Audio-Visual Analysis of Music Performances

Zhiyao Duan *Member, IEEE*, Slim Essid, Cynthia C. S. Liem *Member, IEEE*, Gaël Richard *Fellow, IEEE*, Gaurav Sharma *Fellow, IEEE*

I. Introduction

In the physical sciences and engineering domains, music has traditionally been considered an acoustic phenomenon. From a perceptual viewpoint, music is naturally associated with *hearing*, i.e., the *audio modality*. Moreover, for a long time, the majority of music recordings were distributed through "audio-only" media such as vinyl records, cassettes, CDs, and mp3 files. As a consequence, existing automated music analysis approaches predominantly focus on audio signals that represent information from the acoustic rendering of music.

Music performances, however, are typically multimodal [1], [2]: while sound plays a key role, other modalities are also critical to enhancing the experience of music. In particular, the visual aspects of music—be they disc cover art, videos of live performances or abstract music videos—play an important role in expressing musicians' ideas and emotions. With the popularization of video streaming services over the past decade, such visual representations also are increasingly available with distributed music recordings. In fact, video streaming platforms have become one of the preferred music distribution channels, especially among the younger generation of music consumers.

Simultaneously seeing and listening to a music performance often provides a richer experience than "pure listening". Researchers find that "*the visual component is not a marginal phenomenon in music perception, but an important factor in the communication of meaning*s" [3]. Even for prestigious classical music competitions, researchers find that visually perceived elements of the performance, such as gesture, motion, and facial expressions of the performer, affect the evaluations of judges (experts or novice alike), even more significantly than sound [4].

Symphony music provides another example of visible communicated information where large groups of orchestra musicians play simultaneously in close coordination. For expert audiences familiar with the genre, both the visible coordination between musicians, and the ability to closely watch individuals within the group, adds to the attendee's emotional experience of a concert [5]. Attendees unfamiliar with the genre can also be better engaged via *enrichment*, i.e., offering supporting information in various modalities (e.g., visualizations, textual explanations) beyond the stimuli which the event naturally triggers in the physical world.

In addition to audiences of music performances, others also benefit from information obtained through audio-visual rather than audio-only analysis. In educational settings, instrument learners benefit significantly from watching demonstrations by professional musicians, where the visual presentation provides deeper insight into specific instrument-technical aspects of the performance (e.g., fingering, choice of strings). Generally, when broadcasting audio-visual productions involving large ensembles captured with multiple recording cameras, it is also useful for the producer to be aware of which musicians are visible in which camera stream at each point in time. In order for such analyses to be done, relevant information needs to be extracted from the recorded video signals and coordinated with recorded audio. As a consequence, recently, there has been growing interest in visual analysis of music performances, even though such analysis was largely overlooked in the past.

In this paper, we aim to introduce this emerging area to the music signal processing community and the broader signal processing community. In our knowledge, this paper is the first overview of research in this area. For conciseness, we restrict our attention to *the analysis of audio-visual music performances*, which is an important subset of audio-visual music productions that is also representative of the main challenges and techniques of this field of study. Other specific applications, such as the analysis of music video clips or other types of multi-modal recordings not involving audio and visuals (e.g., lyrics or music score sheets), although important in their own right, are not covered here to maintain a clear focus and a reasonable length.

In the remainder of the paper, we first present the significance and key challenges for audio-visual music analysis in Section II, and survey existing work in Section III. Then we describe notable approaches in three main research lines organized according to how the audio-visual correspondence is modeled: work on static correspondence in Section IV; work on instrument-specific dynamic correspondence in Section V; and work on modeling more general dynamic correspondence for music source separation in Section VI. We conclude the paper with discussions of current and future research trends in Section VII.

II. Significance and Challenges

A. Significance

Figure 1 illustrates some examples of how visual and aural information in a music performance complements each other, and how it offers more information on the performance than what can be obtained by considering only the audio channel

Authors in alphabetical order.

ZD and GS are with the Department of Electrical and Computer Engineering, University of Rochester, NY, USA. e-mail: {zhiyao.duan, gaurav.sharma}@rochester.edu.

SE and GR are with the Department of Images, Data and Signals, Télécom ParisTech, France. e-mail:{slim.essid, gael.richard}@telecom-paristech.fr.

CL is with the Multimedia Computing Group, Delft University of Technology, The Netherlands. e-mail: c.c.s.liem@tudelft.nl.

ZD is partially supported by the National Science Foundation grant No. 1741472.

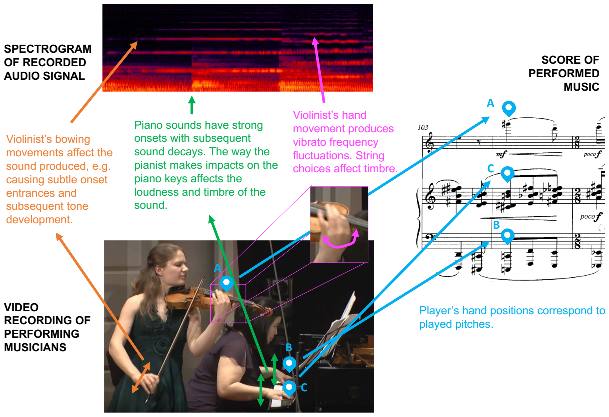

SPECTROGRAM OF RECORDED AUDIO SIGNAL
SCORE OF PERFORMED MUSIC
Violinist's bowing movements affect the sound produced, e.g. causing subtle onset entrances and subsequent tone development.
Piano sounds have strong onsets with subsequent sound decays. The way the pianist makes impacts on the piano keys affects the loudness and timbre of the sound.
Violinist's hand movement produces vibrato frequency fluctuations. String choices affect timbre.
A
C
B
103
mf
poco f
VIDEO RECORDING OF PERFORMING MUSICIANS
A
B
C
Player's hand positions correspond to played pitches.

Figure 1. Examples of information present in three parallel representations of a music performance excerpt: video, audio and score.

and a musical score. In fact, while the musical score is often considered as the "ground truth" of a music performance, significant performance-specific expressive information, such as the use of vibrato, is not indicated in the score, and is instead evidenced in the audio-visual performance signals.

