IEEE Trans. Multimedia, Vol. 12, No. 2, pp. 108-120 (2010)

Onsets Coincidence for Cross-Modal Analysis

Zohar Barzelay and Yoav Y. Schechner

***Abstract*—Cross-modal analysis offers information beyond one extracted from individual modalities. Consider a non-trivial scene, that includes several moving visual objects, of which some emit sounds. The scene is sensed by a camcorder having a *single microphone*. A task for audio-visual analysis is to assess the number of independent audio-associated visual objects (AVOs), pinpoint the AVOs' spatial locations in the video and isolate each corresponding audio component. We describe an approach that helps handling this challenge. The approach does not inspect the low-level data. Rather, it acknowledges the importance of mid-level features in each modality, which are based on significant temporal changes in each modality. A probabilistic formalism identifies temporal coincidences between these features, yielding cross-modal association and visual localization. This association is further utilized in order to isolate sounds that correspond to each of the localized visual features. This is of particular benefit in harmonic sounds, as it enables subsequent isolation of each audio source. We demonstrate this approach in challenging experiments. In these experiments, multiple objects move simultaneously, creating motion distractions for one another, and produce simultaneous sounds which mix.**

I. Introduction

CROSS modal analysis draws a growing interest both computer-vision and in the signal-processing communities. Such analysis aims to deal with scenarios in which the available data is multi-modal by nature. Consequently, co-processing of different modalities is expected to synergize tasks that are traditionally faced separately. Moreover: such co-processing enables new tasks, which cannot be accomplished in a single-modal context, e.g. visually localizing an object producing sound. Indeed, audio-visual analysis [1]-[3] has seen a growing expansion of research directions, including lip-reading [4], [5], tracking [6], and spatial localization [7]-[10]. This also follows evidence of audio-visual cross-modal processing in biology [11].

Let us focus on scenarios that are referred to in the literature as a *cocktail party* [5], [8], [12]. Multiple objects exist simultaneously in multiple modalities. This simultaneity inhibits the interpretation of each component (e.g. sound component). In a simple everyday example, a camera views multiple independent objects, e.g.: lips, music instruments, etc. The objects move simultaneously, and some of them emit sounds. In the microphone that records the scene, all these sounds mix. This paper presents several principles that are very useful for dealing with this kind of complex scenarios. This approach is motivated by both computer-vision studies [13], and studies of the human auditory system [14]. In both fields, studies have shown the importance of *significant synchronous changes*. However, such events had rarely been inspected in a cross-modal context. The principles we describe here yield several notable results. First, the number of independent sources is identified. Second, these principles enable tracking in the video of 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, the audio sources are successfully separated, even though only a *single microphone* is used. This completes the isolation of each contributor in this complex audio-visual scene, as depicted in Fig. 1. Some of the prior methods considered parts of these tasks. Others relied on complex audio-visual hardware, such as an array of microphones that are calibrated mutually and with respect to cameras [1], [6], [15]. This yields an approximate spatial localization of audio sources. A single microphone is simpler to set up, but it cannot, on its own, provide accurate audio spatial localization. Hence, locating audio sources using a camera and a single microphone poses a significant computational challenge. In this context, Refs. [9], [10] spatially localize a single audio-associated visual object (AVO). Refs. [7], [16] localize multiple AVOs if their sound and motion are repetitive. Neither of these studies attempted audio separation. A pioneering exploration of audio separation [8] 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. [3] were mainly of repetitive sounds, without distractions

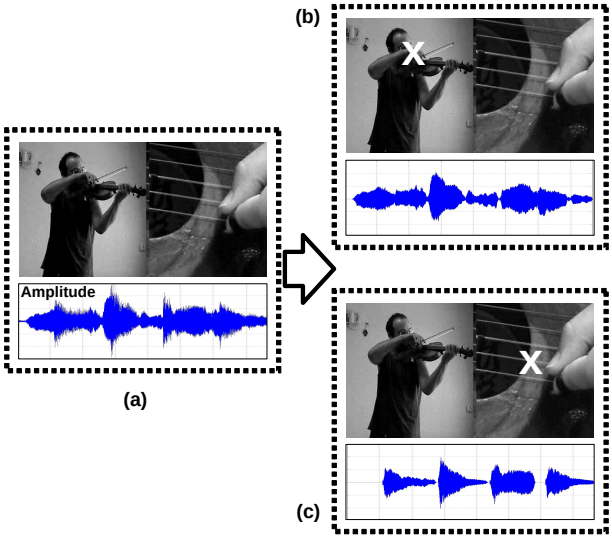

Fig. 1. (a) A frame and the audio of a recorded scene. The single-microphone soundtrack containing a mixture of sources. (b)+(c) Automatic localization of visual objects that correspond to the sound. The audio components of each source are extracted from the soundtrack.

Z. Barzelay and Y. Y. Schechner are with the Department of Electrical Engineering, Technion-Israel Inst. Technology, Haifa 32000, Israel (e-mail: zohar.barzelay@gmail.com; yoav@ee.technion.ac.il)

by unrelated moving objects.[1]

The approach we propose here appears to better manage obstacles faced by prior methods. It can use the simplest hardware: a single microphone and a camera. To match the two modalities, we look for cross-modal temporal coincidences of events having significant change. We formulate a likelihood criterion, and use it in a framework that sequentially localizes independent AVOs. Consequently, continuous audio-visual association and tracking is achieved along with isolation of the sounds produced by each AVO. We present some experimental demonstrations based on real recorded scenes. Current limitations of the algorithm are also discussed. Partial results of this research were published in [19].

II. Perceptual Grouping

Grouping audio and visual components may be considered as a case of *perceptual grouping*. The aim here is to formulate rules, according to which sounds (in audio), and visual events (in vision) are grouped. We now briefly review key observations in perceptual grouping, based on Refs. [14],[20]-[22]. We elaborate on three of them that are utilized in this work. They are:

- The significance of *changes*.
- *Synchronicity* of events.
- The *old-plus-new* auditory heuristic.

The human auditory system is usually able to parse an acoustic input, so that the components belonging to a source are grouped together and form part of a single perceptual stream. The perceptual *separation* of the components arising from different sources is aided by different physical cues. These include differences in fundamental frequency, onset disparities, contrast with previous sounds, drifts in frequency and intensity, Spectro-temporal modulations [23] and sound location [22].

In human vision, perceptual grouping has important functions, such as segmentation. Grouping is affected by various rules [21]: proximity; similarity; continuation; closure; symmetry; familiarity; common fate. Ref. [21] further states that "the probability that a relation did not happen by accident is the most important contributor to its significance". Therefore it is very desirable to determine the statistical significance of a grouping rule in a given scene.

The principle of *common fate* is the basis for our audio-visual association method. This principle was shown to apply to vision, as well as to audition [14], [21]. Common fate prompts grouping of a subset of elements of a scene, whose *changes* are *synchronous*. This segregates them from other elements of the scene which change in a different way. The underlying logic is that it is unlikely that unrelated elements in a scene would undergo parallel changes accidently. Rather, it is more likely that such elements stem from the same physical disturbance [14]. This observation, that *changes* that are *synchronous* are likely to belong together, is a key observation that underlies the work presented here.

[1]Some studies used an approach motivated by computer-vision in order to perform audio-only analysis [17], [18].

When simultaneous sounds co-exist in a scene, auditory scene analysis (ASA) aims to separate them. The ASA principle of *old-plus-new* [14] focuses on instances in which *new sounds begin*, and how these sounds should be interpreted. The principle states: "look for a continuation of what went before, and then pay attention to what has been added to it" [14]. We use this concept in Sec. V, where existing sounds that linger from the past are subtracted from current sounds, in order to identify the new commencing sounds.

III. 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 [8], [9], [24]. That approach is analogous to intensity-based image matching. It implicitly assumes some correlation (possibly nonlinear) between the raw data values in each modality. In this work 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 [25].

Which features are good? In computer-vision, feature-based image registration focuses on sharp spatial changes (edges and corners) [13]. In cross-sensor image matching, Ref. [26] highlighted sharp spatial changes by high-pass filtering. Auditory studies [14] have also shown that one of the major cues for grouping together distinct sounds is a *common change* in their frequency or temporal characteristics. Analogously, in our audio-visual matching problem, we use features having strong *temporal* variations in each of the modalities.

A. Visual Features

We aim for a method that spatially localizes and tracks moving objects, and then isolates the sounds corresponding to these objects. Consequently, we do not rely on pixel data alone. A higher-level representation of the visual modality is sought. Such a higher-level representation should enable tracking of highly non-stationary objects, which move throughout the sequence.

