TEMPORAL SYNCHRONIZATION OF MULTIPLE AUDIO SIGNALS

Julius Kammerl, Neil Birkbeck, Sasi Inguva, Damien Kelly, A. J. Crawford,
Hugh Denman, Anil Kokaram, and Caroline Pantofaru

Google, Inc.
Mountain View, CA, USA

ABSTRACT

Given the proliferation of consumer media recording devices, events often give rise to a large number of recordings. These recordings are taken from different spatial positions and do not have reliable timestamp information. In this paper, we present two robust graph-based approaches for synchronizing multiple audio signals. The graphs are constructed atop the over-determined system resulting from pairwise signal comparison using cross-correlation of audio features. The first approach uses a Minimum Spanning Tree (MST) technique, while the second uses Belief Propagation (BP) to solve the system. Both approaches can provide excellent solutions and robustness to pairwise outliers, however the MST approach is much less complex than BP. In addition, an experimental comparison of audio features-based synchronization shows that spectral flatness outperforms the zero-crossing rate and signal energy.

Index Terms— Multi-signal synchronization, audio feature analysis, minimum spanning tree, belief propagation

1. INTRODUCTION

Due to the popularity of video sharing sites, there is an increasing amount of user-uploaded video content from different vantage points of the same event. For example, sports, concerts, and conferences might all have multiple attendees upload video of the event. Combining these recordings can provide richer user experiences through technologies such as free-viewpoint video [1], overview mash-ups [2, 3], and 3D scene reconstruction [4], but the input signals must first be time synchronized. Unlike the Gen-locked multi-camera rigs used in broadcast or cinema, consumer video is captured adhoc with different devices such as cell-phones, camcorders, or microphones, and must be synchronized after the event.

For video, there exists work on synchronizing two video streams using the geometric consistency of tracked visual features [5, 6, 7, 8]. However, these methods are only applicable when visual features are visible in both videos. Consequently, audio synchronization is widely used for outdoor motion capture [9], mash-ups, identifying video of the same event [10], and is available in commercial editing applications [11].

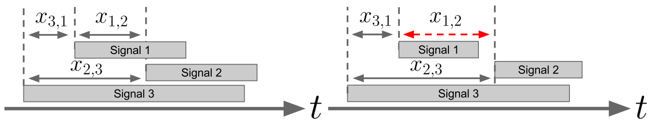


Fig. 1. Left: the inputs and desired solution for three signals. Arrows between the signals indicate a pairwise relationship. Right: an example where two signals do not overlap, so one pairwise offset (red) should not be included in the graph.

Synchronizing content captured of outdoor events on consumer devices is particularly challenging given that the microphones may be far apart or disjoint in time, and hence only partially share audio environments. In addition, common degradation due to compression and noise artifacts impairs the audio quality leading to inconsistencies within pairwise matches. RANSAC-inspired synchronization strategies can help improve pairwise matching [12] in order to produce the most likely temporal offset.

Most previous approaches to this problem performed a bottom-up temporal alignment of multiple signals, first matching signal pairs and then hierarchically merging clusters until a global solution was reached. For example, Bryan *et al.* use audio-fingerprinting [13] to match strongly correlating signal-pairs [14] that are iteratively merged into larger clusters until a global solution is found. Similarly, Shrestha *et al.* [15] and Cremer *et al.* [16] generate multiple audio-fingerprints for small segments in each audio track which are then individually matched against each other. Such bottom-up approaches are sensitive to the propagation of bad initial matching decisions, meaning all non-overlapping or poor candidate pairwise matches must be pruned in advance.

In contrast to these bottom-up approaches, we provide a formulation of the multi-signal synchronization problem as an over-determined graph of pairwise interactions (Fig. 1). Our global approach complements techniques that prune bad or non-overlapping pairwise matches (e.g., by thresholding or fingerprint consistency [14]), and provides additional robustness to any remaining outlier pairwise matches.

To ensure reliable pairwise offsets, we contribute a comparison of three audio features: spectral flatness, zero-crossing rate, and signal energy (§2.1). We then propose two novel graph-based formulations for robust multi-signal synchronization based on a minimum spanning tree approach and belief-propagation (§2.2). A quantitative evaluation of the feature-based matching and the proposed multi-signal synchronization methods is described in §3. Our results indicate that spectral flatness achieve the best performance in terms of robustness and selectivity. In addition, an experimental comparison of our multi-synchronization methods shows that both approaches achieve similar robustness to pairwise outliers and demonstrate resilience with up to 20% outliers. However, the minimum spanning tree solver is preferred since it is less complex than belief propagation.

