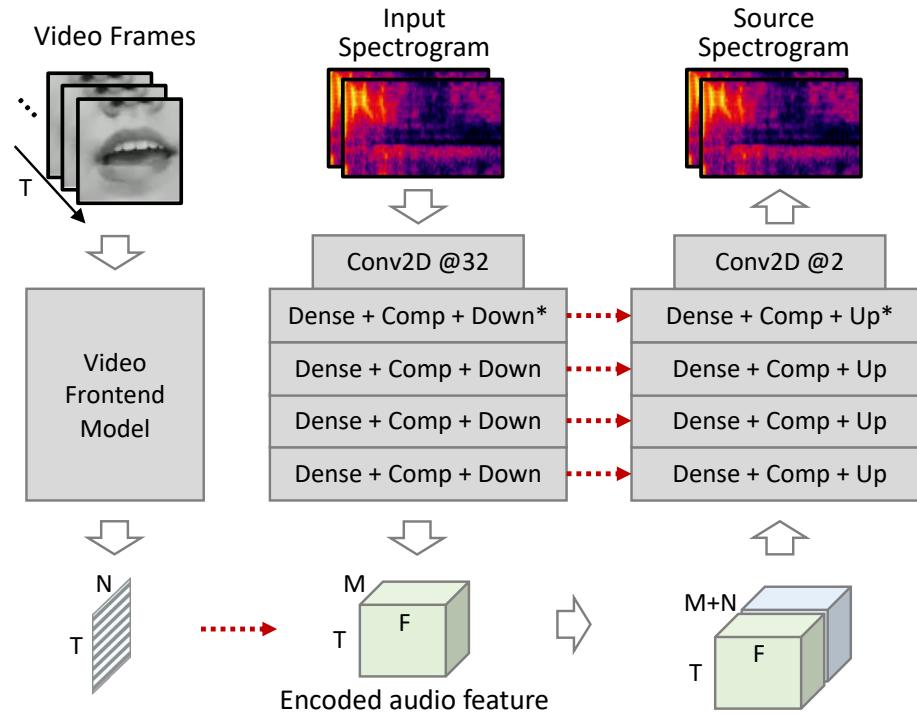


Figure 6.1: The proposed model structure. Dashed arrows denote the concatenation operation. Downsample/upsample are applied to both time and frequency dimensions in the outer layers (marked by *), while they are only applied to the frequency dimension in the inner layers.

6.3 Method

6.3.1 Network Architecture

The proposed system integrates a state-of-the-art audio separation model named MMDenseLSTM (Takahashi, Goswami and Mitsufuji, 2018) with a video frontend model. The MMDenseLSTM model consists of convolutional layers stacked into dense blocks, which alternates downsample/upsample lay-

ers to form a multi-scale structure. It first embeds an input magnitude spectrogram into an encoded feature space and decodes it to recover the separated magnitude spectrogram. Skip connections are added as concatenations on the corresponding layers with the same feature map size. This “encoder-decoder” structure with skip connections is widely applied in several music separation models (Jansson et al., 2017; Stoller, Ewert and Dixon, 2018; Zhao et al., 2019; Liu and Yang, 2018). The video frontend model extracts visual features from mouth movements, which are fused with the encoded audio feature. The network structure is illustrated in Figure 6.1. We explain each part of the model in detail as follows.

6.3.1.1 Dense Block

The densely connected structure is originally proposed in (Huang et al., 2017) for object recognition in computer vision. Within a dense block, the output feature maps of all the convolutional layers³ have the same size and are concatenated with each other along the channel dimension. This structure reuses the feature maps from previous layers and greatly reduces the model size.

6.3.1.2 Compression Layer

We apply a compression layer right after each dense block, which improves the model compactness by reducing the number of feature maps (channels).

³A convolutional layer includes BN+ReLU+Conv2D throughout the chapter.

It is a convolutional layer with 1×1 kernels. The number of output feature maps depends on the compression ratio ranging from 0 to 1. We set it to 0.2, which means that the number of feature maps is reduced by 80% after each compression layer.

6.3.1.3 Downsample-Upsample

Downsample and upsample layers are used to resize the feature maps without changing the the number of channels. For downsample layers, we use average pooling with 2×2 kernels after the first dense block and compression layer, and 1×2 kernels in the following layers. In other words, downsampling is performed along both the time and frequency dimensions in the first layer, but only to the frequency dimension in other layers. Symmetrically, for upsample layers, we apply transposed convolutional layers with 2×2 kernels and strides at the last upsample layer but 1×2 kernels for the other upsample layers. This downsample/upsample strategy processes the frequency dimension in multiple scales, but downsamples the time dimension only once, making the audio stream have the same frame rate as the video stream. The encoded spectrogram feature is denoted as $\mathbf{S} \in R^{M \times T \times F}$, with the channel (M), downsampled time (T), and frequency (F) dimensions.

6.3.1.4 Multi-band

Following (Takahashi and Mitsufuji, 2017), we also equally divide the spectrogram into a low-frequency band and a high-frequency band and apply the

above-mentioned encoder-decoder structure on each sub-band. The dense blocks of low-frequency band have a higher channel number. Detailed parameters can be referred to (Takahashi and Mitsufuji, 2017).

6.3.1.5 Video Frontend Model

We propose to apply a separate input branch to parse the input video stream and fuse it with the encoded audio features. The video stream is a sequence of mouth region images in consecutive video frames. Raw RGB values are normalized to zero mean and unit variance. Note that there exist models that can process mouth region inputs, which are pre-trained on lip reading tasks (Stafylakis and Tzimiropoulos, 2017; Petridis et al., 2018). We do not employ these pre-trained models because we find that they overfit the LRW dataset (Chung and Zisserman, 2016): Each 29-frame data sample contains several words, but only one word around the middle frame is labeled and used in training, causing the pre-trained models to only attend to the middle frames of a video input. Here we use 2D convolutional layers and LSTM layers to extract the visual features from the input RGB frames. Each frame is processed independently sharing the same CNN parameters before being fed into the following LSTM layer. Detailed network architecture is Conv2D@16 (channel number is 16), Conv2D@16, Conv2D@32, Conv2D@32, FC@256, LSTM@128, and FC@ N , where N is the dimension of the encoded feature vector for each video frame. The input video stream with T frames results in a feature map $\mathbf{V} \in R^{N \times T}$. There is no pooling operation along the time

dimension thus the temporal information is preserved.

6.3.1.6 Audiovisual Fusion

The extracted visual feature map from the video branch is fused with the encoded audio spectrogram feature map \mathbf{S}_A . To do so, the visual feature map \mathbf{S}_V is inflated along the third dimension and then concatenated with the audio feature to obtain the audiovisual feature $\mathbf{S}_{AV} \in R^{L \times F \times T}$, where $L = M + N$ is the concatenated channel dimension. Note that the temporal information from both the audio and video branches is correlated during this fusion; This is different from some works where audiovisual fusion is performed on feature maps that aggregate information along time.

In addition to minor structural changes, we also drop the LSTM structure of the original MMDenseLSTM model (Takahashi, Goswami and Mitsufuji, 2018) when we design the audio branch of our proposed model. This follows the observation that the addition of the LSTM structure does not achieve substantial improvement in SiSEC2018 yet the number of parameters would be increased significantly for audiovisual fusion.

6.3.2 Training

We train the model to predict the magnitude spectrogram of the source signal and use the original mixture's phase to recover the time-domain waveform. Many spectral-domain source separation methods, especially those for speech signals, use a spectrogram mask as the training target; This mask is then

multiplied element-wise with the mixture signal’s magnitude spectrogram to recover the source magnitude spectrogram. For music separation, some recent works train networks to directly output the source magnitude spectrogram (Uhlich et al., 2017; Takahashi, Goswami and Mitsufuji, 2018) using a Mean-Squared-Error (MSE) loss. We follow the same way and take the source magnitude spectrogram as the training target. However, we have a mask layer that regularizes the feature maps into the range of [0, 1] using a Sigmoid function and multiplies the mask layer with the input spectrogram. We find that this is necessary for fusing the visual input into the audio feature, as the audiovisual method even degrades the separation performance when the mask layer is not in place. We have a comparative experiment in Section 6.5.5.

The model input is the magnitude spectrogram of the original audio mixture which contains both solo vocals and background music, and the mouth region of the video frames corresponding to the vocals. The output is the magnitude spectrogram of the source audio. Compared to the audio mixture input, the visual input provides much less information about the source signals, therefore, the training loss may not be propagated back sufficiently into the visual branch, making the audiovisual network difficult to train. One way to address this is to explicitly learn audiovisual matching, either through pre-training (Lu, Duan and Zhang, 2018) or early audiovisual fusion (Lu, Duan and Zhang, 2019). Another way might be to add visual reconstruction as another training target, leading to a chimera-like network structure (Hou et al.,

2018).

In this work, we address this problem by adding some extra vocal components to the original mixture, which are not related to the mouth movements and thus are not included in the target vocal spectrogram. This forces the model to learn audiovisual correlations after the fusion and only separate the vocal components that are related to the visual input. Note that in the training samples all of the vocal and accompaniment components are randomly mixed, so neither the extra vocal components or the solo vocal components have harmonic relations with the accompaniment tracks. In the experiments, we show that the strategy of training with randomly generated vocal-accompaniment pairs performs decently on real songs.

6.4 Dataset

Since there is no publicly available audiovisual singing voice dataset containing isolated vocal tracks, we collect our own data for training and evaluating the proposed method.

6.4.1 A Cappella Audition Vocals (AAV)

We curated 491 YouTube videos of solo singing performances by querying the YouTube search API with the keyword “Academic Acappella Audition”. We only selected video excerpts where the singer faces the camera and sings without accompaniment. The total length of these excerpts is about 8 hours.

As it is difficult to find relevant and appropriate accompaniment tracks for these solo singing performances, in our experiments we simply randomly chose instrumental accompaniment tracks (from the “accompaniments” track in the MUSDB18 dataset) and mixed them with the solo singing excerpts to create singing-accompaniment mixtures.

The randomly mixed samples are used for training, validation, and evaluation. Before the mixing process, vocals in AAV are divided into training/validation/evaluation sets roughly as 8:1:1 (50 tracks for evaluation). Accompaniment tracks from MUSDB18 (which contains a wide range of music genres and instrument types) are also divided into the three sets following the official way (also 50 tracks for evaluation). Then mixing is applied on each split independently to form the training/validation/evaluation sets. Volume of each track is normalized using the root-mean-square (RMS) value. For training and validation sets, each track is split into short samples (around 2.5 seconds) for random mixing, resulting in a massive amount of mixed samples. We do not normalize the volume of each individual sample so the mixing may have different SNRs. For evaluation, mixing is performed on a random bijection between the 50 vocals and 50 accompaniments. For each mixing, we pick a 30-second excerpt (with both vocal and accompaniments present) for evaluation, following the same strategy as the MUSDB18 dataset. This set is referred to as “Audition-RandMix” in the following experiments. In addition, we randomly add extra vocals from another vocal track in the test set (with the same RMS value as the target vocal) into these mixings for evaluations,

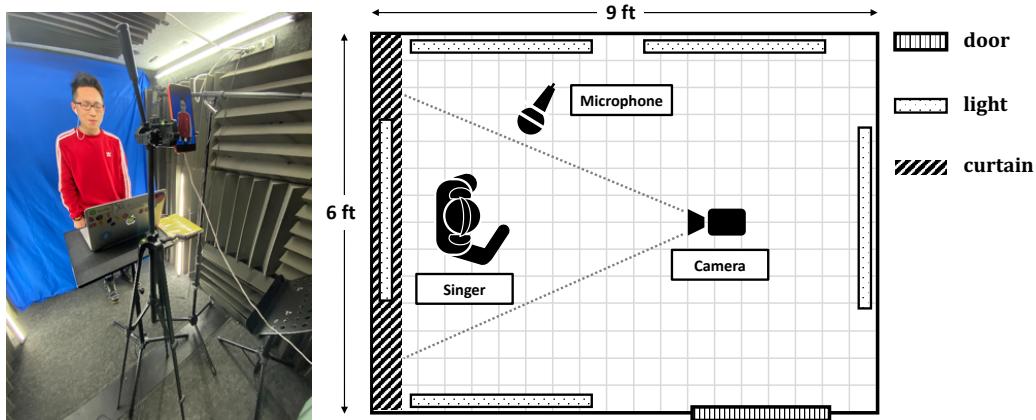


Figure 6.2: A sample photo and floor plan of the sound booth for the recording process of the URSing dataset.

referred as “Audition-RandMix (v+)”, in order to explore the model performance in more challenging cases. Note that all the testing samples in this condition are not musically meaningful and cannot represent real songs.

6.4.2 URSing

To evaluate the proposed method in more realistic singing performances, we create the University of Rochester Multimodal Singing Performance Dataset (URSing). In this chapter, we only use the URSing dataset for evaluation. A brief description of the creation process is described below.

6.4.2.1 Singer Recruiting

Singers are students at the University of Rochester. Audition is performed to filter out unqualified singers who could not sing in tune. Each participant

receives \$5 for recording each song, and is allowed to record up to 5 songs.

6.4.2.2 Piece Selection

To ensure high recording efficiency, the singers pick their own songs and their favorite accompaniment tracks to sing along. We do not put constraints on song genres, but filter out songs of which the accompaniment tracks are of low sound quality.

6.4.2.3 Recording

To ensure synchronization, the singers listen to the accompaniment track through earphones while recording their singing voice. Their voices are recorded using an AT2020 condenser microphone hosted by Logic Pro X, and their videos are recorded using a smartphone. The recording is conducted in a semi-anechoic sound booth. A sample photo and the floor plan of the sound booth are shown in Figure 6.2.

6.4.2.4 Post-processing

For each solo vocal recording we use the following plug-ins to simulate the typical audio production procedure in commercial recordings: a) static noise reduction (*Klevgränd Brusfri* and *Waves X-noise*), b) pitch refinement (*Melodyne*), c) sound compression (*Fabfilter Pro-C 2*), and d) reverberation (*Fabfilter Pro-R*). We also adjust the vocal volume to balance it with the accompaniment track. Beyond this, we do not perform any other editing on



Figure 6.3: Examples of video frames of the URSing dataset and cropped mouth region pictures as the input to the video branch of the proposed method.

the audio recording (e.g., time warping or rhythmic refinement) to preserve the synchronization with the visual performance. To synchronize the audio recording captured by the AT2020 microphone with the video recording captured by the smartphone, we use the audio recording captured by the built-in microphone of the smartphone as the bridge, through cross correlation.

6.4.2.5 Annotation

Since the mouth movements are mostly relevant to the singing performance, we provide the annotations of the mouth regions in the dataset. This is performed using the Dlib library (King, 2009), an automatic tool for facial landmark detection, followed by manual check. The mouth region is represented as a square bounding box with the side length equal to 1.2 times of the maximum horizontal distance for all mouth landmarks.

This results in 65 songs, totaling 4 hours of audiovisual recordings of

singing performance. For each song, we provide the audio recording of the solo singing voice (in WAV, 44.1 KHz, 16 bits, mono) and the corresponding accompaniment audio track (same format, mono or stereo). We also provide the video recording of the soloist (in MP4, 1080P portrait, 29.97 FPS), where the soundtrack is the mixture (direct sum) of the solo vocal and the accompaniment tracks. Note that when we prepare the accompaniment tracks, we do not avoid the tracks containing backing vocals, as they are the challenging and useful cases to study in this chapter. Example video frames and cropped mouth region pictures are provided in Figure 6.3.

