CCIT Report #620 April 2007

Harmony in Motion

Zohar Barzelay and Yoav Y. Schechner
Department of Electrical Engineering
Technion - Israel Inst. Technology
Haifa 32000, ISRAEL
zoharb@tx.technion.ac.il, yoav@ee.technion.ac.il

Abstract

Cross-modal analysis offers information beyond that extracted from individual modalities. Consider a camcorder having a single microphone *in a cocktail-party: it captures several moving visual objects which emit sounds. A task for audio-visual analysis is to identify the number of independent audio-associated visual objects (AVOs), pinpoint the AVOs' spatial locations in the video and isolate each corresponding audio component. Part of these problems were considered by prior studies, which were limited to simple cases, e.g., a single AVO or stationary sounds. We describe an approach that seeks to overcome these challenges. It acknowledges the importance of temporal features that are based on significant changes in each modality. A probabilistic formalism identifies temporal coincidences between these features, yielding cross-modal association and visual localization. This association is of particular benefit in harmonic sounds, as it enables subsequent isolation of each audio source. We demonstrate this in challenging experiments, having multiple, simultaneous highly nonstationary AVOs.*

1. Cross-Modal Analysis

Cross modal analysis is gaining interest in computer vision. Such analysis seeks associations between sources of input data, which have very different natures. Examples of this include registration of images acquired using sensors of different kinds [15], or association of images to text [12], such as in web pages and multimedia subtitles. It also includes audio-visual analysis [23, 25, 29], which has seen a growing expansion of research directions, including lip-reading [7, 13], tracking [24], and spatial localization [6, 9, 17, 18, 22]. This follows evidence of audio-visual cross-modal processing in biology [11].

This work deals with complex scenarios that are sometimes referred to in the literature as a *cocktail party* [9, 13, 26]: multiple sources exist simultaneously in all modalities.

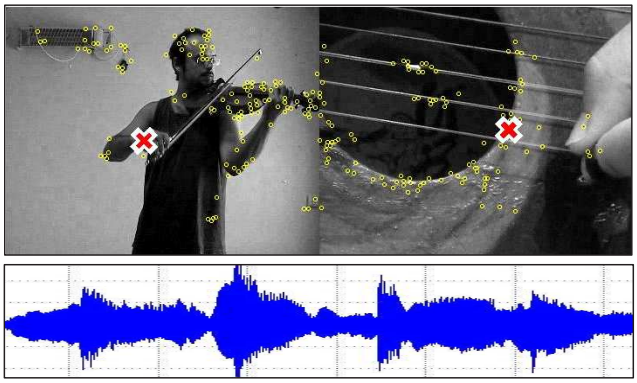

Figure 1. A frame and the audio from the `violin-guitar` movie. A camcorder and a single microphone were used. Two movies were compounded and then processed as a whole. Out of the selected and tracked visual features [Dots], two are automatically associated to the audio [Crosses]: correctly, one per source. The audio mixture is also decoupled to a guitar and a violin. See/hear this via *www.ee.technion.ac.il/∼yoav/research/harmony-in-motion.html*

This inhibits the interpretation of each source. In the domain of audio-visual analysis, a camera views multiple independent objects which move simultaneously, while some of them emanate sounds, which mix. This is depicted in Fig. 1. This paper presents a computer vision approach for dealing with this scenario. The approach has several notable results. First, it automatically *identifies the number of independent sources*. Second, it tracks in the video the multiple *spatial features*, that move in synchrony with each of the (still mixed) sound sources. This is done even in highly non stationary sequences. Third, aided by the video data, it successfully *separates the audio* sources, even though only a *single microphone* is used. This completes the isolation of each contributor in this complex audio-visual scene.

Some of the prior methods considered only parts of these tasks. Others relied on complex audio-visual hardware, such as an array of microphones that are calibrated mu-

tually and with respect to cameras [23, 24]. This yields an approximate spatial localization of audio sources. A single microphone is simpler to set up, but it cannot, on its own, provide audio spatial localization. Hence, locating audio sources using a camera and a single microphone poses a significant computational challenge. In this context, Refs. [17, 22] spatially localize a single audio-associated visual object (AVO). Ref. [6] localizes multiple AVOs if their sounds are repetitive and non-simultaneous. Neither of these studies attempted audio separation. A pioneering exploration of audio separation [9] used complex optimization of mutual information based on Parzen windows. It can automatically localize an AVO if no other sound is present. Results demonstrated in Ref. [29] were mainly of repetitive sounds, without distractions by unrelated moving objects.[1]

Here we propose an approach that better manages obstacles faced by prior methods. It can use the simplest hardware: a single microphone and a camera. Algorithmically, we are inspired by *feature-based* image registration methods, which use spatial *significant changes* (e.g, edges and corners). Analogously, we use as our features the temporal instances of significant changes in each modality. To match the two modalities, we look for cross-modal temporal coincidences of events. Based on a likelihood criterion, the AVOs are then localized and tracked. Following the visual localization of the AVOs, the sound produced by each one is isolated. The algorithm exploits the sparsity of an audio representation we use, and is aided by the essential visual information.

2. Significant Visual and Audio Events

How may we associate two modalities, where each changes in time? Some prior methods use continuous valued variables to represent each modality, e.g., a weighted sum of pixel values. Maximal canonical correlation or mutual information was sought between these variables [9, 14, 17]. That approach is analogous to intensity-based image matching. It implicitly assumes some correlation (possibly nonlinear) between the raw data values in each modality. We do *not* look at the raw data values during the cross-modal association. Rather, here we opt for *feature-based* matching: we seek correspondence between significant features in each modality. Interestingly, there is also evidence that biological neural systems perform cross-modal association based on salient features [10].

Which features are good? Recall a familiar matching problem: that of images. Feature-based image registration focuses on sharp spatial changes (edges and corners) [5], rather than the smooth regions between them. In cross-sensor image matching, Ref. [15] highlighted sharp spatial changes by high-pass filtering. Analogously, in our audio-visual matching problem, we use features having strong *temporal* variations in each of the modalities.

