Audio Fingerprinting for Media Synchronisation and Duplicate Detection

Rolf Bardeli, Jochen Schwenninger and Daniel Stein
Fraunhofer Institute for Intelligent Analysis and Information Systems
Schloss Birlinghoven, 53754 Sankt Augustin, Germany

***Abstract*—With the ever-increasing offer of television content as internet broadcast streams, synchronisation of this material with second screen applications has received considerable interest over the last years. We introduce a novel audio fingerprinting method which can be easily implemented, and offer promising experiments on German news show material. Further, we evaluate possible additional usage for audio fingerprinting in the context of duplicate detection whenever different media shows with the same topic recycle shared media fragments. We give an outlook in how far knowledge drawn from our algorithm can be used to enhance a viewer's experience, by offering automatic skipping and/or recommendation functionality.**

I. Introduction

In Germany, roughly 71% of people under the age of 30 simultaneously use the internet in parallel to watching television [1], studies for America [2] even report that 86% of mobile internet users utilize their mobile device while watching TV. Whenever the television shows are to be enriched with second screen applications specifically designed for this content, the devices typically need to be synchronized, even more so when the original video is offered in an on-demand internet streaming platform or consumed via time-shifted recordings where it cannot be known in advance when a given content is viewed.

While the general topic of media synchronization has received considerable interest over the last years, some techniques developed there fail to translate to internet streaming applications. Also, some of the existing technologies like [3] or [4] are not easily includable in commercial usage due to intellectual property constraints. The problem of audio fingerprinting itself has found ample attention and the field has found a pretty stable algorithmic basis which is described, for example, in [5]. Recent developments in this field — like this work — are typically tied to specific applications, especially in the areas of mobile search and social networks (see [6] for a recent review).

In this paper, we present a novel audio fingerprinting technique which is easy to implement and offers promising results in our experiments. Further, we evaluate possible additional usage for duplicate detection within media shows on the same topic. For a personalized viewer's experience, this knowledge can be used to automatically skip already seen material, recommend similar material or offer more in-depth parts of the program. We evaluate the performance of the fingerprints with German broadcast streams provided by Public Service Broadcaster Rundfunk Berlin-Brandenburg (RBB),[1] as part of the joint LinkedTV project.[2]

II. LinkedTV

The vision of LinkedTV is of a ubiquitously online cloud of Networked Audio-Visual Content decoupled from place, device or source. The aim is to provide an interactive multimedia service for non-professional end-users, with focus on television broadcast content as seed videos. The project workflow can be described as follows: starting from the demands of the use case scenarios, coupled with a description of the targeted multimedia content, the videos are analyzed by various (semi-)automatic ways. The raw data obtained from the single approaches is gathered and further enriched in a second step, by assigning media fragment descriptions and interlinking these with other multimedia information, using knowledge acquired from, e.g., web mining.

At this step, the knowledge of duplicates within the videos is especially valuable in the news show scenario foreseen for the project. Here, each day several different shows have access to shared content material. As opposed to video fingerprinting/visual frame redetection, audio fingerprinting seems most appropriate for the detection algorithm because of three reasons: first, the visual material is not exactly the same but rather adopted to each show, e.g., by customized banner overlays. Second, the same visual information might have been re-used in a different context, e.g., the same scenic background overview of a certain site like the German parliament might be used in totally different topics. The third reason is that audio fingerprints can be expected to have a more favorable real-time factor.

Further down in the LinkedTV workflow, the enriched videos are then shown in a suitably tailored presentation engine which allows the end-user to interact with a formerly linear video, and a recommendation/personalization engine which further gives the possibility to customize this experience.

III. Audio Fingerprints

The audio fingerprinting algorithm presented here relies on detecting characteristic features in the spectrogram of given recordings. For indexing user-defined segments, the corresponding audio stream is continuously converted into a

[1] www.rbb-online.de
[2] www.linkedtv.eu

ISBN - 978-90-5986-410-8

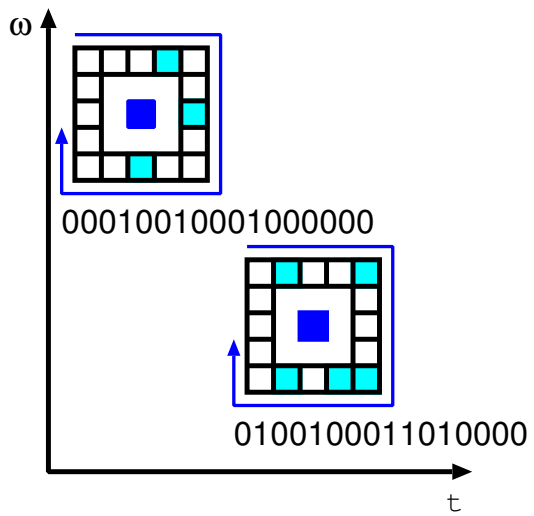

Fig. 1. Diagram depicting the audio fingerprint extraction algorithm. The dark blue box in the center represents the central point of the fingerprint.

sequence of spectrogram windows. The frequency range is mapped into 20 roughly logarithmically spaced bins.

