Leveraging Audio Fingerprinting for Audio Content Synchronization and Replacement

Paolo Casagranda
Rai and University of Torino
Torino, Italy
paolo.casagranda@rai.it

Maria Luisa Sapino
University of Torino
Torino, Italy
mlsapino@di.unito.it

K. Selcuk Candan
Arizona State University
Tempe, USA
candan@asu.edu

MediaSync2015 Workshop, June 3, 2015, Brussels (Belgium)
ISBN/EAN: 978-90-5968-463-4

Abstract
In this paper we describe an innovative synchronization technique based on audio fingerprinting to help create a time base common to source content and receivers. We also introduce, as a relevant use case, the concept of hybrid content radio, a new radio service allowing the enrichment of broadcast radio with personalized audio content. Hybrid content radio is a novel radio service relying on audio replacement, and needing a precise synchronization between the main audio content and the enriching audio. We then discuss a variation of an existing audio fingerprinting algorithm, used to assess the proposed technique using real audio content.

Author Keywords
Hybrid radio; audio fingerprinting; media synchronization

ACM Classification Keywords
H.5.1. [Multimedia Information Systems]: Audio input/output; H.3.1. [Content Analysis and Indexing]: Indexing methods

Introduction
Traditional radio delivers professional and high-quality content through broadcast-based mechanisms: radio programs often target large audiences (this is especially true for national and regional radios) and broadcast modality ensures that the content delivery is cheap for both listeners and con-

The need for a hybrid solution. Traditional broadcast networks reach mass audiences with optimal efficiency in terms of cost and immediacy, but fall short of providing personalized contents that listeners now expect. While, Internet audio appears to excel in terms of personalization opportunities as well as variety of content, using cellular networks to deliver live radio services greatly increases delivery costs in mobile contexts, [1], [8]. Hybrid radio allows to take the best from both worlds.

tent providers. More recently, however, traditional radio is facing harsh competition from Internet audio services (Pandora, Spotify, Deezer, Rdio, Last.fm and others). Unlike broadcast radio, these services provide *personalized content* and listeners can choose from a wider variety of materials.

Hybrid content radio (HCR) aims to fill the gap between broadcast efficiency and content personalization and flexibility allowed by the Internet. Hybrid content radio is a service concept proposed by Rai and other European broadcasters in the Digital Radio Platforms and Services Group of the European Broadcasting Union (EBU) in February 2014. Intuitively, a hybrid content radio service aims to enrich broadcast radio with personalized and context aware fragments of audio mainly from the broadcaster's archives (Figure 1), using for example Internet in a standard and open way, without losing the benefits of the broadcast network. The final purpose of such enhancement is to *improve the service users' listening experience, decreasing their propensity to channel surf and giving them more targeted content*, such as local news and entertainment taken from the broadcaster's archives, music and also relevant advertisements.

In this paper, after introducing the hybrid content radio framework, we detail a novel solution to solve the technical problem of precisely synchronizing recommended "enriching" content with the broadcast content to create hybrid content. The proposed fingerprint-based synchronization scheme works for any kind of live audio service, independently from the encoding and the transport technology: it can be broadcast digital (e.g. DAB+, DRM), analogue (e.g. FM), or Internet streaming. Second, this technique leaves the audio content of the service untouched, without adding watermarks of any kind and leaving the transmission chain

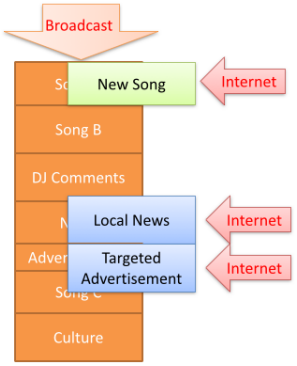

Figure 1: Hybrid content radio concept: enrichment of broadcast audio content

unchanged, a valuable advantage for a broadcaster.

