Synchronization Ambiguity in Audio Content Generated by Users Attending the Same Public Event

Nikolaos Stefanakis
Foundation for Research and Technology – Hellas,
Institute of Computer Science,
70013 Heraklion, Crete, Greece
nstefana@ics.forth.gr

Athanasios Mouchtaris
Foundation for Research and Technology – Hellas,
Institute of Computer Science,
70013 Heraklion, Crete, Greece
mouchtar@ics.forth.gr

.

Abstract
Exploiting correlations in the audio, several works in the past have demonstrated the ability to automatically match and synchronize User Generated Recordings (UGRs) of the same event. The synchronization process is of fundamental importance as it provides the basis for combining the different sources of content in order to improve the audiovisual experience of the captured event. In this paper, we show that depending on the complexity of the sound scene, the time offsets required to synchronize the audio recordings are not unique, and depend on the locations and the activity of the sound sources. We use simulation results to illustrate that this problem is very likely to occur in athletic events and we demonstrate how it may impair the listening experience.

Author Keywords
audio synchronization; audio fingerprinting; content based management; user generated content.

ACM Classification Keywords
H.5.5 Sound and Music Computing

Introduction
Given a collection of UGRs, several approaches have been proposed about how to exploit the available visual

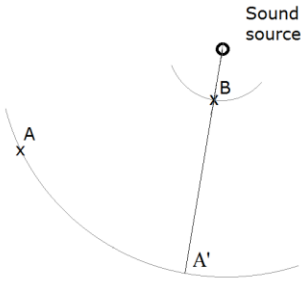

Figure 1: Synchronization of two UGRs acquired at locations A and B under the assumption of a single sound source. The time-offset required for synchronizing the two recordings are uniquely defined.

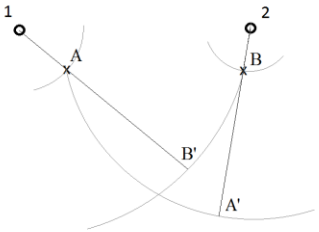

Figure 2: Synchrony ambiguity problem illustrated for the case of two UGRs at locations A and B, under the assumption of a two distant sound sources at locations 1 and 2. The time-offset required for synchronizing one recording with the other are not uniquely defined any more.

and audio content - as well as several types of metadata - in order to identify video clips associated to the same moment of the captured event, to estimate the overlap between these clips and to synchronize them along the same temporal axis. The audio content is a key to solving this problem and several works have shown that the relations between different UGRs can be revealed by exploiting the correlations in their associated audio streams [1,3,4,7,8].

An emerging research challenge now is to investigate different means by which this low-quality but organized content can be synergistically processed and combined, so as to construct a new sequence which improves the audiovisual experience of the captured acoustic event. The potential is particularly interesting with respect to the audio modality, as a multitude of synchronized UGRs essentially provides a multichannel recording of the acoustic event. By combining the available content, it becomes possible to produce a new acoustic sequence of increased duration, enhanced quality and enriched spatial impression. In [9] it is shown that audio streams resulting as a simple monophonic or stereophonic mixture of the different sources of content carry a significant potential for improving the listening experience of the captured event, as opposed to when each UGR is consumed individually. In the same direction, the authors in [5,6] propose collaborative signal processing tools as the means to enhance the most interesting sound components of the sound scene, at the same time suppressing noise and interference which is unwanted.

Successful synchronization of UGRs is thus an essential requirement for supporting the novel forms of content production that have begun to emerge in the context of user generated content. However, under certain conditions, synchronization poses difficulties that have not been addressed so far. In the vast majority of the works dealing with UGR synchronization, it is assumed that the time-offset required in order to time align two or more overlapping UGRs is uniquely defined. However, it is often the case that the fine time-offsets required for perfect synchronization are not unique and depend on the locations and activity of the different sound sources comprising the sound scene. As it will be shown, this may significantly degrade the listening experience transmitted to the user when simple forms of audio mixing are used in order to combine the UGRs.

Ideal case of UGR synchronization

Assume that two users at locations A and B simultaneously record the same event using their smartphones, as shown in Fig. 1. The analysis that follows further assumes that the content of interest originates from a single sound source confined in a small region of space. This scenario is representative of many cultural public events, such as outdoor concerts, where the dominant sound source is the public address system which is used for sound reinforcement. Due to the different propagation distances between sources and sensors, the acoustic waves transmitted from the sound source arrive with different delays at locations A and B. In particular, if c is the speed of sound, the time difference of arrival (TDOA) between locations A and B can be calculated by dividing the length of linear segment (BA') with the speed of sound, i.e., τ_{AB}=(BA')/c. Let now $s_A(t)$ and $s_B(t)$ symbolize the sound streams captured at locations A and B respectively and lets for simplicity assume that the two devices started recording sound simultaneously. A simple procedure to combine the two recordings is to superimpose a delayed version of $s_B(t)$ on $s_A(t)$ to