Compared to audio-only music performance analysis, the visual modality offers extra opportunities to extract musically meaningful cues out of recorded performance signals. In some cases, the visual modality allows for addressing tasks that would not be possible in audio-only analysis, e.g., tracking a musician's fingerings or a conductor's gestures, and analyzing individual players in the same instrumental section of an orchestra. In other cases, the visual modality provides significant help in task-solving, e.g., in source separation, and in the characterization of expressive playing styles. In Section III, we discuss several representative tasks along these lines.

Audio-visual analysis of music performances broadens the scope of music signal processing research, connecting the audio signal processing area with other areas, namely image processing, computer vision, and multimedia. The integration of audio and visual modalities also naturally creates a connection to emerging research areas such as virtual reality and augmented reality, and extends music-related human-computer interaction. It serves as a controlled testbed for research on multimodal data analysis, which is critical for building robust and universal intelligent systems.

B. Challenges

The multimodal nature of audio-visual analysis of music poses new research challenges. First, the visual scenes of music performances present new problems for image processing and computer vision. Indeed, the visual scene is generally cluttered, especially when multiple musicians are involved, who additionally may be occluded by each other and by music stands. Also, musically meaningful motions may be subtle (e.g., fingering and vibrato motion); and camera views may be complex (e.g., musicians not facing to cameras, zoom-in/out and changes of views).

Second, the way to integrate audio and visual processing in the modeling stage of musical scene analysis is a key challenge. In fact, independently tackling the audio and visual modalities to merely fuse, at a later stage, the output of the corresponding (unimodal) analysis modules, is generally not an optimal approach. To take advantage of potential cross-modal dependencies, it is better to combine low-level audio-visual representations as early as possible in the data analysis pipeline. This is, however, not always straightforward: certain visual signals (e.g., bowing motion of string instruments) and audio signals (e.g., note onsets) of a sound source are often highly correlated, yet some performer movements (e.g., head nodding are not directly related to sound [6]. How to discover and exploit audio-visual correspondence in a complex audio-visual scene of music performances is thus a key question.

Third, the lack of annotated data is yet another challenge. While commercial recordings are abundant, they are usually not annotated and also subject to copyright restrictions that limit their distribution and use. Annotated audio datasets of musical performances are already scarce due to the complexities of recording and ground-truth annotation. Audio-visual datasets are even more scarce and their creation requires more effort. The lack of large-scale annotated datasets limits the application of many supervised learning techniques that have proven successful for data-rich problems. We note that available music datasets are surveyed in a recent paper [7] that details the creation of a new multi-track audio-visual classical music dataset. The dataset provided in [7] is relatively small with only 44 short pieces but is richly annotated, providing individual instrument tracks to allow assessment of source separation methods and associated music score information in a machine readable format. At the other end of the data spectrum, the Youtube-8M dataset [8] provides a large-scale labeled video dataset (with embedded audio) that also includes many music videos. However, the Youtube-8M dataset is currently only annotated with overall video labels and therefore suited primarily for video/audio classification tasks.

III. Overview of Existing Research

It is not an easy task to give a well structured overview of an emerging field, yet here we make a first attempt from two perspectives. Section III-A categorizes existing work into different analysis tasks for different instruments, while Section III-B provides a perspective on the type of audio-visual correspondence that is exploited during the analysis.

A. Categorization of audio-visual analysis tasks

Table I organizes existing work on audio-visual analysis of music performances along two dimensions: 1) the type of musical instrument, and 2) the analysis task.

The first dimension is not only a natural categorization of musicians in a music performance, it is also indicative of the types of audio-visual information revealed during the performance. For example, percussionists show large-scale motions that are almost all related to sound articulation. Pianists' hand and finger motions are also related to sound articulation, but they are much more subtle and also indicative

Table I

Categorization of existing research on audio-visual analysis of musical instrument performances according to the type of the instrument and the analysis task. Certain combinations of instruments and tasks do not make sense, and are marked by "N/A". Various techniques and their combinations have been employed, including support vector machine, hidden Markov models, non-negative matrix factorization, and deep neural networks.

Visual	*Is Critical*		*Is Significant*				
Tasks	Fingering	Association	Play/Non-Play	Onset	Vibrato	Transcription	Separation
Percussion	N/A		[9]		N/A	[10]	
Piano	[11], [12]				N/A		
Guitar	[13], [14], [15], [16]					[16]	
Strings	[17]	[18], [19]	[9], [20]	[19]	[21]	[17], [20]	[22]
Wind			[9]	[23]			
Singing	N/A						

of the notes being played (i.e., the musical content). For guitars and strings, the left hand motions are indicative of the notes being played, while the right hand motions tell us how the notes are articulated (e.g., *legato* or *staccato*). For wind instruments, note articulations are difficult to see, and almost all visible motions (e.g., fingering of clarinet or hand positioning of trombone) are about notes. Finally, singers' mouth shapes only reveal the syllables being sung but not the pitch; also their body movements can be correlated to the musical content but are not predictive enough for the details.

The second dimension is about tasks or aspects that the audio-visual analysis focuses on. The seven tasks/aspects are further classified into two categories: tasks in which visual analysis is critical and tasks in which visual analysis provides significant help. *Fingering analysis* is one example of the first category. It is very difficult to infer the fingering purely from audio while it becomes possible by observing the finger positions. There has been research on fingering analysis from visual analysis for guitar [13], [14], [15], [16], violin [17], and piano [11], [12]. Fingering patterns are mostly instrument-specific, however, the common idea is to track hand and finger positions relative to the instrument body. Another task is *audio-visual source association*, i.e., which player in the visual scene corresponds to which sound source in the audio mixture. This problem is addressed for string instruments by modeling the correlation between visual features and audio features, such as the correlation between bowing motions and note onsets [18] and that between vibrato motions and pitch fluctuations [19].

The second category contains more tasks. *Playing/Non-Playing (P/NP) activity detection* is one of them. In an ensemble or orchestral setting, it is very difficult to detect from the audio mixture whether a certain instrument is being played, yet the visual modality, if not occluded, offers a direct observation of the playing activities of each musician. Approaches based on image classification and motion analysis [9], [20] have been proposed. *Vibrato analysis* for string instruments is another task. The periodic movement of the fingering hand detected from visual analysis has been shown to correlate well with the pitch fluctuation of vibrato notes, and has been used to detect vibrato notes and analyze the vibrato rate and depth [21]. *Automatic music transcription* and its subtasks such as multi-pitch analysis are very challenging if only audio signals are available. It has been shown that audio-visual analysis is beneficial for monophonic instruments such as violin [17], polyphonic instruments such as guitar [16] and drums [10], and music ensembles such as string ensembles [20]. The common underlying idea is to improve audio-based transcription results with play/non-play activity detection and fingering analysis. Finally, *audio source separation* can be significantly improved by audio-visual analysis. Motions of players are often highly correlated to sound characteristics of sound sources [6]. There has been work on modeling such correlations for audio source separation [22].

