Analysis and evaluation of audio-similarity algorithms for cover and live song identification

Martin Angelov

MSci Computer Science with Industrial Year, Student ID 1422087

School of Computer Science, University of Birmingham

Supervisor

Prof. Achim Jung

Acknowledgements

Write here your acknowledgements...

Abstract

The abstract goes here. The abstract should be self-contained and:

- clearly state the problem dealt with by the thesis;
- give a synthetic description of the proposed solution;
- highlight the sense in which the proposed solution enhances the state of the art.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Project overview

Music information retrieval (MIR) is an area of analysis dedicated to extracting information from music. It combines many different disciplines of science including psychology, psychoacoustics, signal processing and computer science. One of the main aims when applying MIR techniques solving the task of song identification, i.e. matching an audio stream to a particular song [1]. This is usually achieved through a form of hashing applied on the digital signal and comparing the resulting representation to a reference fingerprint [2], [3]. This approach returns good results for the task, since we can easily quantify a good match between both fingerprints.

We can further modify the original song identification task to apply to cover songs. A cover song is a very creative reinterpretation of a released song usually performed by an artist different than the original. The cover can therefore differ significantly from the origin in tempo, pitch or song structure (*add more*). The amount of variation in a cover strongly depends on the genre of the primary track - Western popular music pieces are for example more likely to be transformed than ones from classical music [4]. Therefore the only remaining common feature between the cover song and the original is the underlying fundamental melody of the piece and potentially the lyrics.

Because of these potential disparities between two versions of a single song, the problem of identifying covers of songs is much more difficult than determing an identical match with the original. The above fingerprinting approach has been attempted [5] and the results are insignificant [6]. Direct comparison between the fingerprints of the song is unable to capture the remaining similarity within two audio files. Other MIR methods need to be considered in order to measure similarity when attempting cover song recognition.

The general advances of technology have allowed companies such as Spotify [7], Apple [8], SoundCloud [9] and more to create large-scale music databases and offer them as commercial services. Proportionally to the increasing availability of large music collections grows the need for managing the volumes of audio information through MIR techniques, with cover song identification being one of them. As a consequence most modern mechanisms to cover song recognition work by comparing an audio track called *query song* against a large database of songs, a *reference database*. Each mechanism is evaluated based on its similarity estimation performance, as well as its scalability as we increase the database size.

This project analyses the principles of a set of non-hashing based cover song identification algorithms and evaluates their performance. Most of the examined algorithms are designed to work with large-scale databases and follow the workflow model described above. The evaluation considers only their similarity estimation results and does not account for scalability. After analysis of the results a hypothesis on the best performing audio similarity technique is established (*or maybe devised?*).

1.2 Report structure

The sections of this report are as follows:

- *Chapter 2* offers a summary of the background information required to understand and implement the audio similarity algorithms
- *Chapter 3* explores other state of the art methods of measuring similarity not examined in detail by the project
- *Chapter 4* provides a description of the evaluation task through which each algorithm is analysed
- *Chapter 5* contains detailed descriptions of each algorithm
- *Chapter 6* expands on implementation details related to the benchmark tool
- *Chapter 7* outlines the best results achieved and offers an analysis on them
- *Chapter 8* summarises potential further contributions to the project
- *Chapter 9* discusses the main challenges related to the project and the task of cover song identification
- *Chapter 10* is a summary of the project management techniques utilised during the project

Chapter 2

Background theory

Each type of information extracted from an audio stream is referred to as an *audio feature*. Audio features are mainly derived using various transformations on the signal based on some basic properties of sound. This section presents low-level theory required to understand the high-level description of how each feature is obtained further in the report. At this level of the project description we only require an understanding of general principles related to audio and signal processing, therefore the explanations are kept concise without going deeper into technical details.

2.1 Basic properties of audio signal

A *digital audio signal* is a representation of the continuous sound wave as a series of binary numbers. This representation helps preserve the *frequency* (the speed of the vibrations), as well as the *amplitude* (the fluctuations of the vibrations) of the sound. The energy that each sound wave emits through vibrations is called *sound energy* and its rate is measured through *sound power*. The majority of audio features use frequency or power as a primary audio property used to define the feature (*modify/change*).

