Rhythmic arrangement from finger-tapping audio signals

Javier Nistal

MASTER THESIS UPF / 2016

Master in Sound and Music Computing

Master thesis supervisor:

Dr. Sergi Jordà

Prof. Perfecto Herrera

Department of Information and Communication Technologies

Universitat Pompeu Fabra, Barcelona

Acknowledgments

First, I would like to thank my supervisors, Perfecto Herrera and Sergi Jordà for their guidance and support throughout this project. Additionally, I would also like to thank Dimitry Bogdanov who helped me out with Essentia and other programing issues. Thanks also to all the participants in the crazy experiments that were undertaken for this project. Without your collaboration, this project would not have been possible. Finally, a big thank to all my folks in the Master, specially to Kosmas Kritsis, for holding my recurrent discussion about the same topic.

Abstract

We present a novel study on the behavior of human finger-tapping. This technique can be understood as the casual and rhythmic hitting of objects for the expression of a musical idea, even when it is unconscious or sometimes just for stress relief or because of nervousness. The idea that underlies this project is the connection of spontaneous finger-tapping with human-computer interaction and automatic arrangement of percussion. An application under this functional concept would certainly be a useful tool for the home-studio producer. Our first step was to study the behavior of spontaneous rhythmic expression as well as the needs of inexperienced users in the context of rhythm expression. To this end, we first collected a dataset by recording spontaneous finger-tapping patterns performed by subjects from different music backgrounds. Then, an online survey gathering information about the recording was submitted to the volunteers. Further analysis of the survey answers and several spectro-temporal features extracted from the audio content allowed to infer meaningful information about this behavior. Results of this experiment suggested that there are two clear ways for finger-tapping depending on the music training of the performer. We demonstrate the former hypothesis by conducting a classification task between onsets from both finger-tapping methods. We achieved a 99% of accuracy in recognizing drumming expertise levels (expert vs. nave) by means of using onset-related acoustic features. This suggested that people with percussion training are more concerned about timbre aspects and, thus, they take advantage of this quality of sound to provide differences to each stroke when finger-tapping, as opposed to non-expertise individuals. Secondly, we aimed to convert all the gathered knowledge into a creative tool for arranging drum patterns. Therefore, we describe a system for finger-tapping transcription as an underlying step in the usage of this behavior as a mean for improving human-computer interaction in the context of computer music creation. The system can be divided into three parts: an onset detection and feature extraction step, in which a set of frame-wise time-dependent features are calculated. These features are fed into a k-Means clustering/classification step, in which the feature representation of the finger-tapped onsets are clustered, assigned to a drum sound class and then translated into a drum MIDI-like symbolic representation.

Keywords: human finger-tapping, rhythm expression analysis, human-computer interaction, interfaces, music creation

Contents

List of Figures

List of Tables

Chapter 1

INTRODUCTION

Finger-tapping, also known as tap picking, is popularly known as a guitar and bass technique in which the strings of the instrument are fretted against the fingerboard in order to produce legato notes [2]. This technique can be understood in a more general sense though, not only in the context of string music instruments but from a more cognitive perspective, defining it as the casual and rhythmic hitting of objects for the expression of a musical idea, even when it is unconscious or sometimes just for stress relief or because of nervousness. Moreover, and given that the hands and feet are human's main limbs for interacting with the environment, it can be assumed that finger-tapping, together with foot-tapping, is one of the most straightforward ways for the spontaneous expression of rhythm. Notice the behavioral connotation of the provided definition; beyond a simple technique for playing a certain instrument, we are referring to it as a human habit and as such, widely extended. In fact, this enjoyable behavior can be clear for all to see in everyday situations such as listening or composing music, killing time in long waits, speaking on the telephone, and many more. Despite being a broad and innate habit granted by our physiology, few music technologies take advantage of this trend for improving human-computer interaction, specially, and surprisingly given the clear suitability of those technologies for the creation of digital drums in computer music production.

Drum composition represents one of the most arduous tasks in computer music production. Particularly for those producers that do not have a deep percussion experience or may not be familiar with digitized drum editing and sequencing, many difficulties appear when it comes to project a conceived and ideal drum pattern, in the performer's mind, to the Digital Audio Workstation (DAW). This is due either to the lack of knowledge from the part of the producer or to the unavailability or limitations of the commercial equipment for interacting with this musical dimension. An example of these devices are drum machines, MIDI (Musical Interface for Digital Instruments) trigger pads, MIDI drums and other user-oriented

technologies such as Air-drums[1] or sensor gloves that take advantage of the ergonomics of the hands[2]. In general, the interface design baseline for this tools is to provide a set of sensitive surfaces that trigger a certain sound in the computer when they are hit. Many drum machines implement a second layer for creating patterns in time (e.g. Roland TR-808, Akai MPC 2000, etc). In other words, there are "pads" that just trigger, but there are also "step sequencers", that trigger in time. In addition to their relative high cost, these technologies are generally constrained to their specific mode of use, which is hard for non-experienced users and, of course, does not provide any further processing other than the translation of the sequence of strokes into a MIDI-like symbolic representation. Any further arrangement in the MIDI pattern is left in the user's hands. Also, the existence of a huge amount of drum sounds and rhythm patterns has added the challenge of choosing style, as the combination and musical interplay of these two aspects of percussion. Deciding the appropriate drum sound set given a certain pattern requires a great music experience and is crucial to the good cohesion of any musical piece.

Music academia has traditionally understood arrangement or transcription as a process aiming to adapt a certain song, composed for an ensemble or a specific instrument, for a different target instrument [3]. From a general perspective, it refers to the "musical reconceptualization of a previously composed work" [4]. Generally, there are two major approaches for music arrangement. One is rewriting a piece of existing music with additional material. Apart from this, score reduction approach reduces the original work from a larger score to a smaller one. In other words, arranging is a process that adds new thematic material for conferring musical variety (e.g. introductions, transitions, or modulations, and endings) to a certain piece, through compositional techniques. In this work we are particularly interested in the arrangement of rhythm. However, rhythm itself has been source of debate, partly because it has often been identified with one or more of its constituent (such as accent, metre, and tempo), but not wholly separated elements. Some theories require periodicity as the foundation of rhythm, other include in it even non-recurrent configurations of movement. Moreover, the concept of 'rhythm varies between cultures (Indian music, Turkish, Arabian, etc). On this basis, it is difficult to strictly define the role of rhythm arrangement. The elements that make rhythm are pulse beat and measure; unit and gesture; alternation and repetition; tempo and duration; metric structure[3]. However, many music producers do not know about rhythm theory and may not stick to this terminology. Thus, we can expect that some of these elements may not be perceived or

[1] http://aerodrums.com/aerodrums-product-page/

[2] https://learn.adafruit.com/midi-drum-glove/overview

[3] For further definitions of rhythm elements you may refer to [5, 4, 6]

be imprecise in the context of spontaneous finger-tapping. For now, as conventional wisdom would understand, the arrangement of rhythm can be understood as a process aiming at bringing over an ideal rhythmic pattern, conceived in a particular subjects mind, to the one actually performed and recorded. The idea that underlies this project is the connection of spontaneous finger-tapping with human-computer interaction and automatic arrangement of percussion. We believe that an application under this functional concept would certainly be a useful tool for the home-studio producer.

