A Robust Audio Fingerprinter Based on Pitch Class Histograms Applications for Ethnic Music Archives

JOREN SIX[*]
OLMO CORNELIS
Royal Academy of Fine Arts & Royal Conservatory, University College Ghent
joren.six@hogent.be
olmo.cornelis@hogent.be

Abstract

In this paper we present a new acoustic fingerprinting system, based on pitch class histograms. The aim of acoustic fingerprinting is to generate a small representation of an audio signal that can be used to identify identical, or recognize similar, audio snippets in a large audio set. A robust fingerprinting system generates similar fingerprints for perceptually similar audio signals. A piece of music with a noise added should generate an almost identical fingerprint as the original. The new system, presented here, has some interesting features which makes it a valuable tool to manage ethnic music archives: the fingerprints are rather robust against pitch shift, tempo changes, several synthetic audio effects, and reversal of the audio. When only part of the audio is used to generate a fingerprint, the system keeps working but retrieval performance degrades.

1. Introduction

In the process of digitizing a large music collection it is possible that the same music is present on different physical media, either as complete copies or as copies of individual tracks. Sometimes it is hard to keep track of which physical media are already digitized and which are still to process. An ability to search for music based on the content of the signal is a valuable tool to prevent duplicates entering the digital version of the music archive. Here we present a system with those capabilities. Another use case for such system is to (re)connect meta-data to an audio fragment without any information, but is present in the digital connection.

For large, historical collections of ethnic music the problems sketched above are almost inevitable. Often, individual collections of recordings or discs –without meta-data– are donated to museums. These collections usually are very diverse and several recordings may already be present in the archive, which is where the need for content based search comes to play. Due to the nature of the original physical media - wax cylinders, wire recordings, magnetic tapes, gramophone records - and the, often abysmal, recording quality a content based search system for ethnic music has to have special features for robustness. The goal of the system, as it is presented here, is to *identify identical audio excerpts even if they were processed or digitized in a different way*. Our research is focused on pitch class histograms which appear to be robust enough for the task of acoustic fingerprinting, even in the context of historical ethnic music collections.

This paper is structured as follows: we start with an overview of the system and then argue why it shows potential. Then details about the implementation are unveiled. The third section describes an experiment with the system. The paper ends with a conclusion.

2. System Overview

Figure 1 shows a general acoustical fingerprinting system. Features are extracted from audio and with these features a fingerprint is constructed. The fingerprint is a small representation of the audio. In the best case, perceptually similar audio generates related fingerprints, identical audio should generate identical fingerprints. With the generated fingerprint and a list of previously generated fingerprints an unknown piece of audio is identified. In an ideal system the fingerprints are small but unique for each piece of audio and searching through a large number of fingerprints is efficient. Alternative systems include the ones described by Haitsma & Kalker

[*] Visit http://tarsos.0110.be/ for more information.

(2002); Wang (2003); Allamanche (2001), there is also a review article on audio fingerprinting by Cano et al. (2005).

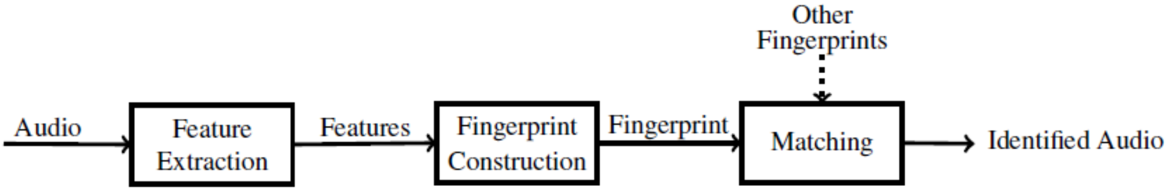


Figure 1: A general audio fingerprinter.

The workflow of our system, which can be seen in Figure 2, is exactly the same as the general acoustic fingerprinting system but it shows which features are extracted and how a fingerprint is created in our system. The first step is to extract features from audio, in this case a pitch extraction algorithm extracts pitch from audio.