2.1.1 Sampling and sample rate

The process of converting an analogue sound wave to a digital one involves a process of extracting points (samples) from the continuous signal and using them to describe the signal into a discrete form. This method is called *sampling* and the amount of samples collected per time frame is *sample rate*. The representation of a song used during feature extraction is a sequence of samples extracted from the digital signal of the song based on its sample rate.

2.1.2 Tones

In order to understand how properties of a digital signal could form audio features we first need to examine how they relate to the ways of how people perceive music and sound. Western music is described using *tones*, steady periodic sounds [10]. They can be pure if the sound has sinusoidal waveform, or complex if they are a combination of pure tones with a periodic repetition pattern. A half tone is called a *semitone* and it is the smallest measure of period between sounds. Semitones are grouped into *octaves* where each octave contains 12 semitones. The acoustic opposite of tone is noise, a disordered sound which is unpleasant to the human brain and is disruptive to hearing [11].

2.1.3 Psychoacoustic properties

Tones are distinguished by several basic perceptual properties of sound - pitch, loudness, timbre and duration [12]. Consequently these properties are also regularly used to describe what is accurately captured by each audio feature. Klapuri et al. [13] form good definitions of the psychoacoustical terms outlined.

They define *pitch* as a perceptual attribute which offers ordering of tones on a frequency-related scale. Different pitches could be labelled through the Helmholz pitch notation [14] using letters, through scientific pitch notation [15] utilising letters and numbers, or directly using numbers representing the closest frequency in hertz (hz). Despite being determined by clear and stable frequencies in sound, pitch is more importantly a subjective auditory sensation, so a strict mathematical relationship between frequency and pitch does not exist [16]. As a standard it is accepted the musical note of A above C (Helmholz notation) or A4 (scientific notation) has a frequency of 440 Hz [17]. The human ear perceives musical intervals on an approximately logarithmic scale with respect to a *fundamental frequency*, the lowest frequency available in a tone and therefore determining the overall pitch of the tone. Using this notion of logarithmic perception there are various mappings of pitches to frequencies, the most famous of which is the MIDI tuning standard [18].

Loudness is another subjective psychoacoustical attribute of sound [13]. Similar to how pitch is related to frequency, loudness is a perception of sound pressure. *Sound pressure* is a measurement of the pressure divergence caused by a sound wave to the ambient atmospheric pressure [19]. Loudness orders sounds on a scale ranging from quiet to loud [16].

Tones also have a "colour" attribute attached to them through the introduction of *timbre*. Timbre allows distinguishing sounds with potentially identical pitch, loudness and duration, but produced by different musical instruments [13]. It is a complex concept which cannot be defined by a single property of sound. There are many attempts at breaking down the attribute into components. Robert Erickson [20] offers one of the most accepted decompositions where timbre relates to the following acoustic parameters of sound:

1. Tonal and noise characters

2. Time envelope - A *time envelope* describes how sounds changes over time [21]. It measures how much time it takes for the sound to reach an amplitude level when a musical instrument is activated (a key on the piano is pressed, for example) and subsequently how long it takes for the sound to go back to its initial level.

3. Spectral envelope and any changes to it - when only the curve of the amplitude out of a time envelope is taken into consideration then a *spectral envelope* is used. Figures 2.1 and 2.2 illustrate the difference of between both envelopes.

4. Changes in the fundamental frequency

5. Onset dissimilar to the dominant vibration - with *onset* being the start of a musical note we look for 'anomalies' in the vibration of the wave compared to the vibrations following the anomaly.

