Fast second screen TV synchronization combining audio fingerprint technique and generalized cross correlation

Ngoc Q. K. Duong, Christopher Howson, Yvon Legallais
Technicolor
1 avenue de Belle Fontaine, Cesson Sevigne, France
Email: {Quang-Khanh-Ngoc.Duong, firstname.lastname}@technicolor.com

***Abstract*—For the implementation of emerging second screen TV applications, there is a need for a technique to assure fast and accurate synchronization of media components streamed over different networks to different rendering devices. One approach of great value is to exploit the unmodified audio stream of the original media, and compare it to a reference version. We consider two major approaches for this purpose, namely fingerprinting techniques and generalized cross correlation, where the former can greatly reduce computational cost and the latter can offer sample-accurate synchronization. We propose an approach combining these two techniques where coarse frame-accurate synchronization positions are first found by fingerprint matching, then a possible accurate synchronization position is verified by generalized cross correlation with phase transform (GCC-PHAT). Experimental results in a real-world setting confirm the accuracy and rapidity of the proposed approach.**

***Index Terms*—Audio synchronization, audio fingerprinting, generalized cross correlation, streaming media, TV.**

I. Introduction

More and more consumers today are using a second screen device (laptop, tablet or smartphone) whilst watching TV. This has opened the door to personalized TV applications [1] where additional services and related content can be accessed on the web to accompany the main TV view. Examples of such emerging services include offering a user the possibility of selecting alternative views and/or audio soundtracks delivered via broadband and rendered on his tablet, in conjunction with the main broadcast content displayed on a TV set [2]. The need for a high level of synchronization accuracy is clear in such cases; an alternative video on the tablet must be frame accurate with the TV content and a user selected soundtrack must not result in any perceptible lip-sync error. In order to implement such applications, the accurate synchronization of the streamed contents, delivered over different networks and rendered on different devices, is a real challenge as we must account for different network delays, protocols and timing models.

For content synchronization in hybrid networking scenarios, several original research directions are applicable such as audio watermarking [3], timeline insertion [2], audio fingerprinting [4], [5], and cross correlation [6]. As approaches using watermarking or a timeline require the insertion of an additional signature into the media content, we seek an approach which does not impact the original content. In this paper, we investigate a solution which directly exploits the streamed audio to achieve accurate synchronization by means of a combination of fingerprinting-based and cross correlation-based audio matching techniques.

An example of our synchronization scenario is that a second screen device records a piece of audio signal from the main content rendering on a TV. This recorded signal may then be matched with a master audio stream to determine where it is in the ongoing program. In this way, the second screen device is able to retrieve associated content from a web server and play it at the appropriate time. For audio matching, the fingerprinting technique [7], [4], [8] is amongst the best choices due, in particular, to its robustness to many kinds of distortion affecting the recording and also to its computational efficiency. Hence, audio fingerprinting has been widely investigated in the literature and already been deployed in many recent commercialized products [9], [10], [11], [12]. However, state-of-the-art fingerprinting techniques generally require several seconds of recording for reliable matching, resulting in a five to ten second wait for users after they start the service. In order to reduce the length of this unwelcome wait, whilst still guaranteeing good matching reliability, we propose an approach which works with a much shorter recording time. By combining a fingerprinting technique, to find coarse synchronization positions, and a cross correlation technique, to validate each coarse position and to further refine the alignment, we are able to achieve sample-accurate synchronization.

The success of this combination has been discussed in a different context of lifelog audio recording [13] where audio segments from the same environments were classified and synchronized. However in [13] the sample cross correlation was directly computed in the time domain whilst, in this paper, we perform faster computation in the time-frequency domain via the generalized cross correlation with phase transform (GCC-PHAT) [6], which is also more robust to noise. Let us also clarify that, since correlation techniques are required to know the original signal, both audio streams and its pre-computed fingerprints, metadata must be accessible in the database.

The structure of the rest of the paper is as follows. We first present two audio coherence measures, namely fingerprinting

and cross correlation, in Section II. We then describe the proposed approach in Section III followed by real-world experimental evaluation in Section IV. Finally, we conclude and discuss possible future work in Section V.