The next step is to create a fingerprint, therefore we use a pitch class histogram. A pitch class histogram contains how many times any pitch class has been annotated in a musical segment or piece. A pitch class is defined here as an integer between 0 and 1200, to correspond with the cent unit introduced by Helmholtz & Ellis (1912). If for example, the value of 880Hz has been detected, this frequency f in Hz can be converted to a cent value c relative to a reference frequency r calculating $c = 1200 \times \log_2(f/r)$. With the standard r=8:176Hz[1] this makes 8100 mod 1200=900 cents. For this block of audio, one value is added to the bin representing 900 cents in the pitch class histogram. If the next block of audio[2] contains for example 220Hz, the exact same thing happens. This is done over and over again for the entire piece. Please note that this approach completely ignores temporal information, the next section explains the advantage of doing this. In theory also timbral information is not included in a pitch class histogram, in practice it has an influence. When the exact same melody is performed on piano and then on flute and subsequently both are analysed, they will generate slightly different pitch class histograms due to the characteristics of imperfect pitch detection algorithms. E.g. some pitch detection algorithms might confuse overtones for fundamental frequencies.

The third step is to match the constructed fingerprint with a list of previously stored fingerprints. In our system this entails calculating a similarity between pitch class histograms. Pitch class histograms are essentially probability density functions, they describe how probable it is a block of audio has a certain pitch. There are different ways to calculate similarity between probability density functions, for an overview, consult the article by Cha (2007).

As the final step in the process, the identified piece of audio is returned.

2.1 Pitch Class Histograms as Acoustic Fingerprints

There has been a lot of research about pitch class histograms, or very similar concepts under sometimes different names e.g. by Sundberg & Tjernlund (1969); Moelants et al. (2009); Gedik & Bozkurt (2010); Six & Cornelis (2011); Tzanetakis et al. (2002), to name a few. Especially the last article is interesting, in the future work section of they mention the following:

> Although mainly designed for genre classification it is possible that features derived from Pitch Histograms might also be applicable to the problem of content-based audio identification or audio fingerprinting (for an example of such a system see Allamanche [2001]). We are planning to explore this possibility in the future.

[1] The MIDI note number standard lets note number 0 correspond with a reference frequency of 8:176Hz, which is

C_{-1} with A_4 tuned to 440Hz. If the same reference frequency is used for cents, then MIDI note numbers and cents differ by a factor 100.

[2] The optimal size of a block of audio depends on the chosen pitch detection algorithm.

As far as we know Tzanetakis *et al.* did not explore the possibility of using pitch histograms for audio fingerprinting any further, this article can be seen as an elaboration on that idea. Both Figure 3 and Figure 4 show why this approach is reasonable. The figures show pitch class histograms of similar but not equal versions of an African song. A pentatonic scale appears to be present. Figure 3 illustrates that pitch class histograms are relatively robust against severe adaptations of the underlying audio: the histogram shape remains more or less the same. Figure 4 shows the result of audio effects which change the pitch. Changing pitch in audio shifts the histogram over the horizontal pitch axis. When calculating a correlation between histograms this needs to be taken into account.