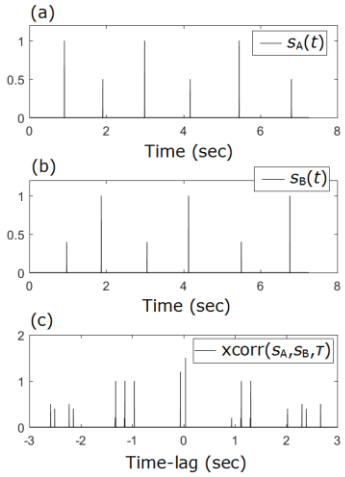

Figure 3: Simulation illustrating the sound signal captured at location A in (a) and location B in (b), with respect to the configuration shown in Fig. 2. The cross-correlation of the two recordings is shown in (c).

derive $s(t)=s_A(t)+s_B(t-\tau_{AB})$. Alternatively, we may use $s_B(t)$ as the reference and superimpose an advanced in time version of $s_A(t)$ to derive $s(t)=s_A(t+\tau_{AB})+s_B(t)$. In any of the two cases, a free from artifacts mixture of the two recordings is derived, as common content will be played back synchronously in both the reference and the delayed component.

Synchronization ambiguity in complex sound scenes

We now present an example considering two sound sources, indexed by 1 and 2 and two recording locations, A and B, as shown in Fig. 2. For reasons of simplicity, we assume that the sound sources and the recording locations are fixed with time and that devices A and B started recording sound simultaneously. Now, as shown in Fig. 2, sound source 1 is closer to location A than in B, while sound source 2 is closer to location B than in A. One can think of this setup as a tennis game taking place in an official size tennis court. The sounds produced by the players when they hit the ball with their rackets are the dominant acoustic components of the sound scene. However, due to the different acoustic paths, sound source 1 arrives earlier and with a higher amplitude in location A than in B, while sound source 2 arrives earlier and with a higher amplitude in B than in A. If we represent the strikes as Diracs, we may easily simulate this problem in the time domain as shown in Fig. 3. Specifically, we assume that, in the studied time interval, each player produces three strikes and that the time difference between successive strikes is random. The signal recorded in location A is shown in Fig. 3(a) while the signal captured in B is shown in Fig. (b).

We would like now to combine the two recordings in order to produce a better acoustic representation of the tennis event in comparison to the case that each recording is reproduced individually. Evidently, if we use recording A (resp. B) only, the strikes of player 2 (resp. 1) will be very weak in comparison to those of player 1 (resp. 2). An obvious choice would be to synchronize and superimpose the two recordings so that both players' actions are perceived equally loud. Using the standard approach, we calculate the cross-correlation of signals $s_A(t)$ and $s_B(t)$ (shown in Fig. 3(c)), hoping to observe a clear indication about the time-lag which is required to synchronize the two recordings. Indeed, the highest peak is close to $\tau=0$, which is the time-lag that we would expect since the two devices started recording simultaneously. However, instead of single clear peak at zero time-lag, we observe two weaker peaks slightly before and after $\tau=0$. This demonstrates the **synchronization ambiguity** problem which is inherent to the acoustic coverage of large acoustic scenes, involving spatially distributed sound sources and distant recording locations. A consequence of synchronization ambiguity is that the time-offset required to perfectly synchronize the audio recordings depends on the location of the sound sources and their activity along time. To see that, lets assume that we rely on the left cross-correlation peak at $\tau_L=-(AB')/c$ to synchronize and superimpose the two recordings. The mixing process can be represented though the new sound stream $s(t)=s_A(t+\tau_L)+s_B(t)$, so that the two recordings are perfectly aligned with respect to the strikes of player 1. Unfortunately, as shown in Fig. 4(a), player 1 strikes are indeed correctly aligned, but the strikes of player 2 appear duplicated, in the form the direct sound and an echo. Had we used the right cross-correlation peak at

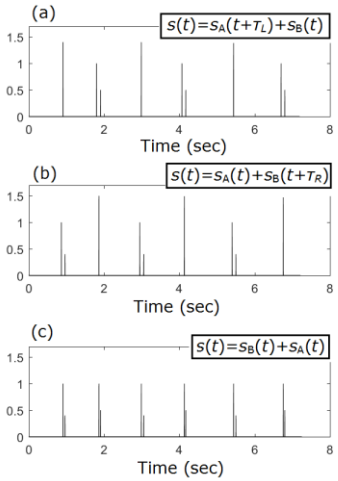

Figure 4: The sound signal produced by mixing the two recordings when synchronizing with respect to the left cross-correlation peak in (a) and the right cross-correlation peak in (b). The result of mixing the two recordings based on clock synchrony is shown in (c).

τ_R=-(BA')/c to synchronize the two recordings, then the opposite problem would happen, as shown in Fig. 4(b); in this case, the unwanted echoes are associated to the strikes of player 1.