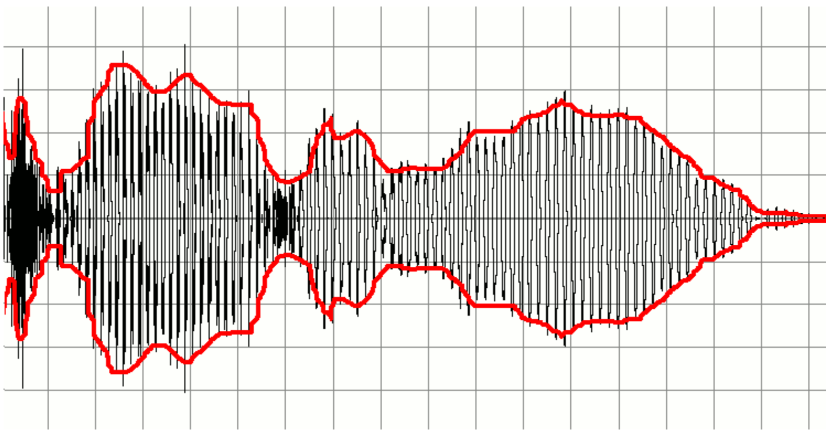

Figure 2.1: Time envelope *add citations and description*

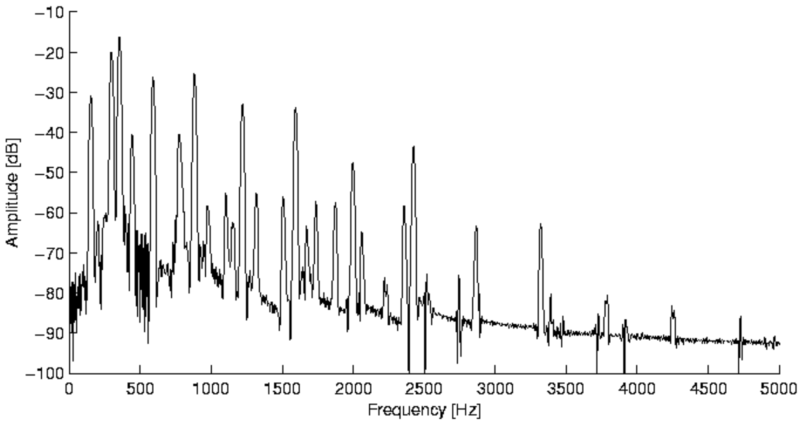

Figure 2.2: Spectral envelope *add citations and description*

Out of all psychoacoustical properties introduced *duration* is the one which is the easiest to directly measure. It is an indicator of a length of any part of a musical composition - tone, pitch, the whole piece, etc [22]. The measurement is expressed using a base unit of time.

2.1.4 Beat

One final basic concept that we need to introduce is the *beat*. It signifies repeating portions in a song which define the overall rhythm of a music piece. Rhythm is formed of "strong" and "weak" beats, with the former signifying a suitable moment for melody change.

2.2 Audio transformation techniques

2.2.1 Fourier transform

Any waveform including the audio ones could be presented as a sum of sinusoids of different frequencies [23]. In music each of the constituting frequencies represents a pure tone. In order to be able to work with the tonal representation of a song we need to find a way to separate the different frequencies within the audio signal. We achieve that using *Fourier transform* - one of the most widely used audio transformations.

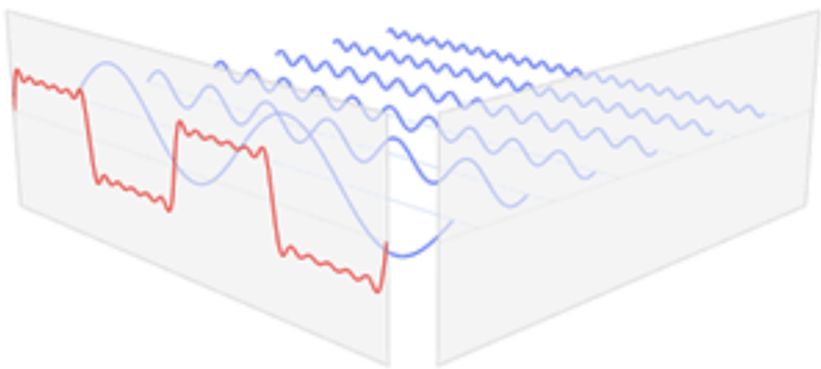

Figure 2.3: Fourier transform applied on a periodical function (in red). The transform distinguishes six sine functions (in blue) represented as an amplitude-frequency relationship [24]

