FINE LANDMARK-BASED SYNCHRONIZATION OF AD-HOC MICROPHONE ARRAYS

Tsz-Kin Hon, Lin Wang, Joshua D. Reiss and Andrea Cavallaro

Centre for Intelligent Sensing, Queen Mary University of London, UK
{tsz.kin.hon, lin.wang, joshua.reiss, a.cavallaro}@qmul.ac.uk

ABSTRACT

We use audio fingerprinting to solve the synchronization problem between multiple recordings from an ad-hoc array consisting of randomly placed wireless microphones or hand-held smartphones. Synchronization is crucial when employing conventional microphone array techniques such as beamforming and source localization. We propose a fine audio landmark fingerprinting method that detects the time difference of arrivals (TDOAs) of multiple sources in the acoustic environment. By estimating the maximum and minimum TDOAs, the proposed method can accurately calculate the unknown time offset between a pair of microphone recordings. Experimental results demonstrate that the proposed method significantly improves the synchronization accuracy of conventional audio fingerprinting methods and achieves comparable performance to the generalized cross-correlation method.

Index Terms— Synchronization, audio fingerprinting, microphone array

1. INTRODUCTION

Ad-hoc acoustic sensor networks composed of randomly distributed wireless microphones or hand-held smartphones have been attracting increased interest due to their flexibility in sensor placement [1]. A challenge in such ad-hoc arrays is that the locations of the microphones are generally unknown and there is no precise temporal synchronization between the microphones. Traditional microphone-array techniques, such as beamforming and sound source localization, which rely on the knowledge of microphone positions and assume sample-synchronized audio channels, cannot be applied directly [2,3]. The synchronization problem between multiple audio channels has been addressed using generalized cross-correlation [2, 4, 5] and audio fingerprinting [5–8].

The generalized cross-correlation (GCC) method [9], which calculates the delay that maximizes the correlation coefficient between two audio channels, is well known for its accurate time delay estimation and synchronization. However, due to high computational cost, GCC is more suitable for audio channels that are already coarsely synchronized [4, 5]. Another synchronization approach is based on audio fingerprinting, which has been originally applied to music information retrieval [6], and clustering and synchronizing multi-camera videos [7, 8]. By matching the audio fingerprints extracted from the sound track, the audio channels can be synchronized. Well-known audio fingerprinting algorithms include onset, audio landmark, and Philips Robust Hash (PRH). Onset is based on detecting the increase of the signal energy and consists of audio features extracted for example from 24 individual frequency bands [10, 11]. PRH extracts more detailed features including all the energy changes in consecutive bands and frames, and is more robust than onset to ambient noise and audio compression distortion [12]. Audio landmark operates in the time-frequency domain, and the extracted features consist of a set of time-frequency pairs [6–8]. With low computational cost and robustness against noise, audio fingerprinting techniques have been widely used in movie and video synchronization [13]. However, the synchronization accuracy achieved by conventional audio fingerprinting methods is limited by the time-frequency analysis hop size, with typical values between a few and tens of milliseconds [10], which is enough for video applications but far below the sample-wise synchronization requirement in microphone array signal processing.

In this paper, we investigate audio-landmark-based synchronization, aiming at sample-wise alignment within an ad-hoc microphone array. We show that by reducing the time-frequency analysis hop size, the classical landmark audio fingerprinting algorithm is able to detect the time difference of arrival (TDOA) of nearby sound sources that are captured by different microphones. Existing audio fingerprinting methods usually use a large hop size, e.g. tens to hundreds milliseconds, in the application of music information retrieval or video synchronization. The TDOA information of multiple sources becomes ambiguous with such a hop size. We further exploit this property to detect the maximum and minimum TDOAs around the microphone array, which can be used to further improve the synchronization accuracy. A sample-accuracy synchronization can be achieved with the proposed algorithm if a sufficient number of sound sources is located around the array.