2. MULTI-SIGNAL SYNCHRONIZATION

As input, we have N audio signals of the same event, $\{s_i\}_{i=1}^N$, where each signal s_i is a single-dimensional vector of length N_i. The multi-signal synchronization problem is to recover a consistent solution of temporal offsets, $x_{1:N} = (x_1, x_2, \cdots, x_N)$, such that the signals are brought into temporal alignment (Fig. 1). This is challenging due to the possible occurrence of temporally non-overlapping signal pairs, as well as noisy signals that hinder pairwise matching.

2.1. Pairwise matching of input signals

The first step is to obtain robust and accurate offset estimates for each signal pair. This is done by extracting a set of time-indexed audio features for each input signal and then cross-correlating the feature sets.

2.1.1. *Audio features*

A set of time-indexed audio feature coefficients $f\{s\}(t)$ is calculated for each input signal. The feature extraction emphasizes the descriptive audio events and increases robustness to noise, volume differences, etc. In the following, three popular audio features are described.

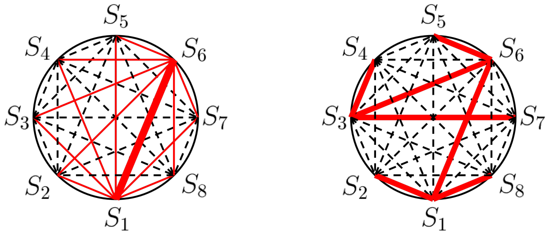
Spectral Flatness: The spectral flatness feature (also known as *Wiener entropy*) describes the variation of tonality over time. It is defined by the ratio of the geometric mean and arithmetic mean of the frequency domain coefficients:

$$f^{\mathrm{sf}}\{s\}(t) = \frac{\sqrt[\Omega]{\Pi_\omega F_s(t,\omega)}}{\frac{1}{\Omega}\sum_\omega F_s(t,\omega)} = \frac{\exp\left(\frac{1}{\Omega}\sum_\omega \log F_s(t,\omega)\right)}{\frac{1}{\Omega}\sum_\omega F_s(t,\omega)}, \tag{1}$$

where $F_s(t,\omega)$ is the power of wavelength $\omega \in \Omega$ at time t.

Zero-crossing Rate: The zero crossing rate is a popular feature in speech recognition for distinguishing between voiced and unvoiced speech segments. It counts the number of sign changes along a signal within a time window, T:

$$f^{\mathrm{zc}}\{s\}(t) = \frac{1}{T-1}\sum_{\tau=1}^{T} \mathbb{I}\left\{s(t+\tau)s(t+\tau-1) < 0\right\}, \tag{2}$$



(a) Calculation of weight $e_{1,6}$ (b) Minimal-spanning tree

Fig. 2. Minimum spanning tree solver approach. Fig. 2(a): the estimation of consistency weight e_{16} based on all overlapping 3-cliques. Fig. 2(b): a minimum spanning tree solution based on the most consistent correlations.

where $\mathbb{I}\{A\}$ is 1 if argument A is *true* and 0 otherwise.

Signal Energy: The signal energy feature computes the root mean square of the signal energy in a window of time T:

$$f^{\mathrm{nrg}}\{s\}(t) = \sqrt{\frac{\sum_{\tau=0}^{T-1} |s(t+\tau)|^2}{T}}. \tag{3}$$

2.1.2. *Pairwise Correlation*

To apply these features for synchronization, consider signals s_i and s_j that yield feature sequences f_i and f_j. The candidate alignment offset is given by the time offset x_{ij} of the maximum peak in the cross-covariance function of f_i and f_j.

$$x_{ij} = \arg\max_t \sum_{\tau \in T_{ij}} (f_i(\tau) - \bar{f}_i)(f_j(\tau+t) - \bar{f}_j(t)) \tag{4}$$

where $T_{ij}(t) = [\max(0,t), \min(T_i - 1, t + T_j - 1)]$ is the region of overlap.

2.2. Multi-signal synchronization

In this section, two independent techniques are proposed to reconcile any inconsistencies in the pairwise offset measurements. The first approach seeks to select a minimal set of consistent pairwise offset measurements to establish the global solution using a minimum spanning tree search. The second approach uses all pairwise hypotheses to define the marginal posteriors of the offset variables.

2.2.1. *Minimum spanning tree solver*

We can formulate the problem as a fully connected graph where each node is a signal and each edge weight is the temporal offset. This complete graph construction produces an over-determined system of $N(N-1)/2$ edges. We need to solve for only $N-1$ edges that form a spanning tree to produce a global synchronization solution.