Fourier transform could be applied to either periodic (resulting in *Fourier series*) or non-periodic functions. As we are working in the domain of music we are focussed on the Fourier workings on functions with periodicity. Using the transform our goal is finding an approximation for function $f(t)$ with period $T = 2L$ using a sinusoid functions each with period a multiple of T [24]. Figure 2.3 shows the destructuring that a Fourier transformation performs on a function. The resulting function $g(t)$ which is the transform of $f(t)$ takes the form of:

$$g(t) = \frac{a_0}{2} + \sum_{n=1}^{\infty}(a_n \cos\frac{n\pi x}{L} + b_n \sin\frac{n\pi x}{L}) \tag{2.1}$$

Figure 2.4: Fourier series representation of a periodic function [25].

The coefficients a_n and b_n determine the relative weights of each of the sinusoids [24]. They are calculated using sine and cosine integrals over the function period:

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{n\pi x}{L}dx,\ n = 1, 2, 3, \ldots \tag{2.2}$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{n\pi x}{L}dx,\ n = 1, 2, 3, \ldots \tag{2.3}$$

Figure 2.5: Derivation of the parameters for the Fourier transform [25].

From a frequency spectrum perspective the relative weights of the sinusoids also represents the amount of frequency present in the original function at a point of time [26]. This information is very valuable when determining sound properties such as pitch, power, timbre and more.

2.2.2 Filters

During the processing of the audio signal we want to detect the frequency maxima and minima of the wave, or possibly attenuate certain frequencies. In order to do that we use *audio filters* - tools that pass certain frequencies and blocks others [27]. Filters are defined in terms of *bands*, the range of frequencies they pass. A combination of filters which helps us produce the required frequency cut-off is called a *filter bank*.

2.2.3 Mel scale

The *mel scale* presents a perceptual ordering of pitches which are determined to be equally far away from each other [28]. The scale introduces a unit of perceptual pitch called *mel*, where 1000 $mel = 1000\ Hz$. The rest of the mel-frequency mapping is displayed in figure 2.6.

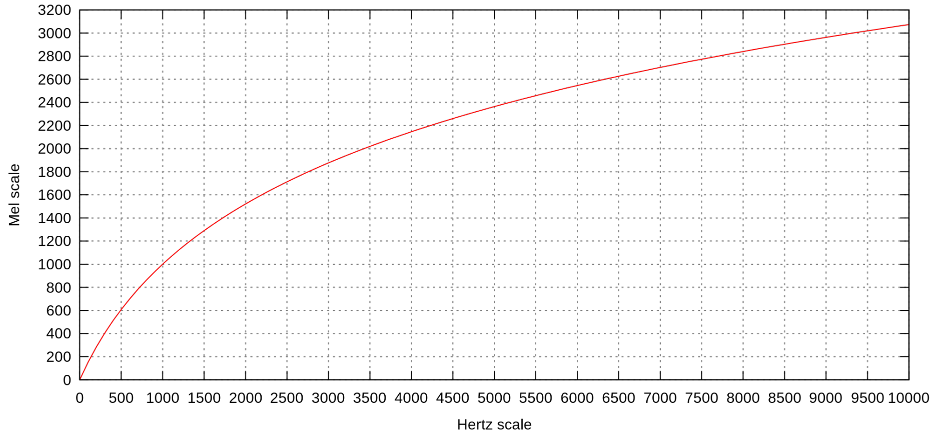

Figure 2.6: Perceived pitches on a Mel scale versus the Hertz frequency scale [29]

2.3 Machine learning techniques

Some cover song identification algorithms create machine learning models based on which they measure similarity when a query song is provided. This section outlines the principles of each method used later in the algorithms.

2.3.1 Random forests

Random forests are a form of ensemble learning for classification and regression [30]. Ensemble methods train multiple learners in an attempt to solve the same problem [31], and in the case of random forests the type of learner is a decision tree. The method works by building a predefined number of decision trees during training, and return a result combining their individual outcomes. In the case of classification the mode of all returned class predictions is taken, while the mean of the predictions is used during regression. Random trees help avoid the tendency of individual decision trees to overfit to their training data [32].