A natural way to track exclusive objects in a scene is to perform feature tracking. The method we used is that of the Ref. [27], the implementation of which is given by Ref. [28]. The method automatically locates image features in the scene. It then tracks their spatial positions throughout the sequence. The result of the tracker is a set of $N_{\rm v}$ visual features.[2] Each visual feature is indexed by $i \in [1, N_{\rm v}]$. Each feature has a spatial trajectory $\mathbf{v}_i(t) = [x_i(t), y_i(t)]^T$, where t is the temporal index (in units of frames), x, y are the image

[2]As we experimented on relatively short scenes, this tracker suffices. All of the features that are successfully tracked throughout the scene are used as an input to the audio-visual association stage. In our experience, the features were tracked reliably. Occasionally, some depict a 'drift' in the location (e.g a feature located on the guitar string - see video [38]).

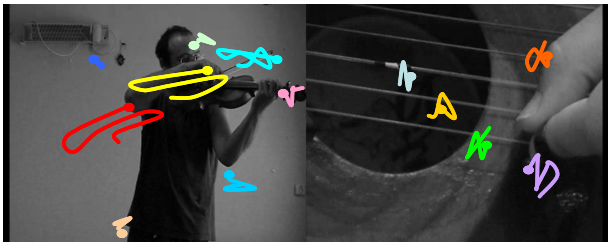

Fig. 2. An illustration of the feature tracking process. Features are automatically located. Their spatial trajectories are tracked. Typically, hundreds of features are tracked.

coordinates, and T denotes transposition. An illustration for tracking results is shown in Fig. 2. Typically, the tracker successfully tracks hundreds of moving features.

We now aim to determine if any of the trajectories is associated with the audio. To do this, significant features are first extracted from each trajectory. These features should be informative, and correspond to significant events in the motion of the tracked feature. We assume that such features are characterized by instances of *strong temporal variation* [29], [30], which we term *visual onsets*. Each visual feature is ascribed a binary vector $\mathbf{v}_i^{\mathrm{on}}$ that compactly summarizes its visual onsets. Each of its elements is set as

$$v_i^{\mathrm{on}}(t) = \begin{cases} 1 & \text{if at } t \text{ feature } i \text{ has a visual onset} \\ 0 & \text{otherwise} \end{cases} . \quad (1)$$

For all features $\{i\}$, the corresponding vectors $\mathbf{v}_i^{\mathrm{on}}$ have the same length N_{f}, which is the number of frames. Next, we describe how the visual onsets corresponding to a visual feature are extracted.

We are interested in locating instances of significant temporal variation in the motion of a visual feature. An appropriate measure is the magnitude of the acceleration of the feature, since it implies a significant change in the motion speed or direction of the feature. Formally, denote the velocity and the acceleration of feature i at instance t by:

$$\dot{v}_i(t) = v_i(t) - v_i(t-1) \ , \quad \ddot{v}_i(t) = \dot{v}_i(t) - \dot{v}_i(t-1), \quad (2)$$

respectively. Then

$$o_i^{\mathrm{visual}}(t) = \|\ddot{v}_i(t)\| \quad (3)$$

is a measure of significant temporal variation in the motion of feature i at time t. From the measure $o_i^{\mathrm{visual}}(t)$, we deduce the set of discrete instances in which a visual onset occurs. Roughly speaking, visual onsets are located right after instances in which $o_i^{\mathrm{visual}}(t)$ has local maxima. The process of locating the visual onsets is summarized in Table I. Next we go into further details.

1) Detection of Visual Features: This section explains how we locate visual onsets for each visual feature. For each feature i, pre-processing normalizes $o_i^{\mathrm{visual}}(t)$ to the range $[0, 1]$:

$$\hat{o}_i^{\mathrm{visual}}(t) = \frac{o_i^{\mathrm{visual}}(t)}{\max_t o_i^{\mathrm{visual}}(t)}. \quad (4)$$

This is done in order to avoid a possible bias for visual features highly accelerating.

TABLE I
DETECTION OF VISUAL ONSETS.

Input: the trajectory of feature i: $\mathbf{v}_i(t)$
Initialization: null the output onsets vector $\mathbf{v}_i^{\mathrm{on}}(t) \equiv \mathbf{0}$
Pre-Processing: Smooth $\mathbf{v}_i(t)$. Calculate $\hat{o}_i^{\mathrm{visual}}(t)$ from Eq. (4)
1. Perform adaptive thresholding on $\hat{o}_i^{\mathrm{visual}}(t)$ (Eq. 5)
2. Temporally prune candidate peaks of $\hat{o}_i^{\mathrm{visual}}(t)$
3. For each of the remaining peaks t_i do
4. while there is a sufficient decrease in $\hat{o}_i^{\mathrm{visual}}(t_i)$
5. set $t_i = t_i + 1$
6. The instance $t_{\mathrm{v}}^{\mathrm{on}} = t_i$ is a visual onset;
Consequently, set $v_i^{\mathrm{on}}(t_{\mathrm{v}}^{\mathrm{on}}) = 1$
Output: The binary vector $\mathbf{v}_i^{\mathrm{on}}$ of visual onsets corresponding to feature i.

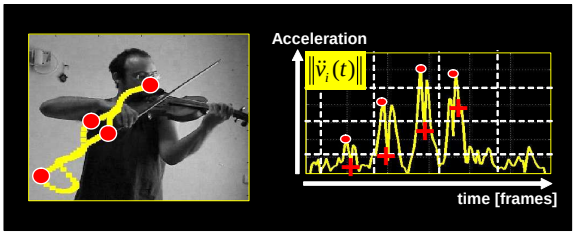

Fig. 3. Detection of visual onsets. [Left] The trajectory corresponds to a feature on the violinist's hand. [Right] The instantaneous magnitude of acceleration of the feature. [Circles] Instances of high acceleration magnitude are detected. [Crosses] Visual onsets.

Next, we look for instances in which $\hat{o}_i^{\mathrm{visual}}(t)$ has a strong local peak. Such a peak hints at the existence of a visual onset. Local peaks are found by adaptively thresholding $\hat{o}_i^{\mathrm{visual}}(t)$ within a temporal window of 2ω frames. Following Ref. [31], the adaptive threshold is given by

$$\widetilde{\delta}_{\mathrm{video}}(t) = \delta_{\mathrm{fixed}} + \delta_{\mathrm{adapt}} \cdot \mathrm{median}_{t\in[t-\omega,\ldots,t+\omega]}\{\hat{o}_i^{\mathrm{visual}}(t)\}. \quad (5)$$

Here δ_{fixed} and δ_{adapt} are positive constants. The first term in Eq. (5) requires $\hat{o}_i^{\mathrm{visual}}(t)$ to exceed a minimal level. The second term requires $\hat{o}_i^{\mathrm{visual}}(t)$ to exceed the local value of $\hat{o}_i^{\mathrm{visual}}(t)$. Here the median provides a robust estimate to this value [31]. The instances in which $\hat{o}_i^{\mathrm{visual}}(t)$ exceeds $\widetilde{\delta}_{\mathrm{video}}(t)$ provide a discrete set of *candidate* visual onsets for object i. We denote this set of temporal instances by V_i^{on}.

This set of candidate onsets may contain false positives. Therefore, it is temporally pruned. The pruning process is based on the assumption that the natural motion of an object is piecewise temporally-coherent [32]-[34]. Hence, the analyzed motion trajectory should have visual onsets only rarely. Thus, pruning removes candidate onsets if they are closer than $\delta_{\mathrm{visual}}^{\mathrm{prune}}$ to another onset candidate having a higher peak of $\hat{o}_i^{\mathrm{visual}}(t)$. Typically in our experiments, $\delta_{\mathrm{visual}}^{\mathrm{prune}} = 10$ frames in video. Hence, this implementation effectively can detect up to 2.5 visual events of a feature per second. To recap, the process is illustrated in Fig. 3.

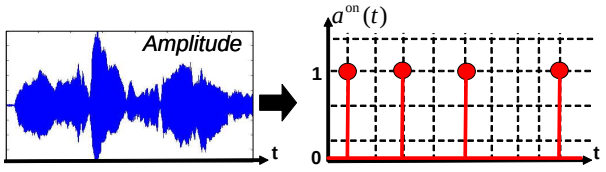

Fig. 4. Detection of audio onsets. The dots mark instances in which a new sound commences in the soundtrack. The detection method is explained in Sec. V-E.

B. Audio Features

We now aim to extract significant temporal variations from the auditory data. We focus on *audio onsets* [14]. These are time instances in which a sound commences (over a possible background).[3] Audio onset detection is well studied [31], [35]. Consequently, we briefly discuss it in Sec. V-E, where we describe the audio peak measurement function $o^{\rm audio}(t)$. We further extract binary peaks from $o^{\rm audio}(t)$, in a manner we describe in Sec. V-E. Similarly to the visual features, the audio onsets instances are eventually expressed by a binary vector $\mathbf{a}^{\rm on}$ of length $N_{\rm f}$. Each of its elements is set as

$$a^{\rm on}(t) = \begin{cases} 1 & \text{if an audio onset takes place at time } t \\ 0 & \text{otherwise} \end{cases} . \tag{6}$$

A new sound begins at instances in which $a^{\rm on}$ equals 1. This is illustrated in Fig. 4.