This work was supported by the U.K. Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/K007491/1.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. PRELIMINARIES

Consider an anechoic environment with an ad-hoc microphone array consisting of M independent microphones and unknown number of N sources randomly distributed around the array. The locations of the microphones and sources are unknown and are denoted as $\boldsymbol{M} = [\boldsymbol{m}_1, \cdots, \boldsymbol{m}_i, \cdots, \boldsymbol{m}_M]$ and $\boldsymbol{R} = [\boldsymbol{r}_1, \cdots, \boldsymbol{r}_b, \cdots, \boldsymbol{r}_N]$, respectively. The recording $a_i(t)$ at each microphone is asynchronous with unknown start time $T_i, i = 1, \cdots, M$. Let the pairwise time offset between two microphones i and j be denoted as

$$T_{ij} = T_i - T_j. \tag{1}$$

The signal recorded at each microphone is denoted as $a_i(t) = \sum_{b=1}^{N} g_{ib} s_b(t - t_{ib}), \quad i = 1, \cdots, M$, where $s_b(\cdot)$, t_{ib} and g_{ib} are the b-th source signal, the propagation time and the attenuation from the b-th source to the i-th microphone, respectively. The propagation time of arrival (TOA) t_{ib} from the b-th sound source to the i-th microphone is $t_{ib} = \frac{||\boldsymbol{r}_b - \boldsymbol{m}_i||}{c}$, where c is the speed of sound and $||\cdot||$ denotes the Euclidean distance. The TDOA of the b-th source between the i-th and j-th microphones, τ_{ijb}, can be expressed as

$$\tau_{ijb} = \frac{||\boldsymbol{r}_b - \boldsymbol{m}_i|| - ||\boldsymbol{r}_b - \boldsymbol{m}_j||}{c} + T_{ij}. \tag{2}$$

The goal is to blindly estimate the time offset T_{ij} between each microphone pair from the microphone recordings. Without loss of generality, we consider only two microphones i and j.

In [2], a GCC-based framework is proposed to estimate the extreme (maximum and minimum) TDOAs of the sources around the microphone array, which are further utilized to estimate the time offset between two asynchronous recordings. It is assumed in [2] that, among the N sources, there are always two sources locating at the end-fire positions with respect to the microphone pair. The end-fire positions are defined as the points on a line that connects the two microphones excluding the ones between the two microphones (Fig. 1). Based on the reverse triangle inequality ($||\boldsymbol{r}_b - \boldsymbol{m}_i|| - ||\boldsymbol{r}_b - \boldsymbol{m}_j|| \leq ||\boldsymbol{m}_i - \boldsymbol{m}_j||$), the maximum or minimum TDOAs among the N sources can be identified once the sources are located at the end-fire locations, and can be expressed as

$$\tau_{\max} = \frac{||\boldsymbol{m}_i - \boldsymbol{m}_j||}{c} + T_{ij}; \quad \tau_{\min} = -\frac{||\boldsymbol{m}_i - \boldsymbol{m}_j||}{c} + T_{ij}. \tag{3}$$

Naturally, the time offset T_{ij} is obtained as

$$T_{ij} = \frac{\tau_{\max} + \tau_{\min}}{2}. \tag{4}$$

In this way, the goal of time offset estimation becomes the task of extreme TDOA estimation. In this paper, we use the same assumption in [2], and show how the maximum and minimum TDOAs can be estimated with the proposed landmark-based audio fingerprinting algorithm.

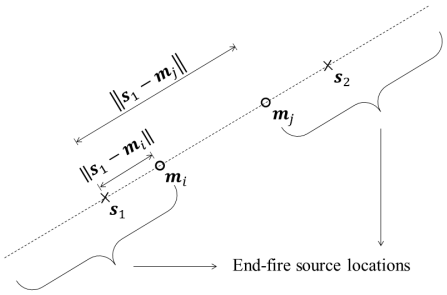

Fig. 1. Illustration of the end-fire source locations.

3. AUDIO LANDMARK AND SINGLE-SOURCE TDOA

The audio landmarks are usually used for coarsely synchronizing two audio recordings [5–7]. However, the extracted landmark features contain some valuable information about the TDOA information of the sound sources.

The classical audio landmark fingerprinting method operates in the time-frequency domain and converts a time-domain signal $a_i(t)$ into a sparse, high-dimensional, and discrete-time landmark feature set $\mathbb{F}_i(n)$ [6]. At first, the time-domain signal $a_i(t)$ is transformed into the short-time Fourier transform (STFT) $A_i(n, k)$, where n is the frame index and k is the frequency index, which downsamples the time axis via the STFT hop size R. The onsets of local frequency peaks are detected from the STFT and are represented as sparse time-frequency points, $f(n_p, k_q)$, where p and q denotes the frame and frequency indices of the detected local peak, respectively. Landmarks are formed by pairing up each two nearby local peaks, represented as $\boldsymbol{y}_i(n_1, k_1; n_2, k_2)$. Finally, the obtained landmarks associated with the frame n are hashed into a time-indexed feature set represented as $\mathbb{F}_i(n)$.