We also choose a set of 30-sec excerpts where both solo vocal and accompaniment tracks are prominent to form a benchmark evaluation set. Specifically, for each of the 65 songs, we choose one 30-sec excerpt without backing vocal and one with back vocal, if such excerpts are available. We provide this information in the metadata. This results in 54 excerpts with accompaniment tracks that only contain instrumental components (referred as “URSing” in the following experiments) and 26 excerpts with accompaniment tracks that also contain backing vocals (referred as “URSing (v+)”). The latter, presumably, are more challenging for solo vocal separation and more useful for showing advantages of audiovisual methods. In this chapter, since we do not use any songs from URSing for training, we only use these 30-sec excerpts for evaluation.

6.5 Experiments

6.5.1 Implementation Details

For audiovisual singing videos, audio is downsampled to 32 KHz. We use a frame length of 1024 and a hop size of 640 (20 ms) for spectrogram calculation. Video data is converted to 25 FPS (equivalent to 40 ms frame hop size), and the frame size of mouth regions is interpolated into 64×64 . Each data sample is 2.56 seconds long, containing 128 audio frames and 64 video frames. The input/output audio spectrogram has the shape of $2 \times 128 \times 513$ (channels \times frames \times frequency bins), and each input video stream has the shape of $64 \times 64 \times 64$ (frames \times width \times height).

During training, for half of the training/validation samples, we add extra vocal components that are not related to the mouth movements to encourage the model to learn audiovisual correlations. We use batch size of 8 for training on a TITAN X GPU with 11.9 GB graphic memory. It takes about 40 hours to train for 50 epochs. We adopt early stopping when the validation loss does not decrease for 10 consecutive epochs.

6.5.2 Evaluation Metric

6.5.2.1 SDR

We calculate the signal-to-distortion ratio (SDR) between the separated vocal waveforms and the ground-truth ones using the BSS Eval toolbox V4, same

as the evaluation measure applied in SiSEC2018. Specifically, for each 30-sec evaluation excerpts, we calculate the median SDR over all 1-sec audio segments.

6.5.2.2 SI-SDR

We also use the Scale-Invariant SDR as proposed in (Le Roux et al., 2019) as a complimentary evaluation measure. It has been proven to be a more robust measure and has been widely used in speech separation. For each 30-sec song excerpt we also calculate the median over all 1-sec audio segments.

6.5.3 Baselines

We first use the original mixture recording (referred as “MIX” in the experiments) as the separated vocal for evaluation on our dataset. This sets lower bounds of separation results without any separation techniques. Then we apply two oracle filtering techniques that utilize ground-truth source signals: The ideal binary mask (IBM) assigns each time-frequency bin to the predominant source. The ideal ratio mask (IRM) distributes the power of each time-frequency bin into different sources according to the power ratio of the ground-truth sources. IBM and IRM set upper bounds for time-frequency masking-based source separation methods.

We then compare our proposed method with three audio-based music separation methods as baselines. The first is a commercial software *RX7* developed by *iZotope*. We apply batch processing of the “music rebalance”

function with the preset “isolate vocals” on “medium” level. The second is “Spleeter” (Hennequin et al., 2019), which achieves best separation results among all open-source tools on the evalution set of the MUSDB18 dataset by 2019. Note that these ready-to-use methods are not trained on our dataset for a more fair comparison. The third audio-based baseline is our own implementation of MMDenseLSTM (Takahashi, Goswami and Mitsufuji, 2018) that achieved the best results in SiSEC2018. We verify our implementation by achieving similar results to the reported ones on the MUSDB18 dataset, following SiSEC2018’s official train/test split. We then train and evaluate this model on our datasets using the same conditions as those for the proposed audiovisual method.

We also implement an audiovisual speech enhancement method named AVDCNN proposed in (Hou et al., 2018). This method applies 2D CNNs to take noisy speech and the mouth region visual recording as inputs, fuses encoded audio and visual features using fully-connected layers, and outputs the enhanced speech signal as well as reconstructed video frames of mouth movements. We choose audiovisual speech enhancement instead of audiovisual speech separation as the baseline, because we believe that speech enhancement is more relevant to singing voice separation from background music in terms of foreground-background relations of sources.

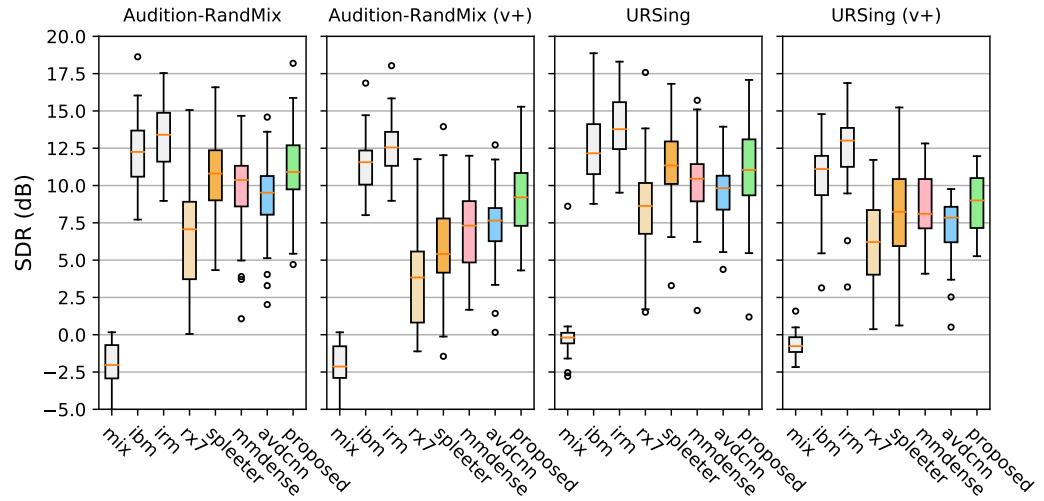


Figure 6.4: The SDR (dB) comparison on separated solo vocals with different methods.

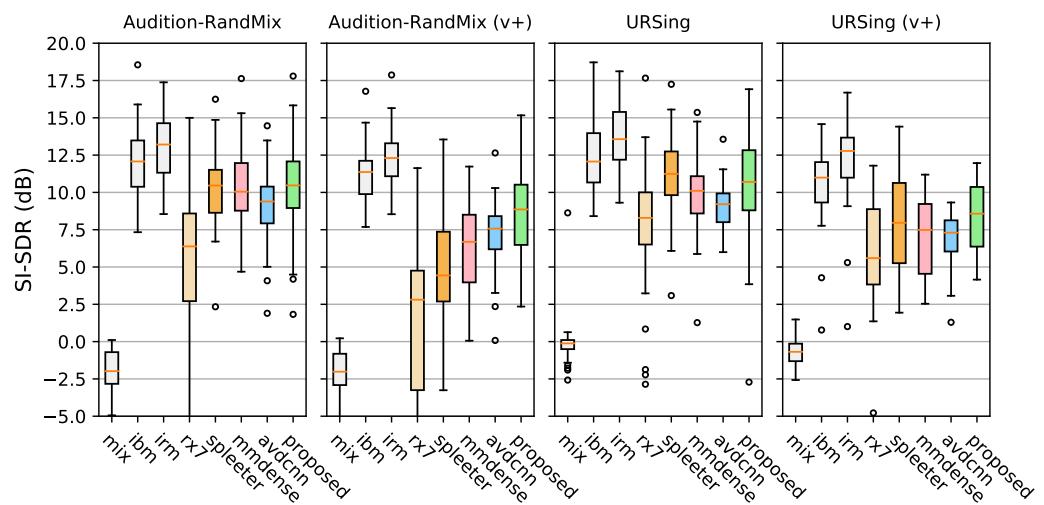


Figure 6.5: The SI-SDR (dB) comparison on the separated solo vocal from different methods.

6.5.4 Overall Results

We evaluate the comparison methods on the four test sets described in Section 6.4: Audition-RandMix, Audition-RandMix (v+), URSing, and URSing (v+). Note that Audition-RandMix sets are randomly mixed samples and URSing sets represent realistic songs. “v+” means that the accompaniments contain vocal components. Boxplots of SDR and SI-SDR results are shown in Figures 6.4 and 6.5, where each data point in the boxplots is the median SDR or SI-SDR of the separated vocal of all 1-sec segments of a 30-sec excerpt. The horizontal line inside each box indicates the median value across all excerpts. Several interesting observations can be made from these figures.

6.5.4.1 Benefits of Visual Information.

First, the proposed method outperforms all audio-based separation baselines in most of the evaluation sets. This shows the advantage of incorporating visual information about the singer’s mouth movement for solo singing voice separation. Among the audio-based baseline methods, MMDenseLSTM is much stronger than iZotope RX7, because MMDenseLSTM is our own implementation and is trained on our dataset while iZotope RX7 is not. However, Spleeter outperforms our proposed system on the URSing set. This is because Spleeter is trained on a much larger internal dataset that contains 24,097 songs totalling 79 hours. Comparing songs with backing vocals (Audition-RandMix (v+) and URSing (v+)) to songs without backing vocals (Audition-RandMix and URSing), we can see that the outperformance

of the proposed method is better pronounced on songs with backing vocals. We argue that this is because audio-only methods tend to assign all the vocal components to the separated singing voice, while the proposed audiovisual method learns to only separate the vocal signals that are correlated to the solo singer’s mouth movements.

Figure 6.6 shows one 10-sec sample as an extreme case to compare the spectrograms of audio-based MMDenseLSTM method and the proposed audiovisual method when backing vocal components are strong (e.g., the middle part of the sample). We also show the mouth movement in several frames throughout this excerpt. It can be seen that MMDenseLSTM recognizes the backing vocal components in the middle frames as the solo vocal, while the audiovisual method suppresses those components significantly.

Wilcoxon signed-rank tests show that the proposed method significantly improves over MMDenseLSTM baseline on Audition-RandMix (v+) and URSing (v+) with a p value of 10^{-3} and 10^{-2} , respectively. The reason that the improvement is more pronounced on Audition-RandMix (v+) than on URSing (v+), we argue, are twofold: 1) backing vocals in URSing (v+) are not as strong as the intentionally added backing vocals in Audition-RandMix (v+), and 2) backing vocals in URSing (v+) often overlap with solo vocals and share the same lyrics, showing high correlations with the mouth movements of the solo singer, while the added backing vocals in Audition-RandMix (v+) are irrelevant to the solo vocal.

On songs without backing vocals, the outperformance of the proposed

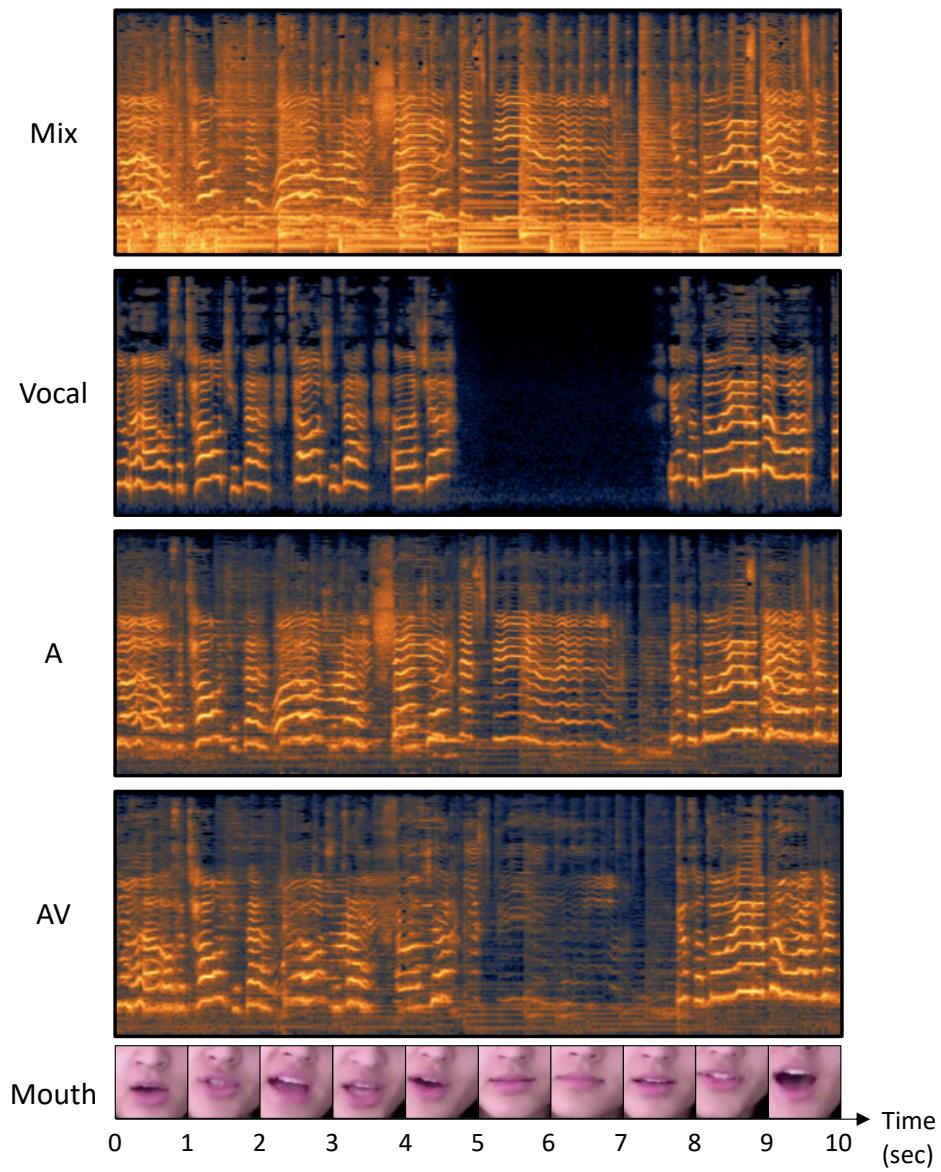


Figure 6.6: One 10-sec example comparing audio-based separation (MM-DenseLSTM) with audiovisual separation (proposed) on a song excerpt with strong backing vocals. The four spectrograms from top to bottom are original mixture, ground-truth vocal, audio-based vocal separation, and audiovisual vocal separation. This sample result has 10-sec long, and one mouth frame of each second is attached.)

method can still be observed. Subjective listening by the authors suggests that the visual information helps to reduce high-frequency percussive sounds (e.g., hi-hats) from the solo vocal, as the former do not correlate with mouth movements well.

6.5.4.2 Superiority of Proposed Audiovisual Architecture

Second, the proposed method outperforms the audiovisual speech enhancement baseline significantly in all evaluation sets. Note that the baseline is trained and evaluated on the same dataset as the proposed method. This shows the superiority of the proposed network architecture on the solo singing voice separation task. In particular, we argue two main reasons. First, the proposed model utilizes the commonly used U-net structure with skip connections, which generally achieves good results in music separation (Jansson et al., 2017; Stoller, Ewert and Dixon, 2018; Takahashi and Mitsufuji, 2017). Second, in our audiovisual fusion scheme we preserve the temporal correspondence, which prevents a substantial increase of the number of trainable parameters in the fusion layer. This is important when the DenseNet-based audio sub-network has a small model size. The variations of different video sub-networks, however, does not make much difference on the separation performance, as we analyzed in Section 6.5.5.