As a pre-processing step, image features (corners etc.) that can be easily locked-on are automatically found [28] and then tracked [3] (see Fig. 1). The result is a set of $N_{\rm v}$ visual features, each indexed by $i \in [1, N_{\rm v}]$. Each feature has a trajectory $\mathbf{v}_i(t) = [x_i(t), y_i(t)]^T$, where t is the temporal index (in units of frames), and x, y are the image coordinates. One of the tasks is to determine if any of these trajectories is of an AVO.

The magnitude of the acceleration $\|\ddot{\mathbf{v}}_i(t)\|$ of feature i is a measure of significant change in its motion speed or direction.[2] We process $\|\ddot{\mathbf{v}}_i(t)\|$ in analogy to the way image gradients are processed to detect edges [28]: we threshold and temporally prune $\|\ddot{\mathbf{v}}_i(t)\|$ to derive a binary vector $\mathbf{v}_i^{\rm on}$

$$v_i^{\rm on}(t) = \begin{cases} 1 & \text{feature } i \text{ has high acceleration at } t \\ 0 & \text{otherwise} \end{cases}, \quad (1)$$

which expresses the *visual onsets* of image feature i. For all features $\{i\}$, the corresponding vectors $\mathbf{v}_i^{\rm on}$ have the same length $N_{\rm f}$, which is the number of frames.

Audio is treated in a similar manner. We focus on *audio onsets* [4]. These are time instances in which a sound commences (over a possible background).[3] Audio onset detection was well studied [2, 19]. It is briefly discussed in Sec. 4.3. This detection process results in a binary vector $\mathbf{a}^{\rm on}$ of length $N_{\rm f}$

$$a^{\rm on}(t) = \begin{cases} 1 & \text{an audio onset takes place at time } t \\ 0 & \text{otherwise} \end{cases}. \quad (2)$$

In the next section, we describe how audio onsets are temporally matched to visual (motion) onsets.

3. A Coincidence-Based Approach

Our cross-modal association is based on a simple assumption. Consider a pair of significant events (onsets): one event per modality. We assume that if both events coincide in time, then they are possibly related. If such a coincidence re-occurs multiple times for the same feature i, then the likelihood of cross-modal correspondence is high. On the other hand, if there are many temporal mismatches, then the matching likelihood is inhibited.

In the specific context of the audio and visual modalities, the choice of audio and visual *onsets* is not arbitrary. These onsets indeed coincide in many scenarios. For example: the sudden acceleration of a guitar string is accompanied by the beginning of the sound of the string; a sudden deceleration

[1] Some studies used an approach motivated by computer-vision in order to make only-audio analysis [16, 27].

[2] A criterion [22] of absolute position $\|\mathbf{v}_i(t)\|$ is sensitive to initialization of the origin of the position coordinates.

[3] We opt not to rely on sound terminations for this purpose, as these are often not sufficiently fast and distinct.

of a hammer hitting a surface is accompanied by noise; the lips of a speaker open as he utters a vowel.

Let us consider for the moment the correspondence of audio and visual onsets in some ideal cases. If just a single AVO exists in the scene, then ideally, there would be a one-to-one audio-visual correspondence, i.e., $\mathbf{v}_i^{\rm on} = \mathbf{a}^{\rm on}$ for a unique feature i. Now, suppose there are several independent AVOs, where the onsets of each object i are exclusive, i.e., they do not coincide with those of any other object. Then, $\sum_{i\in\mathcal{J}} \mathbf{v}_i^{\rm on} = \mathbf{a}^{\rm on}$, where $\mathcal{J}$ is the set of true AVOs. Such ideal cases usually do not occur in practice: there are outliers in both modalities, due to clutter and to imperfect detection of onsets, having false positives and negatives. Thus, we define a matching criterion that is based on a probabilistic argument and enables imperfect matching. It favors coincidences, and penalizes for mismatches. This criterion is then used in a fast iterative algorithm, in the spirit of [21].

3.1. Matching Algorithm

We now describe both the matching criterion, and the iterative algorithm. Define $\mathbf{1}$ as a column vector, all of whose elements equal 1. The criterion we use is

$$\widetilde{L}(i) = 2[(\mathbf{a}^{\rm on})^T \mathbf{v}_i^{\rm on}] - \mathbf{1}^T \mathbf{v}_i^{\rm on} \ . \qquad (3)$$

In Sec. 3.2, we show that Eq. (3) is equivalent to a matching likelihood. Out of all the visual features $i \in [1, N_{\rm v}]$, $\widetilde{L}(i)$ should be maximized by the one corresponding to an AVO. Let us first gain some intuition into Eq. (3). The number of visual onsets of feature i that coincide with audio onsets is $(\mathbf{a}^{\rm on})^T \mathbf{v}_i^{\rm on}$, since $\mathbf{v}_i^{\rm on}$ and $\mathbf{a}^{\rm on}$ are binary. Moreover, $(\mathbf{1} - \mathbf{a}^{\rm on})^T \mathbf{v}_i^{\rm on}$ is the number of visual onsets of i, that are inconsistent with $\mathbf{a}^{\rm on}$. Therefore, Eq. (3) favors coincidences while penalizing for the inconsistencies. We calculate Eq. (3) for each visual feature i. The one corresponding to the highest value of $\widetilde{L}$ is a *candidate* AVO. Let its index be $\hat{i}$. This candidate is classified as an AVO, if its likelihood $\widetilde{L}(\hat{i})$ is above a threshold. Note that by definition, $\widetilde{L}(i) \leq \widetilde{L}(\hat{i})$ for all i. Hence, if $\widetilde{L}(\hat{i})$ is below the threshold, neither $\hat{i}$ nor any other feature is an AVO.