II. Audio coherence measures

A. Fingerprinting technique

Audio fingerprinting, also known as audio hashing, is well-known as a powerful technique to perform audio identification and frame resolution synchronization. It has been widely employed in music recognition systems [7], [9] for matching the observed audio signal with its origin stored in a large database. This technique basically involves two major steps: fingerprint extraction and matching search, as depicted in Fig. 1. The first step derives and encodes a set of relevant audio features which is required to be invariant to various kinds of signal distortion, such as background noise, audio compression, A/D conversion, etc. Fingerprints of the original audio collection and its corresponding metadata (*e.g.* audio ID, name, time frame index) are systematically stored in a database. Then, given a short recording from the user side, its fingerprints are computed in the same way as they were for the original data. Finally, a searching algorithm will find the best match between these fingerprints and those stored in the database so that the recorded audio signal is labeled by the matched metadata.

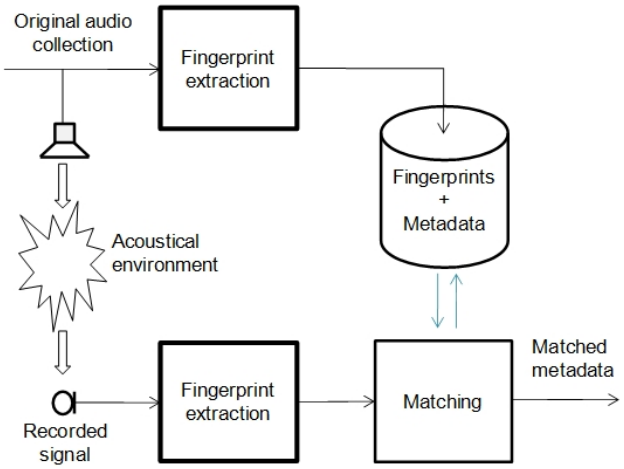

Fig. 1. General architecture of an audio fingerprinting system.

Though the proposed approach can work with any fingerprinting methods, we adapt the well-known Philips' algorithm [7] in our system where the audio signal is first segmented into overlapping frames of 0.37 second length with an overlapping factor of 31/32, and weighted by a Hanning window. The spectral representation is computed by performing a Fourier transform over every time frame. Then only the magnitude of the spectrum is retained and mapped to 33 Bark-scale frequency bands lying in the range from 300 Hz to 2000 Hz. Finally, only the sign of the energy band difference over both time and frequency neighbors is extracted. Let us denote by $E_{n,m}$ the energy of n-th frame and m-th Bark scale frequency band, $m = 1, 2, ..., 33$, the bits of the fingerprint are defined as

$$F_{n,m} = \begin{cases} 1 & \text{if } E_{n,m} - E_{n,m+1} - \left(E_{n-1,m} - E_{n-1,m+1}\right) > 0 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

This strategy results in a very compact 32-bit fingerprint extracted for each interval of an 11.6 millisecond audio signal downsampled to 5 kHz. Note that this very short frameshift interval is indeed relevant to the short recording duration we are targeting and the resulting fingerprint block size will vary depending on the length of the available query signal instead of being fixed by 256 as in [7]. In the searching step, Hamming distances between the query's fingerprint bock and fingerprint blocks of the original audio signals stored in a Hash table are quickly computed in order to specify a match.

Our experiments with one and two second recordings from movie soundtracks showed that the Philips's fingerprint offers better synchronization accuracy than other less compact fingerprints derived directly from well-known audio features such as mel-frequency cepstral coefficients (MFCC), spectral centroid, or spectral flatness.

B. Generalized cross correlation

Cross correlation is probably the simplest and most straightforward measure of the similarity between two audio signals realigned according to a given time lag. It is therefore often adopted to estimate the time delay between two signals. In the presence of background noise and reverberation, generalized cross correlation with phase transform (GCC-PHAT) [6] was shown to be more effective than the conventional time-domain cross correlation implementation and hence widely applied in the source localization community [14].