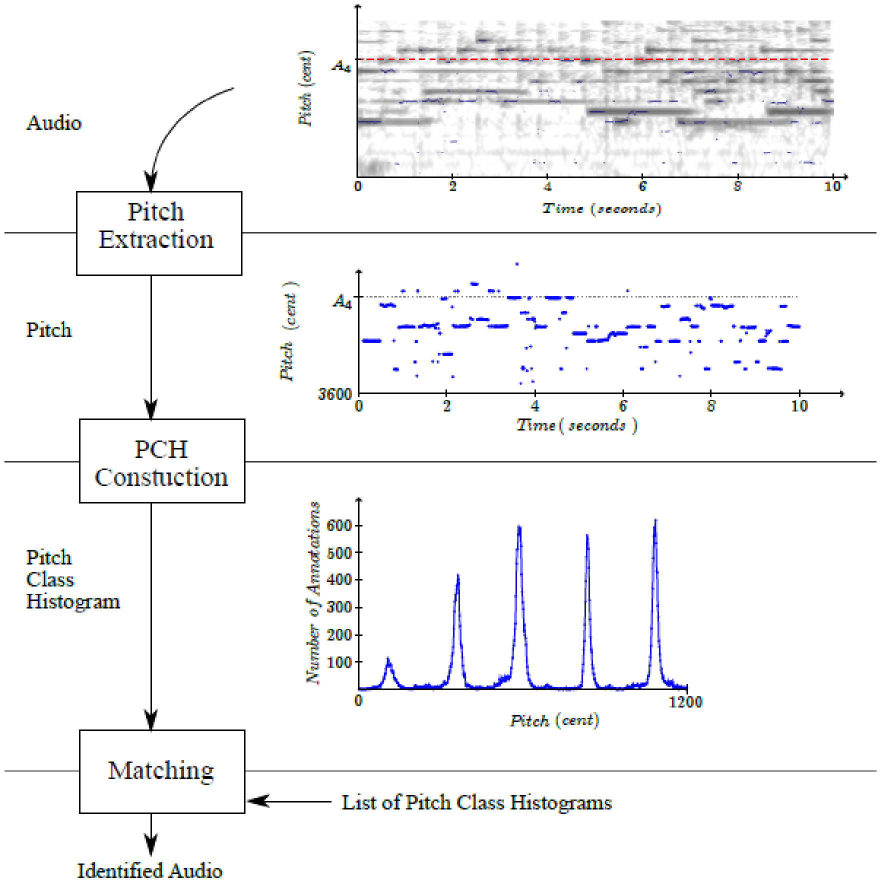


Figure 2: An acoustic fingerprinting scheme based on pitch features and pitch class histograms as fingerprints. Data is processed in an identical fashion as the general audio fingerprinter in Figure 1.

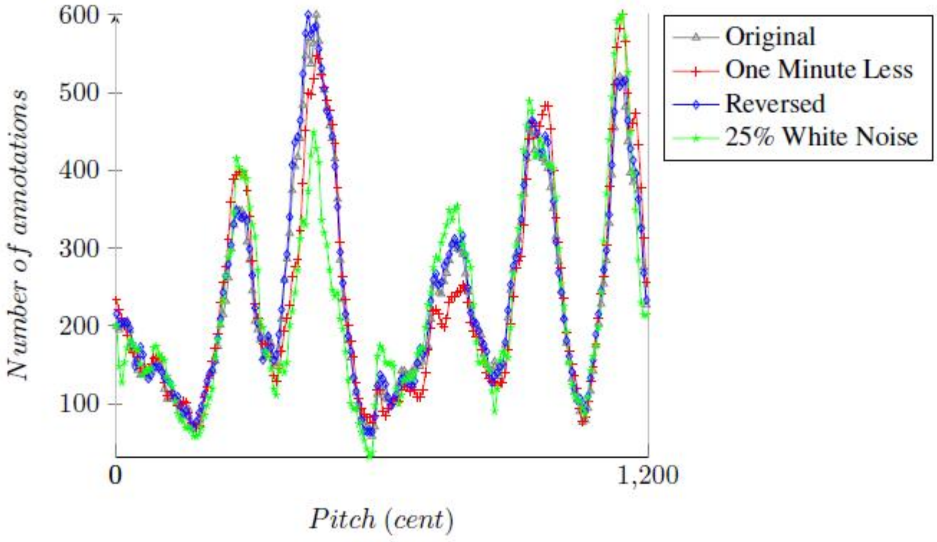


Figure 3: A pitch class histogram of an African song. The histogram of the original song is present, together with a histogram of a reversed, a cropped and noisy rendering of the song. It shows that pitch class histograms are relatively robust against severe mutilations of the underlying audio.