At this stage, a major goal has been accomplished. Once feature $\hat{i}$ is classified as an AVO, it indicates audio-visual association not only at onsets, but for the *entire trajectory* $\mathbf{v}_{\hat{i}}(t)$, for all t. Hence, it marks a specific tracked feature as an AVO, and this AVO is visually traced continuously throughout the sequence. For example, consider the `violin-guitar` sequence, one of whose frames is shown in Fig. 1. It was recorded by a simple camcorder and using a single microphone.[4] Onsets were obtained as described in Sec. 2. Then, the visual feature that maximized Eq. (3) was the *hand of the violin player*. Its detection and tracking were automatic.

Now, the audio onsets that correspond to AVO $\hat{i}$ are given by the vector $\mathbf{m}^{\rm on} = \mathbf{a}^{\rm on} \bullet \mathbf{v}_{\hat{i}}^{\rm on}$, where $\bullet$ denotes the logical-AND operation per element. Let us eliminate these corresponding onsets from $\mathbf{a}^{\rm on}$. The *residual* audio onsets are represented by $\mathbf{a}_1^{\rm on} \equiv \mathbf{a}^{\rm on} - \mathbf{m}^{\rm on}$. The vector $\mathbf{a}_1^{\rm on}$ becomes the input for a new iteration: it is used in Eq. (3), instead of $\mathbf{a}^{\rm on}$. Consequently, a new candidate AVO is found, this time optimizing the match to the residual audio vector $\mathbf{a}_1^{\rm on}$.

This process re-iterates. It stops automatically when a candidate fails to be classified as an AVO. This indicates that the remaining visual features cannot "explain" the residual audio onset vector. The main parameter in this method is the mentioned classification threshold of the AVO. We set it to $\widetilde{L}(\hat{i}) = 0$. If $\widetilde{L}(\hat{i}) < 0$, it means that more than half of the onsets in $\mathbf{v}_{\hat{i}}^{\rm on}$ are not matched by audio ones. In other words, most of the significant visual events of i are not accompanied by any new sound. We thus interpret this object as *not* audio-associated.

To recap, our matching algorithm is

Input: vectors $\{\mathbf{v}_i^{\rm on}\}, \mathbf{a}^{\rm on}$

0. Initalize: $l = 0,\ \mathbf{a}_0^{\rm on} = \mathbf{a}^{\rm on},\ \mathbf{m}_0^{\rm on} = \mathbf{0}$.
1. Iterate
2. $\quad l = l + 1$
3. $\quad \mathbf{a}_l^{\rm on} = \mathbf{a}_{l-1}^{\rm on} - \mathbf{m}_{l-1}^{\rm on}$
4. $\quad \hat{i}_l = \arg\max_i \{2(\mathbf{a}_l^{\rm on})^T \mathbf{v}_i^{\rm on} - \mathbf{1}^T \mathbf{v}_i^{\rm on}\}$
5. $\quad$ If $\{(\mathbf{a}_l^{\rm on})^T \mathbf{v}_{\hat{i}}^{\rm on} \geq \frac{1}{2}\mathbf{1}^T \mathbf{v}_{\hat{i}}^{\rm on}\}$, then
6. $\quad\quad \mathbf{m}_l^{\rm on} = \mathbf{v}_{\hat{i}}^{\rm on} \bullet \mathbf{a}_l^{\rm on}$
7. $\quad$ else
8. $\quad\quad$ quit

Output:

- The estimated number of independent AVOs is $\widehat{|\mathcal{J}|} = l - 1$.
- A list of AVOs and corresponding audio onsets vectors $\{\hat{i}_l, \mathbf{m}_l^{\rm on}\}$.

Here $\mathbf{0}$ is a column vector, all of whose elements are null. Note that the output $\widehat{|\mathcal{J}|}$ accomplishes another goal of this paper: the automatic estimation of the *number of independent AVOs*. This algorithm is fast (linear): $\approx |\mathcal{J}|$ iterations, each having $\mathcal{O}(N_{\rm f} N_{\rm v})$ calculations.

In the `violin-guitar` sequence mentioned above, this algorithm automatically detected that there are two independent AVOs: the *guitar string*, and the hand of the *violin player* (marked as crosses in Fig.1). Note that in this sequence, the sound and motions of the guitar pose a distraction for the violin, and vice versa. However, the algorithm correctly identified the two AVOs.

[4] The sampling parameters of the audio and video are given in Sec. 4.1.

3.2. Likelihood Interpretation

Here we show that Eq. (3) can be interpreted as equivalent to the matching likelihood of feature i. Let $v_i(t)$ be a random variable which follows the probability law

$$Pr\left[v_i^{\rm on}(t)|a^{\rm on}(t)\right] = \left\{ \begin{array}{ll} p & , v_i^{\rm on}(t) = a^{\rm on}(t) \\ 1-p & , v_i^{\rm on}(t) \neq a^{\rm on}(t) \end{array} \right. . \quad (4)$$

Assuming that the elements $a^{\rm on}(t)$ are statistically independent of each other, the matching likelihood of a vector $\mathbf{v}_i^{\rm on}$ is

$$L(i) = \prod_{t=1}^{N_{\rm f}} Pr\left[v_i^{\rm on}(t)|a^{\rm on}(t)\right] . \quad (5)$$

Denote by $N_{\rm agree}$ the number of time instances in which $a^{\rm on}(t) = v_i^{\rm on}(t)$. From Eqs. (4,5),

$$L(i) = p^{N_{\rm agree}} \cdot (1-p)^{(N_{\rm f} - N_{\rm agree})} . \quad (6)$$

Both $\mathbf{a}^{\rm on}$ and $\mathbf{v}_i^{\rm on}$ are binary, hence the number of time instances in which both are 1 is $(\mathbf{a}^{\rm on})^T \mathbf{v}_i^{\rm on}$. The number of instances in which both are 0 is $\left(\mathbf{1}-\mathbf{a}^{\rm on}\right)^T \left(\mathbf{1}-\mathbf{v}_i^{\rm on}\right)$, hence

$$N_{\rm agree} = (\mathbf{a}^{\rm on})^T \mathbf{v}_i^{\rm on} + \left(\mathbf{1}-\mathbf{a}^{\rm on}\right)^T \left(\mathbf{1}-\mathbf{v}_i^{\rm on}\right) . \quad (7)$$