1.1 Motivation and goals

In the above described context, the interest for studying this particular human behavior emerges. As we will see in the next Chapter, research on human finger-tapping is very limited. Also, my experience into computer music production has made me aware of the difficulties that involves drum composition. The selection of the appropriate drum-kit, deciding the number of percussion layers, the interaction with the controller (if you have!), creating coherent variations of the pattern you played, etc. These facts motivate the study of finger-tapping from a perceptual perspective as well as within the framework of human-computer interaction. The purpose is to gather knowledge about this topic and, hopefully, develop a new creative tool for interacting with drums taking advantage of these preexisting human habits. To this end, we have stated the following preliminary goals:

- Study the behavior of spontaneous rhythmic expression
- Study the needs of inexperienced users in the context of rhythm expression
- Convert this knowledge into a creative tool for arranging drum patterns

1.2 Structure of the document

In Chapter two it is provided a deep insight into the state of the art related with finger-tapping or with other general aspects that may contribute to the understanding of rhythm and percussion. In Chapter three the methodology applied in the research is presented, including the study of the user properties and context as well as the development and evaluation of the potential software. In chapters four results are presented and discussed. Chapter five provides a summary of the contributions made in the scope of our research and presents suggestions for future work.

Chapter 2

STATE OF THE ART

Different research fields have undertaken the study of human finger-tapping, in an indirect way though. Most of the work that we found belongs to the following areas:

- In medicine, as mean in the diagnose of hyperactivity [7] or Parkinson's disease [8] and other diseases.
- In the field of music cognition, as a mean for studying human beat tracking [9] and other perceptual aspects of rhythm.
- Few research has been focused towards the understanding of this behavior itself, specially from the perspective of musical expression. It can only be found some psycho-motor studies on the velocity of finger-tapping for percussionists against non-percussionists [10].

General tapping is characterized by a consecutive and rhythmic hitting of a surface or several surfaces producing different sounds. The beater can either be a stick, our extremities or any other object. Thus, we can expect that dynamics (i.e. the strength of each stroke), the relative timbre between strokes (i.e. the acoustical characteristics of the objects involved) and time (i.e. the relative location in time of each stroke within the pattern) play an important role in the description of human finger-tapping. Moreover, it is necessary the symbolic representation of this information for attempting further arrangement processing. For this reason, we find it useful as well as inspiring to get a general scope of the contributions made so far in the field of percussion sound analysis, rhythm classification and other audio-driven technologies that integrate feasible human conducts, with special emphasis in those involving finger-tapping. First we review some of the works focusing on the description of timbre spaces for drum sound characterization and symbolic representation, which will provide a good insight into the

feasible techniques for the classification and transcription of finger-tapping patterns. Following, a deep insight of the different approaches for rhythm description and classification is provided. This topic could be useful for providing the system with rhythm-wise knowledge for further high-level capabilities (symbolic representation, pattern recommendation, event sound recommendation, intelligent drum pattern enrichment, etc). Next, following the thread of applications that make use of finger-tapping and other human common habits, we describe some of the existent techniques for audio content retrieval based on query by tapping and beat-boxing techniques. To conclude this section, it is provided an overview of some high level applications such as sound recommendation and, from a general perspective, automatic music arrangement systems.

2.1 Analysis and classification of timbre in percussion instruments

Research on the classification of percussion sounds provides a good hint of the kind of descriptors that are feasible for describing timbre in finger-tapping. Timbre "is the quality of a musical note, sound, or tone that distinguishes different types of sound production, such as voices and musical instruments, string instruments, wind instruments, and percussion instruments" [11]. To our knowledge, it is not straightforward the definition of a finger-tapping timbre space. We expect a common timbre component between different recordings derived from the use of the hands, which provide a particular texture to the strokes. However, the main contribution to the timbre quality of the generated sound will depend on the surface and shape of the object being hit. Moreover, the transcription of each different stroke to sounds from a drum-set is not clear. On one side, we are interested on classifying finger-tapped strokes into a set of classes from a drum-set family taxonomy, but at the same time, the assignation of a drum class to a given stroke is constrained to perceptual aspects. Thus, a general revision of the methods for percussive sound description and its symbolic representation, both from a perceptual and taxonomic perspective, is provided.

2.1.1 Taxonomic classification of percussive sounds

In 2002 there were already very accurate studies in the classification of drum sounds. Herrera [12] studied the classification of standard isolated drum sounds from a set of 634 drum samples and a taxonomy of up to nine instrument classes. They carried out three category level classification with different machine learning algorithms and feature selection techniques. Results demonstrated the relevance

of Zero Crossing Rate (ZCR), third and fourth moments (Skewness and Kurtosis), spectral and temporal centroid, relative energy in specific bands, and some low-order Mel-Frequency Cepstral Coefficients (MFCC's), in the classification of this kind of sounds. The same author reported very good performance rate at identifying 33 different classes of acoustic and electronic drum sounds from a dataset of 1976 sound samples [13]. Results show that log-transformed spectral features perform considerably better than other combination, achieving a 84% of accuracy and regardless of using or not feature selection. Further results show that it can be described instrumental sounds beyond the class label to some more detailed and idiosyncratic level: in this case, the name of two manufacturers. More recent studies attempted to discriminate sounds produced by the same percussion idiophone instrument [14]. More specifically, sounds produced by different cymbal types such as China, Crash, Hi-hat, Ride and Splash. The authors propose the use of spectral features from non-negative matrix factorization to train an I-Nearest Neighbor algorithm to classify specific combinations of cymbals with a very limited amount of training data. In this sense, another related work is presented in [15] that proposes a more challenging investigation with a two-level classification of cymbal sounds. In the first level corresponds to the cymbal type and the second level classifies how the sound was made. The overall classification rate obtained for three cymbal combinations was 86%. These works try to provide some first steps towards the building of systems for detailed drum transcription from polyphonic music.

2.1.2 Perceptual description of percussive sounds

Research on perceptual similarity of sounds provides useful information for addressing the problem of automatic classification of drum sounds. In perceptual studies, dis-similarity judgments between pairs of sounds are derived from human subjects. With multidimensional scaling techniques, researchers find the dimensions that underlie to the dis-similarity judgments. Even further, with proper comparison between those dimensions and physical features of sounds, it is possible to discover the links between perceptual and physical dimensions of sounds [12]. A three dimensional perceptual space for percussive instruments, depicted in Figure 2.1, has been hypothesized by Lakatos [1]. This percussive perceptual space spans three related physical dimensions: log-attack time, spectral centroid and temporal centroid. These physical dimensions are also used in the MPEG-7[1] description format as descriptors for timbre. However, experiments concluded that

[1]MPEG-7 is a multimedia content description standard. It was standardized in ISO/IEC 15938 (Multimedia content description interface). This description will be associated with the content itself, to allow fast and efficient searching for material that is of interest to the user. MPEG-7 is formally called Multimedia Content Description Interface.

the quality of these descriptors was insufficient to distinguish fine details in the samples [16]. Same authors accomplished the implementation of a hierarchical user interface for efficient exploration and retrieval of drum sound libraries based on a computational model of similarity and Self-Organizing Maps (SOM). The model follows a perceptual approach by considering on its computation the impulse response of the outer and middle ear as well as masking aspects of the auditory system. The similarity method is based on simple distance after an alignment process. The user interface enables to navigate through the drum sound collection by means of a visual representation of the model based on SOM. Herrera et al. [17], presented several music content descriptors that are related to percussion instrumentation and are subjected to semantic aspects of percussive-like sounds perception. The percussion index estimates the amount of percussion that can be found in a music audio file and yields a (numerical or categorical) value that represents the amount of percussion detected in the file. A further refinement is the percussion profile, which roughly indicates the existing balance between drums and cymbals.