6.5.4.3 Limitations and Room for Improvement

Third, compared with reported SDR values in SiSEC2018, the SDR values in Figure 6.4 are much higher. For example, MMDenseLSTM reaches over 10 dB on URSing but only less than 7 dB in SiSEC2018 (method “TAK1” in (Stöter, Liutkus and Ito, 2018)). We argue that the songs used in SiSEC2018 (i.e., the MUSDB18 dataset) are professionally recorded, mastered and mixed vocals. They often contain complex components such as polyphonic vocals, background humming, and strong reverberation. They are mastered and mixed by professional music producers to intentionally make them better fused into the background music. In contrast, the ground-truth vocals in our datasets are solo vocals recorded in controlled environments with limited vocal effects added. It is reasonable to believe that the benefits of visual information can be further demonstrated on these professionally produced songs. In addition, the performance difference between the two Audition-RandMix test sets and the two URSing test sets seems to be small for all methods, including the oracle results. This shows that randomly mixed songs, although lacking harmonic and rhythmic coherence, are not easier to separate than the more realistically mixed songs, suggesting that it is reasonable to use randomly mixed songs to train the methods (Luo et al., 2017). However, whether this is still true for professionally produced songs is still a question.

On the other hand, there is still some gap between the proposed method and the oracle results on both SDR and SI-SDR in our evaluation sets. It is

likely that this gap will be even bigger on professionally produced songs. This suggests that much work can be done to improve the separation performance. For example, time-domain separation for the audio branch may improve the performance significantly (Luo and Mesgarani, 2018).

6.5.5 Different Video Front-End Models

To investigate the effects of the video front-end on the separation performance, we replace the proposed Conv2D+LSTM video front-end with several other widely-used visual feature extraction frameworks:

- No-mask. This experiment has the same video branch, but without a mask layer after the audiovisual fusion.
- Conv3D. The Conv3D model takes all the video frames from each sample as a feature map and a 4-th dimension is added as the channel dimension set as 64. We then apply 2 Conv2D layers (with the channel dimension 128 and 256) on each frame to share the channel dimension with Conv3D. Followed by pool operation and fully-connected layers, we obtain the video feature with the same dimension as $\mathbf{V}_{\text{Conv3D}} \in R^{N \times T}$. Note that in this structure, the temporal information is only parsed at the very first Conv3D structure, since no recurrent network is applied.
- DenseNet. Different from the proposed model, we replace the Conv2D layers with a dense block from the DenseNet structure. Each dense

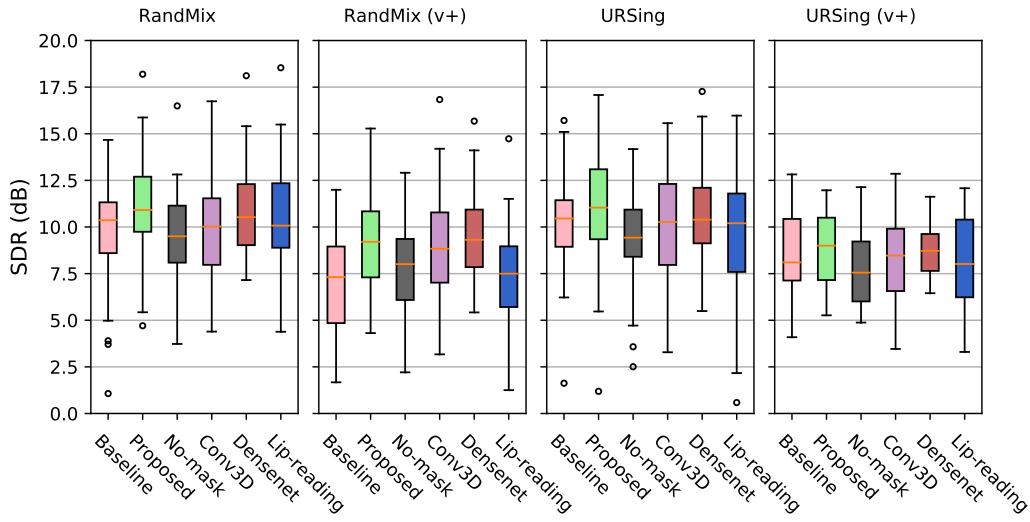


Figure 6.7: The SDR (dB) comparison on the separated solo vocal from the audiovisual method using different video front-end models. “v+” denotes for songs with backing vocals.

block has 2 layers with growth rate of 12. Then a Conv2D layer with 1×1 kernels is applied to compress the channel number to 32, resulting in the same feature dimension as the proposed CNN+LSTM model before feeding into the FC@256.

- **Lip-reading.** This variation uses a pre-trained model proposed in (Petridis et al., 2018) on the lip reading task on the LRW dataset (Chung and Zisserman, 2016). The original model structure consists of Conv3D, ResNet-34, and GRU. We only use the pre-trained model to extract the visual feature to integrate into our proposed audiovisual source separation model.

A comparison of different video front-end models is shown in Figure 6.7.

It can be seen that the proposed (Conv2D+LSTM) model achieves the highest SDR values for most cases, but some video front-end models do not make much difference. Applying a mask layer is critical, otherwise audiovisual method even degrades from the audio-based method. The Conv3D framework slightly degrades the performance, but still outperforms the audio-based baseline method (MMDenseLSTM). One reason for this performance drop may be that in this framework, there is no recurrent structure, and the temporal evolution of visual information is only processed by the Conv3D structure. As the Conv3D structure takes the raw input of mouth frames, it may be sensitive to mouth position changes due to landmark detection errors. The model pre-trained on lip reading ranks the worst among the audiovisual models. This is because the lip reading model was trained on the LRW dataset where for each 29-frame sample containing several words, only one word around the center frames is annotated as the training target. This makes the model only attend to middle frames of a video excerpt, leading to limited guidance for the singing voice separation and even degradation from audio-based methods.

6.5.6 Non-Informative or Misleading Visual Input

To further investigate how the incorporation of visual information affects the separation performance, in this section, we substitute the visual input (i.e., mouth region of the solo singer) with some irrelevant content. Figure 6.8 shows the separation results when we feed the visual branch with

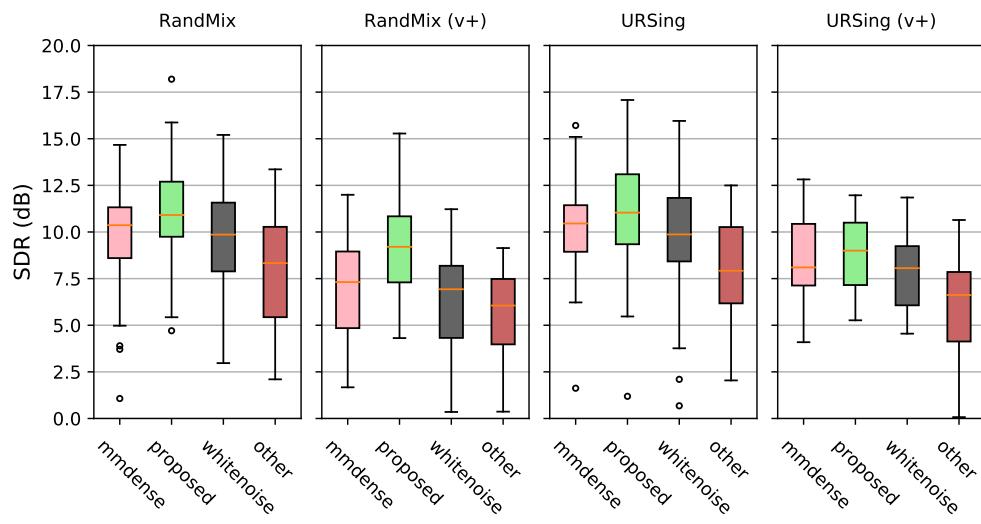


Figure 6.8: The SDR (dB) comparison on the separated solo vocal of MM-DenseLSTM audio-based baseline, the proposed audiovisual method, and the proposed method taking two kinds of irrelevant visual inputs (white noise and mouth region video of another singer). “v+” denotes for songs with backing vocals.



Figure 6.9: One sample frame of an a cappella song for subjective evaluation.

white noise or mouth movement of an irrelevant singer. It can be seen that when the visual input is white noise, the separation performance degrades slightly from the audio-based baseline MMDenseLSTM, suggesting that a non-informative visual input is slightly harmful for separation. When the visual input is the mouth region video of an irrelevant singer to provide misleading information about the singing activity, we can see a significant degradation. This shows the importance of visual information for our proposed model from the opposite side.

6.5.7 Evaluation on A Cappella Songs

In this section, We further evaluate the benefits of visual information incorporated in our proposed method on a cappella songs in the wild. We

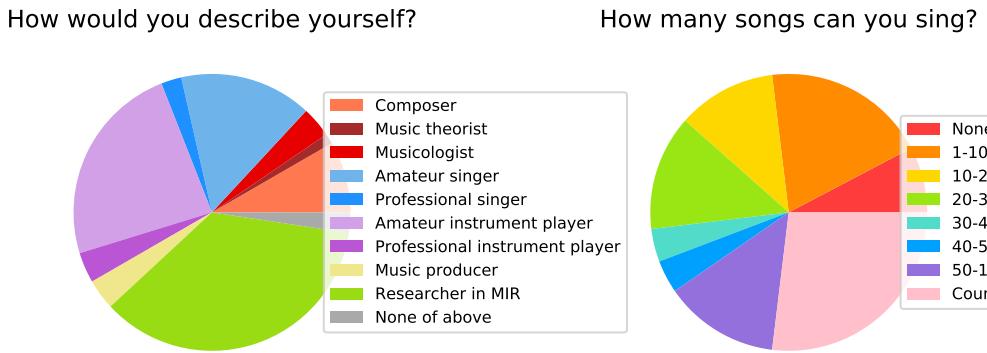


Figure 6.10: Statistics of the 26 subjects' musical background related to the subjective evaluation.

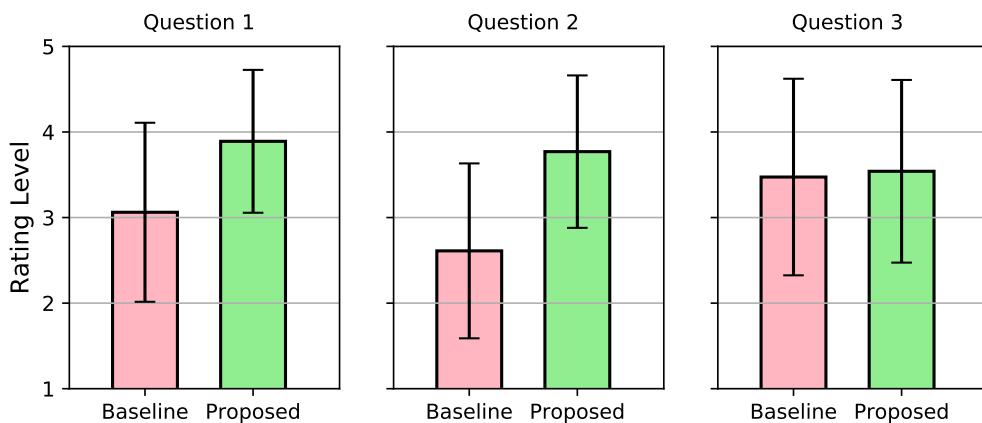


Figure 6.11: The subjective ratings of the separation quality in response to the three questions. Each error bar shows mean \pm standard deviation.

collect 35 audiovisual a cappella recordings from YouTube. These collections represent the extreme cases where all the accompaniment components are vocals (except for several cases where additional percussive instruments are also present), to study how much the proposed audiovisual method is advantageous while the audio-based method is very likely to fail. Most of these songs are chorus performance with a solo singer accompanied by harmonic vocals and/or vocal beatbox, while some are performance with multiple solo singers. We only keep the videos where the solo singer's mouth is visible and clear, without video shot transition for at least 10 seconds. A sample frame of one song is shown in Figure 6.9 with the mouth region of the targeted solo singer highlighted.

As we do not have access to the source tracks, we cannot evaluate the separation performance using common objective evaluation metrics. Instead, we conduct a subjective evaluation over 51 people. Some subjects are students or faculty from the University of Rochester, others are subscribers from the International Society for Music Information Retrieval (ISMIR) community. Statistics of the subjects' music background is shown in Figure 6.10. Each survey asks a subject to rate 7 of the 35 songs, and each subject may take more than one surveys. For each song, the subjects first watch a 10-sec excerpt of the original performance and then watch the same video twice with the solo singing voice separated by two different singing voice separation methods in a random order to rate the separation quality. Due to the variations across these songs, the original recording serves as a reference for

a consistent scoring scheme. For each video we also highlight the mouth region of the target solo singer (see Figure 6.9) to help subjects focus on the corresponding solo voice. The specific evaluation questions are:

- Question 1: *What do you think about the overall separation quality for the targeted singer?*
- Question 2: *What do you think about the separation quality in terms of removing backing vocal accompaniments in the separated solo voice?*
- Question 3: *What do you think about the separation quality in terms of not introducing artifacts into the separated solo voice?*

The subjects answer each question using a scale from 1 to 5, where “1” represents *Very bad* and “5” represents *Very good*. The three questions are related to the common definitions of the three objective source separation evaluation metrics, SDR, SIR, and SAR, respectively.

The results of the subjective evaluations are presented in Figure 6.11. According to the collected responses for Question 1, the proposed audiovisual method is rated significant higher than the baseline audio-based method (Wilcoxon signed-rank test shows a p value of 3.5×10^{-31}); The average rating is raised from 3.1 to 3.9. For Question 2, the difference is even more significant, as the average rating is increased from 2.6 to 3.8 (with a p value of 3.1×10^{-45}), showing that the proposed method is especially beneficial for removing accompaniments from the mixture. Regarding the artifacts introduced into the separated solo vocals in Question 3, both methods achieve

a rating between “neutral” and “good”, and the difference is not statistically significant (with a p value of 0.46).

6.6 Conclusion

In this chapter, we proposed an audiovisual approach to address the solo singing voice separation problem by analyzing both the auditory signal and mouth movement of the solo singer in the visual signal. To evaluate our proposed method, we created the URSing dataset, the first publicly available dataset of audiovisual singing performances recorded in isolation for singing voice separation research. Both objective evaluations on artificially mixed singing music and subjective evaluation on professionally produced a cappella songs showed that the proposed method significantly outperforms state-of-the-art audio-based methods. The advantages of the proposed method is especially pronounced when the accompaniment track contains backing vocals, which have been difficult to separate from solo vocals by audio-based methods.

Chapter 7

Visual Performance Generation

Generating expressive body movements of a pianist for a given symbolic sequence of key depressions is important for music interaction, but most existing methods cannot incorporate musical context information and generate movements of body joints that are further away from the fingers such as head and shoulders. This chapter addresses such limitations by directly training a deep neural network system to map a MIDI note stream and additional metric structures to a skeleton sequence of a pianist playing a keyboard instrument in an online fashion. Experiments show that (a) incorporation of metric information yields in 4% smaller error, (b) the model is capable of learning the motion behavior of a specific player, and (c) no significant difference between the generated and real human movements is observed by human subjects in 75% of the pieces.

7.1 Introduction

Music performance is a multimodal art form. Visual expression is critical for conveying musical expression and ideas to the audience (Davidson, 1993; Dahl and Friberg, 2007). Furthermore, visual expression is critical for communicating musical ideas among musicians in a music ensemble, such as predicting the leader-follower relationship in an ensemble (Tsay, 2014).

Despite the importance of body motion in music performance, much work in automatic music performance generation has focused on synthesizing expressive audio data from a corresponding symbolic representation of the music performance (e.g., a MIDI file). We believe that, however, body motion generation is a critical component that opens door to multiple applications. For educational purposes, for example, replicating the visual performance characteristics of well-known musicians can serve as demonstrations for instrument beginners to learn from. Musicologists can apply this framework to analyze the role of gesture and motion in music performance and perception. For entertainment purposes, rendering visual performances along with music audio enables a more immersive music enjoyment experience as in live concerts. For automatic accompaniment systems, appropriate body movements of machine musicians provide visual cues for human musicians to coordinate with, leading to more effective human-computer interaction in music performance settings.