Let us consider two signals $x(t)$ and $y(t)$, and denote by $X(\omega)$ and $Y(\omega)$ their Fourier transforms, respectively. The time lag $\hat{\tau}$ between two signals estimated by GCC-PHAT is given by

$$\hat{\tau} = \text{argmax}_{\tau}\, R_{xy}(\tau) \quad (2)$$

where

$$R_{xy}(\tau) = \int_{-\infty}^{\infty} \frac{X(\omega)Y^{\star}(\omega)}{|X(\omega)Y^{\star}(\omega)|} e^{j\omega\tau} d\omega \quad (3)$$

is the generalized cross correlation function for a given τ; * denoting the complex conjugate. $R_{xy}(\tau)$ is computed as the inverse Fourier transform of the cross spectrum weighted by the amplitude so that only the phase information is taken into account. In favorable conditions, $R_{xy}(\tau)$ offers a predominant peak with respect to the actual time lag. Fig. 2 shows an example of the time lag estimation by GCC-PHAT where the top two plots depict, respectively, an audio signal and its delayed and distorted version and the lowest plot displays the cross correlation value $R_{xy}(\tau)$ with a predominant peak at the delay time. When the level of noise increases, GCC-PHAT was shown to provide better performance than the Adaptive

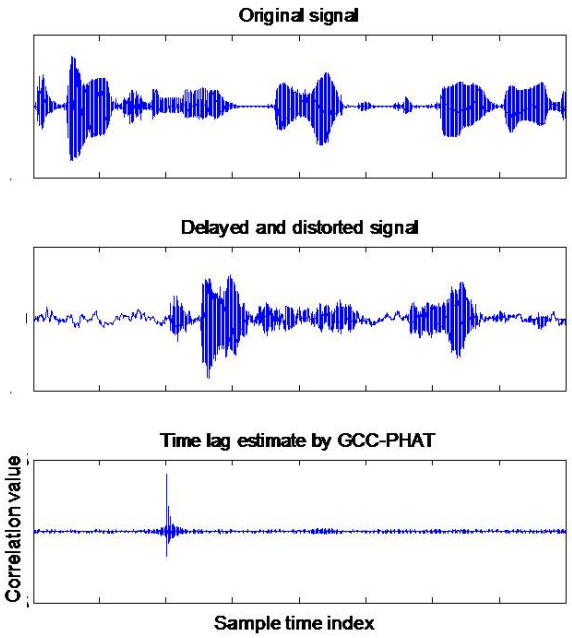

Fig. 2. Time lag estimation by GCC-PHAT.

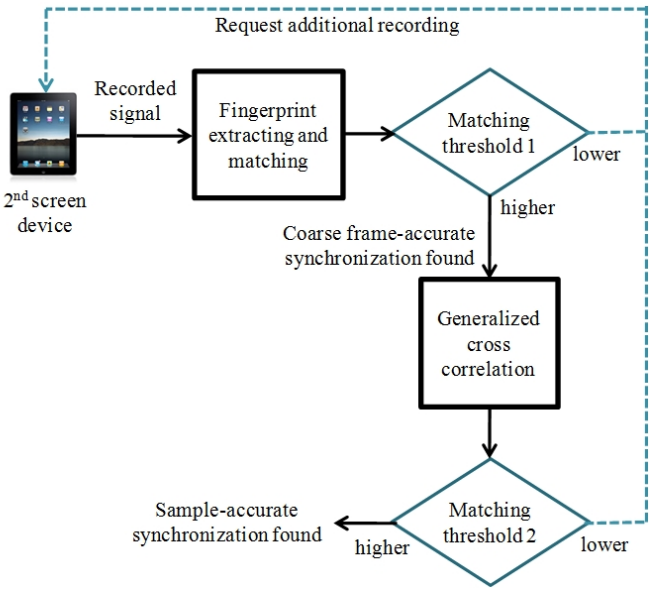

Fig. 3. Workflow of the proposed synchronization approach.

Eigenvalue Decomposition (AED) method while requiring less computational expense [15].