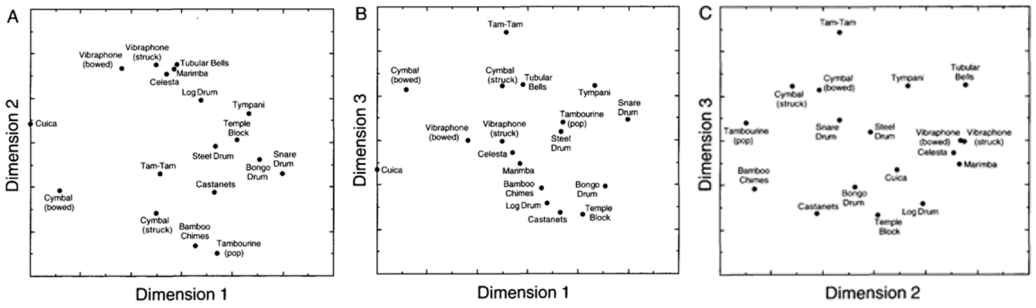

Figure 2.1: Three-dimensional feature representation of the unpitched percussion sound classes [1].

2.1.3 Transcription of percussion sound patterns

Recent research on the description of percussion sounds has been drawn towards the transcription of unpitched percussion instruments found in Western popular music, with special emphasis on standard pop-rock drum kits (snare drum, kick drum, toms, hi-hats, and cymbals). A notable exception to this is the work on tabla and djembe transcription [18]. Drum transcription methods attempt to provide a symbolic representation of a rhythmic content. To this end, it is required to recognize the different kind of sound sources that produce each stroke within a percussion loop. In general, the percussion transcription problem has been addressed through three different strategies [19, 20]: onset detection approach,

segmentation-based approach or combination of these. Systems in the first category first detect multiple streams corresponding to drum types, usually via a signal or spectral decomposition approach, e.g. [21, 22], or simpler sub-band filtering [23, 24], and then identify onsets in the individual streams. Systems in the second category detect a regular or irregular event grid in the signal, segment the signal according to the grid, extract features such as MFCCs [25, 26] or multiple low-level features [27, 18] and then classify the segments using Gaussian Mixture Models [25], k nearest neighbor classification [28], or Support Vector Machines [27, 18]. Other methods combine aspects of both categories, via adaptation [29] or joint detection of onsets and drums [30]. To ensure temporal consistency many approaches make use of high-level statistical models that encode some musical knowledge by means, for instance, of hidden Markov models [30, 31, 18, 26]. The methods greatly differ in terms of the number of instruments they are capable of detecting; most detect only bass drum, snare drum and hi-hat [30, 27].

Apart from the mentioned studies focusing on drum or other percussion instruments, we dedicate some space to those focusing particularly on the transcription of body percussion sounds. These works are more connected with the goal of this project.

Classification and transcription of body percussion sounds

In 2005, Hazan [32] approaches the transcription of voice generated percussive sounds in a similar way to previous mentioned methods. The system consists of a simple energy based onset detection system which segments the input into percussive events from which spectral and temporal descriptors are computed. Finally, a machine learning component assigns to each of the segmented sounds of the input stream a symbolic class. This approach achieves a classification accuracy of 90% in a test using performers which where not in the training set and with a taxonomy of four different drum sound classes. Different approaches are based on Autonomous Classification Engine (ACE) and incorporate one more category in the classification taxonomy [33]. A total of five voiced percussion sounds were recorded and manually segmented. ACE was used to compare various classification techniques, both with and without feature selection. The best result was 95.55% accuracy using AdaBoost with C4.5 decision tress. Continuing this path, a comparative study of human beat-box with speech [34] together with other studies regarding the mechanisms for producing beat-boxing [35], lead to the work described by Hipke [36]. The former article attains the implementation of an end-user interactive interface for recognizing beat-box sounds, enabling to control or trigger, for instance, a drum kit sample.

There exist few commercial applications that make use of finger-tapping for

providing interactivity with drums. 'Table-drum'[2] is an augmented audio application for iphone which enables the user to play drums from the mobile phone when hitting the objects that are surrounding, without the need of touching the screen. The user can train on the fly a machine learning algorithm for detecting the sounds captured through the microphone of the mobile, which are then synchronized with real drum sounds. This technology works as an audio-driven MIDI controller but does not provide further processing over the generated pattern and has no intelligent understanding about the rhythm played by the user. Moreover, is the user the one in charge of defining the association between a certain finger-tapping sound and its correspondent drum instrument that is triggered from the mobile. So, the user must have good percussion skills in order to achieve a good representation of the rhythm.

2.2 Description of rhythm

In this section we review some of the most meaningful work towards the description of rhythm. The goal is to study the feasibility of the existent state of the art techniques in this topic for providing to the pursuit system with rhythm-wise capabilities for the arrangement of drum patterns. Several proposals have been made so far for attempting the description of this musical dimension. The main differences between them are the type of data that is used to capture the rhythm information, the way of representing it or the algorithm used to compare rhythm [37].

2.2.1 Similarity Matrix-Based Approach

Self-Similarity Matrix (SSM) is a Data Analysis technique that enables to identify similar sequences within an audio data series. The main difference between the methods that follow this approach is the type of information (STFT coefficients, MFCCs, etc) and the way for computing similarity (Distance Matrix, Auto-Correlation, histograms or spectral properties comparison). Foote proposes the use of the amplitude coefficients of the short-time Fourier transform (STFT) or Mel-Frequency Cepstral Coefficients (MFCCs) of the audio signal [38]. SSM is computed using either Matrix Distance methods (Euclidean or cosine distance) and summing the values along diagonals at specific time lags or computing Auto-Correlation. Following this path, the same author accomplish the implementation of a system for retrieving rhythm based on the previous proposal [39]. Similar approach considers chroma-based MFCC features, extracted either from the whole

[2]http://www.appsafari.com/music/17007/tabledrum/

signal or from an estimated segment [40]. The resulted rhythmic signatures are compared using Dynamic time warping to compute the similarity distance. Evaluation is performed on Greek Traditional Dance and African music.