Plugging Eq. (7) in Eq. (6) and re-arranging terms,

$$\log\left[L(i)\right] = N_{\rm f} \log(1-p) + \\ + \left[(\mathbf{a}^{\rm on})^T \mathbf{v}_i^{\rm on} + \left(\mathbf{1}-\mathbf{a}^{\rm on}\right)^T \left(\mathbf{1}-\mathbf{v}_i^{\rm on}\right) \right] \log\left(\frac{p}{1-p}\right) . \quad (8)$$

We seek the feature i whose vector $\mathbf{v}_i^{\rm on}$ maximizes $L(i)$. Thus, we eliminate terms that do not depend on $\mathbf{v}_i^{\rm on}$. This yields an equivalent objective function of i,

$$\widetilde{L}(i) = \left\{ 2\left[(\mathbf{a}^{\rm on})^T \mathbf{v}_i\right] - \mathbf{1}^T \mathbf{v}_i^{\rm on} \right\} \log\left(\frac{p}{1-p}\right) . \quad (9)$$

It is reasonable to assume that if feature i is an AVO, then it has more onset coincidences than mismatches. Consequently, we may assume that $p > 0.5$. Hence, $\log\left[p/\left(1-p\right)\right] > 0$. Consequently, $\widetilde{L}(i)$ is maximized when Eq. (3) is maximized.

4. Audio Processing and Isolation

Up to now, we derived the method for establishing the AVOs in the scene. The described matching algorithm outputs a set of AVOs, each with its vector of corresponding audio onsets: $\{\hat{i}_l, \mathbf{m}_l^{\rm on}\}$. This vector of audio onsets points to the time instances in which the sounds of the AVO commence. In order to isolate the soundtrack of the AVO, we need to isolate each of these sounds. How do we isolate a single sound from a mixture, given only its onset time? This is described next.

4.1. Binary Masking

Audio isolation is based on Fourier analysis. Let $s(n)$ denote the recorded sound signal, typically sampled much faster than the video. Here n is a discrete sample index of the sound. This signal is analyzed in short temporal windows w, each being $N_{\rm w}$-samples long. Consecutive windows are shifted by M samples. In our experiments, the audio was sampled at 16 kHz, and analyzed with a Hamming window of $80msec$, equivalent to $N_{\rm w} = 1280$. Our use of $M = N_{\rm w}/2$ ensured synchronicity of the windows with the video frame rate ($25Hz$). Recalling that t is the frame (time) index, the short-time Fourier transform of $s(n)$ is

$$F(t,f) = \sum_{n=0}^{N_{\rm w}-1} s(n+tM) w(n) e^{-j(2\pi/N_{\rm w})nf} , \quad (10)$$

where f is the frequency index. The spectrogram is $|F(t,f)|^2$. See for example the spectrograms in Fig. 2.

As seen in Fig. 2, the energy of each distinct sound lies in a set Γ of time-frequency bins $\{(t,f)\}$. A common assumption [1, 26, 31] is that if there are other sound sources, then the energy distribution in $\{(t,f)\}$ of these disturbances has only little overlap with the bins in Γ. This assumption is based on the sparsity of typical sounds, particularly *harmonic* ones, in spectrograms. Consequently, a sound of interest can be enhanced by maintaining the values of $F(t,f)$ in Γ, while nulling the other bins. This *binary masking* forms the basis for many methods [1, 26, 31]. The masked $F(t,f)$ is then transformed back [26] to a sound signal $\tilde{s}(n)$.

How is the set Γ of a sound characterized? In an harmonic sound, the acoustic energy lies in a *pitch frequency* f_0 and in integer multiples of this frequency (*harmonies*). This is seen in the spectrogram of a violin at the bottom-right of Fig. 2. We note that f_0 of a distinct sound may drift in time, i.e., $f_0 = f_0(t)$, as shown in the left panel of Fig. 3 (speech). Thus,

$$\Gamma = \{(t\,, f_0(t)k)\} , \quad (11)$$

where $k \in \mathbb{N}^+$. The set defined in Eq. (11) is bounded temporally by $t \in [t^{\rm on}, t^{\rm off}]$, where $t^{\rm on}$ is the *onset* of this sound. Here $t^{\rm off}$ is the offset instance, in which the sound is considered as terminated or effectively faded. Consequently, *given only the onset instance* $t^{\rm on}$, we determine Γ by detecting $f_0(t^{\rm on})$, and then tracking $f_0(t)$ in $t \in [t^{\rm on}, t^{\rm off}]$. The detection and tracking procedures are described next.

4.2. Directional Derivative of the Spectrogram

Ref. [8] describes a method for estimating $f_0(t)$ of a *single* sound using the amplitude $A(t,f) = |F(t,f)|$ as input.