Assume that only the b-th source is active and the time offset between two microphones is zero. The STFT of $a_i(t)$ and $a_j(t)$ can be expressed as

$$\begin{cases} A_i(n, k) = g_{ib} S_b(n - n_{ib}, k) \\ A_j(n, k) = g_{jb} S_b(n - n_{jb}, k) \end{cases}, \tag{5}$$

where $S_b(n, k)$ is the STFT of $s_b(t)$, $n_{ib} = \lfloor t_{ib}/R \rfloor$, and $n_{jb} = \lfloor t_{jb}/R \rfloor$, where the operator $\lfloor \cdot \rfloor$ denotes the integer part.

By matching landmarks between the two channels [6], the landmarks corresponding to the same time-frequency peak pairs can be extracted. This is expressed as

$$\boldsymbol{y}_i(n_1 - n_{ib}, k_1; n_2 - n_{ib}, k_2) = \boldsymbol{y}_j(n_1 - n_{jb}, k_1; n_2 - n_{jb}, k_2), \tag{6}$$

and consequently

$$\mathbb{F}_i(n) = \mathbb{F}_j(n - \lfloor \frac{\tau_{ijb}}{R} \rfloor), \tag{7}$$

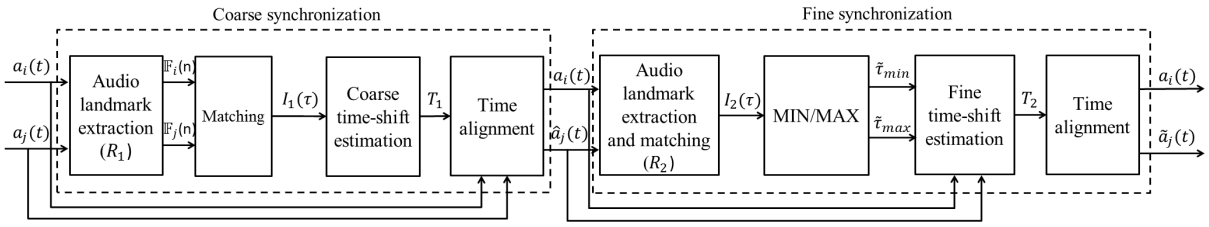

Fig. 2. Block diagram of the proposed coarse-to-fine synchronization method.

where $\mathbb{F}_i(\cdot)$ and $\mathbb{F}_j(\cdot)$ denote the extracted audio fingerprints of $a_i(\cdot)$ and $a_j(\cdot)$, respectively, and $\boldsymbol{y}_i(\cdot)$ and $\boldsymbol{y}_j(\cdot)$ denote two matched local peak pairs in the two channels.

Let

$$I_{ij}(\tau) = I(\mathbb{F}_i(n), \mathbb{F}_j(n-\tau)) \tag{8}$$

be the correlation function denoting the matching score between $\mathbb{F}_i(n)$ and a time-shifted version $\mathbb{F}_j(n-\tau)$. The TDOA of the b-th source is obtained by matching the audio landmarks, i.e.

$$\hat{\tau}_{ijb} = \arg\max_{\tau}\{I_{ij}(\tau)\} \cdot R. \tag{9}$$

The correlation function (matching score) $I_{ij}(\cdot)$ is calculated from the number of matched landmarks between two channels, with more details provided in [6]. The TDOA is detected as the peak of the correlation function. This is similar to the GCC method, but here the correlation function is calculated by matching the audio landmarks.

4. TIME OFFSET ESTIMATION

4.1. Extreme TDOA estimation

To estimate the maximum and minimum TDOAs in a multi-source environment, we employ a fine landmark strategy. As shown in (7), the resolution of the landmark-based TDOA estimation is confined by the STFT hop size R. An improved resolution can be achieved by reducing the hop size. A hop size as small as $R = 1$ can be used so that a sample-wise resolution can be achieved. In this way, the TDOAs of different sources can be distinguished from each other in the correlation function $I_{ij}(\tau)$, appearing as different peaks.