For generating visual music performance, i.e., body position and motion

data of a musician, it is important to create an expressive and natural movement of the *whole body* in an online fashion. To consider both expressiveness and naturalness, the challenge is to maintain some common principles in music performance constrained by the musical context being played. Most previous work formulates it as an inverse kinematics problem with physical constraints, where the generated visual performance is limited to hand shapes and finger positions. Unfortunately, this kind of formulation fails to address the two challenges; specifically, (1) it fails to generate the whole body movements that are relevant to music expression, such as the head and body tilt, and (2) it fails to take into account the musical context constraints for generation, which do not contribute to ergonomics.

7.2 Related Work

There has been work on cross-modal generation, mostly for speech signals tracing back to the 1990s (Bregler, Covell and Slaney, 1997), where a person's lips shown in video frames are warped to match the given phoneme sequence. Given the speech audio, similar work focuses on synthesizing photo-realistic lip movements (Suwajanakorn, Seitz and Kemelmacher-Shlizerman, 2017), or landmarks of the whole face (Eskimez et al., 2018). Some other work focuses on the generation of dancers' body movements (Seo et al., 2013; Li, Zhou, Xiao, He and Li, 2018) and behaviors of animated actors (Perlin and Goldberg, 1996).

Similar problem settings for music performances have been rarely studied. When the visual modality is available, the system proposed in (Li, Dinesh, Duan and Sharma, 2017) explores the correlation between the MIDI score and visual actions, and is able to target the specific player in an ensemble for any given track. Purely from the audio modality, Chen et al. (2017) propose to generate images of different instrumentalists in response to different timbres using cross-modal Generative Adversarial Networks (GAN). Regarding the generation of videos, related work generates hand and finger movements of a keyboard player from an MIDI input (Yamamoto et al., 2010) through inverse kinematics with appropriate constraints. All of the above-mentioned works, however, do not model musicians' creative body behavior in expressive music performances.

Given the original MIDI score, Widmer, Flossmann and Grachten (2009) propose to predict three expressive dimensions (timing, dynamics, and articulations) on each note event using a Bayesian model trained on a corpus of human interpretations of piano performances. It further gives a comprehensive analysis of computer's creative ability in generating expressive music performances, and proves that certain aspects of personal styles are identifiable and even learnable from MIDI performances. Regarding to the expressive performance generation in visual modality, Shlizerman et al. (2018) propose to generate expressive body skeleton movements and adapt them into textured characters for pianists and violinists. Different from our proposed work, they take the input of audio waveforms rather than MIDI performances. We ar-

gue that MIDI data is a more scalable format to carry context information, regardless of recording conditions and piano acoustic characteristics. And most of piano pieces have the sheet music in MIDI format, which can be aligned with a waveform recording.

7.3 Method

The goal of our method is to generate a time sequence of body joint coordinates, given a live data stream of note events from the performer’s actions on the keyboard (MIDI note stream), and synchronized metric information. We seek to create the motion at 30 frames-per-second (FPS), a reasonable frame-rate to ensure a perceptually smooth motion. In this section, we introduce the technical details of the proposed method, including the network design and training conditions. We first use two CNN structures to parse the raw input of the MIDI note stream and the metric structure, and feed the extracted feature representations to an LSTM network to generate the body movements, as a sequence of upper-body joint coordinates forming a skeleton. The network structure is shown in Figure 7.1.

7.3.1 Feature Extraction by CNN

In contrast to traditional methods, our goal is to model expressive body movements that are associated with the keyboard performance. In this sense, the system should be aware of the general phrases and the metric structure

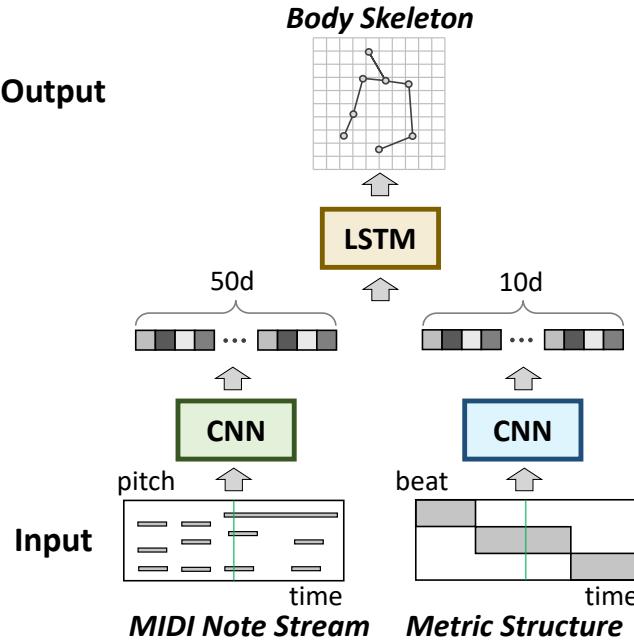


Figure 7.1: The proposed network structure.

in addition to each individual note event. Instead of designing hand-crafted features, we use CNNs to extract features from the raw input of the MIDI note stream and the metric structure, respectively.

7.3.1.1 MIDI Note Stream

We convert the MIDI note stream into a series of two-dimensional representations known as the *piano-roll matrix*, and for each of them extract a feature vector ϕ_x as the *piano-roll feature*.

To prepare the piano roll, the MIDI note stream input is sampled at 30 frames-per-second (FPS) to match the target frame rate. This quantizes the time resolution into the unit of 33 ms, as a video frame. Then for each time

frame t we define a binary piano-roll matrix $\mathbf{X} \in R^{128 \times 2\tau}$, where element (m, n) is 1 if there is a key depression action at pitch m (in MIDI note number) and frame $t - \tau + n - 1$, and 0 otherwise. We set $\tau = 30$. The key depression timing is quantized to the closest unit boundary. Note that the sliding window covers both past τ frames and future $\tau - 1$ frames, and the note onset interval in \mathbf{X} captures enough information for motion generation to “schedule” its timing. Looking into the future is necessary for the generation of proper body movements, which is also true for human musicians: to express natural and expressive body movements, a human musician should either look ahead on the sheet music, or be acquainted with it beforehand. Later in Section 7.3.2 we will introduce in which cases we can avoid the potential delays in real-time applications.

We then use a CNN to extract features from the binary piano-roll matrix \mathbf{X} , as CNNs are capable of capturing local context information. The design of our CNN structure is illustrated in Figure 7.2.a. The input is the piano-roll matrix \mathbf{X} and the output is a 50-d feature vector ϕ_x as the piano-roll feature. There are two convolutional layers followed by max-pooling layers, and we use leaky rectified linear units (ReLU) for activations. The kernel spans 5 semitones and 5 time steps, assuming that the whole body movement is not sensitive to detailed note occurrence. Overall, it is thought that in addition to generating expressive body movements, the MIDI note stream constrains the hand positions on the keyboard.

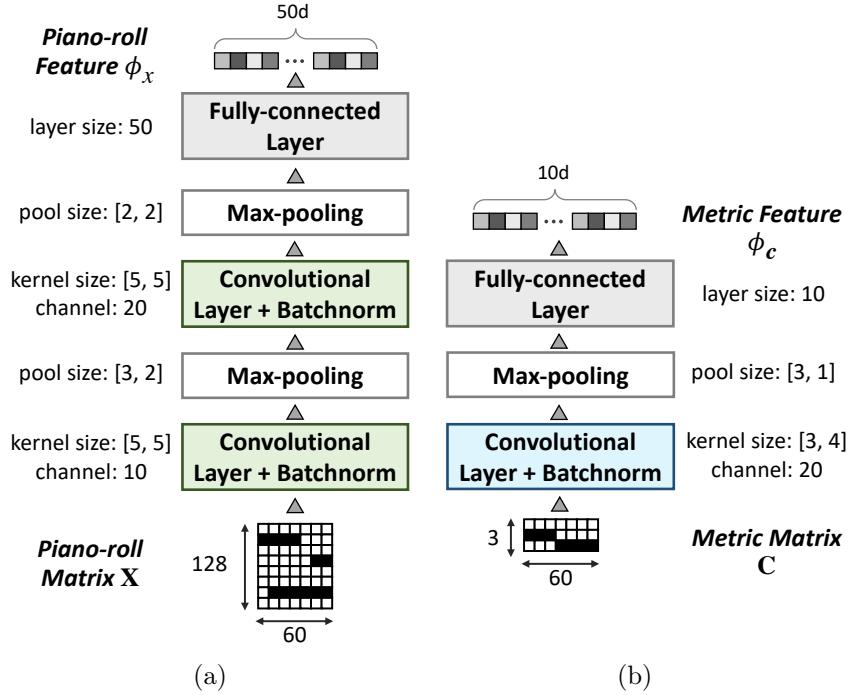


Figure 7.2: The CNN structures and parameters for feature extraction from the (a) MIDI note stream and (b) metric structure information.

7.3.1.2 Metric Structure

Since the body movements are likely to correlate with the musical beats, we also input the metric structure to the proposed system to obtain another feature vector. This metric structure indexes beats within each measure, which is not encoded in the MIDI note stream. The metric structure can be obtained by aligning the live MIDI note stream with the corresponding symbolic music score with explicitly-annotated beat indices and downbeat positions.

Similar to the MIDI note stream feature, we sample them with the same

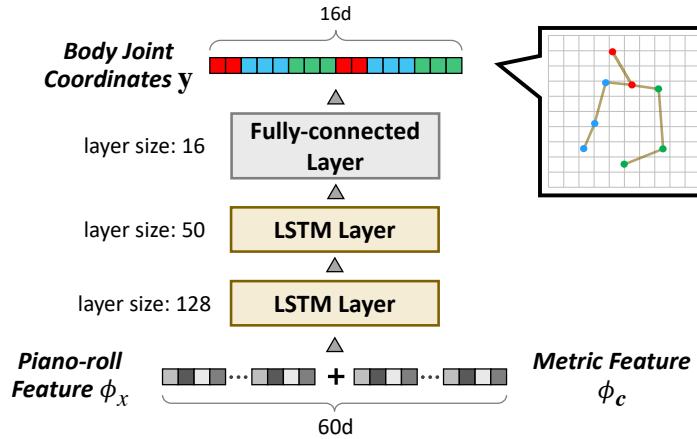


Figure 7.3: The LSTM network structure for body movement generation.

FPS and window length, and, at each frame t , define the metric information as a binary *metric matrix* $\mathbf{C} \in R^{M \times 2\tau}$, with $M = 3$. Here, element (m, n) is a one-hot encoding of the metric information at frame $t - \tau + n - 1$, where the three rows correspond to downbeats, pick-up beats, and other positions, respectively. We then build another CNN to parse the metric matrix \mathbf{C} and obtain a 10-d output vector ϕ_c as the *metric feature*, as illustrated in Figure 7.2.b.

7.3.2 Skeleton Movement Generation by LSTM

To generate the skeleton sequence, we apply the LSTM network, which is capable of preserving the temporal coherence of the output skeleton sequence while learning the pose characteristics associated with the MIDI input. The input to the LSTM is a concatenation of the piano-roll feature ϕ_x and the metric feature ϕ_c , and the output is the normalized coordinates of the body

joints \mathbf{y} . Since musical expression of a human pianist is mainly reflected through upper body movements, we model the x - and y - visual coordinates of K joints in the upper body as $\mathbf{y} = \langle y_1, y_2, \dots, y_{2K} \rangle$, where K is 8 in this work, corresponding to nose, neck, both shoulders, both elbows, and both wrists. The first K indices denote the x -coordinates and the remaining denote the y -coordinates. Note that all the coordinate data in \mathbf{y} , for each piece, are shifted such that the average centroid is at the origin, and scaled isotropically such that the average variance along x - and y -axis sums to 1. The network structure is illustrated in Figure 7.3. It has two LSTM layers, and the output layer is fully-connected to get the 16-d vector approximating \mathbf{y} for the current frame. The output skeleton coordinates are temporally smoothed using a 5-frame moving window. We denote the predicted body joint coordinates, given \mathbf{X} , \mathbf{C} and network parameters θ , as $\hat{\mathbf{y}}(\mathbf{X}, \mathbf{C}|\theta)$.

Since the LSTM is unidirectional, the system is capable of generating motion data in an online manner, with a latency of 30 frames (i.e., 1 second). However, feeding the pre-existing reference music score (after aligned to the live MIDI note stream online) to the system enables an anticipation mechanism like human musicians, which makes it applicable in real-time scenarios without the delay.

7.3.3 Training Condition

To train the model, we minimize, over θ , the sum of a loss function $J(\mathbf{y}, \mathbf{C}, \mathbf{X}, \theta)$ evaluated over the entire training dataset. The loss function expresses a mea-

sure of discrepancy between the predicted body joint coordinates $\hat{\mathbf{y}}$ and the ground-truth coordinates \mathbf{y} .

We use different loss functions during the course of training. In the first 30 epochs, we simply minimize the Manhattan distance between the estimated and the ground-truth body joint coordinates with weight decay:

$$J(\mathbf{y}, \mathbf{C}, \mathbf{X}, \theta) = \sum_k |\hat{y}_k(\mathbf{X}, \mathbf{C}|\theta) - y_k| + \beta \|\theta\|^2, \quad (7.1)$$

where k is the index for the body joints and $\beta = 10^{-8}$ is a weight parameter. We call this kind of loss the *body joint constraint* (see Figure 7.4.a). After 30 training epochs, we add another loss to ensure that not only the coordinates are correct but also consistent with the expected limb lengths:

$$\begin{aligned} J(\mathbf{y}, \mathbf{C}, \mathbf{X}, \theta) &= \sum_k |\hat{y}_k(\mathbf{X}, \mathbf{C}|\theta) - y_k| \\ &\quad + \sum_{(i,j) \in E} |\hat{z}_{ij}(\mathbf{X}, \mathbf{C}|\theta) - z_{ij}| + \beta \|\theta\|^2, \end{aligned} \quad (7.2)$$

where $z_{ij} = (y_i - y_j) + (y_{K+i} - y_{K+j})$ is the displacement between two joints i and j on a limb (e.g., elbow-wrist), $E = \{(i, j)\}$ is the set of possible limb connections (i, j) of a human body. We call the added term the *body limb constraint* (see Figure 7.4.b). This is similar to the geometric constraint as described in (Ning, Zhang and He, 2017). There are 7 limb connections in total, given the 8 upper body joints. We then train another 120 epochs using the limb constraint. We use the Adam (Kingma and Ba, 2015) optimizer,

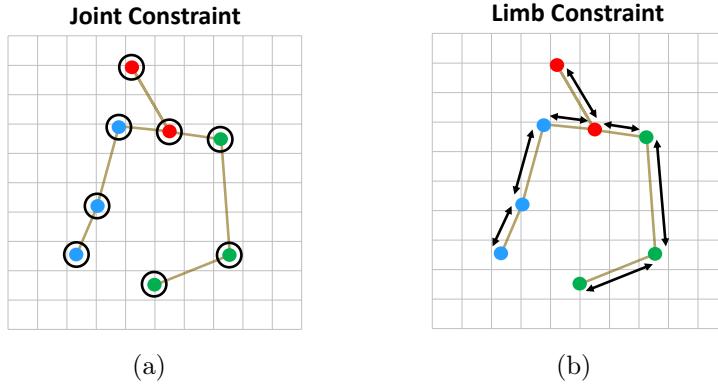


Figure 7.4: The two constraints applied during training.

which is a stochastic gradient descent method, to minimize the loss function.