Let us define a second graph with the same nodes and edges as the original but different edge weights. We will define this graph such that smaller edge weights reflect better correlation between signals. We can solve for the global offsets by finding a minimum spanning tree (MST) of this second graph. Critical to this approach is the definition of the edge weights. Possible choices include measurements on the pairwise match scores, such as the correlation score or peakiness of the correlation. However, such measurements may not be comparable across nodes. For this reason, we define a measure of pairwise matching quality that is independent of the correlation score. Notice that if the offsets could be measured without error, the sum of the offsets along any cycle in the original graph would be zero. Since measurements are never error-free, empirically the sum of a cycle is actually non-zero. So we define the penalty score of a given 3-clique in the graph as

$$z_{ijk} = |x_{ij} + x_{jk} + x_{ki}|. \tag{5}$$

Accordingly, for a given edge in the secondary graph e_{ij}, we define the total edge consistency weight as

$$e_{ij} = \sum_k z_{ijk}, \tag{6}$$

which is illustrated in Fig. 2(a). Prim's MST algorithm is applied to select the $N-1$ most consistent edges that span the nodes in the second weighted graph (Fig. 2(b)). These edges correspond to the most consistent offset hypotheses in the original graph structure.

The algorithmic complexity of the MST approach is $O(N^3)$ due to the calculation of consistency weights along all 3-cliques for each edge in the graph.

2.2.2. *Belief propagation approach*

The belief-propagation approach uses the hypothesis extracted from the pairwise analysis to build pairwise evidence,

$$\phi_{ij}(x) \propto \exp\left(\frac{-(x - x_{ij})^2}{2\sigma^2}\right) + c. \tag{7}$$

with c being a uniform offset prior. We model the joint probability distribution by combining the pairwise evidence, ϕ_{ij}, giving

$$p(x_{1:S}) \propto \prod_{ij} \phi_{ij}(x_j - x_i). \tag{8}$$

This leads to an ambiguity where $p(x_{1:S}) = p(x_{1:S} + t)$, so we fix one node as a reference and set it to $x_1 = 0$, giving

$$p(x_{2:S}) \propto \prod_{i>1,j>1} \phi_{ij}(x_j - x_i) \prod_{i>1} \phi_i(x_i). \tag{9}$$

The marginals of x in (9) are approximated through loopy belief propagation. At iteration $l \geq 1$, the message from node i to j is defined as

$$m_{ij}^l(x_j) = \int \phi_{ij}(x_j - x_i)\, \underbrace{\phi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{ki}^{l-1}(x_i)}_{\text{Partial belief}} \,\mathrm{d}x_i, \tag{10}$$

with the m_{ij}^0 defined either uniformly or randomly, and $\mathcal{N}(i)$ is the neighbors of i.

The *belief* at iteration l approximates the marginals and is defined using the propagated messages,

$$b_i^l(x_i) = \phi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{ki}^{l-1}(x_i). \tag{11}$$

Note that (10) is a convolution of the pairwise factor with the *partial belief*, which allows efficient message computation via the Fourier transform. The final solution after L iterations is $x_i = \arg\max_x b_i^L(x)$.

As loopy belief propagation is not guaranteed to converge, we try all possible nodes as the reference to obtain S hypotheses, $\{x_{1:S}^i\}_{i=1}^S$ and take the final solution as the one that maximizes a consistency score,

$$F(x_{1:S}) = \sum_i \sum_{j\in\mathcal{N}(i)} \phi_{ij}(x_j - x_i). \tag{12}$$

Since the BP algorithm calculates the discretized marginals for each edge in the tree using a FFT-based convolution of length N_i during L iterations, its computational complexity is $O(LN^2 N_i \log(N_i))$.

3. EXPERIMENTAL EVALUATION

In the first experiment, we investigate the influence of audio feature selection (§2.1.1) on the robustness and selectivity of the pairwise cross-correlation function. The second experiment compares the performance of the two proposed multi-signal synchronization techniques.

3.1. Feature comparison

To enable a quantitative comparison of the audio features, we generate a large benchmark data set with known ground truth. It contains multiple sets of signal pairs with varying temporal overlap ($t = 5s$ to $25s$ in $2s$ intervals) for three different recording scenarios: conference, concert and soccer. For each scenario and overlap amount, one hundred 30-second long signal pairs are extracted at random.