2.2.2 Feature-based Approach

Many contributions are based on feature extraction, either from a beat histogram [41, 42, 43] or from the audio content itself [44, 45, 46]. An example of the formers [41] is based on a BPM histogram obtained by collecting over time the contribution of the dominant peaks in the auto-correlation function, computed from a down-sampled and filtered version of the Discrete Wavelet Transform of the audio excerpt. Various features are derived from this histogram providing information about the inter-peak relative energy as well as the periodicity, and used in combination with other timbre and pitch content features for music genre classification. Other mentioned proposals include features derived from the Periodicity Histogram [42], and from an Inter-Onset-Interval Histogram (IOIH) tempo[47]. Recently, in [44] it is provided a set of features that capture not only amplitude, but also tonal and general spectral changes in the signal in order to obtain a complete description of rhythm. A novelty function is then applied to the computed features aiming to identify prominent changes in the temporal evolution and extract the beat histograms. Another study considers tempo estimation errors as part of the estimation process [48]. From those methods based exclusively in features obtained from the audio content, Paulus attempts to characterize rhythm by means of acoustic features that gather information about the fluctuation of loudness and brightness within the pattern [45]. Dynamic Time Warping (DTM) is then applied to align the patterns to be compared. More recently, Pikrakis [46] proposed the use of a model based in deep neural networks architecture consisting of a stack of Restricted Boltzmann Machines (RBM) on top of which lies an associative memory. This model is fed with MFCC rhythmic signatures of music recording samples.

2.2.3 Temporal Pattern Based Approach

A limitation of the feature-based approach based on beat histograms or periodicity distributions, is that these encode information about the relative frequency of various time inter-onset intervals, but discard the information about their sequence in time. Thus, many proposals have been made so far for attempting a representation of rhythm that beholds the above-mentioned criteria. One of the first contributions that addresses this issue, proposes to extract rhythmic patterns directly from the audio rather than from features [49]. This work accomplishes a temporal rhythmic pattern representation obtained from the temporal evolution of the energy inside

each bar of the audio signal. To this end, the authors describe the use of Beat-Root method [50] for finding the first bar, however, some manual corrections had to be introduced. Once the bar positions are calculated, simple segmentation of the energy envelope between start and end points is carried out. Based on this representation, features describing meter, syncopation and swing factor, are also calculated.The authors report, testing in a dance music dataset of a 50% correct recognition using only the pattern, 84% when including other automatically computed features and up to 96% when using also the correct tempo. Another outlook considers the rhythmic pattern an timing in recordings of Afro-Cuban Music, particularly focusing on Clave. Clave, apart from being a percussion instrument, is a repeated syncopated rhythmic pattern that is often explicitly played, but often only implied; it is the essence of periodicity in Cuban music [51]. Using a matched-filtering approach, they first enhance the presence of claves in the audio signal. The derived positions of the discrete onsets are then compared to a set of temporal-templates representing the theoretical positions of claves at various tempi and pattern rotations. A rotation-aware dynamic programming algorithm is then used to find the tempo, beat, and down-beat positions.

2.2.4 Periodicity Measures Approach

Periodicity representations are tempo-dependent measures that provide a measurement of the periodic content in a given rhythmic pattern. Several proposals have been made so far regarding this method for describing rhythm. Some propose the use of dynamic periodicity warping (DPW) over a frame-wise Discrete Fourier Transform (DFT), with frequency resolution sub-multiple of the tempo for independence, and then compute rhythmic similarity using different methods [52]. Similar study proposes the use of Melin Transform (MT) [53] over the Auto-correlation (AC) of the onset-function to provide a theoretically tempo-independent representation. Jensen [54] also proposes the use of the auto-correlation of the onset-function to exponentially group the lags. Each track is represented by the values of 60 exponentially spaced bands representing the lags between 0.1 s and 4 s. While this representation is robust against small tempo changes, it is not completely tempo independent. Another approach is based on the beat histogram computed from the auto-correlation of the onset function [55] . They propose the use of a logarithmically-spaced lag-axis in order to get rid of tempo changes. In order to compute it, they propose an algorithm for the estimation of a reference point. Results on two private test-sets show improvements over the usual linear-lag beat histogram for task of classification and similarity. More recent studies [37] demonstrate that the use of simple rhythm representations such as the DFT or a concatenated version of the DFT and Auto Correlation Function (ACF), allows achieving high recognition rates for a task of genre classification.

2.3 Retrieving audio content from Rhythm

Due to the increasing number of digital audio collections available, both online and in personal libraries, new interfaces for navigating through all this content have to be proposed to users, allowing the retrieval based on musical properties instead of only text. Investigation in content-based music retrieval via acoustic input, focus on the development of retrieval tools that enable users to sing or whistle an excerpt of the musical piece searched. In particular, most people do not have professional music skills and the best way to specify an intended song is to sing or hum it. As a result, this kind of systems are the most natural tools for common people interaction with music. Among such strategies, query by tapping (QBT) and query by beat-boxing are inherently related to rhythm pattern retrieval and are closely related with the endeavor of this project.

2.3.1 Query by tapping

Query by tapping is a mechanism for retrieving audio content based on finger-tapped audio recordings from the user. These techniques generally extract the note onset time from a recording, computes some kind of rhythmic fingerprint which is then compared against a song database to retrieve the correct song. Other techniques for content-based audio retrieval like the above mentioned query-by-singing/humming (QBSH) or whistling, take into consideration just the melody pitch for comparison and no rhythmical information is contemplated [56]. The first step towards QBT [57] allowed a user to clap or tap the rhythm of a requested song and, by recording it with a microphone to the computer, an offline process extracts the notes durations and compares them by Dynamic Programming to a database for retrieval. Another study [58] proposes a similar approach but in which rhythm is tapped on a MIDI keyboard or on an e-drum. The system operates in real time and online, which means that after every tap made by the user, the system presents the actual search result list. The database content is represented in an MPEG-7 compliant manner from which, for this system, only the beat descriptor is evaluated. The Descriptor Beat contains a vector of integers, describing the melodys rhythm. The vector is formed by numbering every note with the integer number of the last full beat. Similarity of two vectors is also compared through Dynamic Programming. Following this work [59], it is introduced the computation of efficient similarity measures (Direct Measure and Wring Measure), which yielded good results for comparing MPEG-7 compliant rhythms. Peters [60] implements an interactive web site where visitors can tap the rhythm of a songs melody using the space bar on their computer keyboard. A Java applet generates a MIDI file, which is sent to their application server for analysis, and the database will be searched.The MIDI file containing a monophonic sequence

of notes is analyzed to generate a rhythmic contour string. Rhythmic characterization is understood as a sequence of note durations and rests in the MIDI file; 'beat' is considered to be the time taken from the start of a note to the beginning of the next note. Normalization of the durations of the beats is done in order to eliminate any global tempo dependence when searching for matches. Then, the approximate string matching algorithm from Sun Manber (1992) is used to calculate the edit distance between the input string and each string in the database.