Besides instrumental players, conductor gesture analysis has also been investigated in audio-visual music performance analysis. Indeed, conductors do not directly produce sounds (besides occasional noises), however, they are critical in music performances. Under the direction of different conductors, the same orchestra can produce significantly different performances of the same musical piece. One musically interesting research problem is comparing conducting behaviors of different conductors and analyzing their influences on the sound production of the orchestra. There has been work on conductor baton tracking [24] and gesture analysis [25] using visual analysis.

B. *Different levels of audio-visual correspondence*

Despite the various forms of music performances and analysis tasks, the common underlying idea of audio-visual analysis is to find and model the correspondence between audio and visual modalities. This correspondence can be *static*, i.e., between a static image and a short time frame of audio. For example, a certain posture of a flute player is indicative of whether the player is playing the instrument or not; a static image of a fingering hand is informative for the notes being played. This correspondence can also be *dynamic*, i.e., between a dynamic movement observed in the video and the fluctuation of audio characteristics. For example, a strumming motion of the right hand of a guitar player is a strong indicator of the rhythmic pattern of the music passage; the periodic rolling motion of the left hand of a violin player well corresponds to the pitch fluctuation of vibrato notes. Due to the large variety of instruments and their unique playing techniques, this dynamic correspondence is often instrument-specific. The underlying idea of dynamic correspondence, however, is *universal* among different instruments. Therefore, it is appealing to build a unified framework for capturing this dynamic correspondence. If such correspondence can be captured robustly, the visual information can be better exploited to stream the corresponding audio components into sources, leading to visually informed source separation.

In the following three sections, we will further elaborate these different levels of audio-visual correspondence by summarizing existing works and presenting concrete examples.

IV. Static audio-visual correspondence

In this section, we first discuss works focusing on the modeling of static audio-visual correspondence in music performances. "Static" here refers to correspondences between sonic realizations and their originating sources that remain stable over the course of a performance, and for which the correspondence analysis does not rely on short-time dynamic variations. After giving a short overview with more concrete examples, a more extended case study discussion will be given on Playing/Non-Playing detection in instrument ensembles.

A. *Overview*

Typical static audio-visual correspondences have to do with *positions* and *poses*: which musician sits where, at what parts of the instrument does the interaction occur that leads to sound production, and how can the interaction with the instrument be characterized?

Regarding musicians' positions, when considering large ensemble situations, it will be too laborious for a human to annotate every person in every shot, especially when multiple cameras record the performance at once. At the same time, due to the typically uniform concert attire worn by ensemble members, and musicians being part of large player groups that will actively move and occlude one another, recognizing individual players purely by computer vision methods is again a non-trivial problem, for which it also would be unrealistic to acquire large amounts of training data. However, within the same piece, orchestra musicians will not change relative positions with respect to one another. Therefore, the orchestra setup can be considered as a quasi-static scene. The work in [26] proposed to identify each musician in each camera over a full recording timeline by combining partial visual recognition with knowledge of this scene's configuration, and a human-in-the-loop approach in which humans are strategically asked to indicate the identities of performers in visually similar clusters. With minimal human interaction, a scene map is built up, and the spatial relations within this scene map assist face clustering in crowded quasi-static scenes.

Regarding positions of interest on an instrument, work has been performed on the analysis of fingering. This can be seen as static information, as the same pressure action on the same position of the instrument will always yield the same pitch realization. Visual analysis has been performed to analyze fingering actions on pianos [11], [12], guitars [13], [14], [15], [16] and violins [16], [17]. Main challenges involve the detection of the fingers in unconstrained situations and without the need to add markers to the fingers.

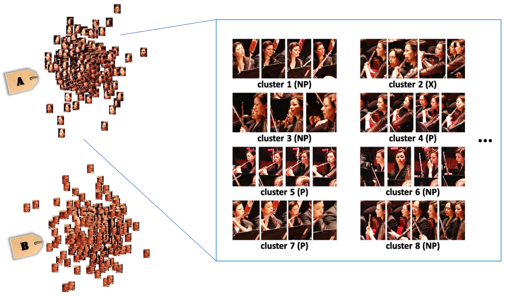


Figure 2. Example of hierarchical clustering steps for Playing/Non-Playing detection: First, diarization is performed on global face clustering results (left) to identify a musician's identity; then, within each global artist cluster, sub-clusters are assigned with a Playing/Non-Playing label (right).

B. *Case study: Playing/Non-Playing detection in orchestras*

Whether individual musicians in large ensembles are playing their instrument or not seems banal information; however, this information can be significant up to critical in audio-visual analysis. Within the same instrument group, not all players may be playing at once. If this occurs, in a multi-channel audio recording, it is not trivial to distinguish which subset of individuals is playing, while this will visually be obvious. Furthermore, having a global overview of what instruments are active and visible in performance recordings provides useful information for audio-visual source separation.

In [9], a method is proposed to detect Playing/Non-Playing (P/NP) information in multi-camera recordings of symphonic concert performances, in which unconstrained camera movements and varying shooting perspectives occur. As a consequence, performance-related movement may not always be easily observed from the video, although coarser P/NP information can still be inferred through face and pose clustering.

A hierarchical method is proposed, that is illustrated in Figure 2, and focuses on employing clustering techniques, rather than learning sophisticated human-object interaction models. First, *musician diarization* is performed to annotate which musician appears when and where in a video. For this, key frames are extracted at regular time intervals. In each keyframe, face detection is performed, including an estimation of the head pose angle, as well as inference of bounding boxes for the hair and upper body of the player. Subsequently, segmentation is performed on the estimated upper body of the musician, taking into account the gaze direction of the musician, as the instrument is expected to be present in the same direction.