2.3.2 K-means clustering

K-means clustering is an unsupervised machine learning approach which aggregates a collection of data points together because of some similarity between them. The end result is a separation of the points into k distinct clusters. The outcome of the clustering is evaluated through the sum of squares defined as:

$$Sum\ of\ squares = \sum_{i=1}^{n} (x_i - \overline{x})^2 \tag{2.4}$$

Figure 2.7: The sum of squares for n items where x_i is the value of the i-th item and $\overline{x}$ is the mean of the set n.

The intent is to reduce the sum of squares as much as possible. Clustering is complete when either the sum of squares does not significantly change any more, or the algorithm runs a pre-determined number of iterations.

In the area of digital signal processing K-means clustering is used to perform *quantisation* - mapping a large (possibly continuous) set of values to a countable smaller set which is easier to work with [33].

Chapter 3

Related work

Summary

short summary of the chapter...

One or more chapters should be devoted to the description of the proposed approach...

In particular, this chapter describes the design adopted by this research to achieve the aims and objectives stated in the Introduction.

3.1 Examination of other audio similarity techniques and algorithms not analysed by the project

3.2 Scientific paper rewritten

Chapter 4

The task

Summary

Discuss here the methodology used in the study, the stages by which the methodology was implemented, and the research design; For examples, one section details the participants in the study, another section lists all the instruments used in the study and justifies their use; another section outlines the procedure (algorithms, code,..) used; a section discusses how the data was analysed, etc..

4.1 Design

4.2 Evaluation methods and metrics

4.3 Datasets used for evaluation

Chapter 5

The algorithms

Summary

Details all the results of your study here (exploits graphics for results visualization). This chapter should also contain a full discussion, interpretation and evaluation of the results.

5.1 Osmalskyj big algorithm

5.2 Osmalskyj weak features

5.3 Ellis cross-correlation algorithm

5.4 Osmalskyj quantisation algorithm

5.5 Tralie timbre algorithm

5.6 Rank aggregation techniques

5.7 Rafii audio fingerprinting algorithm

Chapter 6

The benchmark

Summary

Details all the results of your study here (exploits graphics for results visualisation). This chapter should also contain a full discussion, interpretation and evaluation of the results.

6.1 Implementation details

6.2 Brief usage information

6.3 Algorithm structure in the benchmark

6.4 Result format produced by benchmark

Chapter 7

Results

Summary

Details all the results of your study here (exploits graphics for results visualisation). This chapter should also contain a full discussion, interpretation and evaluation of the results.

7.1 Best results

7.2 Comparison to results from papers

7.3 Result analysis

Chapter 8

Further work

Details all the results of your study here (exploits graphics for results visualisation). This chapter should also contain a full discussion, interpretation and evaluation of the results.

Chapter 9

Challenges

Summary

Details all the results of your study here (exploits graphics for results visualisation). This chapter should also contain a full discussion, interpretation and evaluation of the results.

9.1 Lack of datasets

9.2 Lack of universal comparison metric?

9.3 Academic papers algorithm description

Chapter 10

Project management

Summary

Details all the results of your study here (exploits graphics for results visualisation). This chapter should also contain a full discussion, interpretation and evaluation of the results.

10.1 Using GitLab

10.2 Canvas logs

10.3 other? Gantt chart?

Chapter 11

Conclusion

Conclusions should summarize the problem, the solution and its main innovative features, outlining future work on the topic or application scenarios of the proposed solution.

References

[1] E. Weinstein and P. Moreno, "Music identification with weighted finite-state transducers," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 2, pp. II–689, IEEE, 2007. 1

[2] A. Wang *et al.*, "An industrial strength audio search algorithm.," in *Ismir*, vol. 2003, pp. 7–13, Washington, DC, 2003. 1

[3] J. Haitsma, T. Kalker, and J. Oostveen, "Robust audio hashing for content identification," in *International Workshop on Content-Based Multimedia Indexing*, vol. 4, pp. 117–124, Citeseer, 2001. 1

[4] D. P. Ellis and G. E. Poliner, "Identifyingcover songs' with chroma features and dynamic programming beat tracking," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–1429, IEEE, 2007. 2