Here we propose to combine the two kinds of constraints in our training epochs. The body limb constraints are important because the loss of joint positions are minimized *independently of each other* in the body joint constraint. Figure 7.5 demonstrates several generated skeleton samples on the normalized plane, where the limb constraint is not applied in the following 120 epochs. Limb constraint adds dependencies between the loss among different joints, encouraging the model to learn a natural movement that considers the consistency of limb lengths. We only use this constraint at later epochs, however, because the body joint constraint is an easier optimization problem; if we optimize with body limb constraints from the very beginning, the training sometimes fails and remains a state of what seems a local optima, perhaps because the loss function wants to minimize the body joint errors but the gradient must pass through regions where the limb constraint increases. In this case, the arrangements of the body joints tend to

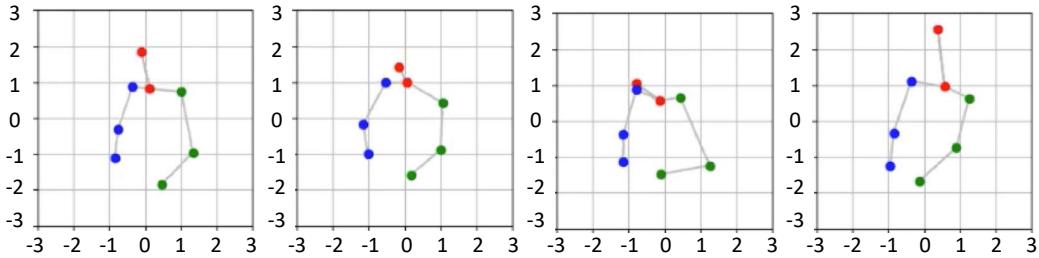


Figure 7.5: Several generated unnatural skeleton samples without the limb constraint.

be arbitrary and not ergonomically reasonable.

7.4 Experiments

We perform objective evaluations to measure the accuracy of the generated movements, and subjective evaluations to rate their expressiveness and naturalness.

7.4.1 Dataset

As there is no existing dataset for the proposed task, we recorded a new audio-visual piano performance dataset with synchronized MIDI stream information on a MIDI keyboard. The dataset contains a total of 74 performance recordings (3 hours and 8 minutes) of 16 different tracks (8 piano duets) played by two pianists, one male and one female. The two players were respectively assigned the primo and the secondo parts of 8 piano duets. Each player then played the 8 tracks multiple times (1-7 times) to render

different expressive styles, e.g., normal, exaggerated, etc. At each time the primo and secondo are recorded together to ensure enough visual expressiveness on the players for interactions. The key depression information (pitch, timing, and velocity) is automatically encoded into the MIDI format by the MIDI keyboard. For each recording, the quantized beat number and the downbeat positions were annotated by semi-automatically aligning the MIDI stream and the corresponding MIDI score data. The camera was placed on the left-front side of the player and the perspective was fixed throughout all of the performances. The video frame rate was 30 FPS. The 2D skeleton coordinates were extracted from the video using a method based on OpenPose (Cao et al., 2017). The video stream and the MIDI stream of each recording were manually time-shifted to align with the key depression actions. Note that we extract the 2D body skeleton data purely from computer vision techniques instead of capturing 3D data using motion sensors, which makes it possible to use the massive online video recordings of great pianists (e.g., Lang Lang) to train the system.

7.4.2 Objective Evaluations

We conduct two experiments to assess our method. Since there is no similar previous work to model the players' whole body pose from MIDI input, we set different experimental conditions for the proposed model as baselines and compare them. First, we investigate the effect of incorporating the metric structure information, which is likely to be relevant for expressive motion

generation but does not directly affect the players' key depression actions on the keyboard. Second, we compare the performance of the network when training on a specific player versus training on multiple players. To numerically evaluate the quality of the system output, we use the mean absolute error (MAE) between the generated and the ground-truth skeleton coordinates at each frame.

7.4.2.1 Effectiveness of the Metric Structure

The system takes as the inputs the MIDI note stream and the metric information. Here we investigate if the latter one can help in the motion generation process, by setting a baseline system that takes the MIDI note stream as the input, ignoring the metric structure by fixing ϕ_c to 0. We evaluate the MAE of the two models, using piece-wise leave-one-out testing over all the 16 tracks.

Results show that adding the metric structure information into the network can decrease the MAE from **0.180** to **0.173**. The unit is in the scale of the normalized plane, where the length of an arm-wrist limb is around 1.2 (see Figure 7.5). The result is significant because it not only demonstrates that our proposed method can effectively model the metric structure, but also that features that are not indirectly related to physical placement of the hand *does* have an effect on expressive body movements. Although our dataset for evaluation is small, we argue that overfit should not exist since the pieces are quite different.

On the other hand, we also observe that even without the metric structure information, the system output is still reasonable by learning the music context from the MIDI note stream. This setting broadens the use scenarios of the proposed system, such as when the MIDI note stream is from an improvised performance without corresponding metric structure information. Nevertheless, including a reference music score is beneficial for the system not only because it improves the MAE measure, but it also enables an anticipation mechanism to favor real-time generation without potential delays.

7.4.2.2 Training on A Specific Player

In this experiment, we evaluate the model’s performance when fixing the same player for training and testing. Now the experiments are carried out on the two players separately. We first divide the dataset into two subsets, each obtaining the 8 different tracks performed by the two players respectively. On each subset we use the leave-one-out testing for the 8 tracks and calculate the MAE between the generated and ground-truth coordinates of body skeletons. The average of the MAE from the two subsets is **0.170**. Comparing the MAE of 0.173 in Section 7.4.2.1 and the MAE of 0.170 in this experiment, we see that training on a generic model only on a target player is slightly better than training over different players. This slight improvement may not be statistically significant. The marginal difference also suggests that even when trained on multiple players as in Section 7.4.2.1, the system is capable of remembering the motion characteristic of each player.



Figure 7.6: One sample frame of the assembled video for subjective evaluation.

7.4.3 Subjective Evaluation

Although the objective evaluation using MAE reflects the system's capability of reproducing the players' body movements on a new MIDI performance stream, this measure is still limited. There can be multiple creative ways on body motions to expressively interpret the same music, and the ground-truth body motion is just one possibility. In addition, from MAE we cannot infer the naturalness of the generated body movements, which is even more important than simply learning to reproduce the motion. In this section, we conduct subjective tests to evaluate the quality of the generated body movements, addressing both expressiveness and naturalness. The strategy is to mix the ground-truth body movements with the generated ones and let the testers to tell if each sample is real (ground-truth from human) or fake (generated).

7.4.3.1 Arrangements

In the subjective evaluation, we mix the two players together and cross-validate on the 16 tracks, as in Section 7.4.2.1. Here we do not add the metric structure input because positive feedbacks on the generation results purely from the keyboard actions will promise broader use cases of the system, i.e., improvised performance without a reference music score.

From the generated skeleton coordinates, we recover them to the original pixel positions on real video frames using the same scaling factor when normalizing the ground-truth skeleton before training. Then we generate an animation showing body joints as circles and limb connections as straight lines on the background environment image taken by the camera from the same perspective. In the same generated video, we also render a dynamic piano-roll that covers a rolling 5-second segment around the current time frame together with the synthesized audio. For a fair comparison, instead of using the original video recordings of real human performances, we generate human body skeletons by repeating the same process using the ground-truth skeletal data. Figure 7.6 shows one sample frame of the assembled video as a visualization.

We arrange 16 pairs of the generated and ground-truth skeleton motions on all the 16 tracks, and randomly crop a 10-second excerpt from each one (excluding several chunks containing long silence parts or page turning motions). This results in 32 video excerpts. We shuffle the 32 excerpts before

showing them to subjects for evaluation.

We recruit 18 subjects from Yamaha employees, who are in their 20's to 50's, all with rich experience in musical acoustics or music audio signal processing. 17 subjects have instrument performance experiences (15 on keyboard instruments). This guarantees that most of them have a general knowledge of how a human pianist performance may look like based on a given MIDI stream, considering different factors such as hand positions on the keyboard according to pitch height, dominant motions for leading onsets, etc. Based on expressiveness and naturalness they rated the videos on a 5-point scale: absolutely generated (1), probably generated (2), unsure (3), probably human (4), and absolutely human (5).

7.4.3.2 Results

Figure 7.7 shows the average subjective ratings as bar plots and their standard deviations as whiskers. A Wilcoxon signed rank test on each piece shows that no significant difference is found in 12 out of the 16 pairs ($p = 0.05$). This suggests that for 3/4 of the observation videos, the generated body movements achieve the same level of expressiveness and naturalness as the real human videos.

In Figure 7.7, the pieces with significant differences in the subjective ratings between generated and real human videos are marked with “*”. On the 1st piece, we observe an especially significant difference. Further investigation reveals that this piece is in a fast tempo (130 BPM), where the eighth notes

are alternatively played by the right and left hand with an agile motion, as shown in Figure 7.8.a. The generated performance lacks this kind of dexterity. In addition, the physical body motions from the human players are distinct and exaggerated around the phrase boundaries, but the generated ones tend to create more conservative motions. Figure 7.8.b gives an example, where in the real human’s performance the head moves forward extensively on the leading bass note (marked in red), whereas the generated one does not. Another observed drawback is the improper wrist positioning of a resting hand; a random position is often predicted in these cases. This is because the left/right hand information is not encoded in the MIDI file, and when only one hand is used, the system does not know which hand to use and how to position the other hand. Generally speaking, the generated movements that are rated significantly lower than real human movements tend to be somewhat dull, which might provide the subjects a cue to discriminate between human and generated movements. We present all of the generated videos online¹.

7.5 Conclusions

In this chapter, we proposed a system for generating a skeleton sequence that corresponds to an input MIDI note stream. Thanks to data-driven learning between the MIDI note stream and the skeleton, the system is capable of generating natural playing motions like a human player with no explicit

¹<http://www.ece.rochester.edu/projects/air/projects/skeletonpianist.html>

constraints on the physique or fingering, reflecting musical expressions, and attuning the generated motion to a particular performer.

For future work, we will apply more music contextual features to generate richer skeleton movements, and extend our method to the generation of 3D joint coordinates. Generating textured characters based on these skeletons is another future direction.

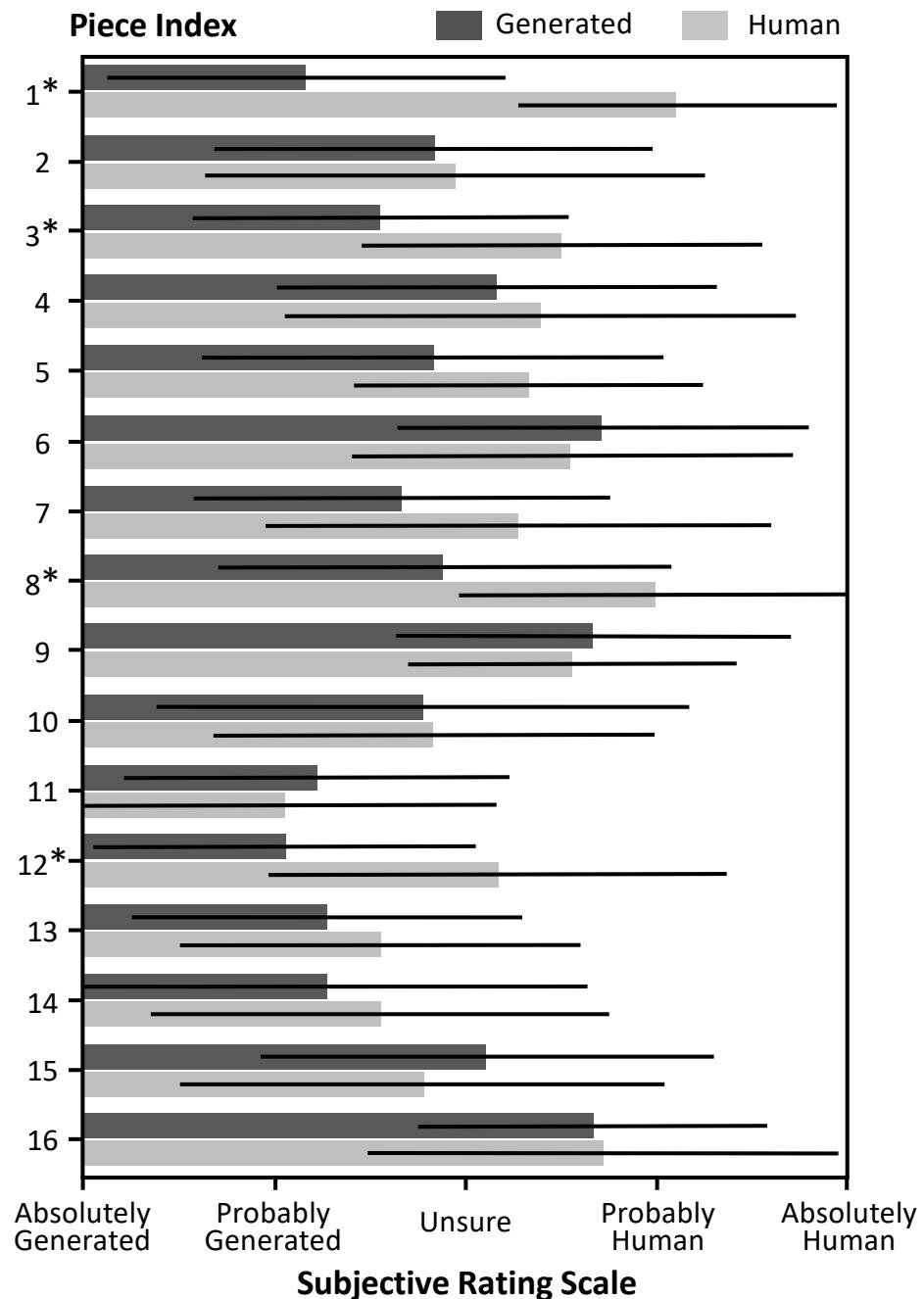
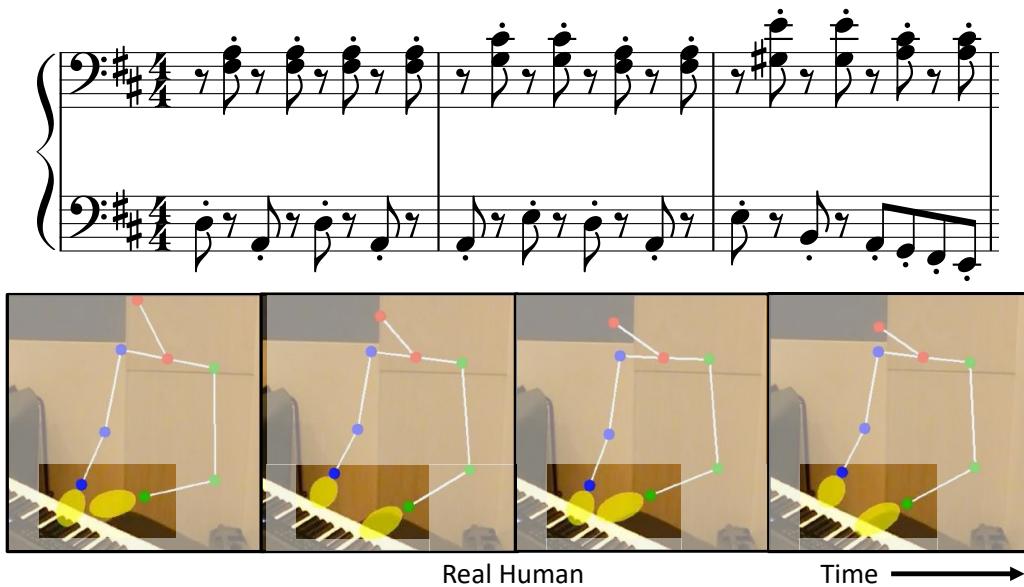
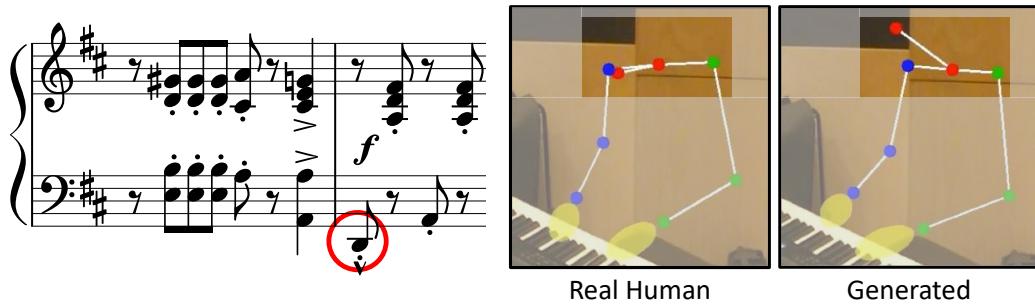


Figure 7.7: Subjective evaluation on expressiveness and naturalness of the generated and human skeleton performance videos. The tracks with significant different ratings are marked with “*”.