III. Proposed Synchronization Approach

Though each of the two audio coherence measures presented in Section II could be used on its own for synchronization, our experiments have shown that error rate increases significantly with the reduction of the recording duration, which is what we are targeting to enhance the user experience. We therefore propose a combination workflow as shown in Fig. 3 where a fingerprinting technique is first applied to quickly find potential coarse synchronization positions in the original audio stream, *i.e.* audio track indexes and elapsed time from the beginning of the tracks, that may match with the recorded signal. The use of very short recordings typically results in multiple synchronization positions and so, in the second step, the similarity between each piece of the original audio stream around these positions and the recorded signal is verified by GCC-PHAT, thereby determining a confident match. This process is intuitively explained in Fig. 4. The upper plot shows Hamming distances between a Philips' fingerprint block of a one second recording query and the fingerprint blocks extracted from an original movie sound track where any of the three minimum peak positions A, B, and C can be roughly specified as potential matches. However, when computing cross correlation between the query samples and pieces of the original signal centered on those three positions, only position B results in a predominant peak, as shown in the lower plot, and therefore it is finally determined as a correct match while positions A and C are eliminated. Note that in this two step combination, since GCC-PHAT applies only to a small number of the candidate positions resulting from fingerprint matching, its computation is very small and can be negligible.

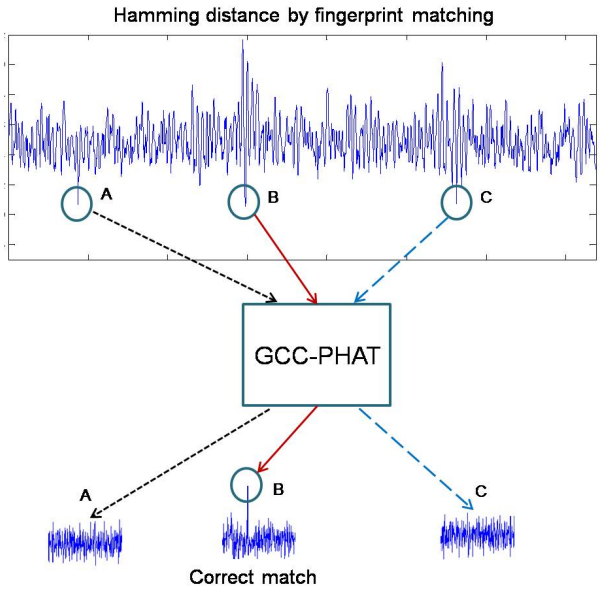

Fig. 4. An example of the correct synchronization (position B) determined by the cross correlation when fingerprint matching offers multiple local minimum distances.

In a small number of cases, when the similarity measure by either fingerprint matching or GCC-PHAT is lower than a predefined threshold, an additional duration of recording from the second screen device could be requested until synchronization is found. Such rare cases are likely to happen when the recording is performed during a silent period of the audio stream or when the background noise level is unusually high.

IV. Experiment

In order to validate the performance of the proposed system, we made 300 real-world recordings in a living room by means of an iPad placed at a distance ranging from one to three metres from a TV set. The room reverberation time is about 250 ms and the average signal-to-noise (SNR) ratio is approximately 15 dB. Each recording was started at a random elapsed time from the beginning of a 75 minute movie and lasted for 3 seconds. Fingerprints of the original movie soundtrack were pre-computed and stored in a memory together with the signal itself. The synchronization performance is evaluated in terms of *precision*, *i.e.* the fraction of detected synchronization positions that are correct, and *recall*, *i.e* the fraction of correct synchronization positions that are detected. In order to avoid false positive detection, in the considered application, we target the highest precision when the first piece of recording is used while recall can be improved when using additional pieces of recording.

TABLE I
Synchronization performance with one second recordings

Approach	Precision	Recall
Philips' fingerprint	0.96	0.94
GCC-PHAT	0.87	0.81
Proposed approach	1.0	0.92

When the full 3 second recordings were used to find the synchronization positions with the original audio stream, we observed that either fingerprinting matching or cross correlation method can find all the correct matching positions, *i.e.* both precision and recall are equal to 1. When only the first second of each recording is used, the synchronization performance resulting from fingerprint matching, cross correlation method, and the proposed combination approach are shown in Table I. It is not surprising that the Philips's fingerprint matching offers better performance than cross correlation method since fingerprints are specifically designed to be highly robust against various types of signal distortion. However, fingerprint matching alone still results in 4 percent false positive detection meaning that 4 percent of cases user may get undesired content rendering on their second screen. These harmful errors are eliminated with the proposed approach where its resulting precision is 1.