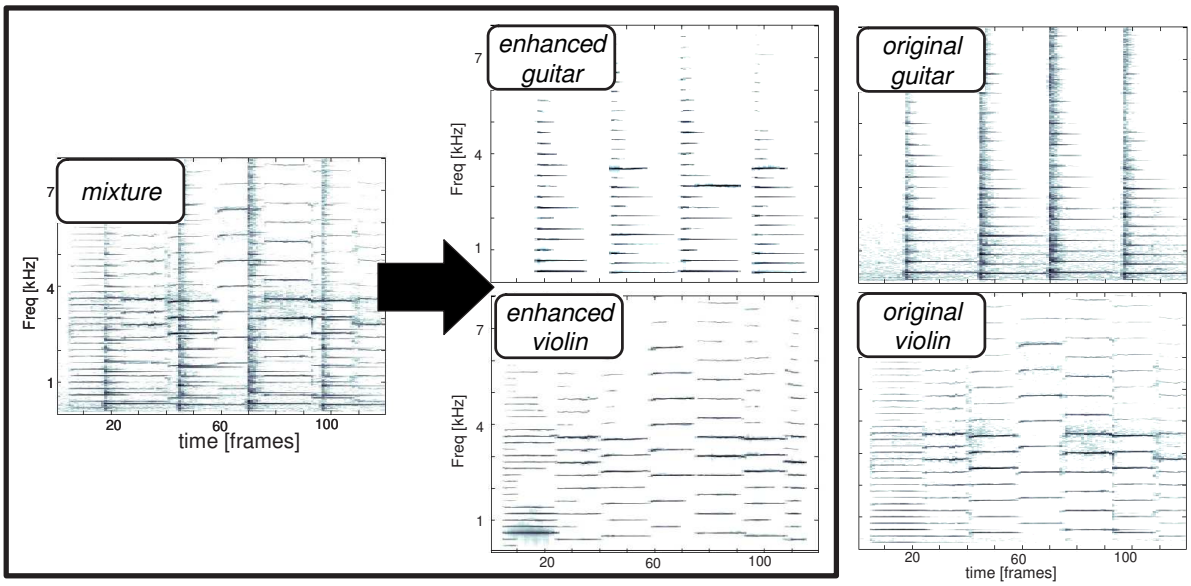

Figure 2. Spectrograms corresponding to the `violin-guitar` sequence. Darker points in each plot indicate a higher energy content, as a function of t and f. Based on *visual* data, the audio components of the violin and guitar were automatically separated from a soundtrack, which had been recorded by a *single* microphone. [Right] The true components of each instrument, acquired separately. You may listen to the results via the link *www.ee.technion.ac.il/∼yoav/research/harmony-in-motion.html*

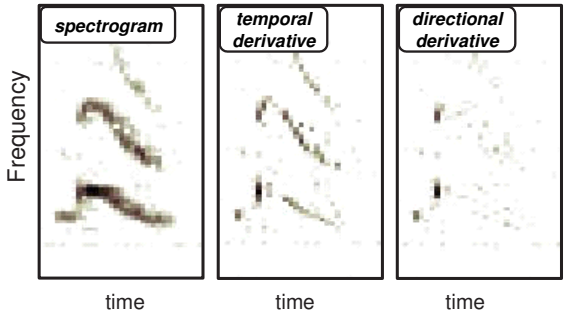

Figure 3. [Left] A section of a spectrogram (female speaker) exhibiting a frequency drift. [Middle] A temporal derivative (Eq. 12) results in high values through the entire sound duration. [Right] The directional derivative (Eq. 14) handles the frequency drift well. Resulting high values occur mainly at the onset.

In our case, however, multiple sounds coexist.[5] We would be able to use Ref. [8] per sound source, if we remove most of the energy of the other sounds. How can we achieve this?

Here we exploit again the onsets that had been detected. The sound of interest is the one commencing at $t^{\rm on}$. Thus, the disturbing audio at $t^{\rm on}$ is assumed by us to have commenced *prior* to $t^{\rm on}$. These disturbing sounds linger from the past, and hence, they can be eliminated by comparing the audio components at $t = t^{\rm on}$ to those at $t < t^{\rm on}$, particularly at $t = t^{\rm on} - 1$. Specifically, the relative temporal derivative

$$D(t,f) = \frac{A(t,f) - A(t-1,f)}{A(t-1,f)} \tag{12}$$

emphasizes an increase of amplitude in frequency bins that have been quiet (no sound) just before t.

As a practical criterion, however, Eq. (12) is not robust. The reason is that sounds which have commenced prior to t may have a slow frequency *drift* (Fig. 3). This poses a problem for Eq. (12), which is based solely on a temporal comparison per frequency channel. Drift results in high values of Eq. (12) in some frequencies f, even if no new sound actually commences around (t, f), as seen in Fig. 3. To overcome this, we perform a *directional* derivative in the time-frequency (spectrogram) domain.[6] It fits neighboring bands at each instance, hence tracking the drift. Consider a small frequency range Ω around f. In analogy to image alignment, *frequency alignment* at time t is obtained by

$$f^{\rm aligned} = \arg\min_{f_z \in \Omega} |A(t^{\rm on}, f) - A(t^{\rm on} - 1, f_z)| \ . \tag{13}$$

Then $f^{\rm aligned}$ at $t-1$ corresponds to f at t, partially correcting the drift. The map

$$\widetilde{D}(t,f) = \frac{A(t,f) - A(t-1, f^{\rm aligned})}{A(t-1, f^{\rm aligned})} \tag{14}$$

is indeed much less sensitive to drift, and is responsive to true onsets (Fig 3). The map $\widetilde{D}_+(t,f) = \max\{0, \widetilde{D}(t,f)\}$

[5] We do not know the number of sound sources in the scene: in addition to the visual AVOs there can be audio sources of objects out of view. Hence, we cannot use [1, 20, 30].

[6] Treating the spectrogram as a two-dimensional signal (image) was suggested in [16].

maintains the onset response, while ignoring amplitude decrease caused by fade-outs.

We may now use $\widetilde{D}_+(t^{\rm on}, f)$ as input to the algorithm of Ref. [8]. This yields the pitch f_0 at $t^{\rm on}$. Following the detection of $f_0(t^{\rm on})$, it is tracked during $t \geq t^{\rm on}$, until $t^{\rm off}$. This procedure for tracking $f_0(t)$ and for determining $t^{\rm off}$ is described below, in Sec. 4.4.

As described earlier in Sec. 4, this procedure and binary masking are repeated for *each* of the onsets of the AVO. The isolated sounds per onset are then concatenated to a single soundtrack. This effectively yields the isolated soundtrack of the AVO. As an example, Fig. 2 illustrates the results obtained in the `violin-guitar` sequence.