Related Work

Hybrid Media

The concept of hybrid content radio is directly related to research on personalized audio/video services on one side, and to RadioDNS hybrid radio on the other. Audio and video personalization and media synchronization have been widely studied. As for television content personalization, the European iMedia Project [3] addressed advertisements personalization, using both broadcast TV and the Internet, enabling a kind of hybrid content TV. The concept of *virtual channel* was explored in [6], as an abstraction of the traditional TV channel. RadioDNS [13] has created a set of standards allowing to effectively link broadcast audio and additional content (text, images, service and program information, on-demand content) for a listening experience enriched with visuals and information. Hundreds of Euro-

Hybrid Content Radio

Who, Where and When. Hybrid content radio (HCR) is a service concept proposed by Rai and other European broadcasters in the Digital Radio Platforms and Services Group of the European Broadcasting Union in Geneva (February 2014), aiming to unify similar experimental services.

What. Hybrid content radio aims to enrich broadcast radio with personalized and context aware fragments of audio mainly from the broadcaster's archives (Figure 1), using for example Internet in a standard and open way, without losing the benefits of the broadcast network.

pean radios now support at least some of the RadioDNS applications. RadioDNS doesn't directly support hybrid content radio, however the Service and Programme Information (SPI) detailed in [11] enables it.

Content Fingerprinting
Content based search in audio and especially music has been used for several applications. With the ubiquitous presence of smartphones, content based retrieval has become very popular. [5] discusses audio retrieval algorithms specifically for mobile usage. Clausen et al. [7] proposed a robust algorithm based on spectral features to make searches on audio content. A review of audio fingerprinting algorithms can be found in [4].

Audio fingerprints to effectively retrieve music from a database of indexed songs have been used by the popular application Shazam. The adopted technique, detailed in [14], allows music retrieval in presence of noise. Another algorithm with similar applications has been proposed by Philips Research [12]. In [2], a new fingerprinting algorithm was presented, specially targeted to second screen TV synchronization and duplicate detection. In the present work, the target is not to identify duplicate audio (which requires techniques robust to a large variety of transformations), but only sense the presence of synchronization points in live audio.

Audio Fingerprinting Based Synchronization for Content Replacement

Broadcasters make live audio content available across different transports. In Europe, FM Radio, DAB Radio, and live streaming over the Internet are three widely used transport technologies, and broadcasters often take advantage of all of them. The hybrid-content radio application aims to enrich radio content provided by these live audio services by enabling the replacement of fragments of contents in the linear broadcast audio, constrained by a tight schedule to be respected for seamless integration.

Synchronization Challenge
Synchronization is an especially significant challenge in this context: *hybrid content radio requires a receiver time base precisely synchronized with the live content.* If a morning talk show is announced on the SPI as starting at 8:00:00AM, and the hybrid radio service has to replace it with a news summary chosen by the recommender system, a precise link between the original content and the received content has to be created, in such a way that the news summary begins to play at the correct time instant. On the other hand, *time bases at the receiver and at the content provider are not synchronized*. For example, Digital Radio based on DAB/DAB+ [9] protocol provides date and time information in a location independent format using Coordinated Universal Time (UTC), in the Service Information, but it is meant to be a date and time reference for radios and has no reference to the precise time when the content is played on the receiver. Audio content reaching the receiver can have a delay of several seconds compared to the time when it was broadcast, depending on the audio encoding buffer and the DAB frame and other parameters [9]. Moreover, the delay is not only due to encoding and transmission operations, but it depends also on the receiver, which has to buffer DAB frames and decode them. This is especially true if the receiver is "service-following" enabled as several seconds of buffering are required to allow a smooth audio join and to avoid discontinuities. Service-following between Digital and Analogue Radio [10] or Internet Radio [11] allows a receiver to maintain the same audio or data content in spite of varying reception conditions, changing between services with the same content on different transports.

Consequently, synchronization between metadata about

Why fingerprinting? The choice of relying on audio fingerprints is due to two main reasons. First, the proposed synchronization should work for any kind of live audio service, independently from the encoding and the transport technology: it can be broadcast digital (e.g. DAB+, DRM), analogue (e.g. FM), or Internet streaming. Second, it should leave the audio content of the service untouched, without adding watermarks of any kind.