IV. A Coincidence-Based Approach

In the previous section, we explained that *visual onsets* and *audio onsets* are extracted from the visual and auditory modalities. In this section we describe how the audio onsets are temporally matched to visual onsets. 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 of a hammer hitting a surface is accompanied by noise; the lips of a speaker open as he utters a vowel. This may be seen as a generalization of the grouping principle of *common fate* that we reviewed in Sec. II.

Our approach for 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. We formulate this principle in the following sections.

A. General Approach

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 temporal 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,

$$\sum_{i \in \mathcal{J}} \mathbf{v}_i^{\rm on} = \mathbf{a}^{\rm on}, \tag{7}$$

where $\mathcal{J}$ is the set of the indices of the true AVOs. features that satisfies Eq. (7). If one assumes that the number of prominent visual features is small, the solution to Eq. (7) may be established by seeking a sparse set of visual features (See [9], [36]).

However, cases of perfect correspondence usually do not occur in practice. There are outliers in both modalities, due to clutter and imperfect detection of onsets, having false positives and negatives. We may detect false audio onsets, which should be overlooked, and on the other hand miss true audio onsets. This is also true for detection of onsets in the visual modality. Thus, the path we take is different. It is a sequential approach, motivated in spirit by *matching pursuit* [37]. We define a matching criterion that is based on a probabilistic argument and enables imperfect matching. It favors coincidences and penalizes for mismatches, as we describe in Sec. IV-B.

Using a matching likelihood criterion, we *sequentially* locate the visual features most likely to be associated with the audio. First, the best matching visual feature is found. Then, the audio onsets corresponding to this feature are removed from $\mathbf{a}^{\rm on}$. This results in the vector of the residual audio onsets. We then continue to find the next best matching visual feature. This process re-iterates, until a stopping criterion is met. In the next sections, we first derive a matching criterion that quantifies which visual feature has the highest likelihood to be associated with the audio. We then incorporate this criterion in the sequential framework.

B. Matching Criterion

Here we derive the likelihood of a visual feature i, which has a corresponding visual onsets vector $\mathbf{v}_i^{\rm on}$, to be correlated to the audio onsets vector $\mathbf{a}^{\rm on}$. Assume that $v_i(t)$ is a random variable which follows the probability law

$$\Pr\left[v_i^{\rm on}(t) | a^{\rm on}(t)\right] = \begin{cases} p & , v_i^{\rm on}(t) = a^{\rm on}(t) \\ 1-p & , v_i^{\rm on}(t) \neq a^{\rm on}(t) \end{cases} . \tag{8}$$

In other words, at each instance, $v_i(t)$ has a probability p to be equal to $a^{\rm on}(t)$, and a $(1-p)$ probability to differ from it. 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

$$L(i) = \prod_{t=1}^{N_{\rm f}} \Pr\left[v_i^{\rm on}(t) | a^{\rm on}(t)\right] . \tag{9}$$

Denote by $N_{\rm agree}$ the number of time instances in which $a^{\rm on}(t) = v_i^{\rm on}(t)$. From Eqs. (8,9),

$$L(i) = p^{N_{\rm agree}} \cdot (1-p)^{(N_{\rm f} - N_{\rm agree})} . \tag{10}$$

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

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 $(\mathbf{1}-\mathbf{a}^{\rm on})^T(\mathbf{1}-\mathbf{v}_i^{\rm on})$, hence

$$N_{\rm agree} = (\mathbf{a}^{\rm on})^T\mathbf{v}_i^{\rm on} + (\mathbf{1}-\mathbf{a}^{\rm on})^T(\mathbf{1}-\mathbf{v}_i^{\rm on}) \ . \qquad (11)$$

Plugging Eq. (11) in Eq. (10) and re-arranging terms,

$$\log[L(i)] = N_{\rm f}\log(1-p) + \\ + \left[(\mathbf{a}^{\rm on})^T\mathbf{v}_i^{\rm on} + (\mathbf{1}-\mathbf{a}^{\rm on})^T(\mathbf{1}-\mathbf{v}_i^{\rm on})\right]\log\left(\frac{p}{1-p}\right) \ . \qquad (12)$$

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,

$$\widetilde{L}(i) = \{2\left[(\mathbf{a}^{\rm on})^T\mathbf{v}_i\right] - \mathbf{1}^T\mathbf{v}_i^{\rm on}\}\log\left(\frac{p}{1-p}\right) \ . \qquad (13)$$

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, $\log[p/(1-p)] > 0$. Thus, we may omit the multiplicative term $\log[p/(1-p)]$ from Eq. (13). We can now finally rewrite the likelihood function as

$$\widetilde{L}(i) = (\mathbf{a}^{\rm on})^T\mathbf{v}_i^{\rm on} - (\mathbf{1}-\mathbf{a}^{\rm on})^T\mathbf{v}_i^{\rm on} \ . \qquad (14)$$

Eq. (14) has an intuitive interpretation. Let us begin with the second term. Recall that, by definition, $a^{\rm on}$ equals 1 when an audio onset occurs, and equals 0 otherwise. Hence, $(1-a^{\rm on})$ equals 1 when an audio onset does not occur. Consequently, the second term of Eq. (14) effectively counts the number of the visual onsets of feature i that do *not* coincide with audio onsets. This mismatch acts as a penalty term in Eq. (14). On the other hand, the first term counts the number of the visual onsets of feature i that *do* coincide with audio onsets. Overall, Eq. (14) favors coincidences (which should increase the matching likelihood of a feature), and penalizes inconsistencies (which should inhibit this likelihood). In the next section we describe how this criterion is embedded in a framework which sequentially extracts the prominent visual features.

C. Sequential Matching

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. The visual feature that corresponds 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, available online at [38], and one of whose frames is shown in Figs. 1, 2. It was recorded by a simple camcorder and using a single microphone.[4] Note that in this sequence, the sound and motions of the guitar pose a distraction for the violin, and vice versa. Onsets were obtained as we describe in Sec. V-E. Then, the visual feature that maximized Eq. (14) was the *hand of the violin player*. Its detection and tracking were automatic.

TABLE II
CROSS-MODAL ASSOCIATION ALGORITHM.

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 $\{\hat{i}_l, \mathbf{m}_l^{\rm on}\}$.

Now, the audio onsets that correspond to AVO $\hat{i}$ are given by the vector

$$\mathbf{m}^{\rm on} = \mathbf{a}^{\rm on} \bullet \mathbf{v}_{\hat{i}}^{\rm on}, \qquad (15)$$

where $\bullet$ denotes the logical-AND operation per corresponding element pair. Let us eliminate these corresponding onsets from $\mathbf{a}^{\rm on}$. The *residual* audio onsets are represented by

$$\mathbf{a}_1^{\rm on} \equiv \mathbf{a}^{\rm on} - \mathbf{m}^{\rm on}. \qquad (16)$$

The vector $\mathbf{a}_1^{\rm on}$ becomes the input for a new iteration: it is used in Eq. (14), 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 framework is the mentioned classification threshold of the AVO. We set it to $\widetilde{L}(\hat{i}) = 0$. Based on Eq. (14),

$$0 > (\mathbf{a}^{\rm on})^T\mathbf{v}_i^{\rm on} - (\mathbf{1}-\mathbf{a}^{\rm on})^T\mathbf{v}_i^{\rm on}. \qquad (17)$$

Failure to pass the threshold occurs when

$$(\mathbf{a}_l^{\rm on})^T\mathbf{v}_{\hat{i}}^{\rm on} < \frac{1}{2}\mathbf{1}^T\mathbf{v}_{\hat{i}}^{\rm on}. \qquad (18)$$

Consequently, when $\widetilde{L}(\hat{i}) < 0$, 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 given in Table II (here $\mathbf{0}$ is a column vector, all of whose elements are null).

The output $|\widehat{\mathcal{J}}|$ estimates the number of independent AVOs. This algorithm is fast (linear in the number of AVOs): $\approx |\mathcal{J}|$

[4]The sampling parameters of the audio and video are given in App. -B.

iterations, each having $\mathcal{O}(N_{\rm f}N_{\rm v})$ calculations. In the above mentioned `violin-guitar` sequence, this algorithm automatically detected two independent AVOs: the *guitar string*, and the hand of the *violin player* (marked as crosses in Fig.1).

D. Temporal Resolution

The previous sections derived a framework for establishing audio-visual association. It implies perfect temporal coincidences between audio and visual onsets: an audio onset is assumed to be related to a visual onset, if both onsets take place *simultaneously* (Table II, step 4). However, in practice, the temporal resolution of our system 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, in practice, 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.