4.3. Audio Onsets Detection

Past sections relied on prior detection of audio onsets. Methods for this detection have been extensively studied [2]. Here we describe our particular method. Our criterion for significant signal increase is simply $o(t) = \sum_f \widetilde{D}_+(t, f)$. It is similar to a criterion used in Ref. [19], but is more robust, since it suppresses lingering sounds. As in Ref. [2], the binary onset vector $\mathbf{a}^{\rm on}$ is a result of thresholding of $o(t)$.

4.4. Pitch Tracking

In Sec. 4.2 we described how the pitch frequency $f_0(t^{\rm on})$ of a sound commencing at $t^{\rm on}$ is detected. We now describe how we track $f_0(t)$, and how the instance of its termination, namely $t^{\rm off}$, is established.

Given the detected pitch frequency at $f_0(t)$, we wish to establish $f_0(t+1)$. It is known to lie in a frequency neighborhood Ω of $f_0(t)$, since the pitch frequency changes slowly through time [30]. Recall that an harmonic sound contains multiples of the pitch frequency (the "harmonies"). Denote the set of harmonies at time t by $\mathcal{K}(t) = [1, \ldots, K]$. The estimated frequency $f_0(t+1)$ may be found as the one whose harmonies capture most of the energy of the harmonic signal

$$f_0(t+1) = \arg\max_{f \in \Omega} \sum_{k \in \mathcal{K}(t)} \left[A\left(t+1, f \cdot k\right)\right]^2 , \qquad (15)$$

where $A(t, f) = |F(t, f)|$.

Eq. (15), however, does not take into account the existence of other sources in the mixture. Disrupting sounds of high energy may be present around the harmonies $(t+1, f \cdot k)$ for $f \in \Omega$, and may distort the detection of $f_0(t+1)$. To reduce the effect of these sounds, we do not use the amplitude of the harmonies $A(t+1, f \cdot k)$ in Eq. (15). Rather, we use $\log[A\left(t+1, f \cdot k\right)]$. This effectively causes the estimate in Eq. (15) to use many frequency bins for the estimation of $f_0(t+1)$, and significantly reduces the error induced by a few noisy ones.

Recall that the pitch is tracked in order to identify the set Γ of time-frequency bins in which an harmonic sound lies. We now go into the details of how to establish Γ. According to Eq. (11), Γ should contain all of the harmonies of the pitch frequency, for $t \in [t^{\rm on}, t^{\rm off}]$. However, we may make Γ tighter, by not including all of the harmonies at each instance. Harmonies may be removed due to two reasons: First, there may be some harmonies in which a strong interference exists. Second, some harmonies may fade out. To identify these cases, we inspect the relative amplitude temporal change

$$\frac{A\left[t+1, f_0(t+1) \cdot k\right]}{A\left[t, f_0(t) \cdot k\right]} \qquad (16)$$

for each of the harmonies $k \in \mathcal{K}(t)$. When Eq. (16) for some k_0 exceeds a threshold , we deduce that a disrupting sound has now entered the frequency bin $[t+1, f_0(t+1) \cdot k_0]$. We therefore remove k_0 from $\mathcal{K}(t+1)$. Similarly, when Eq. (16) goes below a threshold, we deduce that the harmony k_0 has faded and set $\mathcal{K}(t+1) = \mathcal{K}(t) \backslash k$.

We initialize the tracking process with $f_0(t^{\rm on})$ and $\mathcal{K}(t^{\rm on}) = [1, \ldots, K]$, and iterate it through time. When the number of active harmonies $|\mathcal{K}(t)|$ drops below a certain threshold, this indicates the termination of the signal at time $t^{\rm off}$. The domain Γ that the tracked sound occupies in $t \in [t^{\rm on}, t^{\rm off}]$ is composed from the active harmonies at each instance t. Formally :

$$\Gamma = \{(t, f_0(t) \cdot k\} , \qquad (17)$$

where $t \in [t^{\rm on}, t^{\rm off}]$ and $k \in \mathcal{K}(t)$.

5. Results

In our experiments we compound separately-recorded movies (e.g., a violin sequence and a guitar sequence) into a single video.[7] Such a procedure is a common practice in single-micrhopone audio-separation studies [1, 13, 26], since it provides access to the audio ground-truth data. This allows quantitative assessment of the quality of audio isolation, as we describe below.

The cross-modal method has several parameters, such as the spectrogram window size (Sec. 4.1) and the temporal-resolution of coincidences, discussed below in Sec. 6. Other parameters are derived from the analogy of our approach to image edge-detection. Such a detection usually involves setting of an edge scale, a threshold of significant change,

[7] Compounding individiual scenes does *not* simplify the experiments relative to a simultaneous recording of AVOs. The reverberations of each source are preserved after sampling and compounding, since these are linear operations. For the same reason, the individual sources still interfere with each other, regardless of whether they are recorded separately or simultaneously.

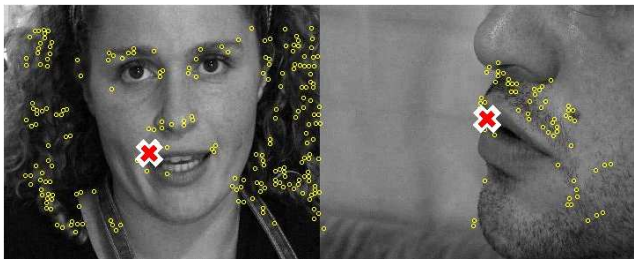

Figure 4. A frame from the speakers sequence. Out of the selected and tracked visual features [Dots], two are automatically associated to the audio [Crosses]: correctly, one per source. The audio mixture is also decoupled to two separate speakers.

and a proximity parameter for pruning [28]. Such parameters influence the results, and thus should be tuned.