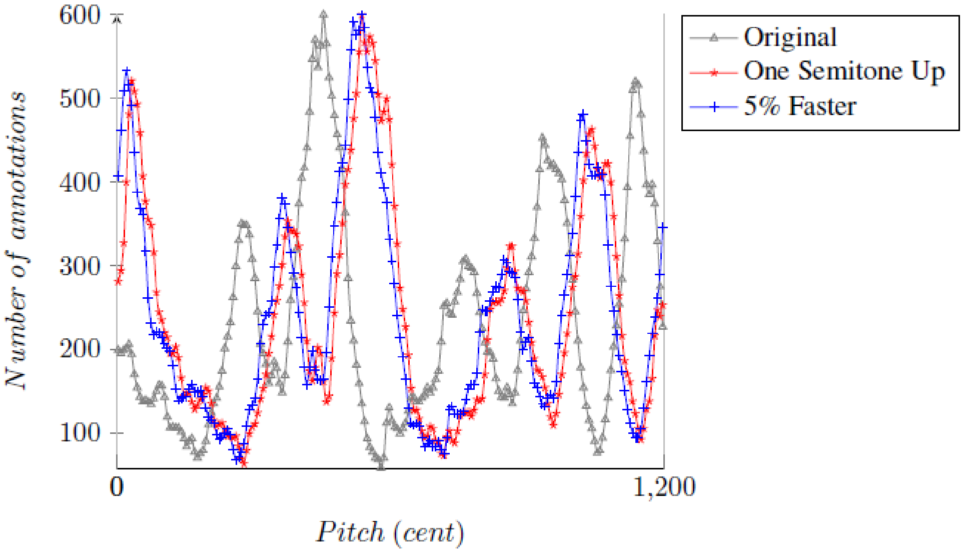


Figure 4: A pitch class histogram of an African song together with a histogram of a version played 5% faster and a pitch shifted version (without affecting the duration). It is clear that almost the same histogram is present three times, only shifted slightly over the horizontal pitch axis.

Figure 4 also shows why ignoring temporal information can be a good idea. Changing the playback speed of a song –with corresponding pitch shift– results only in a horizontal shift of the histogram, as can be seen in the illustration. In the context of analogue media this means that magnetic tape digitized on an incorrect speed can be matched with the same content digitized on the correct speed.

Then histogram overlap or intersection is used as a distance measure because Gedik & Bozkurt (2010) show that this measure works best for pitch class histogram retrieval tasks. The overlap

$c(h_1, h_2)$ between two histograms h_1 and h_2 with K classes is calculated with equation 1. To calculate the correlation with a pitch shift n equation 2 is used. To make sure that the bin k remains within the bounds of the histogram a mod K calculation is done. In our application this means that the octave relation is respected, e.g. with n equal to 50 cent[3], the bin at 1170 cent[3], the bin at 1170 cent of h_1 is compared with the bin at (1170+50) mod 1200=20 cent of h_2.

	5% Faster	One Minute Less	One Semitone Up	Original	Reversed	25% White Noise
5% Faster	1.00	0.92	0.97	0.95	0.97	0.87
One Minute Less	0.92	1.00	0.92	0.94	0.93	0.89
One Semitone Up	0.97	0.92	1.00	0.95	0.96	0.87
Original	**0.95**	**0.94**	**0.95**	**1.00**	**0.96**	**0.89**
Reversed	0.97	0.93	0.96	0.96	1.00	0.88
25% White Noise	0.87	0.89	0.87	0.89	0.88	1.00

Table 1: Similarity between different pitch class histograms of several adapted versions of a song. It shows that the histogram of the song with white noise added differs the most from the original histogram (89%).

To find the pitch shift n with maximum correlation, an exhaustive search is done by simply calculating the correlation for each possible shift. A possible significant performance increase would be to detect peaks on each histogram and then compare the histograms on only those positions (shifts), this is similar to 'tonic detection' in Gedik & Bozkurt (2010).