Let $t_v^{\rm on}$ denote the temporal location of a visual onset. Let $t_a^{\rm on}$ denote the temporal location of an audio onset. Then the visual onset may be related to the audio onset if

$$|t_v^{\rm on} - t_a^{\rm on}| \leq \delta_1^{\rm AV}. \tag{19}$$

In our experiments, we set $\delta_1^{\rm AV} = 3$ frames. The frame rate of the video recording is 25 frames/sec. Consequently, an audio onset and a visual onset are considered to be coinciding if the visual onset occurred within $3/25 \approx 1/8sec$ of the audio onset.

V. Audio Processing and Isolation

Section IV described a procedure for finding the visual features that are associated with the audio. This resulted in a set of AVOs, each with its vector of corresponding audio onsets: $\{\hat{i}_l, \mathbf{m}_l^{\rm on}\}$. We now describe how the sounds corresponding to each of these AVOs are extracted from the single-microphone soundtrack.

Let $s_{\rm desired}$,$s_{\rm interfere}$ and s denote the amplitudes of the source of interest, the interfering sounds, and the mixture, respectively. Then

$$s = s_{\rm desired} + s_{\rm interfere}. \tag{20}$$

Out of the soundtrack s, we wish to isolate the sounds corresponding to a given desired AVO $\hat{i}$. To do this, we utilize the audio-visual association achieved. Recall that AVO $\hat{i}$ is associated with the audio onsets in the vector $\mathbf{m}^{\rm on}$. In other words, $\mathbf{m}^{\rm on}$ *points to instances in which a sound associated with the AVO commences*. We now need to extract from the audio mixture only the sounds that begin at these onsets. We may do this sequentially: isolate each distinct sound, and then concatenate all of the sounds together to form the isolated soundtrack of the AVO. How may we isolate a single sound commencing at a given onset instance $t^{\rm on}$? To do this, we employ the method of *binary masking* [12], [40], [41], which we review next.

A. Binary Masking

Let $s(n)$ denote a sound signal, where n is a discrete sample index of the sampled sound. This signal is analyzed in short temporal windows w, each being $N_{\rm w}$-samples long. Consecutive windows are shifted by $N_{\rm sft}$ samples. The short-time Fourier transform (STFT) of $s(n)$ is

$$S(t,f) = \sum_{n=0}^{N_{\rm w}-1} s(n) w(tN_{\rm sft} - n) e^{-j(2\pi/N_{\rm w})nf}, \tag{21}$$

where f is the frequency index and t is the temporal index of the analyzed instance. Let us denote the amplitude of the STFT by $A(t,f) = ||S(t,f)||$. The *spectrogram* is defined as $A(t,f)^2$.

To re-synthesize a discrete signal given its STFT $S(t,f)$, the overlap-and-add (OLA) method may be used [39]:

$$\hat{s}(n) = \frac{N_{\rm sft}}{W(0)} \sum_{t=-\infty}^{\infty} \left[\frac{1}{N_{\rm w}} \sum_{f=0}^{N_{\rm w}-1} S(t,f) e^{j(2\pi/N_{\rm w})nf} \right]. \tag{22}$$

Here

$$W(0) = \sum_{n=-\infty}^{\infty} w[n]. \tag{23}$$

If for all n

$$\sum_{r=-\infty}^{\infty} w[rN_{\rm sft} - n] = \frac{W(0)}{N_{\rm sft}}, \tag{24}$$

then $\hat{s}(n) = s(n)$ following [39].

Through *binary masking* [12], [40], [41], this re-synthesis process is modified. Only a subset of the time-frequency (T-F) bins of $s(n)$ is maintained. For instance, assume that the STFT-amplitude of $s_{\rm desired}$ is non-zero in a finite set $\Gamma_{\rm desired}$ of T-F bins $\{(t,f)\}$. Define a mask

$$M_{\rm desired}(t,f) = \begin{cases} 1 & (t,f) \in \Gamma_{\rm desired} \\ 0 & \text{otherwise} \end{cases}. \tag{25}$$

Then by modifying Eq. (22) into

$$\hat{s}_{\rm desired}(n) = \frac{N_{\rm sft}}{W(0)} \sum_{t=-\infty}^{\infty} \left[\frac{1}{N_{\rm w}} \sum_{f=0}^{N_{\rm w}-1} M_{\rm desired}(t,f) S(t,f) e^{j(2\pi/N_{\rm w})nf} \right] \tag{26}$$

we re-synthesize only the components lying in $\Gamma_{\rm desired}$.

Audio-isolation methods utilizing this process of binary masking focus on identifying the T-F bins that should be included in $M_{\rm desired}$ [12], [40], [41]. These methods assume that the set $\Gamma_{\rm desired}$ should very rarely contain energy of the other sources in the scene.[5] This assumption is based [43] on the sparsity of typical sounds, particularly *harmonic* ones, in the spectrogram. The frequency contents of an harmonic sound contain a fundamental frequency f_0 (the *pitch*), along with integer multiples of this frequency (the *harmonies*). Since typical sounds are sparsely distributed across the T-F plane,

[5]Sources may overlap in a T-F bin. Binary-masking methods [12] then assign the bin to the source whose estimated amplitude in the bin is the strongest. To simplify our approach, however, here we allow a T-F bin to be assigned to several sources.

independent sounds mixed together should rarely overlap. This is the main motivation for the binary-masking method.

In this work we assume that underlying sources are harmonic. Consequently, a sound of interest can be enhanced by maintaining the values of $S(t,f)$ in $\Gamma_{\rm desired}$, while nulling the other bins. This should maintain the components of the desired sound, while leaving only little of the interfering sounds. In the next section we explain how we establish $\Gamma_{\rm desired}$ that corresponds to a sound commencing at $t^{\rm on}$.

B. Principles for Building the Binary Mask

We are given an audio onset instance $t^{\rm on}$, and wish to identify the set of T-F bins $\Gamma_{\rm desired}$ that belong to this sound. Around the instance of the audio onset, several frequency bins undergo a simultaneous amplitude increase. By the principal of *common fate* (Sec. II), we assume that such frequency bins that have just become active all belong to the desired commencing sound. It is this sound which we wish to isolate. To identify the desired frequency bins, we utilize the harmonic nature of the sound. Hence, the sounds contains a pitch-frequency and the integer multiples of the frequency (harmonies). Therefore:

1) We may identify the frequency bins belonging to the commencing sound, by detecting the pitch f_0 of the sound commencing at $t^{\rm on}$.
2) Since the sound is assumed to be harmonic, we may track the pitch frequency $f_0(t)$ through time.
3) When the sound fades away, at $t^{\rm off}$, the tracking is terminated.

This process provides the required mask, corresponding to the sound that commences at $t^{\rm on}$:

$$\Gamma^{t^{\rm on}}_{\rm desired}(t,f) = \{(t\ , f_0(t)k)\}. \tag{27}$$

Here $t \in [t^{\rm on}, t^{\rm off}]$ and $k \in [1 \dots K]$, K being the number of considered harmonies. To conclude: *given only an onset instance* $t^{\rm on}$, we determine $\Gamma^{t^{\rm on}}_{\rm desired}$ by detecting $f_0(t^{\rm on})$, and then tracking $f_0(t)$ in $t \in [t^{\rm on}, t^{\rm off}]$.

The following sections provide the details for this procedure. Sec. V-C explains how we first emphasize commencing sounds in the spectrogram over existing ones. This eases the detection of the pitch frequency at $t^{\rm on}$. The pitch-detection is then described in Sec. V-D. Once the pitch frequency is established at the onset instance, it is tracked until the sound fades out. This is described in App. -A. These steps provide all that is needed in order to isolate the sound of interest.

C. Elimination of Prior Sounds

The sound of interest is the one commencing at $t^{\rm on}$. We wish to identify its pitch-frequency $f_0(t^{\rm on})$. However, other sounds in the mixture may also be present at $t^{\rm on}$, interfering with the pitch-detection procedure. Therefore, before detecting $f_0(t^{\rm on})$, we first emphasize the components of the commencing sound of interest over these interfering sounds. This section explains the method for achieving it.

The sound of interest is the one commencing at $t^{\rm on}$. Thus, the disturbing audio at $t^{\rm on}$ is assumed to have commenced *prior* to $t^{\rm on}$. These disturbing sounds linger from the past.

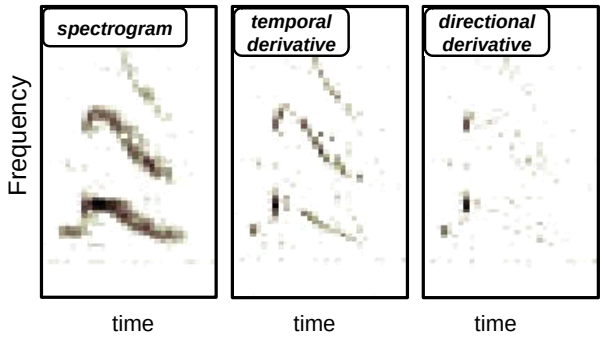

Fig. 5. Effects of frequency drift on the STFT temporal derivative. [Left] A section of a spectrogram (female speaker) exhibiting a frequency drift. [Middle] A temporal derivative (Eq. 28) results in high values through the entire sound duration. [Right] The directional derivative (Eq. 30) handles the frequency drift well. High output values occur mainly at the onset.