2.3.2 Query by Beat-boxing

Beat-boxing (also beat boxing or b-boxing) is a form of vocal percussion originated in 1980s within Hip-hop culture. Primarily involves the vocal imitation of drum machines, drums and other percussion, using mouth, lips, tongue, and voice. The term beatboxing" is sometimes used to refer to vocal percussion in general. It may also involve vocal imitation of turn-tablism, and other musical instruments such as bass-lines, melodies, and vocals, to create an illusion of polyphonic music. Beat-boxing developed outside academia and separated from other vocal styles commonly studied by universities and conservatories. Therefore, there is very few academic work on the topic. A remarkable study on Human Beat-boxing (HBB) [34] describes the acoustic properties of some sounds used in HBB compared to speech sounds based on authors observations. Also, some investigations focus on the use of HBB as a query mechanism for music information retrieval [61], the automatic classification of HBB sounds amongst kick/hi-hat/snare categories [62] as well as interactive tools for creating drums by beat-boxing [36, 63, 64, 32]. Further experiments, analyzed the repertoire of a human beat-boxer by real-time magnetic resonance imaging [35], where the articulatory phonetics involved in HBB performance were formally described. The vocal tract behavior in HBB was analyzed through fiberscopic imaging [65], to understand how they manage instrumental, rhythmic and vocal sounds at the same time. More recently, various pitch tracking and onset detection methods are compared and assessed against an annotated HBB audio dataset [66]. Moreover, Hidden Markov Models are evaluated, together with an exploration of their parameters space, for the automatic recognition of different types of vocal sounds.

2.4 Sound recommendation

The research so far in sound and, particularly, music recommendation systems has been mainly focused on its application to music web pages with large-scale audio data-bases. The goal of this technology is to help users find music content or information in accordance with their interests in a personalized manner. The

main difference between the existing approaches is the type of information considered. Common recommender approaches are mainly based on collaborative [67, 68] or item-based filtering algorithms [69] or hybrid combinations [70] with textual information enrichment [71]. We are specially interest in those based on acoustic features extracted from the audio content [72]. This approach generates high quality meta-data based on the similarity of the extracted features. Following this work, we are interested in the use of features extracted from the finger-tapping signals for providing further recommendation related with sound or even time information.

2.5 Automatic arrangement of music

In the music industry, there are many applications of music arrangement. Already in 1995, Aoki and Maruyama [73] released a patent concerning an automatic arrangement apparatus and an electronic musical instrument for performing an automatic operation while arranging notes in real time and, more particularly, to a technique for automatically generating additional notes such as contrapuntal notes, countermelody notes and the like on the basis of melody notes. Nagashima and Kawashima [74] employed chaotic neural networks to create variations on melodies. The examples of the variations of an original music object are sent to train chaotic neural networks. The networks model the characteristics of the variations and make a new variation of the original music. Berndt et al. [75] presented the strategies to synchronize and adopt the game-music with player interaction behaviors. The approach to arrange music in the context of the interaction of applications is to vary the rhythmic, harmonic, and melodic elements of the basic theme. Chung [76] proposed a real-time music-arranging system that reacts to the affective cues from a listener. The system re-assembles a set of music segments according to the inferred affective state of a listener. Based on a probabilistic state transition model, the target affective state can also be induced. Ka-Hing [77] proposes a music arrangement engine for games and interactive applications, in which orchestral music can be automatically arranged from musical materials, rules and parameters provided by composers, subjected to the emotional requirements in a game or an interactive application. Since the research on guitar fingering became mature [78, 79] presented an approach for guitar arrangement. The main concept is to choose a set of important notes by a search algorithm, with the constraint on the playability of the guitar. However, this approach is dedicated to a solo guitar and cannot arrange for various roles in music. Huang [80] proposes the use of score reduction for implementing a system for piano score arrangement.

2.6 Conclusions

We have provided a general overview of the different state of the art techniques for classification of percussion timbre, with special emphasis in human body percussion timbre. Also, we reviewed some of the work on drum transcription and rhythm description as well as other high level applications such as sound recommendation and automatic arrangement of music. Through this knowledge, our main goal is to study the feasibility of finger-tapping behavior for the development of new creative tools for interacting with digital drums. We understand that the first step to accomplish, previous to implementing any kind of rhythm arrangement, is to generate a MIDI-like symbolic representation of human finger-tapping signals. In the following Chapter, we present the undertaken methodology for attempting the study of this human behavior.

Chapter 3

METHODOLOGY

In this chapter we describe the methodology applied in our research. Efforts encompassed four main targets: obtaining the dataset, the study of the properties and context of the average music producer when composing drums, the implementation of a creative and interactive tool for the arrangement of digitized drum patterns via finger-tapping input, and the evaluation of the same.

3.1 Audio collection acquisition

A group of 47 western people were asked to record a rhythmic pattern by finger-tapping in a given surface. Recording was carried out using a "Yamaha pocketrak PR7" portable stereo recorder at 44100 Hz sampling rate. The only constraint demanded to the performer was to tap a repetitive pattern and to only use hands and fingers so as to preserve spontaneity. Also, in order preserve timbre as constant as possible between different recordings, it was provided an empty cardboard box for tapping. Audio recordings where then segmented in a fix number of phrase repetitions.

3.2 User study

The study of the user properties and context has the purpose of understanding how people finger-tap and, foremost, how do they project a conceived percussion rhythm into their hands or fingers. To this end it was submitted an online survey to the same people that recorded their finger-tapping. This survey gathered general information concerning the respondent's experience with music, the conception of the performed finger-tapping and the potential capabilities that users would demand to a hypothetical expert software for the arrangement of drums. The goal is to understand the musical context of the user, what exactly is the subject

thinking when he is finger-tapping, what he actually performs in comparison to the former, and what kind of processing would like to be done in the performed rhythm. The whole test was designed based on the Likert scale, which is answered in a five-level option of agreement or disagreement, so as to correlate the answers to different questions. For gathering this information, the survey was divided into three parts:

1. The first part formulated questions concerning the musical training of the subject, specially in the percussion domain, as well as some demographic details (Fig 3.1).

2. The second part of the survey was aimed to gather information about the subject's conceived rhythm and the way he actually tapped it (Fig 3.2). The formulated sentences are:

    - I used my hands for tapping the rhythm
    - I used my fingers for tapping the rhythm
    - I was thinking on a percussive instrument when tapping
    - I was particularly thinking on a real drum instrument when tapping
    - There are overlapping voices/strokes in my tapping
    - How many instrument voices does your rhythm encode
    - There was a straight match between the hand or finger used for each stroke and the voice that it represented
    - I took advantage of the timbral possibilities of the box to encode different percussive voices within the rhythm
    - In general, I used my palm or thumb to play the role of a kick or other bass percussion instrument
    - In general I used my fingers to play the role of bright percussion instruments such as a snare or hats

3. The last part was dedicated to collect information about the potential capabilities that users would demand to a hypothetical drum-expert software for the arrangement of percussion. The purpose is to understand the real problems that average users have to overcome when producing rhythm in the computer 3.3. In addition, it was provided a free section for suggesting new capabilities. The formulated sentences in this part are:

    - I would like to be able to convert my tapped rhythm into a real drum sound pattern.