[3] Half a semitone, not to be confused with the American rapper.

$$c(h_1, h_2) = \frac{\sum_{k=0}^{K-1} min(h_1(k), h_2(k))}{max(\sum_{k=0}^{K-1} h_1(k), \sum_{k=0}^{K-1} h_2(k))} \tag{1}$$

$$c(h_1, h_2) = \frac{\sum_{k=0}^{K-1} min(h_1(k), h_2((k+n) \mod K)}{max(\sum_{k=0}^{K-1} h_1(k), \sum_{k=0}^{K-1} h_2(k))} \tag{2}$$

Table 1 shows the correlation, as defined by equation 2, between the different histograms shown in Figure 3 and 4, with optimal pitch shift n. It shows that the histogram based on the original version is, for this song, very much alike histogram based on the reversed audio (96%). The version with added noise differs the most from the original (89%). Cropping one minute from the song, which is 7 minutes and 20 seconds long, results in correlation of 94%. The 97% similarity between the 5% faster and pitch shifted version can be explained by the fact that a 5% speed increase translates to a pitch shift of 84 cents which is almost 100 cents[4]. The only difference then is the length of the song, i.e. the number of elements in the histogram, which can be normalized. Section 3 shows if the described behaviour is unique for this one song or not.

The implementation of the system is done in Java and uses the pitch estimator described in McLeod (2009). For testing purposes, a platform independent version can be downloaded here http://tarsos.0110.be/tag/FMA2012. There you can also find scripts and data used for this paper.

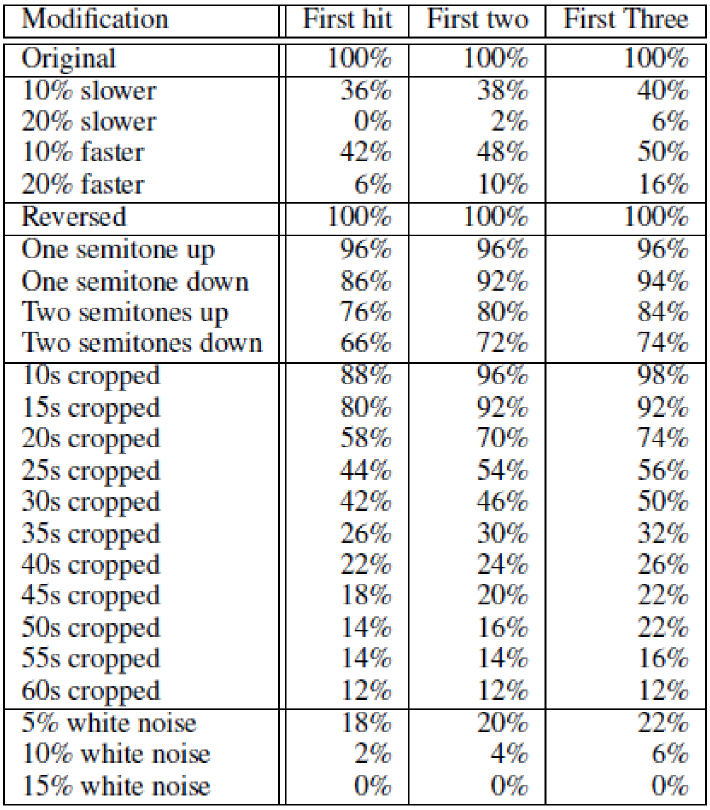
3. Experimental Results

To show that a fingerprinting scheme based on pitch class histograms has potential, an experiment was done on a data set of 10272 songs from Central Africa (see appendix A for more info on the data set). The experiment was constructed as follows: from the data set 50 randomly selected files were copied. A number of modifications and effects –27 in total– were applied to these 50 files[5], generating 1350 modified songs. The goal of the experiment was correctly match those 1350 songs to the original in the data set. Cropping was done at the beginning of the file, the average length of the 50 selected files was about 4 minutes, the shortest was one minute in length.

Table 2 shows the results of the experiment. From those results some conclusions can be drawn. 1) Since the retrieval of the original song always succeeded, it stands to reason that fingerprints for songs are, at least, unique within this data set. An important property of a fingerprint. 2) The reversed audio is also retrieved always, which shows that the pitch estimator used generates almost identical estimations on reversed audio. This is a good sanity check when using autocorrelation based pitch estimators. When using pitch detectors based on ear-models this might be less trivial. 3) Pitch shifting works reasonably well. 4) The performance when leaving out the first number of seconds degrades quickly between 15 and 20 seconds. 5) The method does not handle white noise that well. The 20%, 25% and 30% white noise versions were left out of the table since no matches were found. 6) The data set contains monophonic and polyphonic music. The results show that pitch estimators which generate one pitch per block of audio might suffice for this task, even with polyphonic music.