After this segmentation step, face clustering methods are applied, including several degrees of contextual information (e.g., on the scene and upper body), and different feature sets, the richest feature set consisting of a Pyramid of Histograms of Oriented Gradients, the Joint Composite Descriptor, Gabor texture, Edge Histogram and Auto Color Correlogram.

Upon obtaining per-musician clusters, a renewed clustering is performed per musician, aiming to generate sub-clusters that only contain images of the same musician, performing one particular type of object interaction, recorded from one particular camera viewpoint. Finally, a human annotator action completes the labeling step: an annotator has to indicate who the musician is, and whether a certain sub-cluster contains a Playing or Non-Playing action. As the work in [9] investigates various experimental settings (e.g., clustering techniques, feature sets), yielding thousands of clusters, expected annotator action at various levels of strictness is simulated by setting various thresholds on how dominant a class within a cluster should be.

An extensive discussion of evaluation outcomes per framework module is given in [9]. Several takeaway messages can be taken from this work. First of all, face and upper body regions are most informative for clustering. Furthermore, the proposed method can effectively discriminate Playing vs. Non-Playing action, while generating a reasonable amount of sub-clusters (i.e., enough to yield informative sub-clusters, but not too many, which would cause high annotator workload). Face information alone may already be informative, as it indirectly reveals pose. However, in some cases, clustering cannot yield detailed relevant visual analyses (e.g., subtle mouth movement for a wind player), and the method has a bias towards false positives, caused by playing anticipation movement. The application of merging strategies per instrumental part helps in increasing timeline coverage, even if a musician is not always detected. Finally, high annotator rejection thresholds (demanding for clear majority classes within clusters) effectively filter out non-pure clusters.

One direct application of P/NP activity detection is in automatic music transcription. In particular, for multi-pitch estimation (MPE), P/NP information can be used to improve the estimation of instantaneous polyphony (i.e., the number of pitches at a particular time) of an ensemble performance, assuming that each active instrument only produces one pitch at a time. Instantaneous polyphony estimation is a difficult task from the audio modality itself, and its errors constitute a large proportion of music transcription errors. Furthermore, P/NP is also helpful for multi-pitch streaming (MPE), i.e., assigning pitch estimates to pitch streams corresponding to instruments: a pitch estimate should only be assigned to an active source. This idea has been explored in [20] and it is shown that both MPE and MPS accuracies are significantly improved by P/NP activity detection for ensemble performances.

V. Dynamic audio-visual correspondence

In a music performance, a musician makes many movements [6]. Some movements (e.g., bowing and fingering) are the articulation sources of sound, while others (e.g., head shaking) are responses to the performance. In both cases, the movements show a strong correspondence with certain feature fluctuations in the music audio. Capturing this dynamic correspondence is important for the analysis of music performances.

A. Overview

Due to the large variety of musical instruments and their playing techniques, the dynamic audio-visual correspondence shows different forms. In the literature, researchers have investigated the correspondence between bowing motions and note onsets of string instruments [18], between hitting actions and drum sounds of percussion instruments [10], and between left-hand rolling motions and pitch fluctuations of string vibrato notes [21], [19]. On the visual modality, object tracking and optical flow techniques have been adopted to track relevant motions, while on the audio modality, different audio features have been considered.

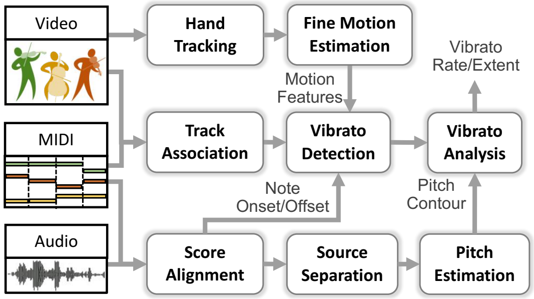


Figure 3. System overview of an audio-visual vibrato detection and analysis system for string instruments in ensemble performances, proposed in [21].

The main challenge lies in determining *what/where to look* for the dynamic correspondence. This is challenging not only because the correspondence is instrument- and playing technique-dependent, but also because there are many irrelevant motions in the visual scene [6] and interferences from multiple simultaneous sound sources in the audio signal. Almost all existing methods rely on the prior knowledge of instrument type and playing techniques to attend to relevant motions and sound features. For example, in [18] for the association between string players and score tracks, the correspondence between bowing motions and some note onsets are captured. This is informed by the fact that many notes of string instruments are started with a new bow stroke and that different tracks often show different onset patterns. For the association of wind instruments, the onset cue is still useful, but the motion capture module would need to be revised to capture the more subtle and diverse movements of fingers.

B. Case study: vibrato analysis of string instruments

Vibrato is an important musical expression, and vibrato analysis is important for musicological studies, music education, and music synthesis. Acoustically, vibrato is characterized by a periodic fluctuation of pitch with a rate between 5-10 Hz. Audio-based vibrato analysis methods rely on the estimation of the pitch contour. In an ensemble setting, however, multi-pitch estimation is very challenging due to the interference of other sound sources. For string instruments, vibrato is the result of periodic change of the length of the vibrating string, which is effectuated by the rolling motion of the left hand. If the rolling motion is observable, then vibrato notes can be detected and analyzed with the help of visual analysis. Because visual analysis does not suffer from the presence of other sound sources (barring occlusion), audio-visual analysis offers a tremendous advantage for vibrato analysis of string instruments in ensemble settings.

In [21], an audio-visual vibrato detection and analysis system is proposed. As shown in Figure 3, this approach integrates audio, visual and score information, and contains several modules to capture the dynamic correspondence among these modalities.

The first step is to detect and track the left hand for each player using the Kanade-Lucas-Tomasi (KLT) tracker. This results in a dynamic region of the tracked hand, shown as the green box in Figure 4. Optical flow analysis is then performed to calculate motion velocity vectors for each pixel in this region in each video frame. Motion vectors in frame t 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. It is noted that these motion vectors may also contain the slower large-scale body movements that are not associated with vibrato. Therefore, to eliminate the body movement effects, the moving average of the signal $\mathbf{u}(t)$ is subtracted from itself to obtain a refined motion estimation $\mathbf{v}(t)$. The right subfigure of Figure 4 shows the distribution of all $\mathbf{v}(t)$ across time, from which the principal motion direction can be inferred through Principal Component Analysis (PCA), which aligns well along the fingerboard. The projection of the motion vector $\mathbf{v}(t)$ onto the principal direction is defined as the 1-d *motion velocity curve* $V(t)$. Taking an integration over time, one obtains a 1-d *hand displacement curve* $X(t) = \int_0^t V(\tau)d\tau$, that corresponds directly to the pitch fluctuation.