In summary, with the proposed combination approach, we observed a 92 percent rate of correct detection when only the first second of each recording was used. An additional one and two seconds of recording was requested for the remaining 5 and 3 percent of cases, respectively, so that correct detection was then achieved. These results demonstrate that in over 90 percent of cases a user could start to enjoy synchronized personalized TV content on their second screen device in as little as one second after the start of the service.

V. Conclusion

We have described an accurate second screen TV component synchronization system which does not require the addition of time reference data to the original media content. The proposed approach relies on only the audio stream in the original content and has been shown to work with very short recording lengths, thereby minimizing the user wait. Our approach combines the characteristics of the fingerprinting technique, which scales better with long contents, and of the cross correlation technique, which scales better for the synchronization resolution. More importantly, the coherent matching offered by both methods can improve the precision. Our preliminary experimental results indicate that a high level of synchronization accuracy can be achieved for a recording period as short as one second.

Future work will be devoted to a more complete experimental evaluation with a large original audio stream database and a larger number of queries recorded in different noisy conditions. We also envisage the implementation of the proposed synchronization approach in real-world scenarios, *e.g.* personalized audio service and second screen multi-view video application [2], where a subjective measure of the user experience (*i.e.* QoE) can be conducted.

References

[1] M. Fink, M. Covell, and S. Baluja, "Social- and interactive-television. applications based on real-time ambient-audio identification," in *Proc. European Interactive TV Conference (Euro-ITV)*, 2006.

[2] C. Howson, E. Gautier, P. Gilberton, A. Laurent, and Y. Legallais, "Second screen TV synchronization," in *Proc. IEEE Int. Conf. on Consumer Electronics - Berlin (ICCE-Berlin)*, 2011, pp. 361–365.

[3] H.J.Kim, Y.H.Choi, J.W.Seok, and J.W.Hong, "Audio watermarking techniques," in *Intelligent Watermarking Techniques*, chapter 8, pp. 185–218. 2004.

[4] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of algorithms for audio fingerprinting," in *IEEE Workshop on Multimedia Signal Processing*, 2002, pp. 169–173.

[5] V. Chandrasekhar, M. Sharifi, and D. A. Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications," in *12th International Society for Music Information Retrieval Conference (ISMIR)*, 2011, pp. 801–806.

[6] M.S. Brandstein and H.F. Silverman, "A robust method for speech signal time-delay estimation in reverberant rooms," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 1997, pp. 375–378.

[7] J. Haitsma and T. Kalker, "A highly robust audio fingerprinting system," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, 2002, pp. 107–115.

[8] Y. Ke, D. Hoiem, and R. Sukthankar, "Computer vision for music identification," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 597–604.

[9] A. L-C. Wang, "An industrial-strength audio search algorithm," in *Proc. Int. Sym. on Music Information Retrieval (ISMIR)*, 2003, pp. 1–4.

[10] R. Macrae, X. Anguera, and N. Oliver, "Muvisync: Realtime music video alignment," in *Proc. IEEE Int. Conf. on Multimedia and Expo (ICME)*, 2010.

[11] TVplus website, "http://www.tvplus.com/," .

[12] Into-Now application from Yahoo, "http://www.intonow.com/ci/soundprint," .

[13] A. B. Nielsen and L. K. Hansen, "Synchronization and comparison of lifelog audio recordings," in *Proc. IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, 2008, pp. 474–479.

[14] C. Blandin, A. Ozerov, and E. Vincent, "Multi-source tdoa estimation in reverberant audio using angular spectra and clustering," *Signal Processing*, pp. 1928–1936, 2012.

[15] A. Brutti, M. Omologo, and P. Svaizer, "Comparison between different sound source localization techniques based on a real data collection," in *Proc. IEEE Wokshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, 2008.