[4] Since $2^{(84/1200)} = 1.05$ a shift of 84 cents translates to a shift in frequency (Hz) of five percent.

[5] SoX - Sound Exchange, a command line utility, was used to apply effects to the original file. Following command line instructions were used: pitch, speed, reverse, trim, and synth whitenoise. For more information on SoX, and the
exact meaning of the effects, see http://sox.sf.net.



Modification	First hit	First two	First Three
Original	100%	100%	100%
10% slower	36%	38%	40%
20% slower	0%	2%	6%
10% faster	42%	48%	50%
20% faster	6%	10%	16%
Reversed	100%	100%	100%
One semitone up	96%	96%	96%
One semitone down	86%	92%	94%
Two semitones up	76%	80%	84%
Two semitones down	66%	72%	74%
10s cropped	88%	96%	98%
15s cropped	80%	92%	92%
20s cropped	58%	70%	74%
25s cropped	44%	54%	56%
30s cropped	42%	46%	50%
35s cropped	26%	30%	32%
40s cropped	22%	24%	26%
45s cropped	18%	20%	22%
50s cropped	14%	16%	22%
55s cropped	14%	14%	16%
60s cropped	12%	12%	12%
5% white noise	18%	20%	22%
10% white noise	2%	4%	6%
15% white noise	0%	0%	0%

Table 2: The results for a retrieval task on a data set of 10272 files. 27 effects were applied to 50 songs, generating 1350 modified versions. The goal of the task was to find the original version of the song. The percentages show much of the modified versions were correctly identified in the first, first two, and first three hits. The original and reversed version are retrieved always. The performance on pitch shift and cropping is reasonable, white noise and large tempo changes are problematic.

4. Conclusion & Future Work

In this paper a new approach for acoustic fingerprinting, based on pitch class histograms, was presented. After the introduction, which sketched the applications for the system, an overview of the working principles of acoustic fingerprinting in general and our system in particular were given. The second section also explained why pitch class histograms can be used as fingerprints. Some details about the implementation are also given. In section three experimental evaluation was done.

This paper has shown that an acoustic fingerprinting system based on pitch class histograms is rather robust and has potential but a lot of questions remain open. The experiment in this paper only discusses a retrieval task for complete songs and for a limited number of audio effects. Some future work includes:

1. Expand the retrieval task to include more audio (Western music) and apply more audio effects: echo, digital analogue/analogue digital conversions, low bit rate encoding, band pass filtering …Testing for the robustness against pitch instability, often observed in old tape recordings and testing with realistic environmental noise from a noise database - e.g. a noisy audience.
2. Document the performance decrease of the system better by using standard information retrieval measures (Precision, Recall, ROC-curves …). E.g. to do failure analysis when adding more and more noise.
3. Investigate what happens when percussive songs - without much pitch information - are fingerprinted.
4. Comparing this system with similar systems on the same data set, using the same measures.

5. See if the system can be applied to identify small fragments of music instead of complete songs. How small is the minimum fragment? Which adaptations need to be done for broadcast monitoring, processing streams?
6. Experiment with pitch estimators or chroma estimation algorithms. If the pitch estimator is replaced, is there a significant impact on the results?
7. Handle scalability and performance issues. Can the fingerprint size be reduced, without loss of accuracy? Is it possible to speed up the matching step significantly?

The system described here shows similarities with some cover song detection systems, it is very similar to the one by Serrà & Gómez (2008). This is remarkable because the goal of both systems is different. Here we want to identify the same audio, with some modifications and in the other system the goal is to identify invariant musical material (cover songs), using similar features. The similarities of the systems make sense if you look at identical audio, with some modifications, as 'the most similar cover song'. This statement results in new questions: can cover song detection systems be used to identify almost identical audio, for acoustic fingerprinting? Or the inverse: how well does the fingerprinting system described here perform at cover song detection? Currently, those questions are left as future work. The data set tested in section 3 does not include different versions - covers - of the same song.

