Automatic Identification of Samples in Hip Hop Music

Jan Van Balen[1], Martín Haro[2], and Joan Serrà[3] *

[1] Dept of Information and Computing Sciences, Utrecht University, the Netherlands
[2] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Barcelona, Spain
j.m.h.vanbalen@uu.nl, martin.haro@upf.edu, jserra@iiia.csic.es

Abstract. Digital sampling can be defined as the use of a fragment of another artist's recording in a new work, and is common practice in popular music production since the 1980's. Knowledge on the origins of samples holds valuable musicological information, which could in turn be used to organise music collections. Yet the automatic recognition of samples has not been addressed in the music retrieval community. In this paper, we introduce the problem, situate it in the field of content-based music retrieval and present a first strategy. Evaluation confirms that our modified optimised fingerprinting approach is indeed a viable strategy.

Keywords: Digital Sampling, Sample Detection, Sample Identification, Sample Recognition, Content-based Music Retrieval

1 Introduction

Digital sampling, as a creative tool in composition and music production, can be defined as the use of a fragment of another artist's recording in a new work. The practice of digital sampling has been ongoing for well over two decades, and has become widespread amongst mainstream artists and genres, including hip hop, electronic, dance, pop, and rock [11]. Information on the origin of samples holds valuable insights in the inspirations and musical resources of an artist. Furthermore, such information could be used to enrich music collections, e.g. for music recommendation purposes. However, in the context of music processing and retrieval, the topic of automatic sample recognition seems to be largely unaddressed [5, 12].

The Oxford Music Dictionary defines sampling as "the process in which a sound is taken directly from a recorded medium and transposed onto a new recording" [8]. As a tool for composition, it first appeared when *musique concrète* artists of the 1950's started assembling tapes of previously released music recordings and radio broadcasts in musical collages. The phenomenon reappeared when

* This research was done between 1/2011 and 9/2011 at the Music Technology Group at Universitat Pompeu Fabra in Barcelona, Spain. The authors would like to thank Perfecto Herrera and Xavier Serra for their advice and support. JS acknowledges JAEDOC069/2010 from Consejo Superior de Investigaciones Científicas and 2009-SGR-1434 from Generalitat de Catalunya. MH acknowledges FP7-ICT-2011.1.5-287711.

9th International Symposium on Computer Music Modelling and Retrieval (CMMR 2012)
19-22 June 2012, Queen Mary University of London

DJ's in New York started using their vinyl players to repeat and mix parts of popular recordings to provide a continuous stream of music for the dancing crowd. The breakthrough of sampling followed the invention of the digital sampler around 1980, when producers started using it to isolate, manipulate, and combine portions of others' recordings to obtain entirely new sonic creations [6, 13]. The possibilities that the sampler brought to the studio have played a role in the appearance of several new genres in electronic music, including hip hop, house music in the late 90's (from which a large part of electronic dance music originates), jungle (a precursor of drum&bass music), dub, and trip hop.

1.1 Motivations for Research on Sampling

A first motivation to undertake the automatic recognition of samples originates in the belief that the musicological study of popular music would be incomplete without the study of samples and their origins. Sample recognition provides a direct insight into the inspirations and musical resources of an artist, and reveals some details about his or her composition methods and choices made in the production. Moreover, alongside recent advances in folk song [16] and version identification [14] research, it can be applied to trace musical ideas and observe musical re-use in the recorded history of the last two decades.

Samples also hold valuable information on the level of genres and communities, revealing cultural influences and dependence. Researchers have studied the way hip hop has often sampled 60's and 70's African-American artists [6] and, more recently, Bryan and Wang [2] analysed musical influence networks in sample-based music, inferred from a unique dataset provided by the WhoSampled web project. Such annotated collections exist indeed, but they are assembled through hours of manual introduction by amateur enthousiasts. It is clear that an automated approach could both widen and deepen the body of information on sample networks.