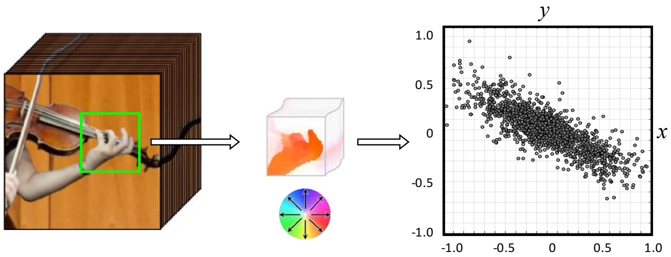


Figure 4. Motion capture results from left hand tracking (left), color encoded pixel velocities (middle), and scatter plot of frame-wise refined motion velocities (right).

In order to use the motion information to detect and analyze vibrato notes, one needs to know which note the hand motion corresponds to. This is solved by audio-visual source association and audio-score alignment. In this work, audio-visual source association is performed through the correlation between bowing motions and note onsets, as described in [18]. Audio-score alignment [27] synchronizes the audio-visual performance (assuming perfect audio-visual synchronization) with the score, from which onset and offset times of each note are estimated. This can be done by comparing the harmonic content of the audio and the score and dynamic time warping. Score-informed source separation is then performed and the pitch contour of each note is estimated from the separated source signal.

Given the correspondence between motion vectors and sound features (pitch fluctuations) of each note, vibrato detection is performed with two methods. The first method uses a Support Vector Machine (SVM) to classify each note as vibrato or non-vibrato using features extracted from the motion vectors. The second method simply sets a threshold on the auto-correlation of the 1-d hand displacement curve $X(t)$.

For vibrato notes, vibrato rate can also be calculated from the autocorrelation of the hand displacement curve $X(t)$. Vibrato extent (i.e., dynamic range of the pitch contour), however, cannot be estimated by capturing the motion extent. This is because it varies upon the camera distance and angle, as well as the vibrato articulation style, hand position, and the instrument type. To address this issue, the hand displacement curve is scaled to match the estimated noisy pitch contour from score-informed audio analysis. Specifically, assuming $F(t)$ is the estimated pitch contour (in MIDI number) of the detected vibrato note from audio analysis after subtracting its DC component, the vibrato extent v_e (in musical cents) is estimated as $\hat{v}_e$ as

$$\hat{v}_e = \arg\min_{v_e} \sum_{t=t^{on}}^{t^{off}} \left| 100 \cdot F(t) - v_e \frac{X(t)}{\hat{w}_e} \right|^2, \qquad (1)$$

where $100 \cdot F(t)$ is the pitch contour in musical cents; $\hat{w}_e$ is the dynamic range of $X(t)$.

VI. Music source separation using dynamic correspondence

Audio source separation in music recordings is a particularly interesting type of task where audio-visual matching between visual events of a performer's actions and their audio rendering can be of great value. Notably, such an approach enables addressing audio separation tasks which could not be performed in a unimodal fashion (solely analyzing the audio signal), for instance when considering two or more instances of the same instruments, say a duet of guitars or violins, as done in the work of Parekh *et al.* [22]. Knowing whether a musician is playing or not at a particular point in time gives important cues for source allocation. Seeing the hand and finger movements of a cellist helps us attend to the cello's section sound in an orchestral performance. The same idea applies to visually informed audio source separation.

A. Overview

There is a large body of works in multimodal (especially audio-visual) source separation for speech signals but much less effort has been dedicated to audio-visual music performance analysis for source separation.

It was however shown in the work of Godoy *et al.* [6] that there are certain players' motions that are highly correlated to sound characteristics of audio sources. In particular, the authors highlighted the correlation that may exist between music and hand movements or the sway in the upper body, by analyzing a solo piano performance. An earlier work by Barzelay and Shechner [28] has exploited such a correlation in introducing an audio-visual system for individual musical source enhancement in violin-guitar duets. The authors isolate audio-associated visual objects (AVO) by searching for cross-modal temporal incidences of events and then use these to perform musical source separation.

B. Case study: motion-driven source separation in a string quartet

The idea that motion characteristics obtained from visual analysis encode information about the physical excitation of a sounding object is also exploited in more recent studies. As an illustrative example, we detail below a model in which it is assumed that the characteristics of a sound event (e.g., musical note) is highly correlated with the speed of sound-producing motion [22]. More precisely, the proposed approach extends the popular Non-negative Matrix Factorization (NMF) framework using visual information about objects' motion. Applied to string quartets, the motion of interest is mostly carried by bow speed. The main steps of this method are the following (see Figure 5).

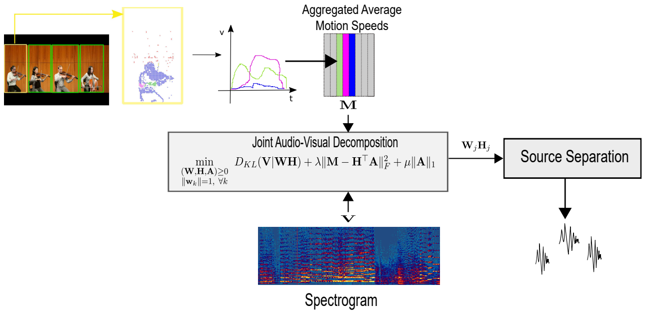

Aggregated Average
Motion Speeds
v
t
$\mathbf{M}$
Joint Audio-Visual Decomposition
$\min_{\substack{(\mathbf{W},\mathbf{H},\mathbf{A})\geq 0\\ \|\mathbf{w}_k\|=1,\ \forall k}} D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda\|\mathbf{M}-\mathbf{H}^\top\mathbf{A}\|_F^2 + \mu\|\mathbf{A}\|_1$
$\mathbf{W}_j\mathbf{H}_j$
Source Separation
$\mathbf{V}$
Spectrogram

Figure 5. A joint audio-visual music source separation system.