(a) The agile fashion in left-right hand alternative playing is not learned.



(b) The exaggerated head nodding on the leading bass note (in red mark) is not learned.

Figure 7.8: The two typical failure cases.

Chapter 8

Conclusion

In this dissertation, I proposed the task of multi-modal analysis for music performance. In Chapters 2, 3 and 4, I introduced the URMP dataset and addressed the coordination of different modalities, e.g., the temporal alignment between auditory signal and symbolic representation of music, and the spatial association between different players in a video recording of ensemble performance and the corresponding audio/score tracks. In Chapters 5, 6, and 7, I conducted research to demonstrate how multi-modal analysis benefits traditional MIR topics, e.g., multi-pitch analysis, performance expressiveness analysis, and source separation, and opens new frontier of emerging MIR topics, e.g., body movement generation from symbolic music.

Multi-modal analysis of music performance is an extremely rich area of research. My preliminary research activities have just opened a crack of the door to this area. For future work, it will be interesting to explore a wider variety of topics. Take an example of cross-modal generation, a system that could interactively generate music from other visual modality

(e.g., facial expressions, body movements, scene transitions) would have large impact on video sharing products, e.g., TikTok. Other scenarios include music generation from plain lyrics, and visual performance generation for instrumentalists and dancers, etc.

One trend in multi-modal analysis nowadays is the use of deep learning techniques. It greatly advances the state of the art for most MIR tasks, but the performances of these data-driven approaches usually depend on the training data and are not ideal for generalization purpose. For example, an audiovisual singing separation system trained on vocals and human faces to separate the solo vocal signals cannot generalize well on separating a solo violin signal from a violin-piano duet. However, traditional statistics-based methods are more robust at generalized scenarios. Music performances, considering all instruments and genres, contain really rich visual scenes from orchestra concerts to commercial music videos. Proposing a multi-modal system with promising performances in its domain as well as generalization capability would be a great contribution.

Last but not the least, with the rapid progress of deep learning applications in computer vision, there seems to be a trend of borrowing successful models developed for computer vision to address computer audition and MIR problems. For example, the idea of DenseNet proposed in 2017 to originally address object recognition tasks has been further adapted to music source separation and won the Champion of SiSEC2018. One important reason might be due to the huge discrepancy on the sizes of these two communi-

ties. Nevertheless, I really would like to witness more techniques originally proposed to address MIR problems also contribute to other areas in the near future.

Bibliography

- Abeßer, Jakob, Estefanía Cano, Klaus Frieler, Martin Pfleiderer and Wolf-Georg Zaddach. 2015. Score-informed analysis of intonation and pitch modulation in jazz solos. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 823–829.
- Abeßer, Jakob, Klaus Frieler, Estefanía Cano, Martin Pfleiderer and Wolf-Georg Zaddach. 2017. “Score-informed analysis of tuning, intonation, pitch modulation, and dynamics in jazz solos.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25(1):168–177.
- Abeßer, Jakob, Olivier Lartillot, Christian Dittmar, Tuomas Eerola and Gerald Schuller. 2011. Modeling musical attributes to characterize ensemble recordings using rhythmic audio features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 189–192.
- Abu-El-Haija, Sami, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan and Sudheendra Vijayanarasimhan. 2016. “YouTube-8M: A large-scale video classification benchmark.” *arXiv preprint arXiv:1609.08675* .
- Acar, Esra, Frank Hopfgartner and Sahin Albayrak. 2014. Understanding affective content of music videos through learned representations. In *Proceedings of the International Conference on Multimedia Modeling*. Springer pp. 303–314.
- Afouras, Triantafyllos, Joon Son Chung and Andrew Zisserman. 2018. The conversation: Deep audio-visual speech enhancement. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*.

- Akbari, Mohammad and Howard Cheng. 2015. “Real-time piano music transcription based on computer vision.” *IEEE Transactions on Multimedia* 17(12):2113–2121.
- Arandjelović, Relja and Andrew Zisserman. 2018. Objects that sound. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 1 pp. 451–466. DOI: https://doi.org/10.1007/978-3-030-01246-5_27.
- Arora, Vipul and Laxmidhar Behera. 2015. “Multiple F0 estimation and source clustering of polyphonic music audio using PLCA and HMRFs.” *IEEE Transactions on Audio, Speech, and Language Processing* 23(2):278–287.
- Arzt, Andreas and Gerhard Widmer. 2010. Simple tempo models for real-time music tracking. In *Proceedings of the Sound and Music Computing Conference (SMC)*.
- Arzt, Andreas, Gerhard Widmer and Simon Dixon. 2008. Automatic page turning for musicians via real-time machine listening. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*. pp. 241–245.
- Arzt, Andreas, Gerhard Widmer and Simon Dixon. 2012. Adaptive distance normalization for real-time music tracking. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*.
- Askenfelt, Anders. 1989. “Measurement of the bowing parameters in violin playing. II: Bow–bridge distance, dynamic range, and limits of bow force.” *The Journal of the Acoustical Society of America* 86(2):503–516. DOI: <https://doi.org/10.1121/1.398230>.
- Baker, Simon, Daniel Scharstein, JP Lewis, Stefan Roth, Michael J Black and Richard Szeliski. 2011. “A database and evaluation methodology for optical flow.” *International Journal of Computer Vision* 92(1):1–31. DOI: <https://doi.org/10.1007/s11263-010-0390-2>.
- Barbancho, Isabel, Cristina de la Bandera, Ana M Barbancho and Lorenzo J Tardon. 2009. Transcription and expressiveness detection system for violin music. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 189–192.

- Barzelay, Zohar and Yoav Y Schechner. 2007. Harmony in motion. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 1–8. DOI: <https://doi.org/10.1109/CVPR.2007.383344>.
- Barzelay, Zohar and Yoav Y Schechner. 2010. “Onsets coincidence for cross-modal analysis.” *IEEE Transactions on Multimedia* 12(2):108–120. DOI: <https://doi.org/10.1109/TMM.2009.2037387>.
- Bay, Mert, Andreas F Ehmann and J Stephen Downie. 2009. Evaluation of multiple-F0 estimation and tracking systems. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 315–320.
- Bay, Mert, Andreas F Ehmann, James W Beauchamp, Paris Smaragdis and J Stephen Downie. 2012. Second fiddle is important too: pitch tracking individual voices in polyphonic music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 319–324.
- Bazzica, Alessio, Cynthia CS Liem and Alan Hanjalic. 2014. Exploiting instrument-wise playing/non-playing labels for score synchronization of symphonic music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 201–206.
- Bazzica, Alessio, Cynthia CS Liem and Alan Hanjalic. 2016. “On detecting the playing/non-playing activity of musicians in symphonic music videos.” *Computer Vision and Image Understanding* 144:188–204.
- Bazzica, Alessio, JC van Gemert, Cynthia CS Liem and Alan Hanjalic. 2017. “Vision-based detection of acoustic timed events: a case study on clarinet note onsets.” *arXiv preprint arXiv:1706.09556* .
- Bello, Juan Pablo, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies and Mark B. Sandler. 2005. “A tutorial on onset detection in music signals.” *IEEE Transactions on Speech and Audio Processing* 13(5):1035–1047. DOI: <https://doi.org/10.1109/TSA.2005.851998>.
- Benetos, Emmanouil, Anssi Klapuri and Simon Dixon. 2012. Score-informed transcription for automatic piano tutoring. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*. pp. 2153–2157.
- Berenzweig, Adam, Daniel PW Ellis and Steve Lawrence. 2002. Using voice segments to improve artist classification of music. In *Proceedings of the*

- AES 22nd International Conference: Virtual Synthetic and Entertainment Audio.*
- Bittner, Rachel M, Justin Salamon, Mike Tierney, Matthias Mauch, Chris Cannam and Juan Pablo Bello. 2014. MedleyDB: A multitrack dataset for annotation-intensive MIR Research. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 155–160.
- Black, Michael. J. and Padmanabhan Anandan. 1993. A framework for the robust estimation of optical flow. In *Proceedings of the International Conference on Computer Vision (ICCV)*. pp. 231–236.
- Böck, Sebastian, Florian Krebs and Markus Schedl. 2012. Evaluating the online capabilities of onset detection methods. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Bonada, Jordi, Robert Lachlan and Merlijn Blaauw. 2016. Bird song synthesis based on hidden markov models. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*.
- Bregler, Christoph, Michele Covell and Malcolm Slaney. 1997. Video rewrite: Driving visual speech with audio. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques*. pp. 353–360.
- Bresin, Roberto and Giovanni Umberto Battel. 2000. “Articulation strategies in expressive piano performance analysis of legato, staccato, and repeated notes in performances of the andante movement of Mozart’s Sonata in G major (K 545).” *Journal of New Music Research* 29(3):211–224.
- Burkholder, J Peter and Donald Jay Grout. 2014. *A history of Western music: Ninth international student edition*. WW Norton & Company, Inc.
- Burns, Anne-Marie and Marcelo M Wanderley. 2006. Visual methods for the retrieval of guitarist fingering. In *Proceedings of the International Conference on New Interfaces for Musical Expression (NIME)*. pp. 196–199.
- Cao, Zhe, Tomas Simon, Shih-En Wei and Yaser Sheikh. 2017. Realtime multi-person 2D pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Vol. 1 pp. 7291–7299. DOI: <https://doi.org/10.1109/CVPR.2017.143>.

- Carabias-Orti, Julio J., Francisco J. Rodriguez-Serrano, Pedro Vera-Candeas, Nicolás Ruiz-Reyes and Francisco J. Canadas-Quesada. 2015. An audio to score alignment framework using spectral factorization and dynamic time warping. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Caramiaux, Baptiste, Marcelo M Wanderley and Frédéric Bevilacqua. 2012. “Segmenting and parsing instrumentalists’ gestures.” *Journal of New Music Research* 41(1):13–29.
- Cartwright, Mark, Bryan Pardo and Josh Reiss. 2014. Mixploration: Rethinking the audio mixer interface. In *Proceedings of the International Conference on Intelligent User Interfaces (IUI)*. pp. 365–370.
- Casanovas, A. Llagostera, Gianluca Monaci, Pierre Vandergheynst and Rémi Gribonval. 2010. “Blind audiovisual source separation based on sparse redundant representations.” *IEEE Transactions on Multimedia* 12(5):358–371. DOI: <https://doi.org/10.1109/TMM.2010.2050650>.
- Casanovas, A. Llagostera and Pierre Vandergheynst. 2010. “Nonlinear video diffusion based on audio-video synchrony.”. Unpublished. <https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.187.4688>.
- Chan, Tak-Shing, Tzu-Chun Yeh, Zhe-Cheng Fan, Hung-Wei Chen, Li Su, Yi-Hsuan Yang and Roger Jang. 2015. Vocal activity informed singing voice separation with the iKala dataset. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 718–722.
- Chandna, Pritish, Marius Miron, Jordi Janer and Emilia Gómez. 2017. Monoaural audio source separation using deep convolutional neural networks. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation*. Springer pp. 258–266.
- Chen, Lele, Sudhanshu Srivastava, Zhiyao Duan and Chenliang Xu. 2017. Deep cross-modal audio-visual generation. In *Proceedings of the ACM Thematic Workshops of Multimedia*. pp. 349–357.
- Chung, Joon Son and Andrew Zisserman. 2016. Lip reading in the wild. In *Proceedings of the Asian Conference on Computer Vision*. Springer pp. 87–103.

- Cont, Arshia. 2006. Realtime audio to score alignment for polyphonic music instruments, using sparse non-negative constraints and hierarchical HMMs. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Cutler, Ross and Larry Davis. 2000. Look who's talking: Speaker detection using video and audio correlation. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. Vol. 3 pp. 1589–1592. DOI: <https://doi.org/10.1109/ICME.2000.871073>.
- Dahl, Sofia. 2004. “Playing the accent – comparing striking velocity and timing in an ostinato rhythm performed by four drummers.” *Acta Acustica united with Acustica* 90(4):762–776.
- Dahl, Sofia and Anders Friberg. 2007. “Visual perception of expressiveness in musicians’ body movements.” *Music Perception: An Interdisciplinary Journal* 24(5):433–454.
- Davidson, Jane W. 1993. “Visual perception of performance manner in the movements of solo musicians.” *Psychology of Music* 21(2):103–113.
- Davy, Manuel, Simon Godsill and Jerome Idier. 2006. “Bayesian analysis of polyphonic Western tonal music.” *The Journal of the Acoustical Society of America* 119(4):2498–2517.
- Dinesh, Karthik, Bochen Li, Xinzha Liu, Zhiyao Duan and Gaurav Sharma. 2017. Visually informed multi-pitch analysis of string ensembles. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. pp. 3021–3025. DOI: <https://doi.org/10.1109/ICASSP.2017.7952711>.
- Dixon, Simon. 2000. On the computer recognition of solo piano music. In *Proceedings of the Australasian Computer Music Conference*. pp. 31–37.
- Dixon, Simon. 2005. Live tracking of musical performances using on-line time warping. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*. pp. 92–97.
- Dixon, Simon. 2006. Onset detection revisited. In *Proceedings of the International Conference on Digital Audio Effects (DAFx)*.

- Dixon, Simon and Gerhard Widmer. 2005. MATCH: A music alignment tool chest. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Driedger, Jonathan, Stefan Balke, Sebastian Ewert and Meinard Müller. 2016. Template-based Vibrato analysis of music signals. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Duan, Zhiyao and Bryan Pardo. 2011a. “Soundprism: An online system for score-informed source separation of music audio.” *IEEE Journal of Selected Topics in Signal Processing* 5(6):1205–1215. DOI: <https://doi.org/10.1109/JSTSP.2011.2159701>.
- Duan, Zhiyao and Bryan Pardo. 2011b. A state space model for online polyphonic audio-score alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 197–200. DOI: <https://doi.org/10.1109/ICASSP.2011.5946374>.
- Duan, Zhiyao, Bryan Pardo and Changshui Zhang. 2010. “Multiple fundamental frequency estimation by modeling spectral peaks and non-peak regions.” *IEEE Transactions on Audio, Speech, and Language Processing* 18(8):2121–2133. DOI: <https://doi.org/10.1109/TASL.2010.2042119>.
- Duan, Zhiyao, Jinyu Han and Bryan Pardo. 2014. “Multi-pitch streaming of harmonic sound mixtures.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 22(1):138–150.
- Duan, Zhiyao, Slim Essid, Cynthia Liem, Gael Richard and Gaurav Sharma. 2019. “Audiovisual Analysis of Music Performances: Overview of an Emerging Field.” *IEEE Signal Processing Magazine* 36(1):63–73. DOI: <https://doi.org/10.1109/MSP.2018.2875511>.
- Ellis, Daniel PW and Graham E. Poliner. 2007. Identifying ‘cover songs’ with chroma features and dynamic programming beat tracking. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Emiya, Valentin, Roland Badeau and Bertrand David. 2010. “Multipitch estimation of piano sounds using a new probabilistic spectral smoothness principle.” *IEEE Transactions on Audio, Speech, and Language Processing* 18(6):1643–1654.