As a final remark, we would like to note that this article is rather unique because it presents a generally applicable algorithm that is tested on ethnic music first. Only later it will be applied to western music. This is partly due to the fact that we only have access to a large data set with African music but is also a philosophical statement: instead of adapting techniques used on Western music for applications with ethnic music, why not, for once, do it the other way around?

References

Allamanche, E. (2001). Content-based identification of audio material using mpeg-7 low level description. In *Proceedings of the 2nd international symposium on music information retrieval* (ISMIR 2001).

Cano, P., Batlle, E., Kalker, T., & Haitsma, J. (2005). A review of audio fingerprinting. *The Journal of VLSI Signal Processing*, 41, 271-284.

Cha, S.-h. (2007). Comprehensive survey on distance / similarity measures between probability density functions. *International Journal of Mathematical Models and Methods in Applied Sciences*, 1(4), 300–307.

Gedik, A. C. & Bozkurt, B. (2010). Pitch-frequency histogram-based music information retrieval for turkish music. Signal Processing, 90(4), 1049–1063.

Haitsma, J. & Kalker, T. (2002). A highly robust audio fingerprinting system. In Proceedings of the 3th International Symposium on Music Information Retrieval (ISMIR 2002).

Helmholtz, H. von & Ellis, A. J. (1912). On the sensations of tone as a physiological basis for the theory of music (translated and expanded by Alexander J. Ellis, 2nd English dr.) [Book]. Longmans, Green, London.

McLeod, P. (2009). Fast, accurate pitch detection tools for music analysis. Academisch proefschrift, University of Otago. Department of Computer Science.

Moelants, D., Cornelis, O., & Leman, M. (2009). Exploring african tone scales. In Proceedings of the 10th International Symposium on Music Information Retrieval (ISMIR 2009).

Serrà, J. & Gómez, E. (2008, 31/03/2008). Audio cover song identification based on tonal sequence alignment. In Ieee international conference on acoustics, speech and signal processing (icassp) (p. 61-64). Las Vegas, USA. Available from files/publications/jserra ICASSP08.pdf

Six, J. & Cornelis, O. (2011). Tarsos - a Platform to Explore Pitch Scales in Non-Western and Western Music. In Proceedings of the 12th International Symposium on Music Information Retrieval (ISMIR 2011).

Sundberg, J. & Tjernlund, P. (1969). Computer measurements of the tone scale in performed music by means of frequency histograms. STL-QPS, 10(2-3), 33-35.

Tzanetakis, G., Ermolinskyi, A., & Cook, P. (2002). Pitch histograms in audio and symbolic music information retrieval. In Proceedings of the 3th International Symposium on Music Information Retrieval (ISMIR 2002) (pp. 31–38).

Wang, A. L. (2003). An Industrial-Strength Audio Search Algorithm. In Proceedings of the 4th International Symposium on Music Information Retrieval (ISMIR 2003) (pp. 7–13).

A Audio Material

In sectoin 3 a subset of the music collection of the Royal Museum for Central Africa (RMCA, Tervuren, Belgium) was used. The museum focuses on the African culture and treasures all kinds of ethnographic objects. The archive of the Department of Ethnomusicology has a digitized collection of about 50.000 sound recordings, with a total of 3000 hours of music, mostly field recordings made in Central Africa of which the oldest dating back to 1910. The audio archive is one of the biggest and best documented[6] archives worldwide for the region of Central Africa. On specific song, also from the RMCA collection, was used in section 2.1. It has tape number MR.1954.1.18-4 and was recorded in 1954 by missionary Scohy-Stroobants in Burundi.

[6] There is a website about the audio dataset of the Royal Museum for Central Africa featuring complete metadata and some audio fragments. It can be found at http://music.africamuseum.be