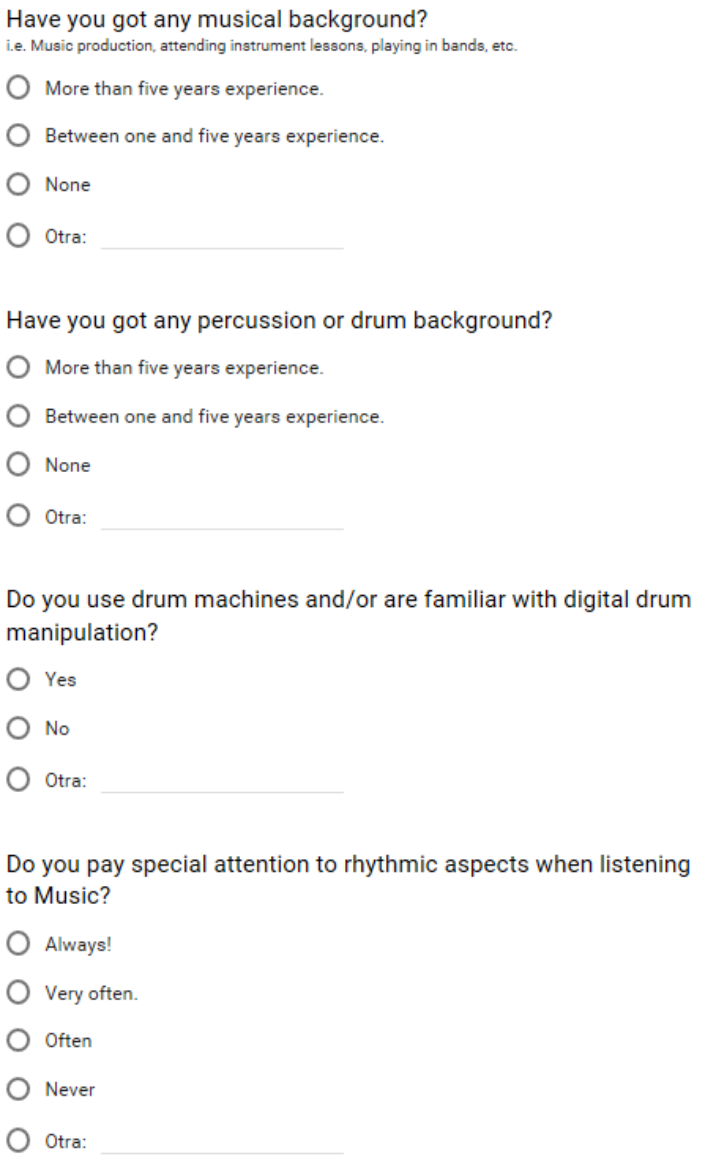

Figure 3.1: First part of the survey.

- I would like the software to recommend drum or percussive sounds for each of the voices in my rhythm.
- I would like the software to automatically enrich the rhythm by adding new drum lines to the basic rhythm I played.
- I would like the software to improve my rhythm by correcting events out of tempo or other clear mistakes.

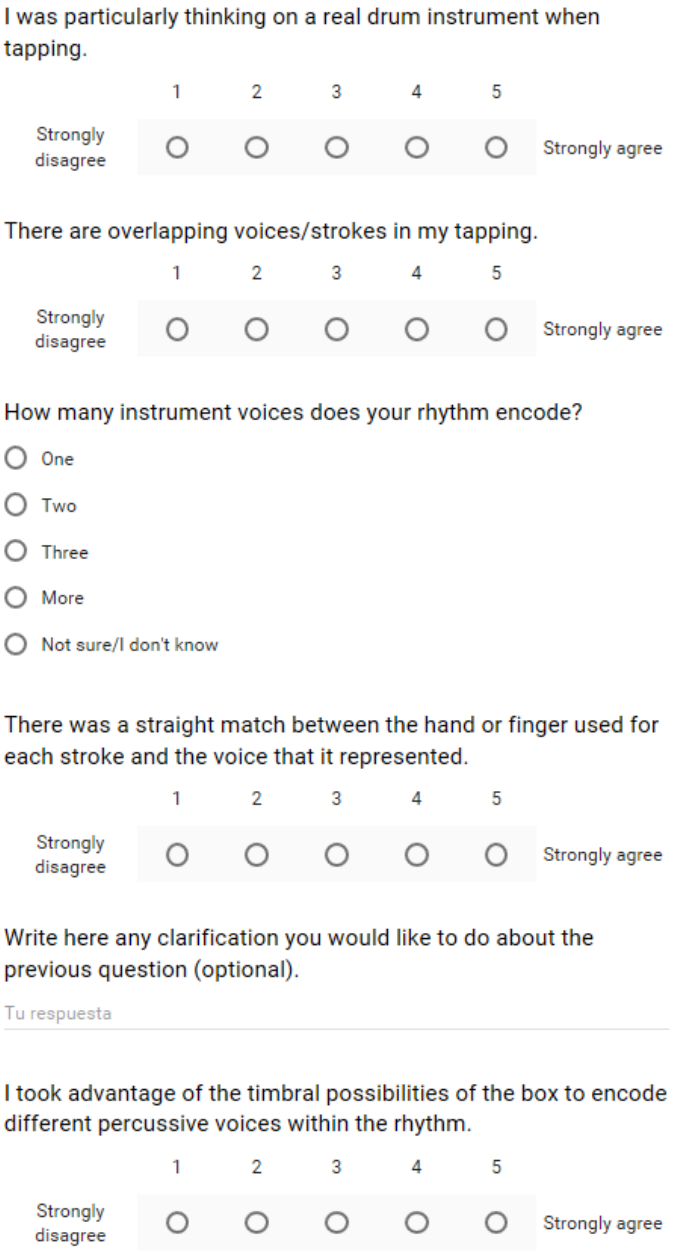

Figure 3.2: Fragment of the second part of the survey.

- I would like the software to recommend new patterns as variations of the rhythm I played.
- I would like the software to recommend an entirely new drum pattern approaching the style I played.