Interestingly, even if we use clock synchrony to align the two recordings, the problem will not disappear. In Fig. 4(c) we illustrate the amplitude as a function of time for the mixture $s(t)=s_A(t)+s_B(t)$. It can be seen that the time separating the two closely spaced Diracs is reduced, but the echoes now appear at both players' strikes. This reveals that, depending on the complexity of the sound scene and the recording locations, **clock synchrony does not** always **imply content synchrony**. Furthermore, the time domain signals in Fig. 4(a), (b) and (c) demonstrate that it is impossible to devise of a Linear Time Invariant (LTI) process to combine the two recordings in a way that avoids the appearance of the unwanted echoes. As explained in the next paragraphs, one must go beyond linear mixing to solve the synchronization ambiguity problem.

Discussion

The presented problem of synchronization ambiguity demonstrates that simple forms of linear combination of the audio content captured in geographically spread public events may lead to audible artifacts. Results from psychoacoustics demonstrate that humans can easily discriminate two similar sounds arriving with a time difference larger than 10 msec. Considering that the speed of sound in normal conditions is approximately 340m/sec, synchronization ambiguity is inevitable to happen in athletic events, where the distance between sound sources and spectators are very large. For example, in a football match, the action in the field is distributed inside a rectangle of length 100 m or more, while the distance between spectators can be even larger. This means that the time difference between the direct sound and the echo can be as high as 100/340=0.29 sec, far above the inaudible limit of 10 msec.

It is evident that one should use more advanced techniques than the previously shown delay and sum approach in order to overcome the synchronization problem. We believe that exploiting topological information is an important prerequisite towards this direction. For example, with respect to the example of Fig. 2, if we could somehow infer that location A (resp. B) is closer to source 1 (resp. 2) than location B (resp. A), then, we could select only the portions of the signal in A (resp.B) which are representative of the activity of sound source 1 (resp. 2) and eliminate the signal portions which originate from source 2 (resp.1). This means that we would end with two different audio streams, each one carrying the content produced from a single sound source. This operation implies the use of clever audio masking (or gating) techniques for removing the unwanted signal components. Alternatively, one can borrow ideas from acoustic echo cancellation techniques [2] in order to jointly process the two recordings to remove the unwanted components.

Conclusion

Synchronization ambiguity problem is a fundamental problem when dealing with user generated audio content acquired in large complex sound scenes such as in the case of athletic events. In such environments, the finer time-offsets required in order to synchronize one or more audio recordings are not unique and depend on the locations of the sound sources and their

activities along time. As a consequence, simple delay and add method to mix the UGRs results to the appearance of audible echoes, which may significantly degrade the listening experience.

Acknowledgements

The project leading to this application has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 687605, Project_COGNITUS.

References

1. Cotton C. and Ellis D. 2010. Audio fingerprinting to identify multiple videos of an event. 2010. in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.* (ICASSP 10), 2386–2389.
2. Hänsler E. and Schmidt G. 2003. Single-channel acoustic echo cancellation. In *Adaptive signal processing: applications to real-world problems*, Benesty J. and Huang Y. (Eds.). Springer Berlin Heidelberg, 59-93.
3. Kammerl J., Birkbeck N., Inguva S., Kelly D., Crawford A., Denman H., Kokaram A. and Pantofaru C. 2014. Temporal synchronization of multiple audio signals. In *Proceedings of the Int. Conf. Acoust., Speech, Signal Process. Process.* (ICASSP 14), 4603–4607.
4. Kennedy L. and Naaman M. 2009. "Less talk, more rock: Automated organization of community-contributed collections of concert videos," in Proceedings of the 18th international conference on World Wide Web, 311-320.
5. Kim M. and Smaragdis P. 2013. Collaborative audio enhancement using probabilistic latent component sharing. In *Proceedings of the Int. Conf. Acoust., Speech, Signal Process.* (ICASSP 13), 896-900.
6. Kim M. and Smaragdis P. 2016. Efficient neighborhood-based topic modelling for collaborative audio enhancement on massive crowdsourced recordings. In *Proceedings of the Int. Conf. Acoust., Speech, Signal Process.* (ICASSP 16).
7. Shrestha P., Barbieri M. and Weda H. 2007. Synchronization of multi-camera video recordings based on audio. In *Proceedings of the 15th ACM international conference on Multimedia*, 545–548.
8. Stefanakis N., Chonianakis S. and Mouchtaris A. 2017. Automatic matching and synchronization of user generated videos from a large scale sport event. In *Proceedings of the Int. Conf. Acoust., Speech, Signal Process.* (ICASSP 17), 3016-3020.
9. Stefanakis N., Viskadouros M. and Mouchtaris A. 2017. A subjective evaluation on mixtures of crowdsourced audio recordings. Submitted to Europ. Sig. Process. Conf. (EUSIPCO 17).