- Ephrat, Ariel, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman and Michael Rubinstein. 2018. “Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation.” *ACM Transactions on Graphics (TOG)* 37(4). DOI: <https://doi.org/10.1145/3197517.3201357>.
- Eskimez, Sefik Emre, Ross K Maddox, Chenliang Xu and Zhiyao Duan. 2018. Generating talking face landmarks from speech. In *Proceedings of the International Conference on Latent Variable Analysis and Signal Separation (LVA-ICA)*.
- Essid, Slim and Gaël Richard. 2012. Fusion of multimodal information in music content analysis. In *Multimodal Music Processing*. Vol. 3 of *Dagstuhl Follow-Ups* pp. 37–52.
- Ewert, Sebastian, Meinard Müller and Peter Grosche. 2009. High resolution audio synchronization using chroma onset features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Eyben, Florian, Sebastian Böck, Björn Schuller and Alex Graves. 2010. Universal onset detection with bidirectional long short-term memory neural networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Fisher, John W and Trevor Darrell. 2004. “Speaker association with signal-level audiovisual fusion.” *IEEE Transactions on Multimedia* 6(3):406–413. DOI: <https://doi.org/10.1109/TMM.2004.827503>.
- Fletcher, Harvey and Larry C. Sanders. 1967. “Quality of violin vibrato tones.” *The Journal of the Acoustical Society of America* 41(6):1534–1544.
- Friberg, Anders, Erwin Schoonderwaldt and Patrik N. Juslin. 2007. “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.
- Fritsch, Joerg and Mark D. Plumbley. 2013. Score informed audio source separation using constrained nonnegative matrix factorization and score synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 888–891.

- Fujihara, Hiromasa and Masataka Goto. 2007. A music information retrieval system based on singing voice timbre. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 467–470.
- Fujihara, Hiromasa, Masataka Goto, Jun Ogata, Kazunori Komatani, Tetsuya Ogata and Hiroshi G Okuno. 2006. Automatic synchronization between lyrics and music CD recordings based on Viterbi alignment of segregated vocal signals. In *Proceedings of the IEEE International Symposium on Multimedia (ISM)*. pp. 257–264.
- Fujishima, Takuya. 1999. Realtime chord recognition of musical sound: A system using common lisp music. In *Proceedings of the International Computer Music Conference (ICMC)*. pp. 464–467.
- Gan, Chuang, Deng Huang, Hang Zhao, Joshua B Tenenbaum and Antonio Torralba. 2020. Music gesture for visual sound separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 10478–10487.
- Gao, Ruohan, Rogerio Feris and Kristen Grauman. 2018. Learning to separate object sounds by watching unlabeled video. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 3 pp. 36–54. DOI: https://doi.org/10.1007/978-3-030-01219-9_3.
- Gemmeke, Jort F., Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal and Marvin Ritter. 2017. Audio Set: An ontology and human-labeled dataset for audio events. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 776–780.
- Geringer, John M., Rebecca B. MacLeod and Michael L. Allen. 2010. “Perceived pitch of violin and cello vibrato tones among music majors.” *Journal of Research in Music Education* 57(4):351–363. DOI: <https://doi.org/10.1177/0022429409350510>.
- Gillet, Olivier and Gaël Richard. 2005. Automatic transcription of drum sequences using audiovisual features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Vol. 3 pp. iii–205.

- Gillet, Olivier and Gaël Richard. 2006. ENST-Drums: An extensive audio-visual database for drum signals processing. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 156–159.
- Gillet, Olivier, Slim Essid and Gaël Richard. 2007. “On the correlation of automatic audio and visual segmentations of music videos.” *IEEE Transactions on Circuits and Systems for Video Technology* 17(3):347–355.
- Godøy, Rolf Inge and Alexander Refsum Jensenius. 2009. Body movement in music information retrieval. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 45–50.
- Goebel, Werner. 2001. “Melody lead in piano performance: Expressive device or artifact?” *The Journal of the Acoustical Society of America* 110(1):563–572.
- Gordon, Stewart. 1996. *A history of keyboard literature: Music for the piano and its forerunners*. Wadsworth Pub Co.
- Gorodnichy, Dmitry and Arjun Yugeswaran. 2006. Detection and tracking of pianist hands and fingers. In *Proceedings of the Canadian Conference on Computer and Robot Vision*. pp. 63–63.
- Goto, Masataka, Hiroki Hashiguchi, Takuichi Nishimura and Ryuichi Oka. 2002. RWC music database: Popular, classical and jazz music databases. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. Vol. 2 pp. 287–288.
- Grubb, Lorin and Roger B. Dannenberg. 1997. A stochastic method of tracking a vocal performer. In *Proceedings of the International Computer Music Conference (ICMC)*. pp. 301–308.
- Gu, Hung-Yan and Zheng-Fu Lin. 2014. “Singing-voice synthesis using ANN vibrato-parameter models.” *Journal of Information Science and Engineering* 30(2):425–442.
- Hargreaves, Steven, Anssi Klapuri and Mark Sandler. 2012. “Structural segmentation of multitrack audio.” *IEEE Transactions on Audio, Speech, and Language Processing* 20(10):2637–2647.

- Hennequin, Romain, Anis Khelif, Felix Voituret and Manuel Moussallam. 2019. Spleeter: A fast and state-of-the-art music source separation tool With pre-trained models.
- Horn, Berthold K. and Brian G. Schunck. 1981. Determining optical flow. In *1981 Technical symposium east*. International Society for Optics and Photonics pp. 319–331.
- Hou, Jen-Cheng, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang and Hsin-Min Wang. 2018. “Audio-visual speech enhancement using multi-modal deep convolutional neural network.” *IEEE Transactions on Emerging Topics in Computational Intelligence* .
- Hsu, Chao-Ling and Jyh-Shing Roger Jang. 2010. Singing pitch extraction by voice vibrato/tremolo estimation and instrument partial deletion. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 525–530.
- Hu, Ke and DeLiang Wang. 2013. “An unsupervised approach to cochannel speech separation.” *IEEE Transactions on Audio, Speech, and Language Processing* 21(1):122–131.
- Hu, Ning, Roger B. Dannenberg and George Tzanetakis. 2003. Polyphonic audio matching and alignment for music retrieval. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*.
- Huang, Gao, Zhuang Liu, Laurens Van Der Maaten and Kilian Q Weinberger. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4700–4708.
- Huang, Po-Sen, Scott Deeann Chen, Paris Smaragdis and Mark Hasegawa-Johnson. 2012. Singing-voice separation from monaural recordings using robust principal component analysis. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 57–60.
- Itohara, Tatsuhiko, Kazuhiro Nakadai, Tetsuya Ogata and Hiroshi G. Okuno. 2012. Improvement of audio-visual score following in robot ensemble with

- human guitarist. In *Proceedings of the IEEE-RAS International Conference on Humanoid Robots*.
- Izadinia, Hamid, Imran Saleemi and Mubarak Shah. 2013. “Multimodal analysis for identification and segmentation of moving-sounding objects.” *IEEE Transactions on Multimedia* 15(2):378–390. DOI: <https://doi.org/10.1109/TMM.2012.2228476>.
- Jansson, Andreas, Eric Humphrey, Nicola Montecchio, Rachel Bittner, Aparna Kumar and Tillman Weyde. 2017. Singing voice separation with deep U-Net convolutional networks. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Järveläinen, Hanna. 2002. Perception-based control of vibrato parameters in string instrument synthesis. In *Proceedings of the International Computer Music Conference (ICMC)*.
- Jiang, Yu-Gang, Zuxuan Wu, Jun Wang, Xiangyang Xue and Shih-Fu Chang. 2018. “Exploiting feature and class relationships in video categorization with regularized deep neural networks.” *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40(2):352–364.
- Joder, Cyril and Bjorn Schuller. 2013. Off-line refinement of audio-to-score alignment by observation template adaptation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Kaprykowsky, Hagen and Xavier Rodet. 2006. Globally optimal short-time dynamic time warping, application to score to audio alignment. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.
- Kerdvibulvech, Chutisant and Hideo Saito. 2007. Vision-based guitarist fin-gering tracking using a Bayesian classifier and particle filters. In *Advances in Image and Video Technology*. Springer pp. 625–638.
- Kidron, Einat, Yoav Y. Schechner and Michael Elad. 2007. “Cross-modal localization via sparsity.” *IEEE Transactions on Signal Processing* 55(4):1390–1404. DOI: <https://doi.org/10.1109/TSP.2006.888095>.

- King, Davis E. 2009. “Dlib–ml: A Machine Learning Toolkit.” *Journal of Machine Learning Research* 10(Jul.):1755–1758.
- Kingma, Diederik P. and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*. pp. 1–5.
- Klapuri, Anssi. 2006. Multiple fundamental frequency estimation by summing harmonic amplitudes. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 216–221.
- Kuhn, Harold W. 1955. “The Hungarian method for the assignment problem.” *Naval Research Logistics (NRL)* 2(1-2):83–97. DOI: <https://doi.org/10.1002/nav.3800020109>.
- Kuo, Fang-Fei, Man-Kwan Shan and Suh-Yin Lee. 2013. Background music recommendation for video based on multimodal latent semantic analysis. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. pp. 1–6.
- Le Roux, Jonathan, Scott Wisdom, Hakan Erdogan and John R Hershey. 2019. SDR – half-baked or well done? In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 626–630.
- Lehtonen, Heidi-Maria, Henri Penttinen, Jukka Rauhala and Vesa Välimäki. 2007. “Analysis and modeling of piano sustain-pedal effects.” *Journal of the Acoustical Society of America* 122(3):1787–1797.
- Li, Bochen, Akira Maezawa and Zhiyao Duan. 2018. Skeleton plays piano: online generation of pianist body movements from MIDI performance. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Li, Bochen, Chenliang Xu and Zhiyao Duan. 2017. Audiovisual source association for string ensembles through multi-modal vibrato analysis. In *Proceedings of the Sound and Music Computing (SMC) Conference*. pp. 159–166.
- Li, Bochen, Karthik Dinesh, Chenliang Xu, Gaurav Sharma and Zhiyan Duan. 2019. “Online audio-visual source association for chamber music

- performances.” *Transactions of the International Society for Music Information Retrieval* 2(1).
- Li, Bochen, Karthik Dinesh, Gaurav Sharma and Zhiyao Duan. 2017. Video-based vibrato detection and analysis for polyphonic string music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 123–130.
- Li, Bochen, Karthik Dinesh, Zhiyao Duan and Gaurav Sharma. 2017. See and listen: score-informed association of sound tracks to players in chamber music performance videos. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 2906–2910. DOI: <https://doi.org/10.1109/ICASSP.2017.7952688>.
- Li, Bochen, Xinzha Liu, Karthik Dinesh, Zhiyao Duan and Gaurav Sharma. 2018. “Data from: “Creating a multi-track classical music performance dataset for multi-modal music analysis: challenges, insights, and applications.”” *Dryad Digital Repository* . <https://doi.org/10.5061/dryad.ng3r749>.
- Li, Bochen, Xinzha Liu, Karthik Dinesh, Zhiyao Duan and Gaurav Sharma. 2019. “Creating a music performance dataset for multimodal music analysis: challenges, insights, and applications.” *IEEE Transactions on Multimedia* 21(2):522–535. DOI: <https://doi.org/10.1109/TMM.2018.2856090>.
- Li, Bochen and Zhiyao Duan. 2015. Score following for piano performances with sustain-pedal effects. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Li, Bochen and Zhiyao Duan. 2016. “An approach to score following for piano performances with the sustained effect.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24(12):2425–2438.
- Li, Bochen, Zhiyao Duan and Gaurav Sharma. 2016. Associating players to sound sources in musical performance videos. In *Late Breaking Demo, International Society for Music Information Retrieval (ISMIR)*.
- Li, Kai, Jun Ye and Kien A Hua. 2014. What’s making that sound? In *Proceedings of the ACM International Conference on Multimedia*. pp. 147–156. DOI: <https://doi.org/10.1145/2647868.2654936>.

- Li, Zimo, Yi Zhou, Shuangjiu Xiao, Chong He and Hao Li. 2018. Auto-conditioned recurrent networks for extended complex human motion synthesis. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Liu, Jen-Yu and Yi-Hsuan Yang. 2018. Denoising auto-encoder with recurrent skip connections and residual regression for music source separation. In *Proceedings of the IEEE International Conference on Machine Learning and Applications (ICMLA)*. pp. 773–778.
- Liu, Yuyu and Yoichi Sato. 2008. Finding speaker face region by audiovisual correlation. In *Proceedings of the Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2)*.
- Lluis, Francesc, Jordi Pons and Xavier Serra. 2019. End-to-end music source separation: is it possible in the waveform domain? In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*.
- Lu, Rui, Zhiyao Duan and Changshui Zhang. 2018. “Listen and look: audio-visual matching assisted speech source separation.” *IEEE Signal Processing Letters* 25(9):1315–1319.
- Lu, Rui, Zhiyao Duan and Changshui Zhang. 2019. “Audio-visual deep clustering for speech separation.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27(11):1697–1712.
- Luo, Yi and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*). pp. 696–700.
- Luo, Yi, Zhuo Chen, John R Hershey, Jonathan Le Roux and Nima Mesgarani. 2017. Deep clustering and conventional networks for music separation: Stronger together. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. pp. 61–65.
- Marchini, Marco, Rafael Ramirez, Panos Papiotis and Esteban Maestre. 2014. “The sense of ensemble: a machine learning approach to expressive performance modelling in string quartets.” *Journal of New Music Research* 43(3):303–317.

- Mauch, Matthias, Chris Cannam, Rachel Bittner, George Fazekas, Justin Salamon, Jaijie Dai, Juan Bello and Simon Dixon. 2015. Computer-aided melody note transcription using the Tony software: accuracy and efficiency. In *Proceedings of the International Conference on Technologies for Music Notation and Representation*. pp. 23–31.
- Mauch, Matthias and Simon Dixon. 2014. pYIN: A fundamental frequency estimator using probabilistic threshold distributions. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. pp. 659–663.
- Meyer, Jürgen. 2009. *Acoustics and the performance of music: manual for acousticians, audio engineers, musicians, architects and musical instrument makers*. Springer Science & Business Media.
- Mick, James Paul. 2012. “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.
- Miron, Marius, Julio J Carabias-Orti, Juan J Bosch, Emilia Gómez and Jordi Janer. 2016. “Score-informed source separation for multichannel orchestral recordings.” *Journal of Electrical and Computer Engineering* .
- Miron, Marius, Julio José Carabias-Orti and Jordi Janer. 2014. Audio-to-score alignment at note level for orchestral recordings. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Montecchio, Nicola and Nicola Orio. 2009. A discrete filter bank approach to audio to score matching for polyphonic music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Müller, Meinard, Henning Mattes and Frank Kurth. 2006. An efficient multiscale approach to audio synchronization. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Müller, Meinard and Sebastian Ewert. 2011. Chroma Toolbox: MATLAB implementations for extracting variants of chroma-based audio features. In *Proceedings of the International Conference for Music Information Retrieval (ISMIR)*.