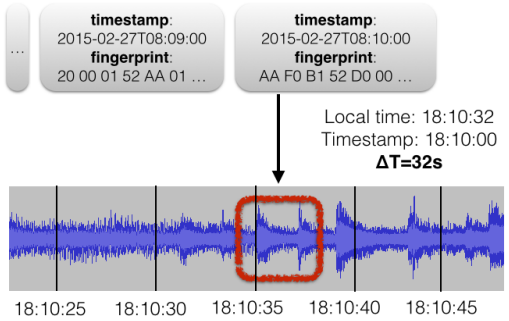

Figure 2: The Receiver reads an updated (fingerprint, timestamp) couple for the service it's listening to, and matches it on the content.

the programs (like the Service and Programme Information [11]) and the live content cannot be precise, and there can be several seconds of delay between the announced program start and the actual audio rendering on the receiver. In this paper we propose a synchronization solution, based on audio fingerprints associated with timestamps and live programme metadata.

Synchronization Mechanism
The synchronization mechanism is as follows:

- The broadcaster, at fixed and known intervals, creates short fingerprints of the audio content it is transmitting. As we discuss in Section Evaluation, in our experiments we evaluated the impact of 5 seconds fingerprints generated every 60 seconds. We refer to Section Audio Fingerprinting Based Synchronization for Content Replacement for a more detailed discussion about the fingerprint creation.
- The set of N fingerprints and its timestamps referring to the chosen time slice T_{slice} (T_{slice} has been set to $5s$ in the following experiments) are then made available to the receiver. Let:
$H_j = \{(h_{bj}, t_j), (h_{bj+1}, t_{j+1}), ..., (h_{bj+N-1}, t_{j+N-1})\}$
denote the $j-th$ fingerprint set made available by the broadcaster, including N fingerprints h_{bj}, and the corresponding timestamps t_j. Alternative ways to make the fingerprint and the timestamp available include publishing them on a public web site, or sending them with the metadata.
- For a time interval of T_{check}, the receiver reads the set H_m of couples (h_{bm}, t_m) associated to the audio service it is currently tuned in and, in parallel, it calculates the set H_n of fingerprints h_{rn} of the audio content of the current audio service at time t_{rn}, looking for a match between h_{bm} and h_{rn}. When it finds a minimum Hamming distance between H_m and the calculated H_n for the time interval T_{check}, the receiver calculates the difference $\Delta T = (t_{rn} - t_{bm})$, where t_{bm} is the timestamp associated with the broadcaster fingerprint h_{bm} and t_{rn} the local receiver time.
- The program start is generally specified in the content metadata, i.e. the Service and Programme Information (SPI) or similar, and the synchronization process is complete (see Figure 2). If there is a delay between the content metadata timestamp and the actual content timestamp, it has to be taken into account adding to ΔT also this delay.

Data. The dataset used in the experiments included audio content from Rai, the Italian public broadcaster. The content included the mixed talks and music from Rai and Italian private radios audio services. The different content allowed to assess the algorithm robustness for different typical live services.

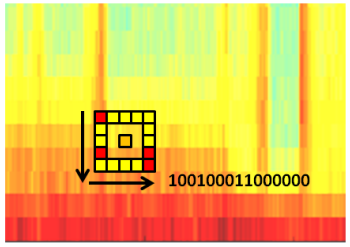

Figure 4: Fingerprints calculation example.

```
<?xml version="1.0" encoding="UTF-8" ?>
<hcr:synchro xmlns:hcr="http://www.rai.it/hcr">
    <hcr:programme>
        rai_radio1
    </hcr:programme>
    <hcr:fingerprint type="base" freqbins="20">
        U3luY2hyb25pemF0aW9uIFRlY2huaXF1ZSBmb3IgSHli...
    </hcr:fingerprint>
    <hcr:utctimestamp>
        2015-02-27T08:10:00
    </hcr:utctimestamp>
</hcr:synchro>
```

Figure 3: An example of the shared information (fingerprint, timestamp) referring to a live service.