For each time position t (with a step size of 20 ms), and each frequency bin b, a fingerprint f is extracted. Each such fingerprint is extracted by comparing the signal's energy at the central point (t, b) to a set of 16 temporal and frequency-wise neighbors as indicated in Figure 1. This results in a fingerprint represented by a bit string f of length 16. If the signal in such a neighboring area of (t, b) has sufficiently higher energy than the central value, the corresponding bit is set to 1, and to 0 otherwise. The fingerprint is kept only if at least one bit is different from zero.

In order to achieve robustness against channel distortions and similar deviations of the signal, each extracted bit sequence is decoded using a linear block code, resulting in a codeword c for each fingerprint f. To this end, we employ the binary $(24, 12, 8)$ Golay code allowing to correct up to 3 bit errors in each fingerprint. The 16 bit fingerprints are zero padded to a length of 24 bits to be used together with this code. The resulting codewords c have a block size of 12 bits. For each time window, a random selection of 500 extracted triples (t, b, c) consisting of time and frequency information as well as the codeword are stored in an index structure which is then persisted. This results in having several hundred descriptors belonging to a segment of five seconds.

During retrieval time, the complete audio signal is processed in the same way as for indexing. Periodically, the extracted descriptors are used as queries for the index structure, which returns all occurences of similar descriptors in the indexed material. If the amount of descriptors belonging to a certain indexed segment is sufficiently high, if their temporal order is correct and if they occur in a comparable duration, a match is declared.

Several measures for detecting distorted versions of the original material have been implemented: The use of a linear block code allows for a certain amount of bit errors in the fingerprints, which correspond to energy fluctuations in the underlying signal. By allowing small timing deviations in the order of the spectrogram's Fast Fourier Transform length, different positions of the spectrogram windows are compensated. And finally, the choice of thresholds for the percentage of matched descriptors and their minimal duration allows the adaptation of the algorithm to different use cases with varying requirements regarding precision and recall.

IV. Experiments

In this section, we offer first results of the fingerprinting technique on the target material.

A. Media Synchronization

First experiments using the fingerprinting algorithm for media synchronisation were carried out using a German TV show. At intervals of ten seconds, one event was indexed, with the goal of detecting these events in potentially distorted material including commercial breaks.

The indexing of 16:30 m of original material with 92 events takes only a couple of seconds on standard PC hardware, the index is smaller than 2 MBytes. The matching runs about 2.5 times faster than real time. Here, the precision of the matching is perfect, i.e., no false positives are detected, while the recall is almost perfect (one miss, i.e., 98.9%). We added white noise, red noise $(1/f)$ and brown noise $(1/f^2)$ on various signal-to-noise (SNR) ratios to the original signal (Figure 2). While there are still no false positives, the recall drops dramatically at around 20 dB SNR for all three types of noises. As could be expected, brown noise which does not distort as much on the whole frequency range performs best, but only slightly.

Artificial noise is a nice way to test the robustness in general, but it still tells little about actual use case environments. In a second screen application setting, presumably taking place at home, we can expect occasional (possibly loud) noise from sources like banging doors, telephone ringing, and other people talking. For an extreme scenario, we decided that noises that can be heard in a passenger train contain all these sounds, both electronically (announcements, automatic doors, engine sounds) and from persons (other travellers talking, giggling, coughing). Thus, one author recorded these environmental sounds during his daily commute, and we again mixed in these sounds at different SNR ratios (Figure 3). Here, the recall starts to drop only at 10 dB SNR, and still is acceptable at 5 dB SNR.

B. Duplicate Detection

For preliminary experiments of the fingerprinting technology, we selected seven different news show scenes from the online portal of RBB, on the topic of Berlin's new airport:

- one scene from *RBB um sechs* (2012/08/14), on the delay of its opening,
- two scenes from the *Abendschau* (2012/08/14), one on the delay of its opening, the other on a terrorist warning,
- one scene from *Brandenburg Aktuell* (2012/08/14), on the delay of its opening

ISBN - 978-90-5986-410-8

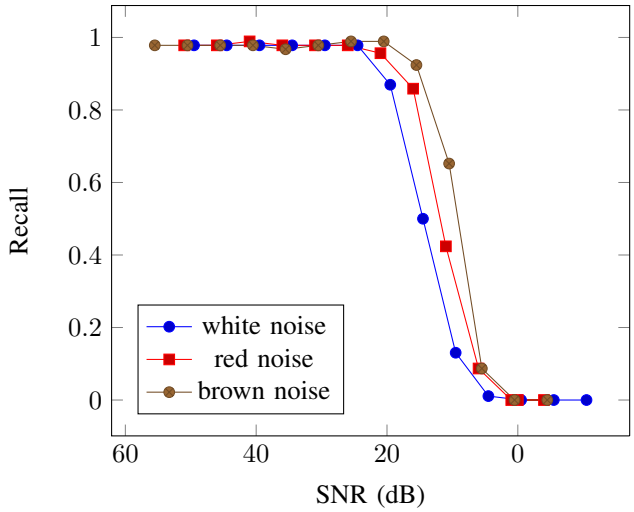

Fig. 2. Recall drop with increasing artificial noise level.