We apply a simple peak detector to $I_{ij}(\tau)$, obtaining a set of P peaks expressed as $\mathbb{T} = \{\tau_1, \cdots, \tau_P\}$. The maximum and minimum TDOAs can be estimated from $\mathbb{T}$ as

$$\hat{\tau}_{\text{min}} = \min(\mathbb{T}) \;\text{ and }\; \hat{\tau}_{\text{max}} = \max(\mathbb{T}). \tag{10}$$

4.2. Coarse-to-fine synchronization

The fine landmark strategy requires a small time-frequency analysis hop size. The computation of landmark feature extraction and matching would become extremely intensive when a large time offset exists between two channels and a large searching area of τ is required to calculate $I_{ij}(\tau)$. To improve the computation efficiency, we propose a coarse-to-fine synchronization scheme (Fig. 2).

In the coarse synchronization stage, the audio landmarks are extracted from the STFT spectra of $a_i(t)$ and $a_j(t)$, with a large hop size R=R_1. After landmark matching, the coarse time offset is estimated from the correlation function $I_1(\tau)$ as

$$T_1 = \arg\max_{\tau}\{I_1(\tau)\} \cdot R_1. \tag{11}$$

The channel $a_j(t)$ is time shifted with T_1 to coarsely align it with $a_i(t)$. This is expressed as

$$\hat{a}_j(t) = a_j(t - T_1). \tag{12}$$

In the fine synchronization stage, the audio landmarks are extracted from the STFT spectra of $a_i(t)$ and $\hat{a}_j(t)$, with a small hop size R=R_2=1. After landmark matching, the maximum and minimum TDOAs can be estimated from $I_2(\tau)$, using (10). The precise time offset is estimated as

$$T_2 = \frac{\tilde{\tau}_{\text{max}} + \tilde{\tau}_{\text{min}}}{2}. \tag{13}$$

Finally, $\hat{a}_j(t)$ is time shifted with T_2, obtaining $\tilde{a}_j(t)$.

The coarse synchronization stage is suitable for recordings with large time offsets, while the fine synchronization stage can improve the synchronization accuracy. Combining the two stages, the time offset between two recordings is

$$\tilde{T}_{ij} = T_1 + T_2. \tag{14}$$

5. EXPERIMENTAL RESULTS

Two experiments are conducted to evaluate the performance of the proposed method. In the first experiment, the audio features extracted by three popular audio fingerprinting algorithms (audio landmark (AL) [1] [6], onset [11], PRH [12]) are compared to show the unique TDOA detection ability of the AL algorithm. In the second experiment, we compare with real recorded data the synchronization performance of the mentioned three audio fingerprinting algorithms, and also

[1]The AL algorithm is implemented using the code from [14].

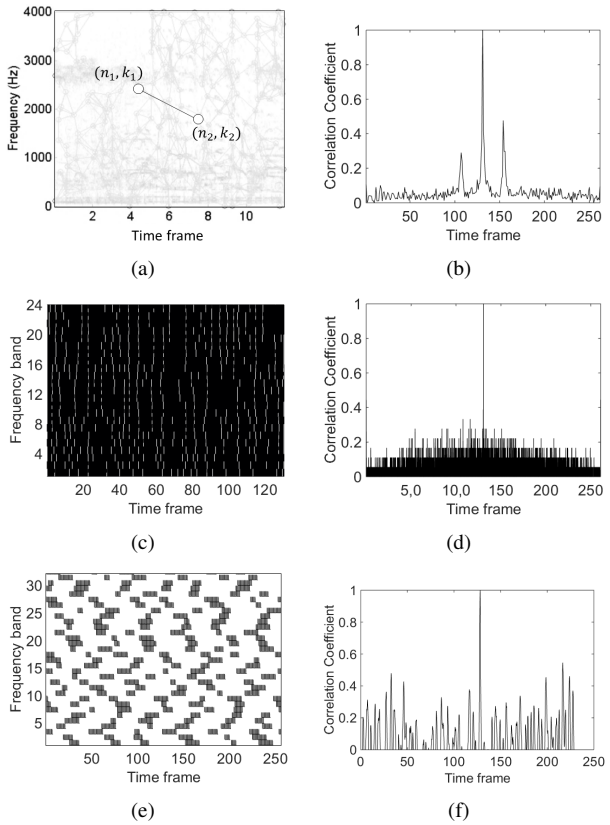

Fig. 3. Feature analysis of three audio fingerprints on TDOA estimation. The left column is the illustration of the features and the right column is the corresponding correlation results for (a)-(b) audio landmark, (c)-(d) onset, (e)-(f) PRH.