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, Ref. [31] suggests the *relative* temporal difference

$$D(t,f) = \frac{A(t,f) - A(t-1,f)}{A(t-1,f)}. \tag{28}$$

Eq. (28) emphasizes an increase of amplitude in frequency bins that have been quiet (no sound) just before t.

As a practical criterion, however, we have found that Eq. (28) lacks robustness. The reason is that sounds which have commenced prior to t may have a slow frequency *drift* (Fig. 5). This poses a problem for Eq. (28), which is based solely on a *temporal* comparison per frequency channel. Drift results in high values of Eq. (28) in some frequencies f, even if no new sound actually commences around (t,f), as seen in Fig. 5. This hinders the emphasis of commencing frequencies, which is the goal of Eq. (28). To overcome this, we compute a *directional* difference in the T-F domain. It fits neighboring bands at each instance, hence tracking the drift. Consider a small frequency range $\Omega_{\rm freq}(f)$ around f. In analogy to image alignment, *frequency alignment* at time t is obtained by

$$f^{\rm aligned}(f) = \arg \min_{f_z \in \Omega_{\rm freq}(f)} |A(t^{\rm on}, f) - A(t^{\rm on}-1, f_z)|. \tag{29}$$

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

$$\widetilde{D}(t,f) = \frac{A(t,f) - A(t-1, f^{\rm aligned}(f))}{A(t-1, f^{\rm aligned}(f))} \tag{30}$$

is indeed much less sensitive to drift, and is responsive to true onsets (Fig 5). The map

$$\widetilde{D}_+(t,f) = \max\{0, \widetilde{D}(t,f)\} \tag{31}$$

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

D. Pitch Detection at $t^{\rm on}$

As described in the previous section, the measure $\widetilde{D}_+(t^{\rm on}, f)$ emphasizes the amplitude of frequency bins that

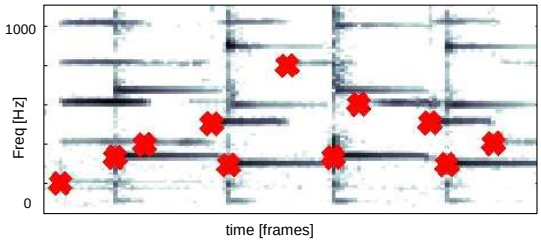

Fig. 6. The STFT-amplitude corresponding to the **violin-guitar** sequence. The horizontal position of the overlaid crosses indicates instances of audio onsets. The vertical position of the crosses indicates the pitch-frequency of the commencing sounds.

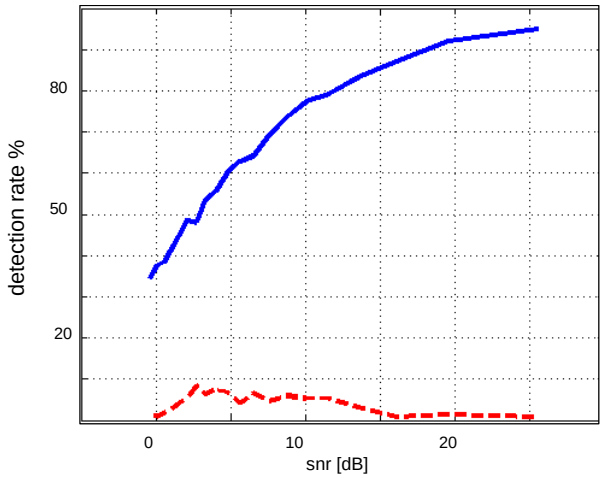

Fig. 7. Onsets detection vs. added white noise in the **violin-guitar** sequence. [bold] Average rate of correct detections. [dot] Average rate of false detections.

correspond to a commencing sound. We may now use $\widetilde{D}_+(t^{\rm on}, f)$ as an input to a pitch-detection algorithm in order to detect the pitch frequency at $t^{\rm on}$. The algorithm we choose to use is the harmonic-product-spectrum(HPS) [44]. An example for the detected pitch-frequencies at audio onsets in the **violin-guitar** sequence is given in Fig. 6. Following the detection of $f_0(t^{\rm on})$, the pitch-frequency needs to be tracked during $t \geq t^{\rm on}$, until $t^{\rm off}$. This procedure is described in App. -A.

E. Detection of Audio Onsets

Methods for audio-onset detection have been extensively studied [31], and are used in a variety of audio-processing applications [47]. Here we describe our particular method for onsets detection. Our criterion for significant signal increase is simply

$$o^{\rm audio}(t) = \sum_f \widetilde{D}_+(t, f), \tag{32}$$

where $\widetilde{D}_+(t, f)$ is defined in Eq. (31). The criterion is similar to a criterion suggested in Ref. [31], which was used to detect the onset of a single sound, rather than several mixed sounds. However, Eq. (32) is more robust in a setup of several mixed sources, as it suppresses lingering sounds (Eq. 31). The extraction of the audio onsets is done in the spirit of Ref. [31].

The onset measure of Eq. (32) relies on a synchronous amplitude increase in several frequency bins together. Therefore, it is relatively robust to background noise, keeping a low rate of false detections (Fig. 7).

VI. Experiments

In this section we present experiments based on real recorded video sequences. In our experiments we compound separately-recorded movies (e.g., a violin sequence and a guitar sequence) into a single video.[6] Such a procedure is a common practice in single-micrhopone audio-separation studies [5], [12], [40], since it provides access to the audio ground-truth data. This allows quantitative assessment of the quality of audio isolation, as we describe below.

We first describe the experiments and the association results. The video sequences presented here are available online [38]. We then provide a quantitative evaluation of the audio isolation for some of the analyzed scenes. Implementation details and typical parameters values are given in App. -B.

A. Results

The violin-guitar sequence: This sequence features a close-up on a hand playing a guitar. At the same time, a violinist is playing. The soundtrack thus contains temporally-overlapping sounds. The algorithm automatically detected that there are two (and only two) independent visual features that are associated with this soundtrack. The first feature corresponds to the violinist hand. The second is the correct string of the guitar (Fig 1). Following the location of the visual features, the audio components corresponding to each of the features are extracted from the soundtrack.

The speakers #1 sequence: This movie has simultaneous speech by male and female speakers. The female is viewed frontally, while the male is viewed from the side. The algorithm automatically detected that there are two visual features that are associated with this soundtrack (Fig. 8). Following the location of the visual features, the audio components corresponding to each of the speakers are extracted from the soundtrack As can be seen, there is indeed a significant temporal overlap between independent sources. Yet, the sources are separated successfully.

The dual-violin sequence: This experiment is very challenging. It contains two instances of the same violinist, which uses the *same* violin to play *different* tunes. Listeners who had observed this mixed scene found it difficult to correctly group the different notes into a coherent tune. However, our algorithm managed to do so. First, it located the relevant visual features . These are exploited for isolating the correct audio components (Fig. 10). This example demonstrates a problem

[6] 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.

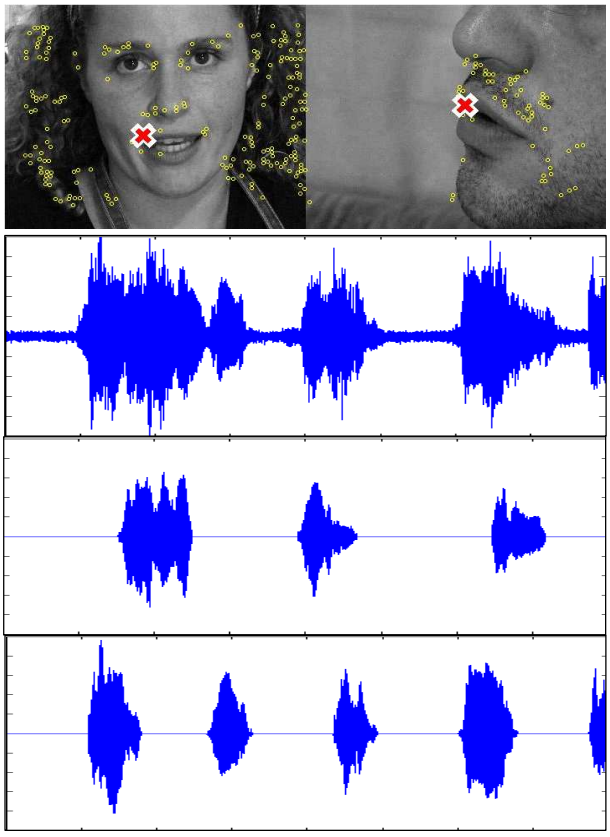

Fig. 8. A frame from the `speakers #1` movie. Out of the selected and tracked visual features [Dots], two are automatically associated to the audio [Crosses]: correctly, one per source. The audio components of each source are extracted from the mixed soundtrack.