To investigate the impact of the feature selection on a cross-correlation-based synchronization process, each audio feature is used to generate a set of time-indexed feature coefficients for each signal pair. The extracted feature coefficients are normalized and the cross-correlation is computed. In this experiment, the peak-to-average-power-ratio (PAPR) is used as a measure for "peakiness" to evaluate the correlation characteristics at the simulated offset t in the cross-correlation function within a time window of 0.5s:

$$\mathrm{PAPR}_{f_t} = \frac{\max|f_t|^2}{P_f}, \tag{13}$$

where P_f is the average power of the discrete feature vector.

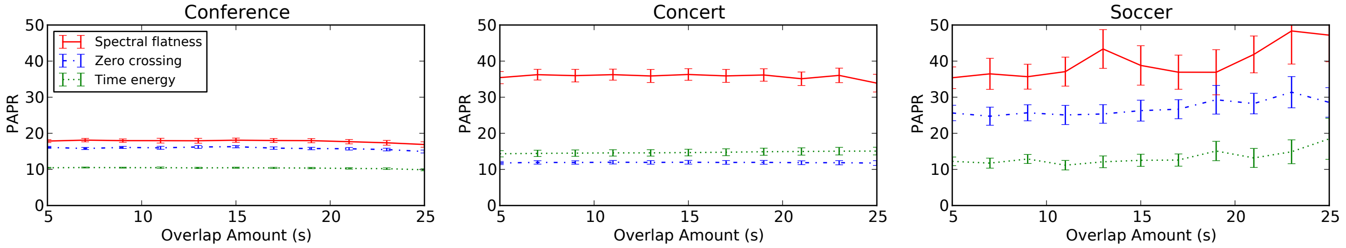


Fig. 3. Experimental results of the audio feature comparison for three different recording scenarios. Each plot shows the peakiness in the cross correlation function determined by the peak-to-average-power ratio (PAPR) as a function of signal overlap. The spectral flatness is the most selective feature as it has the strongest cross correlation peaks in all three scenarios.

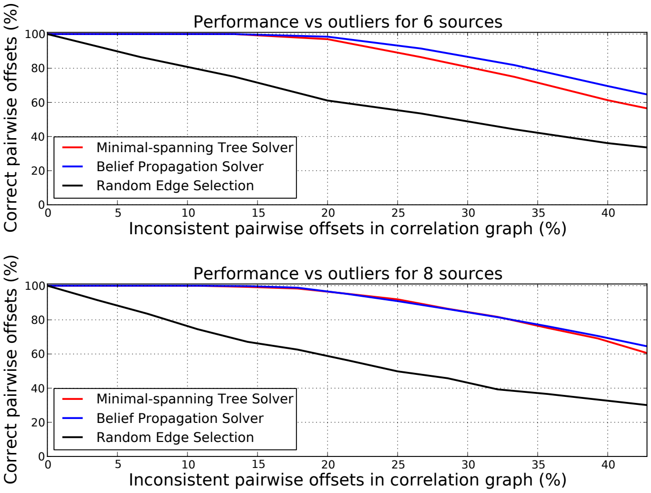


Fig. 4. Performance comparison between MST, BP, and random spanning trees for 6 and 8 source signals with varying outliers. The plots show the percentage of correctly identified offsets as a function of inconsistent signal pairs.

The experimental results of audio feature comparison are illustrated in Fig. 3 for the three recording scenarios, conferences, concerts and soccer games. Each plot shows the peak-to-average-power ratio (PAPR). The results indicate that the spectral flatness feature produces the best selectivity. The zero-crossing and time-energy features show similar performance on the conference and concert scenario. Interestingly in the soccer scenario, the zero-crossing feature outperforms the time-energy feature, which might be due to the strong interplay between background noise and player voices.

3.2. Multi-signal synchronization

To evaluate and compare the synchronization approaches discussed in Section 2.2, correlation graphs are generated based on randomized temporal offset values x. The effects of poorly matched pairs are simulated by selecting random graph edges and replacing them with uniformly distributed random values in the range of -7 to 7s, leading to inconsistencies in the graph. This way we obtain objective benchmark data with known ground truth that can be used to investigate the synchronization performance in terms of robustness against poorly matched signal pairs.

In the experiment, 1000 synthesized correlation graphs with 6 and 8 source nodes are used to evaluate the MST and BP algorithms, followed by comparing their output to the ground truth data. In order to obtain an objective measure of robustness that is independent from the absolute offset values, we calculate the percentage of correctly identified offset estimates with an estimation error $\Delta x \leq \frac{1}{64}s$, which is the sample rate used to represent the densities in the BP.