As the amount of accessible multimedia and the size of personal collections continue to grow, sample recognition from raw audio also provides a new way to bring structure in the organization of large music databases, complementing a great amount of existing research in this direction [5, 12]. Finally, sample recognition could serve legal purposes. Copyright considerations have always been an important motivation to understand sampling as a cultural phenomenon; a large part of the academic research on sampling is focused on copyright and law [11].

1.2 Requirements for a Sample Recognition System

Typically observed parameters controlling playback in samplers include filtering parameters, playback speed, and level envelope controls ('ADSR'). Filtering can be used by producers to maintain only the most interesting part of a sample. Playback speed may be changed to optimise the tempo (time-stretching), pitch (transposition), and/or mood of samples. Naturally, each of these operations complicates their automatic recognition. In addition, samples may be as short as one second or less, and do not necessarily contain tonal information. Moreover,

given that it is not unusual for two or more layers to appear at the same time in a mix, the energy of the added layers can be greater than that of the sample. This further complicates recognition. Overall, three important requirements for any sample recognition system should be: (1) The system is able to identify heavily manipulated query audio in a given music collection. This includes samples that are filtered, time-stretched, transposed, very short, tonal and non-tonal (i.e. purely percussive), processed with audio effects, and/or appear underneath a thick layer of other musical elements. (2) The system is able to perform this task for large collections. Finally, (3) the system is able to perform the task in a reasonable amount of time.

1.3 Scientific Background: Content-based Music Retrieval

Research in content-based music retrieval can be characterised according to *specificity* [5] and *granularity* [9]. Specificity refers to the degree of similarity between query and match. Tasks with a high specificity mean to retrieve almost identical documents, low specificity tasks look for vague matches that are similar with respect to some musical properties. Granularity refers to the difference between fragment-level and document-level retrieval. The problem of automatic sample recognition has a mid specificity and very low granularity (i.e. very short-time matches that are similar with respect to some musical properties). Given these characteristics, it relates to audio fingerprinting.

Audio fingerprinting systems attempt to identify unlabeled audio by matching a compact, content-based representation of it, the fingerprint, against a database of labeled fingerprints [3]. Just like fingerprinting systems, sample recognition systems should be designed to be robust to additive noise and several transformations. However, the deliberate transformations possible in sample-based music production, especially changes in pitch and tempo, suggest that the problem of sample recognition is in fact a less specific task.

Audio matching and version identification systems are typical mid specificity problems. Version identification systems assess if two musical recordings are different renditions of the same musical piece, usually taking changes in key, tempo and structure into account [14]. Audio matching works on a more granular level and includes remix recognition, amongst other tasks [4, 9]. Many of these systems use chroma features [5, 12]. These descriptions of the pitch content of audio are generally not invariant to the addition of other musical layers, and require the audio to be tonal. This is often not the case with samples. We therefore believe sample recognition should be cast as a new problem with unique requirements, for which the existing tools are not entirely suitable.

2 Experiments

2.1 Evaluation Methodology

We now present a first approach to the automatic identification of samples [15]. Given a query song in raw audio format, the experiments aim to retrieve a ranked list of candidate files with the sampled songs first.

To narrow down the experiments, only samples used in hip hop music were considered, as hip hop is the first and most famous genre to be built on samples [6] (though regarding sample origins, there were no genre restrictions). An evaluation music collection was established, consisting of 76 query tracks and 68 candidate tracks [15]. The set includes 104 sample relations (expert confirmed cases of sampling). Additionally, 320 'noise' files similar to the candidates in genre and length were added to challenge the system. Aiming at representativeness, the ground truth was chosen to include both short and long samples, tonal and percussive samples, and isolated samples (the only layer in the mix) as well as background samples. So-called 'interpolations', i.e. samples that have been re-recorded in the studio, were not included, nor were non-musical samples (e.g. film dialogue). This ground truth was composed using valuable information from specialized internet sites, especially WhoSampled[4] and Hip Hop is Read[5]. As the experiment's evaluation metric, the mean average precision (MAP) was chosen [10]. A random baseline of 0.017 was found over 100 iterations, with a standard deviation of 0.007.