All the video/audio material described here is available in the supplementary material, and through *www.ee.technion.ac.il/∼yoav/research/harmony-in-motion.html.* The first experiment is the violin-guitar sequence already described in Figs. 1 and 2. The second experiment is the speakers sequence, which has simultaneous speech by people. The pitch of each speaker drifts significantly in time. A sample frame is shown in Fig. 4, where crosses indicate the automatically-detected AVOs. The features detected correspond to the lips of each speaker. The corresponding results of audio isolation for each speaker in this minor "cocktail party" are shown in Fig. 5. An additional experiment contains two *identical* instruments playing different tunes simultaneously (Fig. 6). The data and the separation results are available through the above-mentioned link.

Quantitative Recovery Criteria

We quantify the quality of the audio-isolation in the experiments by criteria described in Ref. [31]. These measures utilize our access to the ground-truth audio data. The first measure evaluates the improvement of the signal-to-interference-ratio (SIR). The second measure calculates the preserved-signal-ratio (PSR), which is the amount of signal energy that is preserved in the isolation process. For further details about these criteria see Ref. [31].

In the violin-guitar sequence, the SIR of the violin is significantly improved by 17.4 dB. The SIR of the guitar improves by 4.4 dB. Some of the harmonies of the violin coincide with those of the guitar. Consequently, the isolated guitar erroneously contains some components of the violin, creating squeaking sounds. The PSRs of the violin and of the guitar are 0.89 and 0.78, respectively. In the speakers sequence, the SIR improvements of the male and of the female speakers are significant: 12.3 dB and 15.6 dB, respectively. The corresponding PSRs are 0.64 and 0.51. Even though The PSR of the female speaker indicates loss of almost 50% of the speech energy, her isolated speech is very intelligible.

6. Limitation: Temporal Resolution

The approach described in this paper has limits. In particular, its temporal resolution is finite. As in any system, the terms *coincidence* and *simultaneous* are meaningful only within a tolerance range of time. In the real-world, coincidence of two events at an infinitesimal temporal range has just an infinitesimal probability. Thus, correspondence between two modalities can be established only up to a finite tolerance range. Our approach is no exception. Specifically, each onset is determined up to a finite resolution, and audio-visual onset coincidence should be allowed to take place within a finite time window. This limits the temporal resolution of coincidence detection. In our experiments, we considered coincidences if a visual onset occurred within $\approx 1/8sec$ of an audio onset.

7. Relation to Audio-Only Methods

This computer vision work yields visual detection and tracking of AVOs. In addition, it utilizes the visual data for audio isolation. This raises the question of how audio-only (unrelated to vision) methods can benefit from such a framework. Some audio-separation methods are based on microphone arrays [31] having a sufficiently wide baseline. Other methods, which use a single microphone, generally separate audio based on training on specific classes of sources, particularly speech and typical potential disturbances [1]. Such methods may succeed in enhancing continuous sounds, but may fail to *group* discontinuous sounds correctly to a single stream. This is the case when the audio-characteristics of the different sources are similar to one another, for instance, two speakers with close by pitch-frequencies. In such a setting, the visual data becomes very helpful, as it provides a complementary cue for grouping of discontinuous sounds. In our framework, sounds are grouped together according to the coincidence of their onsets with visual onsets of an AVO. Consequently, incorporating our approach with traditional audio separation methods may prove to be worthy.

8. Discussion

We presented a novel approach for cross-modal analysis. It is based on instances of significant change in each modality. Our approach handled complex audio-visual scenarios in experiments, where sounds overlapped and visual motions existed simultaneously. The approach yields a set of distinct visual features, with associated isolated sounds. It does *not* require training. Thus, it is applicable to a wide range of AVOs (not limited to speech or specific instruments). We believe that this general capacity is not limited to the audio-visual domain. Rather, it may be applicable to associating between other types of data. We hypothesize

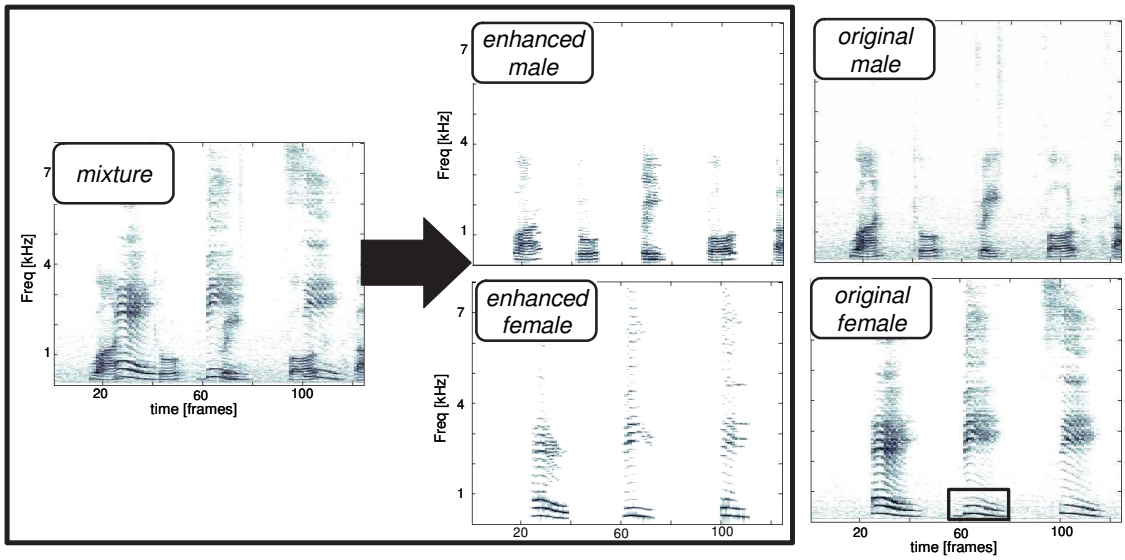

Figure 5. Spectrograms corresponding to the `speakers` sequence. Based on *visual* data, the audio components of each of the speakers were automatically separated from a single soundtrack. A small section in the original spectrogram of the female is marked. It was zoomed-in in Fig. 3.