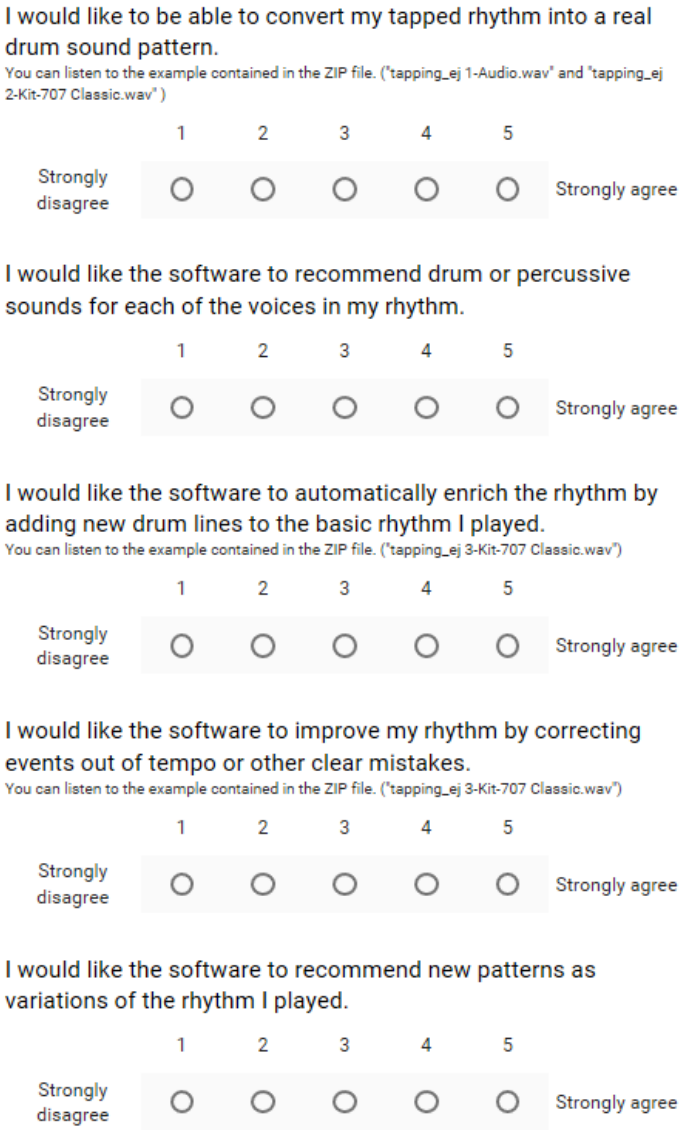

Figure 3.3: Fragment of the third part of the survey.

This survey, together with subsequent analysis of the results, provides the baseline for a solid and substantiated implementation of the proposed algorithm[1].

[1]The full survey can be found in: https://docs.google.com/forms/d/e/1FAIpQLSd8ns0-BIAuHQdfTsoq7aB2HKV5NN5yBhXjBDifhBeEfvsIzw/viewform

3.3 Algorithm development

Analysis of the collected information and the finger-tapping recordings in conjunction with the above described state of the art in the topic, enables to set the basis for the design of the proposed algorithm. The purpose is to determine which information needs to be considered from the part of the user and its relevance for the appropriate functioning of the software.

The proposed method is illustrated in Figure 3.4. It can broadly be divided into three parts: an onset detection and feature extraction step, in which a set of frame-wise features are calculated and fed into a clustering/classification step, in which the feature representation of the finger-tapped onsets are clustered, assigned to a drum sound class and then translated into a drum MIDI-like symbolic representation. For the sake of this study, we assume that finger-tapped recordings start at the beginning of a beat.

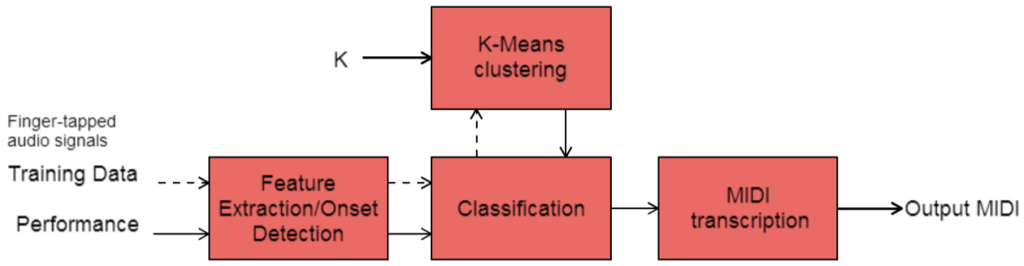

Figure 3.4: Algorithm block diagram

3.3.1 Onset detection and feature extraction

For each of the audio files in the finger-tapping collection, the first step is to detect onsets. This onsets, in the best of the cases, correspond to the times at which a stroke is detected in the audio signal. We used Super Flux algorithm implemented in Essentia Standard library[2]. This algorithm is based on spectral flux feature, which provides a measurement of fast energy changes in the spectrum of an audio signal. The next step is to extract descriptors from each onset. Following Herrera et al. [12], we choose, RMS, spectral centroid, Mel-frequency cepstral coefficients (MFCCs) as basis features for our experiments. Apart from the mentioned features, other descriptors have been considered for experimentation: the energy effective duration of the onset, the Spectral Centroid, 21 Bark-bands, spectral spread, flux and kurtosis. We also considered interesting to include some time

[2]http://essentia.upf.edu/

dependent features. A simple way of doing this is by considering, for each onset, the relative energy with respect its previous and next onset within the pattern. We also applied a normalization and quantization step into a set of ten values ranging from 0 to 1[3]. Features are extracted from audio sampled at 44.1 kHz with a frame size of 2048 samples (46ms) and a hop size of 1024 samples (23ms) using Essentia.

- Spectral Centroid Average: Average of the spectral centroid for the whole signal
- Spectral Centroid Variance: Variance of the spectral centroid for the whole signal
- Energy effective duration: effective duration of an envelope signal. The effective duration is a measure of the time the signal is perceptually meaningful. This is approximated by the time the envelope is above or equal to a given threshold and is above the -90db noise floor.
- Quantized relative energy of previous onset
- Quantized relative energy of following onset
- Mel-Frequency Cepstrum Coefficients: MFCCs are the coefficients of the Mel-cepstrum. They can be used as a compact representation of the spectral envelope.

3.3.2 Onset clustering and classification

The next step before transcriptions is to group together types of onsets. As we saw , this fact is in many cases user dependent. For this reason, we considered that K, the number of clusters, may be an input parameter provided by the user. Because of this and the low dimensionality of the features data we found simple and feasible to use k-means algorithm for clustering the data, provided in the sklearn.cluster.KMeans package of the Python machine learning library, scikit-learn. The classification is done by simply assigning the class corresponding to the nearest cluster centroid, this is I-Nearest neighbor classifier. Many contributions in drum sound classification have proven to perform accurately using this algorithm. The next step is to associate each stroke class to a MIDI event, this is the transcription of the sequence. Following General MIDI percussion Key Map[4]

[3]this kind of temporal modeling approaches have no sense with continuous magnitudes. Therefore we need to discretize the energy values.

[4]http://www.onicos.com/staff/iz/formats/midi-drummap.html

we used keys 35, 40 and 42 for the bass drums, the snare and closed Hi-hat respectively. The effective duration, and the overall energy of the onset are used to compute time and velocity of the MIDI event.

3.4 Evaluation

Given that our experiment depends depends in large extent to perceptual aspects, it is necessary the collaboration of a group of subjects for evaluating the performance of the implemented algorithm. To this end it is submitted an online questionnaire. This questionnaire, in first instance, provides to the respondent a recording of a specific finger-tapped example. For this recording, a set of four different transcriptions are provided, from which one is the output of the implemented system and the rest manual transcriptions of the same. The respondent has to rate the degree of subjective accuracy of each of the transcriptions following Likert scale.

Chapter 4

RESULTS AND DISCUSSION

4.1 User study

As it was described in 3.2, we carried out a study on the behavior of finger tapping by recording a group of western people with different musical backgrounds and submitting each participant an online survey inquiring about their performance. The main goal of this experiment was to study the survey answers and spectro-temporal features obtained from the audio dataset and infer meaningful information about finger-tapping behavior. The collected dataset contained 47 sounds. Further analysis over the recordings allowed comparison of the audio content with the provided answers. Lastly, we discuss its feasibility for improving human-computer interaction in the context of computer music by attempting the implementation of a finger-tapping transcription system. In the following pages we present the results of this experiment.

4.1.1 Survey results

Following we analyze the results of the different parts in the survey.