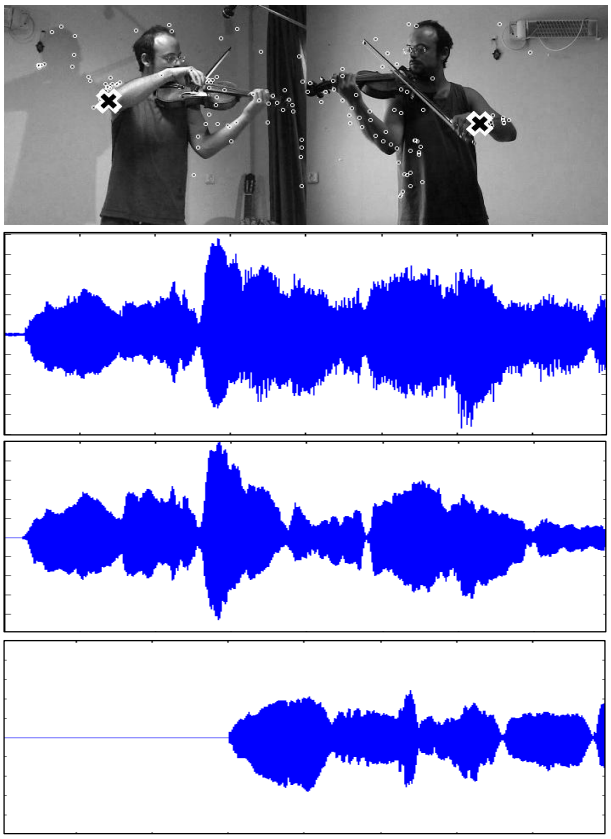

Fig. 10. A frame from the `dual-violin` movie. Out of the selected and tracked visual features [Dots], two are automatically associated to the audio [Crosses]: correctly, one per source.

which is very difficult to solve with audio data alone, but is elegantly solved using the visual modality.

The speakers #2 sequence: This experiment also includes two speakers. It is a recording of a real scene. Two features on the mouth area of each speaker are correctly located (Fig 11). In this sequence, one of the audio onsets of the male coincides with an audio onset of the female. However, our method in its current formulation cannot identify concurrent audio onsets. Thus, no more than one audio onset can detected. In the audio-visual association stage, this concurrent onset was associated to the female speaker. The detected pitch, however, was that of the male. Consequently, in the isolated soundtrack corresponding to the male, one of his words is missing. On the other hand, in the soundtrack corresponding to the female, one of her words is replaced with that of the male. The corresponding spectrograms are shown in Fig 12.

The experiments described above depict the utilization of the perceptual-grouping rules we have described in Sec. II. We focus on *changes*, both auditory and visual ones. This aids in parsing dense audio-visual scenes into sparser events. The *synchronicity* of these events aids in relating otherwise-difficult connections (as in the `dual-violin` sequence). Finally, the *old-plus-new* heuristic aids in differentiating the commencing auditory harmonic components from lingering ones, for subsequent audio enhancement.

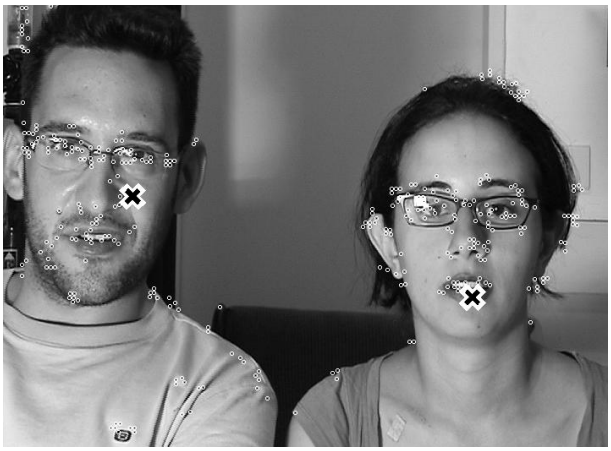

Fig. 11. A frame from the `speakers #2` movie. Out of the selected and tracked visual features [Dots], two are automatically associated with the audio [Crosses]: correctly, one per source.

B. Audio Isolation: Quantitative Evaluation

In this section we provide quantitative evaluation for the experimental separation of the audio sources. The quality measure we use is the Signal-to-Distortion Ratio (SDR) [42] expressed in decibels (dB):

$$SDR = 10log_{10}\frac{\|s\|^2}{\|\hat{s}-s\|^2}. \quad (33)$$

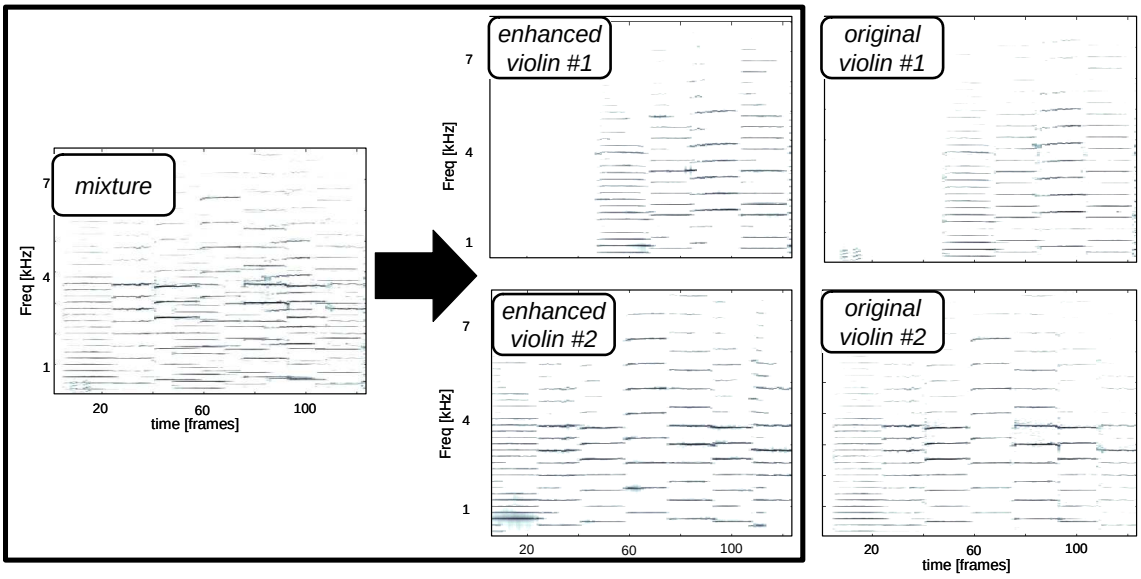

Fig. 9. The spectrograms corresponding to the `dual-violin` sequence. Based on *visual* data, the audio components of each of the violins were automatically separated from a single soundtrack.

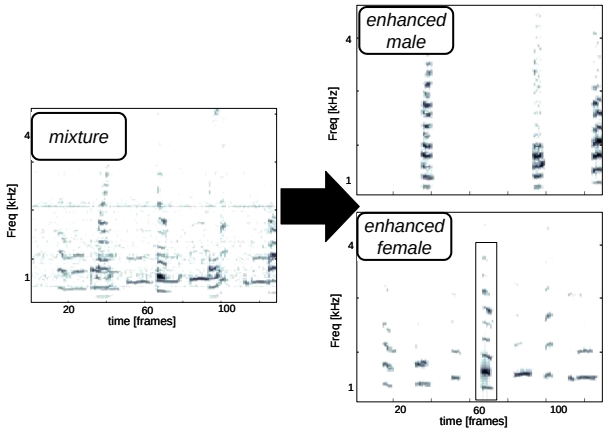

Fig. 12. The log-amplitude STFT images corresponding to the `speakers #2` sequence. Based on *visual* data, the audio components corresponding to each of the speakers were automatically separated from a single soundtrack. The marked box at the bottom-right spectrogram highlights a time-ferquency segment that originally belonged to the male speaker, but was erroneously attributed to the female speaker.

TABLE III
QUANTITATIVE EVALUATION OF THE AUDIO ISOLATION.

sequence	source	SDR improvement [dB]
violin-guitar	violin	6.35
	guitar	3.40
speakers	male	5.22
	female	3.76
dual-violin	violin1	9.42
	violin2	2.26

We choose this measure, since: ”the SDR incorporates all possible kinds of distortion arising from different source separation algorithms, including interference from other sources, gurgling artifacts, filtering distortion and spatial distortion” [42].