1) Gather motion features, namely average motion speeds (further described below), in a data matrix $\mathbf{M} \in \mathbb{R}_+^{N\times C}$ which summarizes the speed information of coherent motion trajectories within pre-defined regions. In the simplest case, there is one region per musician (i.e., per source). $C = \sum_j C_j$ is the number of motion clusters where C_j is the number of clusters per source j and N is the frame size of the Short-Time Fourier Transform (STFT) used for the computation of the audio signal's spectrogram.
2) Ensure that typical motion speeds (such as bow speed) are active synchronously with typical audio events. This is done by constraining the audio spectrogram decomposition obtained by NMF $\mathbf{V} \approx \mathbf{WH}$ and the motion data decomposition $\mathbf{M} \approx \mathbf{H}^\top\mathbf{A}$ to share the same activity matrix $\mathbf{H} \in \mathbb{R}_+^{K\times N}$, where $\mathbf{W} \in \mathbb{R}_+^{F\times K}$ is the matrix collecting the so-called nonnegative audio spectral patterns (column-wise), and where $\mathbf{A} = [\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_C]$ gathers nonnegative linear regression coefficients for each motion cluster with $\boldsymbol{\alpha}_c = [\alpha_{1c}, \ldots, \alpha_{Kc}]^T$.
3) Ensure that only a limited number of motion clusters are active at a given time. This can be done by imposing a sparsity constraint on $\mathbf{A}$.
4) Assign an audio pattern to each source for separation and reconstruction. This is done by assigning the k-th basis vector (column of $\mathbf{W}$) to the j^{th} source if $\operatorname{argmax}_c \alpha_{kc}$ belongs to the j^{th} source cluster. The different sources are then synthesized by element-wise multiplication between the soft mask, given by $(\mathbf{W}_j\mathbf{H}_j)./(\mathbf{WH})$, and the mixture spectrogram followed by an inverse STFT, where "./" stands for element-wise division, $\mathbf{W}_j$ and $\mathbf{H}_j$ are the submatrices of spectral patterns $\mathbf{w}_k$ and their activations $\mathbf{h}_k$ assigned to the j^{th} source (see Figure 6).

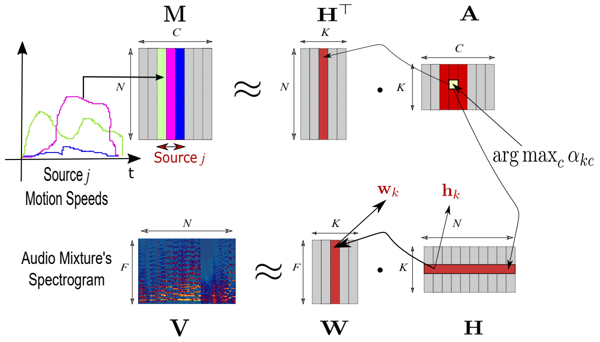


Figure 6. Joint audio-visual source separation: illustration of the audio pattern assignment to source j (example for the k-th basis vector).

A possible formulation for the complete model can then be written as the following optimization problem:

$$\underset{\substack{(\mathbf{W},\mathbf{H},\mathbf{A})\geqslant 0\\ \|\mathbf{w}_k\|=1,\ \forall k}}{\text{minimize}} \quad D_{KL}(\mathbf{V}|\mathbf{WH}) + \lambda\|\mathbf{M}-\mathbf{H}^\top\mathbf{A}\|_F^2 + \mu\|\mathbf{A}\|_1, \quad (2)$$

where D_{KL} is the Kullback-Leibler divergence, λ and μ are positive hyperparameters (to be tuned) and $\|.\|_F$ is the Frobenius norm.

More details can be found in [22], but this joint audio-visual approach significantly outperformed for most situations the corresponding sequential approach proposed by the same authors and the audio-only approach introduced in [29]. For example, for a subset of the URMP dataset [30], the joint approach obtained a Signal-to-Distortion Ratio (SDR) of 7.14 dB for duets and 5.14 dB for trios while the unimodal approach of [29] obtained SDRs respectively of 5.11 dB and 2.18 dB. It is worth mentioning that in source separation a difference of +1 dB is usually acknowledged as significant.

The correlation between motion in the visual modality and audio is also at the core of some other recent approaches. While bearing some similarities with the system detailed above, the approach described in [18] further exploits the knowledge of the MIDI score to well align the audio recording (e.g., onsets) and video (e.g., bow speeds). An extension of this work is presented in [19] where the audio-visual source association is performed through multi-modal analysis of vibrato notes. It is in particular shown that the fine-grained motion of the left hand is strongly correlated with the pitch fluctuation of vibrato notes and that this correlation can be used for audio-visual music source separation in a score-informed scenario.

VII. Current Trends and Future Work

This article provides an overview of the emerging field of audio-visual music performance analysis. We used specific case studies to highlight how techniques from signal processing, computer vision, and machine learning can jointly exploit the information contained in the audio and visual modalities to effectively address a number of music analysis tasks.

Current work in audio-visual music analysis has been constrained by the availability of data. Specifically, the relatively small size of current annotated audio-visual datasets has precluded the extensive use of data-driven machine learning approaches, such as deep learning. Recently, deep learning has been utilized for vision-based detection of acoustic timed music events [23]. Specifically, the detection of onsets performed by clarinet players is addressed in this work by using a 3D convolutional neural network (CNN) that relies on multiple streams, each based on a dedicated region of interest (ROI) from the video frames that is relevant to sound production. For each ROI, a reference frame is examined in the context of a short surrounding frame sequence, and the desired target is labeled as either an “onset” or “not-an-onset”. Although state-of-the-art audio-based onset detection methods outperform the model proposed in [23], the dataset, task setup and architecture setup give rise to interesting research questions, especially on how to deal with significant events in temporal multimedia streams that occur at fine temporal and spatial resolutions. Interesting ideas exploiting deep learning models can also be found in related fields. For example, in [31] a promising strategy in the context of emotional analysis of music videos is introduced. Their approach consists in fusing learned audio-visual mid-level representations using CNNs. Another important promising research direction is transfer learning which could better cope with the limited size of annotated audio-visual music performance datasets. As highlighted in [32], it is possible to learn an efficient audio feature representation for an audio-only application, specifically audio event recognition, by using a generic audio-visual database.