2.2 Optimisation of a State-of-the-Art Audio Fingerprinting System

In a first experiment, a state-of-the-art fingerprinting system was chosen and optimised to perform our task. We chose to work with the spectral peak-based audio fingerprinting system designed by Wang [17]. A fingerprinting system was chosen because of the chroma argument in Section 1.3. The landmark-based system was chosen because of its robustness to noise and distortions and the alleged 'transparency' of the spectral peak-based representation (Table 1): Wang reports that, even with a large database, the system is able to correctly identify each of several tracks mixed together.

Table 1. Strengths and weaknesses of spectral peak-based fingerprints in the context of sample identification.

Strengths	Weaknesses
– High proven robustness to noise and distortions. – Ability to identify music from only a very short audio segment. – 'Transparent' fingerprints: ability to identify multiple fragments played at once. – Does not explicitly require tonal content.	– Not designed for transposed or time-stretched audio. – Designed to identify tonal content in a noisy context, fingerprinting drum samples requires the opposite. – Can percussive recordings be represented by just spectral peaks at all?

[4] http://www.whosampled.com/
[5] http://www.hiphopisread.com/

As in most other fingerprinting systems, the landmark-based system consists of an extraction and a matching component. Briefly summarized, the extraction component takes the short time Fourier transform (STFT) of audio segments and selects from the obtained spectrogram a uniform constellation of prominent spectral peaks. The time-frequency tuples with peak locations are paired in 4-dimensional 'landmarks', which are then indexed as a start time stored under a certain hash code for efficient lookup by the matching component. The matching component retrieves for all candidate files the landmarks that are identical to those extracted from the query. Query and candidate audio segments match if corresponding landmarks show consistent start times [17].

A Matlab implementation of this algorithm has been made available by Ellis[6]. It works by the same principles as [17], and features a range of parameters to control the implementation-level operation of the system. Important STFT parameters are the audio sample rate and the FFT size. The number of selected spectral peaks is governed by the desired density of peaks in the time domain and the peak spacing in the frequency domain. The number of resulting landmarks is governed by three parameters: the pairing horizons in the frequency and time domain, and the maximum number of formed pairs per spectral peak.

A wrapper function was written to slice the query audio into short fixed length chunks, overlapping with a hop size of one second, before feeding it to the fingerprinting system. A distance function is also required for evaluation using the MAP. Two distance functions are used, an absolute distance $d_a = \frac{1}{m+1}$, function of the number of matching landmarks m, and a normalized distance $d_n = \frac{l-m}{l}$, weighted by the number of extracted landmarks l.

Because of constraints in time and computational power, optimising the entire system in an extensive grid search would not be feasible. Rather, we have performed a large number of tests to optimise the most influential parameters. Table 2 summarizes the optimisation process, more details can be found in [15]. The resulting MAPs were 0.228 and 0.218, depending on the distance functions used (note that both are well beyond the random baseline mentioned before). Interestingly, better performance was achieved for lower sample rates. The optimal density of peaks and number of pairs per peak are also significantly larger than the default values, corresponding to many more extracted landmarks per second. This requires more computation time for both extraction and matching, and a requires for a higher number of extracted landmarks to be stored in the system's memory.

2.3 Constant Q Fingerprints

The MAP of around 0.22 is low for a retrieval task but promising as a first result. The system retrieves a correct best match for around 15 of the 76 queries. These matches include both percussive and tonal samples. However, due to the lowering of the sample rate, some resolution is lost. Not only does this discard valuable data, the total amount of information in the landmarks also goes down

[6] `http://labrosa.ee.columbia.edu/matlab/fingerprint/`

Table 2. Some of the intermediate results in the optimisation of the audio fingerprinting system by Wang as implemented by Ellis [15]. The first row shows default settings with its resulting performance.