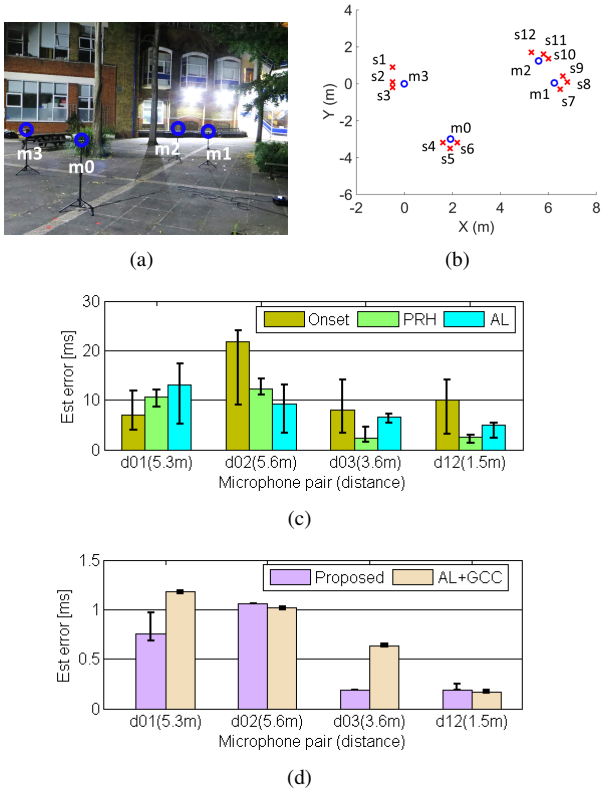

Fig. 4. Performance comparison with real data. (a) Public square where the data is collected. (b) Source and microphone locations. (c) Performance of the 3 audio fingerprinting methods. (d) Performance of the proposed method and the AL+GCC method. The results are represented by median values of 10 independent realizations. The error bars are the first and third quartiles.

compare the synchronization performance of the proposed method and the GCC-based method [2].

The GCC-based synchronization method [2] also calculates the extreme TDOAs for the estimation of the time offset. When implementing this algorithm, we use the AL algorithm as a preprocessing step to coarsely synchronize the data. This preprocessing step is the same as the coarse synchronization stage of the proposed method. The absolute error $\epsilon = |T_{ij} - \tilde{T}_{ij}|$ is used as the objective measure for performance comparison, where T_{ij} is the ground-truth time offset. The time offset between two sound tracks is uniformly randomly selected in the range $[-1, 1]s$.

In the first experiment, we used simulated sound tracks recoded by a microphone pair placed $5m$ apart from each other. Three speech sources are talking concurrently, with two placed at two end-fire locations and one placed in the middle between two microphones. We apply AL, onset, and PRH to perform fingerprint extraction and matching on the two channels. The same hop size of $0.001s$ is used for the three algorithms. Fig. 3 shows the audio feature extraction and matching results by the three algorithms, with left column depicting the extracted features, the right column depicting the matching results, and each row representing one algorithm. As can be seen from the 2nd-3rd rows, both onset and PRH extract a large amount of audio features, and obtain one peak in the correlation functions. In contrast, as shown in the first row, the features extracted by AL consist of sparse peak pairs. Since the matched peak pair comes from the same source, multiple peaks can be observed in the correlation function, with each peak denoting one source. In Fig. 3(b), three peaks are clearly observed, which is equal to the number of sources. With this experiment, the unique TDOA estimation ability of audio landmarks is demonstrated.

In the second experiment, we recorded sound tracks using four Samsung Galaxy III smartphones at $8kHz$ sampling in a $12m \times 12m$ public square. The square and locations of the smartphones are shown in Fig. 4 (a) and (b), respectively. Twelve source positions were set according to the end-fire locations of each microphone pair in Fig. 4(b) and two sources (1 male and 1 female speech) about $20s$ were played simultaneously by a loudspeaker in each recording. We selected four microphone pairs for the comparison with microphone

distances $d_{01} = 5.31m$, $d_{02} = 5.60m$, $d_{03} = 3.55m$ and $d_{12} = 1.47m$.