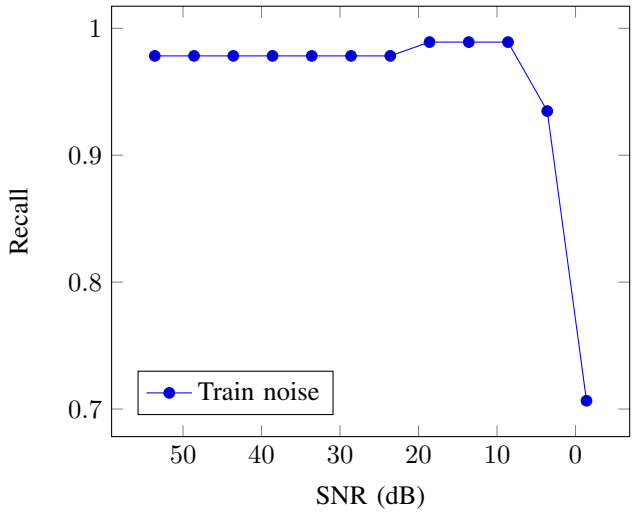

Fig. 3. Recall drop with increasing passenger train noise level.

- two scenes from *RBB Aktuell* (2012/08/14), one on the delay of its opening, the other on a terrorist warning,
- one scene from *Kontraste* (2012/08/12), on a general progress of the airport.

Focussing on the four scenes about the opening delay, we found several interesting differences and commonalities: the host is different for all shows, but the opening material for the report is at times similar. All four reports feature an interview with the politician Anton Hofreiter, but only three of them have the same content; the *RBB um sechs* show takes different quotes from the same material. The show *Brandenburg Aktuell* is the longest, and has a live interview with the politician Jörg Vogelsänger, whereas some shots of Vogelsänger's interview are recycled in *RBB Aktuell*. Both *Abendschau* and *RBB Aktuell* have the same final conclusion drawn by reporter Boris Hermel. See Figure 4 for an overview.

We use the *RBB Aktuell* video as the seed video for the fingerprint algorithm. While the data used is far too small to draw conclusions about the significance, the algorithm was able to detect Hofreiter in the three shows which broadcast the identical utterances, and did not detect a duplicate where other parts of the interview were used. Also, both instances of Vogelsänger and Hermel were detected correctly. As a sanity check, we also ran the fingerprint detection on the other videos which do not contain any duplicates but are on the same topic, and the algorithm correctly reported no matches.

V. Conclusion

In this paper, we presented a novel audio fingerprinting algorithm. Our experiments indicate that it can robustly synchronize a second screen application with broadcast media content. Especially for a non-artificial noise scenario, the algorithm still performs reasonably at a SNR of 5 dB. We therefore conclude that one-time events such as, e.g., banging of doors or coughing of the viewer, do not hamper a second screen application.

While established methods with a similar application exist, our method does not have any intellectual property constraints and can be implemented easily, while still providing good results and using reasonable processing power and storage. Further, for the given data we have seen another promising use case, as the technique can also detect duplicates within an internet broadcast archive.

Currently, the timestamp results for the fingerprint detections might be misleading since an identified smaller chunk could also appear at the end of the larger interview. As future work, we want to further strengthen the ties between the videos, by using diarization and speaker detection so that, e.g., the full interview can be linked right from its beginning. On a larger scale, using automatic topic segmentation on the basis of, e.g., shot segmentation, crawled subtitles and/or automatic speech recognition, we can also link to the beginning of larger reports, a feature that might be especially interesting for news summaries at the end of the day where a user wants to hear the full story.

Acknowledgment This work is partially supported by the Integrated Project LinkedTV (www.linkedtv.eu), funded by the European Commission through the 7th Framework Programme (FP7-287911).

References

[1] G. Neumüller *et al.*, "Seven one top box," Seven One Media, Tech. Rep., Oct. 2011.
[2] Yahoo! and Nielsen, "Mobile shopping framework – the role of mobile devices in the shopping process," Yahoo! and The Nielsen Company, Tech. Rep., Nov. 2010.
[3] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification," in *International Workshop on Content-Based Multimedia Indexing*, 2001, p. 117124.
[4] A. Wang, "An industrial strength audio search algorithm," in *International Conference on Music Information Retrieval (ISMIR)*, 2003.
[5] P. Cano, E. Batlle, T. Kalker, and J. Haitsma, "A review of audio fingerprinting," *The Journal of VLSI Signal Processing-Systems for Signal, Image, and Video Technology*, vol. 41, no. 3, pp. 271–284, Nov. 2005.
[6] V. Chandrasekhar, M. Sharifi, and D. A. Ross, "Survey and evaluation of audio fingerprinting schemes for mobile query-by-example applications," in *ISMIR*, 2011, pp. 801–806.

SHOW	RBB um sechs	Abendschau	Brandenburg Aktuell	RBB Aktuell
TIME	6 pm	7.30 pm	7.30 pm	9.45 pm
LENGTH	1:51 m	1:54 m	3:33 m	2:18 m

Fig. 4. Content of four news shows containing material about Berlin's new airport.

ISBN - 978-90-5986-410-8