pairs/pk	pk density (s^{-1})	pk spacing (bins)	sample rate (Hz)	FFT size (ms)	MAP_n (d_n)	MAP_a (d_a)
3	10	30	8,000	64	0.114	0.116
10	10	30	8,000	64	0.117	0.110
10	36	30	8,000	64	0.118	0.133
10	36	30	2,000	64	0.176	0.162
10	36	30	2,000	128	0.228	0.218

as the range of possible frequency values decreases. We now did a number of tests using a constant Q transform (CQT) [1] instead of a Fourier transform. We would like to consider all frequencies up to the default 8,000 Hz but make the lower frequencies more important, as they contributed more to the best performance so far. The constant Q representation, in which frequency bins are logarithmically spaced, allows us to do so. The CQT also suits the logarithmic representation of frequency in the human auditory system.

We used another Matlab script by Ellis[7] that implements a fast algorithm to compute the CQT and integrated it in the fingerprinting system. A brief optimisation of the new parameters returns an optimal MAP of 0.21 at a sample rate of 8,000 Hz. This is not an improvement in terms of the MAP, but loss of information in the landmark is now avoided (the amount of possible frequency values is restored), amending the system's scalability.

2.4 Repitching Fingerprints

In a last set of tests, a first attempt was made to deal with repitched samples. Artists often time-stretch and pitch-shift samples by changing their playback speed. As a result, the samples' pitch and tempo are changed by the same factor. Algorithms for independent pitch-shifting and time-stretching without audible artifacts have only been around for less than a decade, after phase coherence and transient processing problems were overcome. Even now, repitching is still popular practice amongst producers, as inspection of the ground truth music collection confirms. In parallel to our research [15], fingerprinting of pitch-shifted audio has been studied by Fenet et al. [7] in a comparable way, but the approach does not consider pitch shifts greater than 5%, and does not yet deal with any associated time-stretching.

The most straightforward method to deal with repitching is to repitch query audio several times and perform a search for each of the copies. Alternatively, the extracted landmarks themselves can also be repitched, through the appropriate scaling of time and frequency components (multiplying the time values

[7] See `http://www.ee.columbia.edu/~dpwe/resources/matlab/sgram/` and `http://labrosa.ee.columbia.edu/matlab/sgram/logfsgram.m`

Table 3. Results of experiments using repitching of both the query audio and its extracted landmarks to search for repitched samples.

N	ΔR (st)	r (st)	MAP_n	MAP_a
-	-	0	0.211	0.170
0	-	0.5	0.268	0.288
5	1.0	0.5	0.341	0.334
9	0.5	0.5	0.373	0.390

and dividing the frequency values, or vice versa). This way the extraction needs to be done only once. We have performed three tests in which both methods are combined: all query audio is resampled several times, to obtain N copies, all pitched ΔR semitones apart. For each copy of the query audio, landmarks are then extracted, duplicated and rescaled to include all possible landmarks repitched between $r = 0.5$ semitones up and down. This is feasible because of the finite resolution in time and frequency.

The results for repitching experiments are shown in Table 3. We have obtained a best performance of MAP_n equal to 0.390 for the experiment with $N = 9$ repitched queries, $\Delta R = 0.5$ semitones apart every query. This results in a total searched pitch range of 2.5 semitones up and down, or $\pm 15\%$. Noticeably, a MAP of 0.390 is low, yet it is in the range of some early version identification systems, or perhaps even better [14].

3 Discussion

To the best of our knowledge, this is the first research to address the problem of automatic sample identification. The problem has been defined and situated in the broader context of sampling as a musical phenomenon and the requirements that a sample identification system should meet have been listed. A state-of-the-art fingerprinting system has been adapted, optimised, and modified to address the task. Many challenges have to be dealt with and not all of them have been met, but the obtained performance of 0.39 is promising and unmistakably better than the precision obtained without taking repitching into account [15]. Overall, our approach is a substantial first step in the considered task.