XML, JSON or other formats can be used to specify the (h_{bm}, t_{bm}) information. An example of XML fragment achieving the task is in Figure 3. This file has to be periodically updated, so that when a receiver is turned on, it can synchronize in a reasonable amount of time. During the tests an update time of 60 seconds and a time interval T_{check} of 120 seconds have been chosen.

Note that he above process relies on audio fingerprints created by the broadcaster, to be used by the receiver while reconstructing a common time base. In the next section, we describe a specific audio fingerprint extraction algorithm, which we also later use in our evaluations.

Audio Fingerprint Extraction and Use

Audio fingerprints are used to index the broadcast content and to achieve the synchronization. In this paper, we propose a variant of the fingerprinting algorithm proposed in [2] for this purpose, however other algorithms can be used as well.

Fingerprint Extraction

As in [7] and [14], we use spectral features to build the fingerprint. First, the audio signal spectrogram is calculated. For the purpose, we use a Fast Fourier Transform (FFT) with 1024 time samples and an overlap of 512 samples. We reduce the number of frequency binstaking only 40 frequencies, averaging the logarithmic amplitude in each bin.

Then, for each frequency bin i with center frequency f_i and time bin and j with center time t_j, a 16-bit fingerprint is calculated [2]: we compare the signal energy in the cell (i, j) with each of the 16 neighboring cells, placing a “1” only if the signal in the cell is greater than the corresponding neighboring cell, as shown in figure 4. The 16 bits fingerprint, h_{ij}, is then stored with the associated frequency and time. In the next discussion we’ll mainly refer to the couple (h_{ij}, t_j). Note that this calculation can be done by the service provider at fixed time intervals, for fixed time slices (5 seconds in the following tests).

Each fingerprint block is then shared with the receivers, along with its associated timestamp.

Fingerprint Use

While tuned on the audio service, the receiver tries to synchronize with the service provider’s time base. To reach this goal it computes the spectrogram of the audio content with the same parameters described before, taking only one random fingerprint for each frequency bin. Then, for a given fixed time interval, it performs a comparison with the updated fingerprints (h_{bj}, t_j) created by the broadcaster and made available to the receiver, for example on the broacaster’s web site. The fixed time interval during which the receiver has to calculate the fingerprints depends on the actual delay between the audio service time base and the local receiver time base. Typically, for the observed audio services, this time interval is less than 60 seconds. The

Noise classification

White Noise: it has equal power spectral density in the whole spectrum.
Pink Noise: its power density is inversely proportional to the frequency, i.e. it follows a $1/f$ law.
Brownian Noise: its power density follows a $1/f^2$ law, affecting lower frequencies more than higher frequencies.

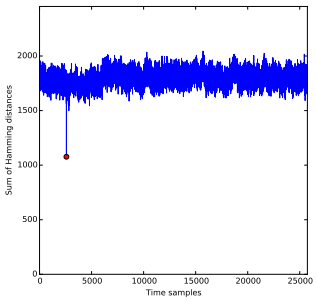

Figure 5: Hamming distances for each time step (SNR is about 1dB, with music content).

minimum Hamming distance between the two fingerprint sets will identify the match (see Figure 5) which will allow to compare the local receiver time with the source content time, and to calculate the delay ΔT.

Evaluation

In this section, we present experimental evaluations aimed at demonstrating the feasibility of the synchronization technique, and to evaluate its performance.

Setup

In the experimental setup the fingerprint lookup was done on an audio sequence of 10 minutes to better assess the algorithm. The audio sample rate was the commonly used 44100Hz. The audio slices used to calculate the fingerprint to be shared were of 5s: this is quite a short time compared to values we can find in the literature ([7], [14], [2], [5]).