- Murphy, Declan. 2003. Tracking a conductor's baton. In *Proceedings of the Danish Conference on Pattern Recognition and Image Analysis*.
- Nakano, Tomoyasu, Masataka Goto and Yuzuru Hiraga. 2006. An automatic singing skill evaluation method for unknown melodies using pitch interval accuracy and vibrato features. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*.
- Niedermayer, Bernhard, Sebastian Böck and Gerhard Widmer. 2011. On the importance of "real" audio data for MIR algorithm evaluation at the note-level – a comparative study. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Ning, Guanghan, Zhi Zhang and Zhiqian He. 2017. "Knowledge-Guided Deep Fractal Neural Networks for Human Pose Estimation." *IEEE Transactions on Multimedia* 20(5):1246–1259.
- Obata, Satoshi, Hidehiro Nakahara, Takeshi Hirano and Hiroshi Kinoshita. 2009. Fingering force in violin vibrato. In *Proceedings of the International Symposium on Performance Science*. Vol. 429.
- Oka, Akira and Mime Hashimoto. 2013. Marker-less piano fingering recognition using sequential depth images. In *Proceedings of the Korea-Japan Joint Workshop on Frontiers of Computer Vision (FCV)*. pp. 1–4.
- Orio, Nicola and Diemo Schwarz. 2001. Alignment of monophonic and polyphonic music to a score. In *Proceedings of the International Computer Music Conference (ICMC)*.
- Orio, Nicola and François Déchelle. 2001. Score following using spectral analysis and hidden Markov models. In *Proceedings of the International Computer Music Conference (ICMC)*.
- Owens, Andrew and Alexei A Efros. 2018. Audio-visual scene analysis with self-supervised multisensory features. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 6 pp. 639–658. DOI: https://doi.org/10.1007/978-3-030-01231-1_39.
- Ozerov, Alexey, Pierrick Philippe, Frédéric Bimbot and Rmi Gribonval. 2007. "Adaptation of Bayesian models for single-channel source separation and

- its application to voice/music separation in popular songs.” *IEEE Transactions on Audio, Speech, and Language Processing* 15(5):1564–1578.
- Paleari, Marco, Benoit Huet, Antony Schutz and Dirk Slock. 2008. A multimodal approach to music transcription. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*. pp. 93–96. DOI: <https://doi.org/10.1109/ICIP.2008.4711699>.
- Palmer, Caroline, Christine Carter, Erik Koopmans and Janeen D Loehr. 2007. Movement, planning, and music: Motion coordinates of skilled performance. In *Proceedings of the International Conference on Music Communication Science*. University of New South Wales Sydney, NSW pp. 119–122.
- Papich, George and Edward Rainbow. 1974. “A pilot study of performance practices of twentieth-century musicians.” *Journal of Research in Music Education* 22(1):24–34.
- Pardo, Bryan and William Birmingham. 2005. Modeling form for on-line following of musical performances. In *Proceedings of the National Conference on Artificial Intelligence*.
- Parekh, Sanjeel, Slim Essid, Alexey Ozerov, Ngoc Duong, Patrick Perez and Gaël Richard. 2017. Motion informed audio source separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 6–10. DOI: <https://doi.org/10.1109/ICASSP.2017.7951787>.
- Park, Jeongsoo and Kyogu Lee. 2015. Harmonic-percussive source separation using harmonicity and sparsity constraints. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 148–154.
- Parncutt, Richard and Gary McPherson. 2002. *The science and psychology of music performance: Creative strategies for teaching and learning*. Oxford University Press. DOI: <https://doi.org/10.1177/1321103X020190010803>.
- Pätynen, Jukka, Ville Pulkki and Tapio Lokki. 2008. “Anechoic recording system for symphony orchestra.” *Acta Acustica united with Acustica* 94(6):856–865.

- Perez-Carrillo, Alfonso, Josep-Lluis Arcos and Marcelo Wanderley. 2015. Estimation of guitar fingering and plucking controls based on multimodal analysis of motion, audio and musical score. In *Proceedings of the International Symposium on Computer Music Multidisciplinary Research (CMMR)*. pp. 71–87.
- Perlin, Ken and Athomas Goldberg. 1996. Improv: A system for scripting interactive actors in virtual worlds. In *Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques*. pp. 205–216.
- Petridis, Stavros, Themos Stafylakis, Pingehuan Ma, Feipeng Cai, Georgios Tzimiropoulos and Maja Pantic. 2018. End-to-end audiovisual speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 6548–6552.
- Platz, Friedrich and Reinhard Kopiez. 2012. “When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance.” *Music Perception: An Interdisciplinary J.* 30(1):71–83.
- Poliner, Graham E. and Daniel PW Ellis. 2007. “A discriminative model for polyphonic piano transcription.” *EURASIP Journal on Applied Signal Processing* 2007(1):154–154.
- Poria, Soujanya, Erik Cambria, Newton Howard, Guang-Bin Huang and Amir Hussain. 2016. “Fusing audio, visual and textual clues for sentiment analysis from multimodal content.” *Neurocomputing* 174:50–59.
- Prockup, Matthew, David Grunberg, Alex Hrybyk and Youngmoo E. Kim. 2013. “Orchestral performance companion: Using real-time audio to score alignment.” *IEEE MultiMedia* 20(2):52–60.
- Puckette, Miller. 1995. Score following using the sung voice. In *Proceedings of the International Computer Music Conference (ICMC)*.
- Puckette, Miller and Cort Lippe. 1992. Score following in practice. In *Proceedings of the International Computer Music Conference (ICMC)*.
- Radicioni, Daniele, Luca Anselma and Vincenzo Lombardo. 2004. A segmentation-based prototype to compute string instruments fingering. In *Proceedings of the Conference on Interdisciplinary Musicology (CIM)*. Vol. 17 p. 97.

- Rafii, Zafar and Bryan Pardo. 2011. A simple music/voice separation method based on the extraction of the repeating musical structure. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 221–224.
- Raphael, Christopher. 2010. Music plus one and machine learning. In *Proceedings of the International Conference on Machine Learning (ICML)*. pp. 21–28.
- Repp, Bruno H. 1995. “Acoustics, perception, and production of legato articulation on a digital piano.” *The Journal of the Acoustical Society of America* 97(6):3862–3874.
- Rodet, Xavier and Florent Jaillet. 2001. Detection and modeling of fast attack transients. In *Proceedings of the International Computer Music Conference (ICMC)*.
- Scarr, Joseph and Richard Green. 2010. Retrieval of guitarist fingering information using computer vision. In *Proceedings of the International Conference on Image and Vision Computing New Zealand (IVCNZ)*. pp. 1–7.
- Senocak, Arda, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang and In So Kweon. 2018. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 4358–4366. DOI: <https://doi.org/10.1109/CVPR.2018.00458>.
- Seo, Ju-Hwan, Jeong-Yean Yang, Jaewoo Kim and Dong-Soo Kwon. 2013. Autonomous humanoid robot dance generation system based on real-time music input. In *Proceedings of the IEEE International Conference on Robot and Human Interactive Communication*. pp. 204–209.
- Shimoda, S., M. Hayashi and Yasuaki Kanatsugu. 1989. “New chroma-key imagining technique with Hi-Vision background.” *IEEE Trans. Broadcasting* 35(4):357–361.
- Shlizerman, Eli, Lucio M Dery, Hayden Schoen and Ira Kemelmacher-Shlizerman. 2018. Audio to body dynamics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

- Sigg, Christian, Bernd Fischer, Bjorn Ommer, Volker Roth and Joachim Buhmann. 2007. Nonnegative CCA for audiovisual source separation. In *Proceedings of the IEEE Workshop on Machine Learning for Signal Processing*. pp. 253–258. DOI: <https://doi.org/10.1109/MLSP.2007.4414315>.
- Sörgjerd, Maria. 2000. “Auditory and visual recognition of emotional expression in performance of music.” *PhD Thesis* . Uppsala Universitet, Institutionen för Psykologi.
- Stafylakis, Themos and Georgios Tzimiropoulos. 2017. Combining residual networks with LSTMs for lipreading. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*. pp. 3652–3656.
- Stoller, Daniel, Sebastian Ewert and Simon Dixon. 2018. Wave-U-Net: A multi-scale neural network for end-to-end audio source separation. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. pp. 334–340.
- Stöter, Fabian-Robert, Antoine Liutkus and Nobutaka Ito. 2018. The 2018 signal separation evaluation campaign. In *International Conference on Latent Variable Analysis and Signal Separation*. Springer pp. 293–305.
- Stöter, Fabian-Robert, Stefan Uhlich, Antoine Liutkus and Yuki Mitsufuji. 2019. “Open-Unmix - A Reference Implementation for Music Source Separation.” *Journal of Open Source Software* .
URL: <https://doi.org/10.21105/joss.01667>
- Su, Li, Hsin-Ming Lin and Yi-Hsuan Yang. 2014. “Sparse modeling of magnitude and phase-derived spectra for playing technique classification.” *IEEE/ACM Transactions on Audio, Speech and Language Processing* 22(12):2122–2132.
- Su, Li and Yi-Hsuan Yang. 2015. Escaping from the abyss of manual annotation: new methodology of building polyphonic datasets for automatic music transcription. In *International Symposium on Computer Music Multidisciplinary Research*. pp. 309–321.
- Sun, Deqing, Stefan Roth and Michael J. Black. 2010. Secrets of optical flow estimation and their principles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 2432–2439. DOI: <https://doi.org/10.1109/CVPR.2010.5539939>.

- Sundberg, Johan. 1994. "Acoustic and psychoacoustic aspects of vocal vibrato." *Speech Transmission Laboratory. Quarterly Progress and Status Reports. (STL-QPSR)* 35(2-3):045–068.
- Suwajanakorn, Supasorn, Steven M Seitz and Ira Kemelmacher-Shlizerman. 2017. "Synthesizing Obama: learning lip sync from audio." *ACM Transactions on Graphics (TOG)* 36(4).
- Takahashi, Naoya, Nabarun Goswami and Yuki Mitsufuji. 2018. MMDenseLSTM: An efficient combination of convolutional and recurrent neural networks for audio source separation. In *Proceedings of the International Workshop on Acoustic Signal Enhancement (IWAENC)*. IEEE pp. 106–110.
- Takahashi, Naoya, Purvi Agrawal, Nabarun Goswami and Yuki Mitsufuji. 2018. PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*. pp. 2713–2717.
- Takahashi, Naoya and Yuki Mitsufuji. 2017. Multi-scale multi-band densenets for audio source separation. In *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. pp. 21–25.
- Thomas, Verena, Christian Fremerey, David Damm and Michael Clausen. 2009. SLAVE: A score-lyrics-audio-video-explorer. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Tian, Yapeng, Jing Shi, Bochen Li, Zhiyao Duan and Chenliang Xu. 2018. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 2 pp. 252–268. DOI: https://doi.org/10.1007/978-3-030-01216-8_16.
- Tolonen, Tero and Matti Karjalainen. 2000. "A computationally efficient multipitch analysis model." *IEEE Transactions on Speech and Audio Processing* 8(6):708–716.
- Tomasi, Carlo and Takeo Kanade. 1991. Detection and tracking of point features. Technical Report CMU-CS-91-132 School of Computer Science, Carnegie Mellon University.

- Tsay, Chia-Jung. 2013. "Sight over sound in the judgment of music performance." *National Academy of Sciences* 110(36):14580–14585.
- Tsay, Chia-Jung. 2014. "The vision heuristic: Judging music ensembles by sight alone." *Organizational Behavior and Human Decision Processes* 124(1):24–33. DOI: <https://doi.org/10.1016/j.obhdp.2013.10.003>.
- Uhlich, Stefan, Marcello Porcu, Franck Giron, Michael Enenkl, Thomas Kemp, Naoya Takahashi and Yuki Mitsufuji. 2017. Improving music source separation based on deep neural networks through data augmentation and network blending. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 261–265.
- Vembu, Shankar and Stephan Baumann. 2005. Separation of vocals from polyphonic audio recordings. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*. Citeseer pp. 337–344.
- Ventura, José, Ricardo Sousa and Anibal Ferreira. 2012. Accurate analysis and visual feedback of vibrato in singing. In *Proceedings of the IEEE International Symposium on Communications Control and Signal Processing (ISCCSP)*. pp. 1–6.
- Vincent, Emmanuel. 2006. "Musical source separation using time-frequency source priors." *IEEE Transactions on Audio, Speech, and Language Processing* 14(1):91–98.
- Vincent, Emmanuel, Rémi Gribonval and Cédric Févotte. 2006. "Performance measurement in blind audio source separation." *IEEE Transactions on Audio, Speech, and Language Processing* 14(4):1462–1469.
- Vinyes, M. 2008. "MTG MASS database."
<http://www.mtg.upf.edu/static/mass/resources>.
- Von Coler, Henrik and Axel Roebel. 2011. Vibrato detection using cross correlation between temporal energy and fundamental frequency. In *Proceedings of the Audio Engineering Society Convention*.
- Wang, Qiang, Zhiyuan Guo, Gang Liu, Chunguang Li and Jun Guo. 2013. Local alignment for query by humming. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 3711–3715.

- Wang, Siying, Sebastian Ewert and Simon Dixon. 2014. Robust joint alignment of multiple versions of a piece of music. In *Proceedings of the International Society for Music Information Retrieval (ISMIR)*.
- Wang, Xueyang, Ryan Stables, Bochen Li and Zhiyao Duan. 2018. Score-Aligned Polyphonic Microtiming Estimation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 361–365.
- Wang, Yao, Jörn Ostermann and Ya-Qin Zhang. 2002. *Video processing and communications*. Vol. 5 Prentice Hall Upper Saddle River.
- Widmer, Gerhard, Sebastian Flossmann and Maarten Grachten. 2009. “YQX plays Chopin.” *AI magazine* 30(3):35–48.
- Wohlmayr, Michael, Michael Stark and Franz Pernkopf. 2011. “A probabilistic interaction model for multipitch tracking with factorial hidden Markov models.” *IEEE Transactions on Audio, Speech, and Language Processing* 19(4):799–810.
- Yamamoto, Kazuki, Etsuko Ueda, Tsuyoshi Suenaga, Kentaro Takemura, Jun Takamatsu and Tsukasa Ogasawara. 2010. Generating natural hand motion in playing a piano. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. pp. 3513–3518.
- Yang, Luwei, Khalid Z Rajab and Elaine Chew. 2017. “The filter diagonalisation method for music signal analysis: frame-wise vibrato detection and estimation.” *Journal of Mathematics and Music* pp. 1–19.
- Yin, Jun, Ye Wang and David Hsu. 2005. Digital violin tutor: An integrated system for beginning violin learners. In *Proceedings of the ACM International Conference on Multimedia*. pp. 976–985.
- Zhang, Bingjun, Jia Zhu, Ye Wang and Wee Kheng Leow. 2007. Visual analysis of fingering for pedagogical violin transcription. In *Proceedings of the ACM International Conference on Multimedia*. pp. 521–524.
- Zhao, Hang, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott and Antonio Torralba. 2018. The Sound of Pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Vol. 1 pp. 587–604. DOI: https://doi.org/10.1007/978-3-030-01246-5_35.

- Zhao, Hang, Chuang Gan, Wei-Chiu Ma and Antonio Torralba. 2019. The Sound of Motions. In *Proceedings of the International Conference on Computer Vision (ICCV)*. pp. 1735–1744.