The SDR of an isolated source is compared to the SDR of the mixed source. Table III summarizes the SDR improvement in decibels for the conducted experiments. Next we provide some insight into these figures.

in the `violin-guitar` sequence, some of the time-frequency components of the violin were erroneously included in the binary mask corresponding to the guitar: towards the end of one of the guitar’s sounds, a violin sound commences. It has a harmony (a multiple of its pitch) which is close to a harmony of the guitar. Consequently, the isolated soundtrack of the guitar contains artifacts traced to that harmony of the violin. indeed, the SDR measure of the guitar is lower than that of the violin.

In `speakers #1`, the speech of the female contains some non-harmonic sounds, which are lost in the separation process that relies on purely harmonic sounds. Finally, in the `dual-violin sequence`, the 1st violin is mostly quiet. Therefore, successful removal of the the 2nd violin (which is very active), greatly enhances the SDR of the 1st violin.

VII. LIMITATIONS

Here we describe current limitations of the described algorithm. Possible improvements are suggested in Sec. VIII.

Vision-Based Auditory Grouping. This work described principles for associating audio and visual events, that are based on temporal coincidences *alone*. For instance, a sound of the guitar may erroneously be associated with a feature corresponding to the violin, if a visual onset of the violinist took place around the same instance in which a sound of the guitar commenced. In crowded scenes (e.g, three or more people) dense audio onsets exist, and this becomes an acute problem: as the temporal resolution is limited (see

Sec. IV-D), the indicator vector of audio onsets has audio onsets at every location: $\mathbf{a}^{\text{on}} \equiv \mathbf{1}$. Consequently, Eq. (14) is reduced to $\widetilde{L}(i) = (\mathbf{a}^{\text{on}})^T \mathbf{v}_i^{\text{on}}$. In other words: every visual feature has full correspondence to the audio. Consequently, the audio-visual synchronicity is no longer a statistically-significant grouping rule(Sec. II).To deal with dense scenes, our framework requires more robust grouping rules.

Visual Pruning. The principle used here groups audio onsets based on vision only. The temporal resolution of the audio-visual association is also limited (Sec. IV-D). This implies that in a dense audio scene, *any* visual onset has a high probability to be matched by an audio onset. To avoid such an erroneous audio-visual association, we aggressively prune visual onsets. Two onsets of a visual feature may not be closer than 10 frames to each other. This is equivalent to assuming an average event rate of $2.5Hz$. This limits the applicability of our current realization in the case of rapidly-moving AVOs.

System Parameters. Our method requires to tune several parameters when analyzing an audio-visual scene. These parameters are detailed in App. -B. This tuning of parameters makes the analysis more difficult.

Audio Onsets and Pitch Detection. Audio onsets of different sources are assumed not to coincide (Sec. IV-A). This assumption is further utilized in the pitch-detection stage (Sec. V-C). To alleviate this limitation, we may utilize pitch-detection methods that are able to differentiate between sounds that commence simultaneously, and that further detect their individual pitch frequencies [45].[7]. We may then initialize a robust pitch tracker with these initial frequencies [51], [46]. Such trackers reliably track the pitch in noisy mixtures. This is achieved mainly by casting the pitch-tracking task as a maximum-likelihood estimation problem; and by inspecting the signal not only frame-by-frame, but rather at bigger temporal segments.

Audio Enhancement. The binary-masking procedure (Sec. V-A) assumes that independent sources should rarely depict an overlap in the T-F domain. Also, it may cause auditory artifacts [48]. *Soft-masking* [43], [48] would improve the quality of the auditory enhancement, and may deal better with T-F overlap. In order to separate sources with very close-by pitch frequencies (e.g. male-male or female-female) from a single microphone, prior training of models is required [48]. Even then, the results are inferior to those of male-female mixtures [49].

VIII. Conclusions

This paper presented a novel approach for cross-modal audio-visual 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).

[7]This is done by sequentially detecting the dominant pitch-frequency, removing it from the mixture, and repeating the process

Future work should avoid associating audio onsets to incorrect visual onsets. Audio onsets that have been grouped together based on correspondence to visual onsets only, would then be further inspected. Comparing different auditory characteristics [50] of the audio onsets in that particular group may reveal whether any of these audio onsets does not actually belong in that group. This would also alleviate the need to aggressively prune the visual onsets of a feature. Such a framework may also lead to automatically setting of method's parameters.

Our audio-visual correspondence may be incorporated with a more general audio-enhancement framework [43], [48], [47]. Such a framework would alleviate the assumption of harmonic sounds, and would improve the quality of the enhancement.

Finally, 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 that this may be potentially useful, for instance, in associating macro-economic events.

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. The work was supported by the Israeli Science Foundation (grant 1031/08), and conducted at the Ollendorff Center in the Elect. Eng. Dept. at the Technion. Minerva is funded through the BMBF.

A. Pitch Tracking

Given the detected pitch frequency at $f_0(t)$, we wish to establish $f_0(t+1)$. It is assumed to lie in a frequency neighborhood Ω_{freq} of $f_0(t)$, since the pitch frequency of a source typically evolves gradually [51]. Recall that an harmonic sound contains multiples of the pitch frequency (the harmonies). Let the set of indices of active harmonies at time t be $\mathcal{K}(t)$. For initialization we set $\mathcal{K}(t^{\text{on}}) = [1, \ldots, K]$. The estimated frequency $f_0(t)$ may be found as the one whose harmonies capture most of the energy of the signal

$$f_0(t+1) = \arg \max_{f \in \Omega_{\text{freq}}} \sum_{k \in \mathcal{K}(t)} \| A\left(t+1, f \cdot k\right) \|^2 . \quad (34)$$

Eq. (34), however, does not account for the simultaneous existence of other audio sources. Disrupting sounds of high energy may be present around the harmonies $(t+1, f \cdot k)$ for some $f \in \Omega_{\text{freq}}$, and $k \in \mathcal{K}(t)$. This 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. (34). Rather, we use $\log[A\left(t+1, f \cdot k\right)]$. This resembles the approach taken by the harmonic-product-spectrum algorithm [44] for dealing with noisy frequency components. Consequently, the estimation of $f_0(t+1)$ is more effectively dependent on many weak frequency bins. This significantly reduces the error induced by a few noisy components.

Recall that the pitch is tracked in order to identify the set $\Gamma^{t^{\text{on}}}_{\text{desired}}$ of time-frequency bins in which an harmonic sound lies. We now go into the details of how to establish $\Gamma^{t^{\text{on}}}_{\text{desired}}$.

TABLE IV
PITCH-TRACKING ALGORITHM.

Input: $t^{\text{on}}, f_0(t^{\text{on}}), A(t,f)$

0. Initalize: $t = t^{\text{on}}, \mathcal{K}(t) = [1,\ldots,K]$.
1. Iterate
2. $f_0(t+1) = \arg\max_{f\in\Omega_{\text{freq}}} \sum_{k\in\mathcal{K}(t)} \|\log[A\,(t+1, f\cdot k)]\|^2$
3. foreach $k \in \mathcal{K}(t)$
4. $\rho(k,t) = \frac{A[t+1, f_0(t+1)\cdot k]}{A[t, f_0(t)\cdot k]}$
5. if $\rho(k,t) \geq \rho_{\text{interfer}}$ or $\rho(k,t) \leq \rho_{\text{dead}}$ then
6. $\mathcal{K}(t) = \mathcal{K}(t-1) - k$
7. end foreach
8. if $|K(t)| < K_{\min}$ then
9. $t^{\text{off}} = t$
10. quit
11. $t = t+1$

Output:

- The offset instance of the tracked sound t^{off}.
- The pitch frequeny $f_0(t)$, for $t \in [t^{\text{on}}, t^{\text{off}}]$
- The indices of active harmonies $\mathcal{K}(t)$, for $t \in [t^{\text{on}}, t^{\text{off}}]$
- The time-frequency domain $\Gamma^{t^{\text{on}}}_{\text{desired}}$ of the tracked sound: $\Gamma^{t^{\text{on}}}_{\text{desired}} = \{(t, f_0(t)\cdot k\}$, for $k \in \mathcal{K}(t)$, $t \in [t^{\text{on}}, t^{\text{off}}]$

According to Eq. (27), $\Gamma^{t^{\text{on}}}_{\text{desired}}$ should contain all of the harmonies of the pitch frequency, for $t \in [t^{\text{on}}, t^{\text{off}}]$. However, $\Gamma^{t^{\text{on}}}_{\text{desired}}$ may also contain unwanted interferences. Therefore, once we identify the existence of a strong interference at a harmony, we remove this harmony from $\mathcal{K}(t)$. This implies that we prefer to minimize interferences in the enhanced signal, even at the cost of losing part of the acoustic energy of the signal. A harmony is removed from $\mathcal{K}(t)$ also if the harmony has faded out: we assume that it will not become active again. Both of these mechanisms of harmony removal are identified by inspecting the following measure:

$$\rho(k,t) = \frac{A\,[t+1, f_0(t+1)\cdot k]}{A\,[t, f_0(t)\cdot k]}. \tag{35}$$

The measure $\rho(k,t)$ inspects the relative temporal change of the harmony's amplitude. Let ρ_{interfer} and ρ_{dead} be two positive constants. When $\rho(k,t) \geq \rho_{\text{interfer}}$ we deduce that an interfering signal has entered the harmony k. Therefore, it is removed from $\mathcal{K}(t)$. Similarly, when $\rho(k,t) \leq \rho_{\text{dead}}$ we deduce that an the harmony k has faded out. Therefore, it is removed from $\mathcal{K}(t)$. Typically we used $\rho_{\text{interfer}} = 2.5$ and $\rho_{\text{dead}} = 0.5$.