The inherent mismatch between the audio content and the corresponding image frames in a large majority of video recordings remains a key challenge for audio-visual music analysis. For instance, at a given point in time, edited videos of live performances often show only part of the performers' actions (think of live orchestra recordings). In such situations, the audio-visual analysis systems need to be flexible enough to effectively exploit the partial and intermittent correspondences between the audio and visual streams. Multiple instance learning techniques already used for multi-modal event detection in the computer vision community may offer an attractive option for addressing this challenge.

As new network architectures are developed for dealing with such structure in multi-modal temporal signals and as significantly larger annotated datasets become available, we expect that deep learning based data-driven machine learning will lead to rapid progress in audio-visual music analysis, mirroring the deep learning revolution in computer vision, natural language processing, and audio analysis.

Beyond the immediate examples included in the case studies presented in this paper, audio-visual music analysis can be extended toward other music genres including pop, jazz, and world music. It can also help improve a number of applications in various musical contexts. Video based tutoring for music lessons is already popular (for examples, see guitar lessons on YouTube). The use of audio-visual music analysis can make such lessons richer by better highlighting the relations between the player's actions and the resulting musical effects. Audio-visual music analysis can similarly be used to enhance other music understanding/learning activities, including score-following, auto-accompaniment, and active listening. Better tools for modeling the correlation between visual and audio modalities can also enable novel applications beyond the analysis of music performances. For example, in recent work on cross-modal audio-visual generation, sound to image sequence generation, or video to sound spectrogram generation has been demonstrated using deep generative adversarial networks [33]. Furthermore, the underlying tools and techniques can also help address other performing arts that involve music. Examples of such work include dance movement classification [34] and alignment of different dancers' movements within a single piece [35] by using (visual) gesture tracking and (audio) identification of stepping sounds.

References

[1] C. C. S. Liem, M. Müller, D. Eck, G. Tzanetakis, and A. Hanjalic, “The need for music information retrieval with user-centered and multimodal strategies,” in *Proc. International Workshop on Music Information Retrieval with User-Centered and Multimodal Strategies (MIRUM) at ACM Multimedia*, Scottsdale, USA, November 2011, pp. 1–6.

[2] S. Essid and G. Richard, *Multimodal Music Processing*, ser. Dagstuhl Follow-Ups. Dagstuhl, Germany: Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2012, vol. 3, ch. Fusion of Multimodal Information in Music Content Analysis, pp. 37–52. [Online]. Available: http://drops.dagstuhl.de/opus/volltexte/2012/3465

[3] F. Platz and R. Kopiez, “When the eye listens: A meta-analysis of how audio-visual presentation enhances the appreciation of music performance,” *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 1, pp. 71–83, 2012.

[4] C.-J. Tsay, “Sight over sound in the judgment of music performance,” *National Academy of Sciences*, vol. 110, no. 36, pp. 14 580–14 585, 2013.

[5] M. S. Melenhorst and C. C. S. Liem, “Put the concert attendee in the spotlight. a user-centered design and development approach for classical concert applications.” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, Málaga, Spain, 2015, pp. 800–806.

[6] R. I. Godøy and A. R. Jensenius, “Body movement in music information retrieval,” in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2009.

[7] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, “Creating a multi-track classical music performance dataset for multi-modal music analysis: Challenges, insights, and applications,” *IEEE Trans. Multimedia*, 2018, accepted for publication, to appear.

[8] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan, “YouTube-8M: A large-scale video classification benchmark,” *arXiv*, vol. abs/1609.08675, 2016. [Online]. Available: http://arxiv.org/abs/1609.08675

[9] A. Bazzica, C. C. S. Liem, and A. Hanjalic, “On detecting the playing/non-playing activity of musicians in symphonic music videos,” *Computer Vision and Image Understanding*, vol. 144, pp. 188–204, 2016.

[10] K. McGuinness, O. Gillet, N. E. O'Connor, and G. Richard, “Visual analysis for drum sequence transcription,” in *Proc. IEEE European Signal Processing Conference*, 2007, pp. 312–316.

[11] D. Gorodnichy and A. Yogeswaran, “Detection and tracking of pianist hands and fingers,” in *Proc. Canadian Conference on Computer and Robot Vision*, 2006.

[12] A. Oka and M. Hashimoto, “Marker-less piano fingering recognition using sequential depth images,” in *Proc. Korea-Japan Joint Workshop on Frontiers of Comp. Vision (FCV)*, 2013.

[13] A.-M. Burns and M. M. Wanderley, “Visual methods for the retrieval of guitarist fingering,” in *Proc. International Conference on New Interfaces for Musical Expression (NIME)*, 2006.

[14] C. Kerdvibulvech and H. Saito, "Vision-based guitarist fingering tracking using a Bayesian classifier and particle filters," in *Advances in Image and Video Tech.* Springer, 2007, pp. 625–638.

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

[16] M. Paleari, B. Huet, A. Schutz, and D. Slock, "A multimodal approach to music transcription," in *Proc. International Conference on Image Processing (ICIP)*, 2008.

[17] B. Zhang and Y. Wang, "Automatic music transcription using audio-visual fusion for violin practice in home environment," The National University of Singapore, Tech. Rep. TRA7/09, 2009.

[18] B. Li, K. Dinesh, Z. Duan, and G. Sharma, "See and listen: Score-informed association of sound tracks to players in chamber music performance videos," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2906–2910.

[19] B. Li, C. Xu, and Z. Duan, "Audiovisual source association for string ensembles through multi-modal vibrato analysis," in *Proc. Sound and Music Computing (SMC)*, 2017.

[20] K. Dinesh, B. Li, X. Liu, Z. Duan, and G. Sharma, "Visually informed multi-pitch analysis of string ensembles," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 3021–3025.

[21] B. Li, K. Dinesh, G. Sharma, and Z. Duan, "Video-based vibrato detection and analysis for polyphonic string music," in *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, 2017, pp. 123–130.

[22] S. Parekh, S. Essid, A. Ozerov, N. Q. K. Duong, P. Perez, and G. Richard, "Guiding audio source separation by video object information," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2017.

[23] A. Bazzica, J. C. van Gemert, C. C. S. Liem, and A. Hanjalic, "Vision-based detection of acoustic timed events: a case study on clarinet note onsets," *arXiv preprint arXiv:1706.09556*, 2017.