The spectrogram used for the fingerprint calculation was generated using a windowed FFT with 2048 samples and overlap of 1024. A number of frequency bins from 20 to 60 were assessed, and the best trade off between computing time and precision was reached at 40. Also, a time resolution of at least $100ms$ was the target, and the chosen FFT window is compatible with the requirement, allowing a time resolution of around $20ms$. It is remarkable that the bandwidth considered for the results is considerably narrow: between $0Hz$ to $1.6kHz$. The chosen bandwidth proved to be sufficient to achieve a reliable match between the fingerprints delivered by the broadcaster and those calculated by the receiver. For matching, only one fingerprint per time bin was taken, at random.

Accuracy Results

The fingerprint algorithm was first evaluated for the original audio sequences; i.e., where no distortion or noise was present. Under these conditions, the accuracy of matching was good: for every sequence the recall was higher than 0.9. In case of a music sequences, the matching was perfect, and every sequence was recognized. It is important to note that, unlike other application domains (such as duplicate detection), in the synchronization context, the audio fingerprint's *robustness* towards noise is not critical to achieve the synchronization, as the audio is directly received from the broadcast media instead of being received using a microphone, as in other second screen applications where robustness to noise is critical [5, 14]. Still, there may be cases in which robustness may be critical, such as when the fingerprints generated by the broadcaster are used not only for time base synchronization, but also for second screen synchronization.

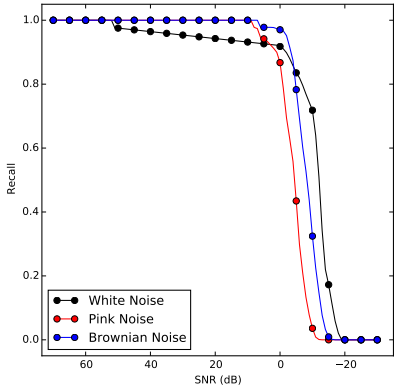

Figure 6: Recall with several levels of white noise (all data)

In Figure 6, we see recall plotted versus the SNR for white, pink, or brownian noises. The algorithm shows similar behavior for three noise types. Recall remains greater than

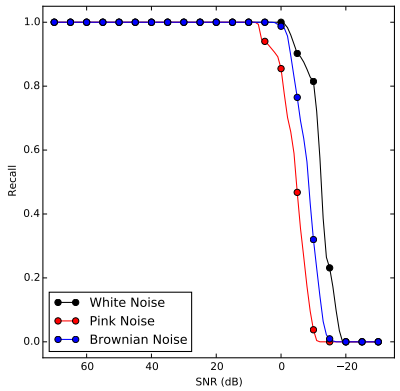

Figure 7: Recall with several levels of white noise (music data).

0.9 for SNR of around $0dB$ and drops by 50% for SNR of about $-10dB$ for all data, and $-12dB$ for music content. This is especially good, given the short time slices used in the tests. Contrast these results with the results reported in [14] where fingerprints based on $5s$ time slices resulted in the recall dropping by 50% at SNR of about $-3dB$.

To better understand the results, we also compared the results with the actual audio content of the files: this showed that the recall was lower in presence of audio streams including talks with pauses. This is expected due to the short time slices used for the match: if in the selected $5s$ fingerprint, there was little audio information, the recall is expected to be low. This conclusion is also supported by examining the recall for music channels: as Figure 7 shows, recall starts dropping later (with the drop in SNR) for music channels than for all data.

Synchronization Accuracy

The last step of the evaluation involves the assessment on the synchronization accuracy. The accuracy of the estimation of the time delay ΔT between the source content time and the receiver local time depends on the accuracy obtained by the fingerprinting match. We used a time window of 2048 samples, so the average error in the estimation will be of 1024 samples. With a sampling frequency of $44.1kHz$, the accuracy is $23ms$ about. Moreover, accuracy can be modified tuning the time window length. The experiments confirm this predicted ΔT accuracy value. The proposed synchronization approach is therefore a viable solution for broadcasters, and especially for HCR services.