Figure 6. A frame from the `dual-violin` movie.

that this may be potentially useful, for instance, in associating subtitles to multimedia (images, movies) databases, or in associating macro-economic events.

Sec. 5 described the need for setting parameters, in analogy to parameters of image edge-detection. It would be preferable to establish methods for automatic adaptation of such parameters to the observed audio-visual scene.

As the number of independent AVOs in the scene increases (a dense cocktail party), it may be expected that our method will eventually break down. It is worth studying the breaking point of our approach. Furthermore, it will be beneficial to construct robust algorithms based on the cross-modal coincidence principle. This would enable the handling of dense scenarios of increased complexity.

Acknowledgements

We thank Danny Stryian, Maayan Merhav and Einav Namer for participating in the experiments. Yoav Schechner is a Landau Fellow - supported by the Taub Foundation, and an Alon Fellow. The work was conducted at the Ollendorff Center in the Elect. Eng. Dept. at the Technion. Minerva is funded through the BMBF.

References

[1] F. R. Bach and M. I. Jordan. Blind one-microphone speech separation: A spectral learning approach. *Proc. NIPS* (2004).

[2] J. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. Sandler. A tutorial on onset detection in music signals. In *IEEE Trans. Speech and Audio Process.*, 5:1035–1047 (2005).

[3] S. Birchfield. An implementation of the Kanade-Lucas-Tomasi feature tracker. Available at http://www.ces.clemson.edu/∼stb/klt/.

[4] A. Bregman. *Auditory Scene Analysis*. Cambridge, USA: MIT Press (1990).

[5] L. S. Brown. Survey of image registration techniques. *ACM Comput. Surv.*, 24:325–376 (1992).

[6] J. Chen, T. Mukai, Y. Takeuchi, T. Matsumoto, H. Kudo, T. Yamamura, and N. Ohnishi. Relating audio-visual events caused by multiple movements: in the case of entire object movement. *Proc. Inf. Fusion*, pp. 213– 219 (2002).

[7] T. Choudhury, J. Rehg, V. Pavlovic, and A. Pentland. Boosting and structure learning in dynamic bayesian networks for audio-visual speaker detection. In *Proc. ICPR.*, vol. 3, pp. 789– 794 (2002).

[8] P. Cuadra, A. Master, and C. Sapp Efficient pitch detection techniques for interactive music using harmonic model. *Proc. ICMI*, (2001).

[9] T. Darrell, J. W. Fisher, , P. A. Viola, and W. T. Freeman. Audio-visual segmentation and the cocktail party effect. In *Proc. ICMI 2000*, pp. 1611-3349 (2000).

[10] W. Fujisaki and S. Nishida. Temporal frequency characteristics of synchrony-asynchrony discrimination of audio-visual signals. *J. Exp. Brain Res.*, 166:455–464 (2005).

[11] Y. Gutfreund, W. Zheng, and E. I. Knudsen. Gated visual input to the central auditory system. *Science* 297:1556 - 1559 (2002).

[12] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16:2639–2664 (2004).

[13] J. Hershey and M. Casey. Audio-visual sound separation via hidden markov models. *Proc. NIPS*, pp. 1173–1180 (2001).

[14] J. Hershey and J. R. Movellan. Audio vision: Using audio-visual synchrony to locate sounds. *Proc. NIPS*, pp. 813–819 (1999).

[15] M. Irani and P. Anandan. Robust multi-sensor image alignment. *Proc. IEEE ICCV*,pp. 959–966 (1998).

[16] Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. *Proc. IEEE CVPR*, vol. 1, pp. 597– 604 (2005).

[17] E. Kidron, Y. Y. Schechner, and M. Elad. Pixels that sound. *Proc. IEEE CVPR*, vol. 1, pp. 88–95 (2005).

[18] E. Kidron, Y. Y. Schechner, and M. Elad. Cross-modal localization via sparsity. *IEEE Trans. Signal Processing*, 55:1390–1404 (2007)

[19] A. Klapuri. Sound onset detection by applying psychoacoustic knowledge. *Proc. IEEE ICASSP*, vol. 6, pp. 3089–3092 (1999).

[20] A. Klapuri. A perceptually motivated multiple-f0 estimation method. *Proc. IEEE Worksh. App. Sig. Proc. to Audio & Acoustics*, pp. 291- 294, (2005).

[21] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Proc. IEEE Trans. Sig. Process.*, 41:3397–3415 (1993).

[22] G. Monaci and P. Vandergheynst. Audiovisual gestalts. *Proc. IEEE Worksh. Percept. Org. in Comp. Vis.* (2006).

[23] K. Nakadai, K. Hidai, H. Okuno, and H. Kitano. Real-time speaker localization and speech separation by audio-visual integration. *IEEE Conf. Robotics & Auto.*, vol. 1, pp. 1043–1049 (2002).

[24] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92:495–513 (2004).

[25] S. Rajaram, A. Nefian, and T. Huang. Bayesian separation of audio-visual speech sources. *Proc. IEEE ICAASP*, vol. 5, pp. 657–660 (2004).

[26] S. T. Roweis. One microphone source separation. *Proc. NIPS*, pp. 793–799 (2001).

[27] B. Sarel and M. Irani. Separating transparent layers of repetitive dynamic behaviors. *Proc. IEEE ICCV*, vol. 1, pp. 26–32 (2005).

[28] J. Shi and C. Tomasi. Good features to track. *Proc. IEEE CVPR*, pp. 593–600 (1994).

[29] P. Smaragdis and M. Casey. Audio/visual independent components. *Proc. ICA*, pp. 709–714 (2003).

[30] M. Wu, D. Wang, and G. Brown. A multi-pitch tracking algorithm for noisy speech. *Proc. IEEE ICAASP*, vol. 2, pp. 229–241 (2002).

[31] O. Yilmaz and S. Rickard. Blind separation of speech mixtures via time-frequency masking. *IEEE Trans. Sig. Process.*, 52:1830– 1847 (2004).