[5] T. Bertin-Mahieux and D. P. Ellis, "Large-scale cover song recognition using hashed chroma landmarks," in *2011 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 117–120, IEEE, 2011. 2

[6] D. P. Ellis and B.-M. Thierry, "Large-scale cover song recognition using the 2d fourier transform magnitude," 2012. 2

[7] Spotify, “Music for everyone - spotify,” 2019. [Online; accessed 29-March-2019]. 2

[8] Apple, “Music - apple (uk),” 2019. [Online; accessed 29-March-2019]. 2

[9] SoundCloud, “Soundcloud – listen to free music and podcasts on soundcloud,” 2019. [Online; accessed 29-March-2019]. 2

[10] Wikipedia contributors, “Musical tone — Wikipedia, the free encyclopedia,” 2019. [Online; accessed 31-March-2019]. 5

[11] The Physics Hypertextbook, “Music & noise - the physics hypertextbook,” 2019. [Online; accessed 31-March-2019]. 5

[12] Acoustic Glossary, “Sound power - acoustic glossary - article,” 2019. [Online; accessed 29-March-2019]. 5

[13] A. Klapuri and M. Davy, *Signal processing methods for music transcription*. Springer Science & Business Media, 2007. 5, 6

[14] H. Helmholtz, *On the sensations of tone*. Courier Corporation, 2013. 6

[15] H. Fletcher, “Loudness, pitch and the timbre of musical tones and their relation to the intensity, the frequency and the overtone structure,” *The Journal of the Acoustical Society of America*, vol. 6, no. 2, pp. 59–69, 1934. 6

[16] A. S. of America. Secretariat and A. N. S. Institute, *American National Standard Psychoacoustical Terminology*. American National Standard, American National Standards Institute, 1986. 6

[17] R. W. Young, “Terminology for logarithmic frequency units,” *The Journal of the Acoustical Society of America*, vol. 11, no. 1, pp. 134–139, 1939. 6

[18] MIDI, "The complete midi 1.0 detailed specification," 2019. [Online; accessed 31-March-2019]. 6

[19] Engineering Toolbox, "Sound pressure," 2019. [Online; accessed 31-March-2019]. 6

[20] R. Erickson, *Sound structure in music*. Univ of California Press, 1975. 6

[21] Wikipedia contributors, "Envelope (music) — Wikipedia, the free encyclopedia," 2019. [Online; accessed 31-March-2019]. 7

[22] B. Benward, *Music in Theory and Practice Volume 1*. McGraw-Hill Higher Education, 2014. 8

[23] The Fourier Transform, "Fourier transform," 2019. [Online; accessed 31-March-2019]. 8

[24] The Fourier Transform, "Fourier series," 2019. [Online; accessed 31-March-2019]. 9, 10

[25] Zachary S Tseng, "Second order linear partial differential equations," 2019. [Online; accessed 31-March-2019]. 10

[26] Wikipedia contributors, "Fourier transform — Wikipedia, the free encyclopedia," 2019. [Online; accessed 31-March-2019]. 10

[27] Betty Lise Anderson, "Audio filters," 2019. [Online; accessed 1-April-2019]. 10

[28] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937. 11

[29] Wikipedia contributors, "Mel scale — Wikipedia, the free encyclopedia," 2019. [Online; accessed 1-April-2019]. 11

[30] T. K. Ho, "Random decision forests," in *Proceedings of 3rd international conference on document analysis and recognition*, vol. 1, pp. 278–282, IEEE, 1995. 11

[31] Z.-H. Zhou, *Ensemble methods: foundations and algorithms.* Chapman and Hall/CRC, 2012. 11

[32] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*, vol. 1. Springer series in statistics New York, 2001. 12

[33] Wikipedia contributors, "Quantization (signal processing) — Wikipedia, the free encyclopedia." `https://en.wikipedia.org/w/index.php?title=Quantization_(signal_processing)&oldid=889000626`, 2019. [Online; accessed 1-April-2019]. 12

[34] J. G. Roederer, *The physics and psychophysics of music: an introduction.* Springer Science & Business Media, 2008.

[35] Wikipedia contributors, "Fourier series — Wikipedia, the free encyclopedia," 2019. [Online; accessed 31-March-2019].