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

[25] Á. Sarasúa and E. Guaus, "Beat tracking from conducting gestural data: a multi-subject study," in *Proc. ACM International Workshop on Movement and Computing*, 2014, p. 118.

[26] A. Bazzica, C. C. S. Liem, and A. Hanjalic, "Exploiting scene maps and spatial relationships in quasi-static scenes for video face clustering," *Image and Vision Computing*, vol. 57, pp. 25–43, 2017.

[27] R. B. Dannenberg and C. Raphael, "Music score alignment and computer accompaniment," *Communications of the ACM*, vol. 49, no. 8, pp. 38–43, 2006.

[28] Z. Barzelay and Y. Y. Schechner, "Harmony in motion," in *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[29] M. Spiertz and V. Gnann, "Source-filter based clustering for monaural blind source separation," in *Proc. Int. Conf. on Digital Audio Effects (DAFx)*, 2009.

[30] B. Li, X. Liu, K. Dinesh, Z. Duan, and G. Sharma, "Creating a musical performance dataset for multimodal music analysis: Challenges, insights, and applications," *IEEE Trans. Multimedia*, 2018.

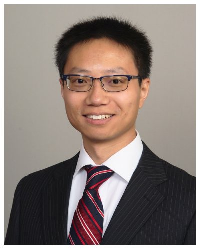
[31] E. Acar, F. Hopfgartner, and S. Albayrak, "Fusion of learned multimodal representations and dense trajectories for emotional analysis in videos," in *2015 13th International Workshop on Content-Based Multimedia Indexing (CBMI)*, 2015, pp. 1–6.

[32] Y. Aytar, C. Vondrick, and A. Torralba, "Soundnet: Learning sound representations from unlabeled video," *Advances in Neural Information Processing (NIPS)*, 2016.

[33] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proc. Thematic Workshops of ACM Multimedia*, 2017, pp. 349–357.

[34] A. Masurelle, S. Essid, and G. Richard, "Multimodal classification of dance movements using body joint trajectories and step sounds," in *Proc. IEEE International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS)*, 2013, pp. 1–4.

[35] A. Drémeau and S. Essid, "Probabilistic dance performance alignment by fusion of multimodal features," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 3642–3646.



Zhiyao Duan (S'09-M'13) is an assistant professor in the Department of Electrical and Computer Engineering and the Department of Computer Science at the University of Rochester. He received his B.S. in Automation and M.S. in Control Science and Engineering from Tsinghua University, China, in 2004 and 2008, respectively, and received his Ph.D. in Computer Science from Northwestern University in 2013. His research interest is in the broad area of computer audition, i.e., designing computational systems that are capable of understanding sounds, including music, speech, and environmental sounds. He co-presented a tutorial on Automatic Music Transcription at ISMIR 2015. 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.

Slim Essid is a Full Professor at Telecom ParisTech's department of Images, Data & Signals and the head of the Audio Data Analysis and Signal Processing team. His research interests are in machine learning for audio and multimodal data analysis. He received the M.Sc. (D.E.A.) degree in digital communication systems from the École Nationale Supérieure des Télécommunications, Paris, France, in 2002; the Ph.D. degree from the Université Pierre et Marie Curie (UPMC), in 2005; and the habilitation (HDR) degree from UPMC in 2015. He has been involved in various collaborative French and European research projects among which are Quaero, Networks of Excellence FP6-Kspace and FP7-3DLife, and collaborative projects FP7-REVERIE and FP-7 LASIE. He has published over 100 peer-reviewed conference and journal papers with more than 100 distinct co-authors. On a regular basis he serves as a reviewer for various machine learning, signal processing, audio and multimedia conferences and journals, for instance various IEEE transactions, and as an expert for research funding agencies.

Cynthia C. S. Liem (M'16) graduated in Computer Science (BSc, MSc, PhD) from Delft University of Technology, and in Classical Piano Performance (BMus, MMus) from the Royal Conservatoire, The Hague. She currently is an Assistant Professor in the Multimedia Computing Group of Delft University of Technology. Her research focuses on search and recommendation for music and multimedia, fostering the discovery of content which is not trivially on users' radars. She gained industrial experience at Bell Labs Netherlands, Philips Research and Google and is a recipient of several major grants and awards, including the Lucent Global Science Scholarship, Google European Doctoral Fellowship and NWO Veni grant.

Gaël Richard (SM'06-F'17) received the State Engineering degree from Télécom ParisTech, France in 1990, the Ph.D. degree from University of Paris XI, in 1994 in speech synthesis, and the Habilitation à Diriger des Recherches degree from University of Paris XI in September 2001. After the Ph.D. degree, he spent two years at Rutgers University, Piscataway, NJ, in the Speech Processing Group of Prof. J. Flanagan, where he explored innovative approaches for speech production. From 1997 to 2001, he successively worked as project manager for Matra, Bois d'Arcy, France, and for Philips, Montrouge, France. In September 2001, he joined Télécom ParisTech, where he is now a Full Professor in audio signal processing and Head of the Image, Data and Signal department. His research interests are mainly in the field of speech and audio signal processing and include topics such as signal representations and signal models, source separation, machine learning methods for audio/music signals, Music Information Retrieval (MIR) or multimodal audio processing. Co-author of over 200 papers, he is now a fellow of the IEEE.

Gaurav Sharma (S'88–M'96–SM'00–F'13) is a professor at the University of Rochester in the Department of Electrical and Computer Engineering, in the Department of Computer Science and in the Department of Biostatistics and Computational Biology. He received the PhD degree in Electrical and Computer Engineering from North Carolina State University, Raleigh in 1996. From Aug. 1996 through Aug. 2003, he was with Xerox Research and Technology, in Webster, NY, initially as a Member of Research Staff and subsequently at the position of Principal Scientist. Dr. Sharma's research interests include multi-media signal processing, media security, image processing, computer vision, and bioinformatics. Dr. Sharma serves as the Editor-in-Chief for the IEEE Transaction on Image Processing. From 2011 through 2015, he served as the Editor-in-Chief for the Journal of Electronic Imaging. He is the editor of the "Color Imaging Handbook", published by CRC press in 2003. He is a fellow of the IEEE, of SPIE, and of the Society of Imaging Science and Technology (IS&T) and a member of Sigma Xi.