As seen in Fig. 4(c), the three algorithms generally achieve bigger estimation errors for larger microphone distances. This is because they only detect the highest peak in the correlation function to estimate the time offset. When the distance between two microphones increases, the highest peak in the correlation might deviate from the true value. For example, the deviation can be as much as $0.017s$ when the microphone distance is $6m$. Among the three algorithms, PRH performs best among the three audio fingerprint algorithms, while onset is the least robust one. PRH combines features in neighbouring frames [12], and thus is more robust to TDOA deviation. Onset only uses the features extracted in individual frames, and thus is least robust. AL uses peak pairs in neighbouring frames, and its performance is in the middle of the other two. Overall, the estimation errors of the three algorithms range between several and tens milliseconds.

As seen in Fig. 4(d), both the proposed method and the AL+GCC method can achieve estimation errors less than $0.001s$ and the accuracy is ten times higher than the conventional audio fingerprinting methods. This is because they both exploit the maximum and minimum TDOAs to estimate the time offset instead of using the highest peak of the correlation function. Both algorithms achieve higher errors when the microphone distance is large. This is because the extracted features become weak when energies of the sound received by the microphones decrease with the microphone distance, leading to deviated estimations of the extreme TDOAs. The proposed method performs slightly better than the AL+GCC method.

6. CONCLUSIONS

In this paper, landmark-based fingerprinting has been refined to detect the TDOAs of the sources in a multi-source environment. Comparison among three fingerprint features demonstrates that only the landmark features can be fined-tuned for TDOA estimation. By estimating the maximum and minimum TDOAs, the proposed method can estimate the time offset between two microphone recordings efficiently. A complete coarse-to-fine synchronization framework is further proposed to deal with recordings with large time offsets. When end-fire sources exist, the proposed method can achieve accuracy comparable to the recent fine synchronization based on GCC using sound tracks recorded in a real environment, and can significantly improve the accuracy of the conventional landmark method.

One assumption of the proposed method is the requirement of a sufficient number of sound sources around the array (i.e. end-fire sources). Further work is required to finely synchronize recordings without the need of end-fire sources.

REFERENCES

[1] A. Bertrand, "Applications and trends in wireless acoustic sensor networks: a signal processing perspective," in *Proc. SCVT*, 2011, pp. 1 – 6.

[2] P. Pertila, M.S. Hamalainen, and M. Mieskolainen, "Passive temporal offset estimation of multichannel recordings of an ad-hoc microphone array," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 11, pp. 2393–2402, Nov. 2013.

[3] N. Ono, H. Kohno, N. Ito, and S. Sagayama, "Blind alignment of asynchronously recorded signals for distributed microphone array," in *Proc. WASPAA*, 2009, pp. 161–164.

[4] A.B. Nielsen and L.K. Hansen, "Synchronization and comparison of lifelog audio recordings," in *Proc. MLSP*, 2008, pp. 474–479.

[5] N.Q.K. Duong, C. Howson, and Y. Legallais, "Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation," in *Proc. ICCE*, 2012, pp. 241 – 244.

[6] A.L. Wang, "An industrial-strength audio search algorithm," in *Proc. ISMIR*, 2003, pp. 7–13.

[7] N.J. Bryan, P. Smaragdis, and G.J. Mysore, "Clustering and synchronizing multi-camera video via landmark cross-correlation," in *Proc. ICASSP*, 2012, pp. 2389–2392.

[8] N.Q.K. Duong and F. Thudor, "Movie synchronization by audio landmark matching," in *Proc. ICASSP*, 2013, pp. 3632–3636.

[9] C.H. Knapp and G.C. Carter, "Generalized correlation method estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 24, no. 4, pp. 320327, Aug. 1976.

[10] P. Shrestha, M. Barbieri, H. Weda, and D. Sekulovski, "Synchronization of multiple camera videos using audio-visual features," *IEEE Trans. Multimedia*, vol. 12, no. 1, pp. 79–92, Jan. 2010.

[11] J. E. Schrader, "Detecting and interpreting musical note onsets in polyphonic music," M.S. thesis, Eindhoven University of Technology, Netherlands, 2003.

[12] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. ISMIR*, 2002, pp. 107–115.

[13] J.S. Seo, "An asymmetric matching method for a robust binary audio fingerprinting," *IEEE Signal Process. Lett.*, vol. 21, no. 7, pp. 844–847, Jul. 2014.

[14] D.P.W. Ellis, *Robust Landmark-Based Audio Fingerprinting*, web resource, 2009, available: http://labrosa.ee.columbia.edu/matlab/fingerprint/.