We initialize the tracking process by setting $f_0(t^{\text{on}})$ and $\mathcal{K}(t^{\text{on}}) = [1,\ldots,K]$. We then iterate the process through time. When the number of active harmonies $|\mathcal{K}(t)|$ drops below a certain threshold $K_{\min}$, termination of the signal at time t^{off} is declared. Typically we used $K_{\min} = 3$. The domain $\Gamma^{t^{\text{on}}}_{\text{desired}}$ that the tracked sound occupies in $t \in [t^{\text{on}}, t^{\text{off}}]$ is composed from the active harmonies at each instance t. Formally :

$$\Gamma^{t^{\text{on}}}_{\text{desired}} = \{(t, f_0(t)\cdot k\}\ ,\ \text{where } t \in [t^{\text{on}}, t^{\text{off}}] \text{ and } k \in \mathcal{K}(t). \tag{36}$$

The tracking process is summarized in Table IV.

B. Implementation Details

This section describes the implementation details of the algorithm described in this paper. It also lists the parameter values used in the implementation. Unless stated otherwise, the parameters required tuning for each analyzed sequence.

Temporal Tolerance: Audio and visual onsets need not happen at the exact same frame. As we explained in Ch. IV, an audio onset and a visual onsets are considered simultaneous, if they occur within 3 frames from one another.

Frequency Analysis: The audio is re-sampled to 16 kHz, and analyzed using a Hamming window of $80msec$. Consecutive windows have 50% overlap. This also ensures synchronicity of the windows with the video frame rate ($25Hz$). A Hamming window with a 50% overlap also realizes constraints of the OLA method [52].

Pitch detection and Tracking: For pitch detection and tracking, the number of considered harmonies is set to $K = 10$. The guidelines of Ref. [53] are taken in order to prevent pitch-halving (erroneously setting the pitch to half its real value).

Visual Processing: Prior to calculating $\ddot{\mathbf{v}}_i(t)$ as described in Sec. III, the trajectory $\mathbf{v}_i(t)$ is filtered to remove tracking noise. The temporal filtering is performed separately on each of the vector components $\mathbf{v}_i(t) = [x_i(t), y_i(t)]^T$. The filtering process consists of performing temporal median filtering (typically window size set to 5 frames) to account for abrupt tracking errors. Consequent filtering consists of smoothing by convolution with a Gaussian kernel with standard deviation of around 1. Finally, the adaptive threshold parameters are tuned in each analyzed scene. We further remove visual onsets whose amplitudes of acceleration and velocity are smaller than specific values. Typically, the velocity and acceleration amplitudes at an onset should exceed the values of 0.2.

In the detection of visual onsets, the instances in V_i^{on} are ones of significant change in motion. However, we have found in our experiments that instances in which these significant change in motion are *over*, and a smooth motion *commences* have better temporal correlation with the audio onsets. Therefore, each temporal location $t_v^{\text{on}} \in V_i^{\text{on}}$ that is currently located at a local maximum of $\hat{o}_i^{\text{visual}}(t)$ is shifted forward in time away from the local maximum, and towards a smaller value of $\hat{o}_i^{\text{visual}}(t)$. The onset is iteratively shifted this way, while there is a significant relative decrease in $\hat{o}_i^{\text{visual}}(t)$. Typically, onsets are shifted by not more than 2 or 3 frames.

REFERENCES

[1] 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).

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

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

[4] 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).

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

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

[7] 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. *Inf. Fusion*, pp. 213–219 (2002).

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

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

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

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

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

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

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

[15] A. O'Donovan, R. Duraiswami and J. Neumann Microphone Arrays as Generalized Cameras for Integrated Audio Visual Processing. *IEEE CVPR* (2007).

[16] S. Ravulapalli and S. Sarkar. Association of Sound to Motion in Video using Perceptual Organization. *Proc. IEEE ICPR*, pp. 1216–1219 (2006).

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

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

[19] Z. Barzelay and Y. Y. Schechner. Harmony in motion. *Proc. IEEE CVPR* (2007).

[20] D. A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach*. Upper Saddle River, N.J. : Prentice-Hall (2003).

[21] D. G. Lowe. *Perceptual Organization and Visual Recognition*. Boston, Mass.: Kluwer Academic Publishers (1985).

[22] B. C. J. Moore. *An introduction to the psychology of hearing*. San Diego, Calif. : Academic Press (1997).

[23] N. Mesgarani, S. V. David, J. B. Fritz, and S. A. Shamma Phoneme representation and classification in the primary auditory cortex *J. Acoust. Soc. Am.*, 123:899–909 (2008).

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

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

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

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

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

[29] C. Rao, M. Shah, T. S. Mahmood. Action Rectionition based on View Invariant Spatio-temporal Analysis. *ACM Multimedia*, (2003).

[30] T. Syeda-Mahmood Segmenting Actions in Velocity Curve Space. *Proc. ICPR*, vol. 4 (2002).

[31] 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).

[32] T. Brox, A. Bruhn, N. Papenberg and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *Proc. ECCV* (2004).

[33] Y. Rui and P. Anandan. Segmenting visual actions based on spatio-temporal motion patterns. *Proc. IEEE CVPR*, vol. 1, pp. 13–15 (2000).

[34] L. Wixson. Detecting salient motion by accumulating directionally-consistent flow. *IEEE Trans. PAMI*, 22:774–780 (2000).

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

[36] S. S. Chen, D. L. Donoho and M. A. Saunders Atomic Decomposition by Basis Pursuit. *SIAM Journal on Scientific Computing*, 20:33–61 (1999).

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

[38] Z. Barzelay and Y. Y.Schechner Experiments data. *www.ee.technion.ac.il/∼yoav/research/harmony-in-motion.html*.

[39] T. F. Quatieri. *Discrete-Time Speech Signal Processing :Principles and Practice*. Upper Saddle River, N.J. : Prentice-Hall PTR (2002).

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

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

[42] E. Vincent, R. Gribnoval and M. D. Plumbly Oracle Estimators for the Benchmarking of Source Separation Algorithms *Signal Processing*, 8:1933–1950 (2007).

[43] L. Benaroya, F. Bimbot and R. Gribonval. Audio source separation with a single sensor. In *IEEE Trans. on Audio, Speech & Language Processing*, 14:191–199 (2006)

[44] L. R. Rabiner and R. W.Schafer *Digital Processing of Speech Signals*. Englewood Cliffs, N.J. : Prentice-Hall (1978).

[45] A. P. Klapuri Multiple fundamental frequency estimation based on harmonicity and spectral smoothness In *IEEE Trans. on Speech & Audio Processing*, 11:804–816 (2003)

[46] J. Tabrikian, S. Dubnov and Y. Dickalov. Maximum a-posteriori probability pitch tracking in noisy environments using harmonic model In *IEEE Trans. on Speech & Audio Processing*, 12:76–87 (2004)

[47] G. Hu and D. L. Wang. Separation of stop consonants. *Proc. IEEE ICASSP*, 749–752 (2003).

[48] A. M. Reddy and B. Raj. Soft Mask Methods for Single-Channel Speaker Separation In *IEEE Trans. on Audio, Speech & Language Processing*, 15:1766–1776 (2007)

[49] M. H. Radfar and R. M. Dansereau. Single-Channel Speech Separation Using Soft Mask Filtering In *IEEE Trans. Speech and Audio Process.*, 8:2299–2310 (2007).

[50] A. Eronen and A. Klapuri. Musical instrument recognition using cepstral coefficients and temporal features. In *IEEE Proc. ICASSP*, vol. 2, pp. 753–756 (2000).

[51] D. Chazan, Y. Stettiner, and D. Malah. Optimal multi-pitch estimation using the EM algorithm for co-channel speech separation. In *Proc. IEEE ICASSP*, vol. 2, pp. 728–731 (1993).

[52] R. E. Crochiere and L. R. Rabiner. *Multirate Digital Signal Processing*. Englewood Cliffs, N.J. : Prentice-Hall (1983).

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