Conclusion and Future Work

In this paper, we introduced the hybrid content radio concept and presented a novel application of audio fingerprinting for synchronization to live audio. The primary application is content replacement in the hybrid content radio framework. The technique allows to precisely rebuild the time base of the content at the receiver side, independently from the transport and encoding of the audio content. Sharing an association of metadata about the audio content, fingerprints and time stamps, the receiver is able to synchronize and insert the enriching content in the correct time slice. While the technique is independent from the fingerprinting algorithm, we presented experimental results using a variant of a fingerprinting algorithm that is easy to implement and reliable also in presence of noise. In the proposed configuration, we achieved a time resolution of $23ms$ about, thus validating the feasibility of the synchronization technique. The match between the fingerprint and the audio content was done in about $1/20$ of real-time, generating lightweight fingerprints of less than $1kB$. We envisage further work in both standardization and optimization of the fingerprinting technique for the live broadcast case.

A standard way to specify the association between content, metadata and synchronization information can be useful to enable more horizontal audio enriching services. Also, we note that the size of the fingerprint is important while sharing it, both for broadcast and internet transports. So further research is needed on the selection of the fingerprinting algorithm and on the tuning of the parameters with the previously described features in mind.

Acknowledgments

We thank Giovanni Vitale and Vittoria Mignone for the valuable suggestions during the writing of the present paper.

References

[1] Teracom AS. 2012. Can the cellular networks cope with linear radio broadcasting? (November 2012). Retrieved April 21, 2015-04-21 from https://www.worlddab.org/public_document/file/470/Teracom_WhitePaper_ENG_Broadcast_vs_cellular_networks_Feb_2014.pdf.

[2] Rolf Bardeli, Jochen Schwenninger, and Daniel Stein. 2012. Audio Fingerprinting for Media Synchronisation and Duplicate Detection. In *Proceedings of the Media Synchronisation Workshop (MediaSync)*.

[3] Theodoros Bozios, Georgios Lekakos, Victoria Skoularidou, and Kostas Chorianopoulos. 2001. Advanced techniques for personalized advertising in a digital TV environment: the iMEDIA system. In *Proceedings of the eBusiness and eWork Conference*. 1025–1031.

[4] Pedro Cano, Eloi Batlle, Ton Kalker, and Jaap Haitsma. 2005. A Review of Audio Fingerprinting. *Journal of VLSI signal processing systems for signal, image and video technology* 41, 3 (2005), 271–284.

[5] Vijay Chandrasekhar, Matt Sharifi, and David A. Ross. 2011. Survey and Evaluation of Audio Fingerprinting Schemes for Mobile Query-by-Example Applications. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, Vol. 20. 801–806.

[6] Konstantinos Chorianopoulos, George Lekakos, and Diomidis Spinellis. 2003. The virtual channel model for personalized television. *Proceedings of the 1st EuroITV conferece: From Viewers to Actors* (2003), 59–67.

[7] M. Clausen and F. Kurth. 2004. A unified approach to content-based and fault-tolerant music recognition. *IEEE Transactions on Multimedia* 6, 5 (Oct 2004), 717–731.

[8] Coutts Communications. 2014. Analysis of use of Mobile Telecommunications Networks to Deliver Broadcast Radio in Australia. (November 2014). Retrieved April 21, 2015 from https://www.worlddab.org/public_document/file/517/Coutts_Report_Australia_November_2014.pdf.

[9] ETSI. 2006. EN 300 401, Radio Broadcasting Systems; Digital Audio Broadcasting (DAB) to mobile, portable and fixed receivers. (2006).

[10] ETSI. 2013. TS 103 176, Digital Audio Broadcasting (DAB); Rules of implementation; Service information features. (2013).

[11] ETSI. 2015. TS 102 818, Hybrid Digital Radio (DAB, DRM, RadioDNS); XML Specification for Service and Programme Information (SPI). (2015).

[12] Jaap Haitsma and Ton Kalker. 2002. A highly robust audio fingerprinting system.. In *Proceedings of the International Symposium on Music Information Retrieval*, Vol. 2002. 107–115.

[13] RadioDNS. 2015. The RadioDNS Project. (2015). Retrieved April 21, 2015 from http://www.radiodns.org.

[14] Avery Wang. 2006. The Shazam Music Recognition Service. *Commun. ACM* 49, 8 (Aug. 2006), 44–48.