Our system retrieved a correct best match for 29 of the 76 queries, amongst which 9 percussive samples and at least 8 repitched samples. A more detailed characterisation of the unrecognised samples is time-consuming but will make a very informative next step in future work. Furthermore, we suggest to perform tests with a more extensively annotated dataset, in order to assess what types of samples are most challenging to identify, and perhaps a larger number of ground truth relations. This will allow to relate performance and the established requirements more closely and lead to better results, paving the road for research such as reliable fingerprinting of percussive audio, sample recognition based on cognitive models, or the analysis of typical features of sampled audio.

References

1. Brown, J. C.: *Calculation of a Constant Q Spectral Transform*, The Journal of the Acoustical Society of America, vol. 89, no. 1, p. 425 (1991)
2. Bryan, N. J. and Wang, G.: *Musical Influence Network Analysis and Rank of Sample-Based Music*, in Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR), pp. 329-334 (2011)
3. Cano, P., Battle, E., Kalker, T. and Haitsma, J.: *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 (2005)
4. Casey, M. and Slaney, M.: *Fast Recognition of Remixed Music Audio*, in Acoustics Speech and Signal Processing 2007 ICASSP 2007 IEEE International Conference on, vol. 4, no. 12, pp. 300-1 (2007)
5. Casey, M., R. Veltkamp, Goto, M., Leman, M., Rhodes, C. and Slaney, M.: *Content-Based Music Information Retrieval: Current Directions and Future Challenges*, Proceedings of the IEEE, vol. 96, no. 4, pp. 668-696 (2008)
6. Demers, J.: *Sampling the 1970s in Hip-Hop*, Popular Music, vol. 22, no. 1, pp. 41-56 (2003)
7. Fenet, S., Richard, G., and Grenier Y.: *A Scalable Audio Fingerprint Method with Robustness to Pitch-Shifting*, in Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), Miami, USA (2011)
8. Fulford-Jones, W.: *Sampling*, Grove Music Online. Oxford Music Online. `http://www.oxfordmusiconline.com/subscriber/article/grove/music/47228` (2011)
9. Grosche, P., Müller, M. and Serrà, J. *Audio Content-Based Music Retrieval*, in Multimodal Music Processing, Dagstuhl Publishing, Schloss Dagstuhl-Leibniz Zentrum für Informatik, Germany. Under Review.
10. Manning, C. D., Prabhakar, R. and Schutze, H.: *An Introduction to Information Retrieval*. Cambridge University Press, Cambridge (2008).
11. McKenna, T.: *Where Digital Music Technology and Law Collide - Contemporary Issues of Digital Sampling, Appropriation and Copyright Law*, Journal of Information Law and Technology, vol. 1, pp. 0-1 (2000)
12. Müller, M., Ellis, D., Klapuri, A. and Richard, G.: *Signal Processing for Music Analysis*, Selected Topics in Signal Processing, IEEE Journal of, vol. 0, no. 99, pp. 1–1 (2011)
13. Self, H.: *Digital Sampling: A Cultural Perspective*, UCLA Ent. L. Rev., vol. 9, p. 347 (2001)
14. Serrà, J., Gómez, E. and Herrera P.: *Audio Cover Song Identification and Similarity: Background, Approaches, Evaluation and Beyond*, in Advances in Music Information Retrieval, Springer, pp. 307–332 (2010)
15. Van Balen, J.: *Automatic Recognition of Samples in Musical Audio*. Master's thesis, Universitat Pompeu Fabra, Spain, `http://mtg.upf.edu/node/2342` (2011)
16. Wiering, F., Veltkamp, R.C., Garbers, J., Volk, A., Kranenburg, P. & Grijp, L.P.: *Modelling Folksong Melodies* Interdisciplinary Science Reviews, vol. 34, no. 2-3, pp. 154-171 (2009)
17. Wang, A.: *An Industrial Strength Audio Search Algorithm*, in Proceedings of the International Conference on Music Information Retrieval (ISMIR) (2003)