The outcome of the multi-signal synchronization for the MST and BP approaches are shown in Fig. 4. The plot shows the percentage of correctly identified offsets versus the number of inconsistent hypotheses. Additionally, the results based on randomly selected spanning trees within the correlation graph are shown as a lower bound on performance. For BP we use constant values of $\sigma = 0.25$, $c = 1$, and $L = 6$. The results show both MST and BP achieve similar performance in terms of robustness to pairwise outliers: for up to 20% outlier measurements in the correlation graph all offsets within the correlation tree are correctly recovered with high probability. In terms of computational complexity, as MST is only dependent on the number of nodes in the graph, it significantly outperforms the BP algorithm (milliseconds vs seconds), making it the preferred method for multi-signal synchronization.

4. CONCLUSION

This paper has presented two techniques for temporally aligning multiple audio signals by exploiting redundancies within an over-determined system of pairwise offset hypotheses. Experiments revealed that robust temporal alignment of signal pairs can be achieved by cross-correlating sequences of spectral flatness feature coefficients. In order to obtain a globally optimized synchronization solution, the minimum spanning tree solver can be applied on multiple offset hypotheses, showing excellent robustness against outliers in our experiments. Future work will focus on the synchronization of multiple disjunct signal sets and the integration of multiple offset hypotheses per signal pair.

5. REFERENCES

[1] M. Tanimoto, M. P. Tehrani, T. Fujii, and T. Yendo, “Free-viewpoint TV.,” *IEEE Signal Processing Magazing*, vol. 28, no. 1, pp. 67–76, 2011.

[2] M. Saini, R. Gadde, S. Yan, and W. T. Ooi, “MoViMash: online mobile video mashup,” in *Proceedings of the 20th ACM International Conference on Multimedia*, 2012, pp. 139–148.

[3] P. Shrestha, H. Weda, M. Barbieri, E. Aarts, et al., “Automatic mashup generation from multiple-camera concert recordings,” in *Proceedings of the International Conference on Multimedia*. ACM, 2010, pp. 541–550.

[4] C. R. Dyer, “Volumetric scene reconstruction from multiple views,” in *Foundations of Image Understanding*, pp. 469–489. Springer, 2001.

[5] A. Elhayek, C. Stoll, K.I. Kim, H.-P. Seidel, and C. Theobalt, “Feature-based multi-video synchronization with subframe accuracy,” in *Pattern Recognition*, vol. 7476 of *Lecture Notes in Computer Science*, pp. 266–275. 2012.

[6] E. Dexter, P. Prez, and I. Laptev, “Multi-view synchronization of human actions and dynamic scenes.,” in *Proceedings of the British Machine Vision Conference*, 2009.

[7] C. Lu and M. Mandal, “An efficient technique for motion-based view-variant video sequences synchronization,” 2011, Proceedings of the Multimedia and Expo (ICME), pp. 1–6.

[8] C. Rao, A. Gritai, M. Shah, and T. Syeda-Mahmood, “View-invariant alignment and matching of video sequences,” in *Proceedings of the Ninth IEEE International Conference on Computer Vision*, 2003, pp. 939–945 vol.2.

[9] N. Hasler, B. Rosenhahn, T. Thormählen, M. Wand, J. Gall, and HP. Seidel, “Markerless motion capture with unsynchronized moving cameras,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 224 – 231.

[10] C. Cotton and D. Ellis, “Audio fingerprinting to identify multiple videos of an event,” in *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 2386–2389.

[11] Adobe, “Adobe premier pro cc,” http://www.adobe.com/products/premiere.html.

[12] F. Schweiger, G. Schroth, M. Eichhorn, E. Steinbach, and M. Fahrmair, “Consensus-based cross-correlation,” in *Proceedings of the 19th International Conference on Multimedia*. 2011, pp. 1289–1292, ACM.

[13] J. Haitsma and T. Kalker, “A highly robust audio fingerprinting system.,” in *Proceedings of the International Symposium on Music Information Retrieval (ISMIR)*, 2002.

[14] N. J. Bryan, P. Smaragdis, and G. J Mysore, “Clustering and synchronizing multi-camera video via landmark cross-correlation,” in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 2389–2392.

[15] P. Shrstha, M. Barbieri, and H. Weda, “Synchronization of multi-camera video recordings based on audio,” in *Proceedings of the 15th International Conference on Multimedia*. ACM, 2007, pp. 545–548.

[16] M. Cremer and R. Cook, “Machine-assisted editing of user-generated content,” *SPIE Electronic Imaging*, vol. 7254, 2009.