Analysis of background questions

The first part of the survey aimed to gather demographic information about the subjects. Table 4.1 contains the percentage of answers to questions 1 through 4 (Q1 to Q4). This information enabled to group individuals according to different aspects, such as their music knowledge or their attention to rhythmic aspects when listening to music, as well as to deduct useful information and establish patterns within the different subgroups.

In Q1: "Have you got any musical background?, over 65% of the subjects ensured to have taken formal musical training for more than five years, which

Table 4.1: Demographic questions and percentages

Questions	Answers	Percentage
Q1: Have you got any musical background?	>5 years	65,2%
	Between 1 and 5 years	0%
	None	34,8%
Q2: Have you got any percussion or drum background?	>5 years	21,7%
	Between 1 and 5 years	8,7%
	None	56,5%
	Other	13,0%
Q3: Do you use drum machines and/or are familiar with digital drum manipulation?	Yes	43,5%
	None	52,2%
	Not sure/not know	4,3%
Q4: Do you pay special attention to rhythmic aspects when listening to Music?	Always	30,4%
	Very often	34,8%
	Often	26,1%
	Never	8,7%

suggests that they are familiar with terms such as timbre, pattern, loop, instrument voice, etc. From the total polled, around 30% were at least one-year experienced in percussion and 13% claim to have taken some other kind of percussion training (minor education or self taught skills), as shown in answers to Q2. In Q3: "Do you use drum machines and/or are familiar with digital drum manipulation? we considered other kinds of percussion experience and, as we can see, 43.5% admited to have used drum machines or manipulated digital drums. In Q4: "Do you pay special attention to rhythmic aspects when listening to Music?" it is revealed that over 65% of the respondents pay attention to rhythmic aspects when listening to music.

Analysis of self-appraisal of tapping behavior

The second part of the survey covered questions 5 to 14 (Q5 to Q14). The demographic information presented in Table 4.1 allowed to group individuals in terms of their experience in percussion instruments, digitized drum and/or drum machine manipulation. From the 24 respondents, 14 turned out to be non-experienced in percussion (NEP), regardless of their music experience in other kinds of instrument, and 10 belonged to the group of experienced in percussion (EP). Considering this group subdivision, we can see in Figure 4.1 that EP clearly

tend to use equally fingers and hands (50%) or mainly their fingers (30%) when tapping, while NEP use mainly, or to a large extent, their hands (50%). Figure 4.2 shows the number of different voices that are encoded in the finger-tapped patterns. The NEP appear not be able to play more than two voices since they mainly use their hands for tapping. Surprisingly, despite using a more complex technique, in Q6 EP subjects claimed to have only played up to three different voices in the pattern. As depicted in 4.3, there is a common trend among EP subjects to use the timbre characteristics of the object being tapped (a simple cardboard box in this particular experiment) to reflect different percussion voices. Conversely, NEP users do not take into account this property of the surface to provide timbre differences to each stroke.

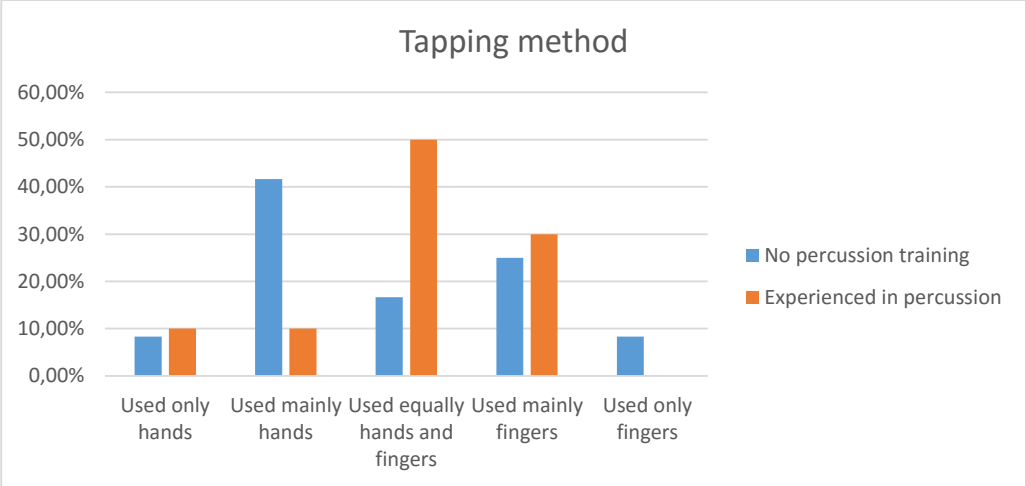

Figure 4.1: Q5: Finger-tapping method. Derived from the answers to the first two sentences of the survey.

As depicted in Figure 4.4, most of EP subjects claimed that there is a match between groups of similar timbre strokes and a particular voice. On the contrary, NEP respondents were neutral (more than 30%) or in disagreement with this statement (more than 50%). Following these results, Figure 4.5 shows that there is a trend among EP subjects to associate strokes played with particular parts of the hand with a certain type of percussion instrument. For instance, we can see in the figure that strokes played with fingers are more commonly associated with high pitch percussion instruments among EP subjects. This fact, in addition to the eyewitness of the author, suggests that EP individuals take advantage of fingers' ergonomics to perform fast and successive strokes as in a snare-roll. In Q10, half of EP respondents seem to be in agreement with the role of the thumb or the palm as a low-pitch percussion instrument and the other half disagree. NEP subgroup

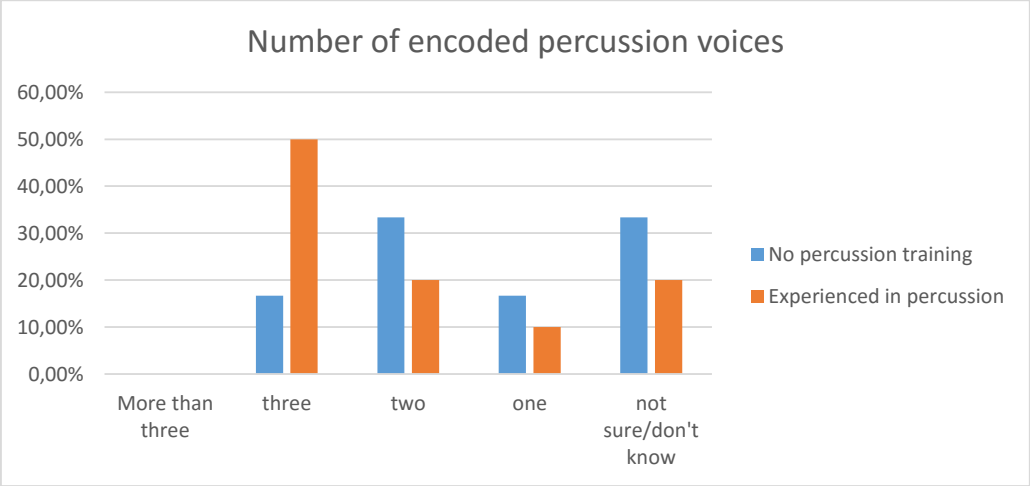

Figure 4.2: Q6: Number of encoded voices in the finger-tapped pattern.

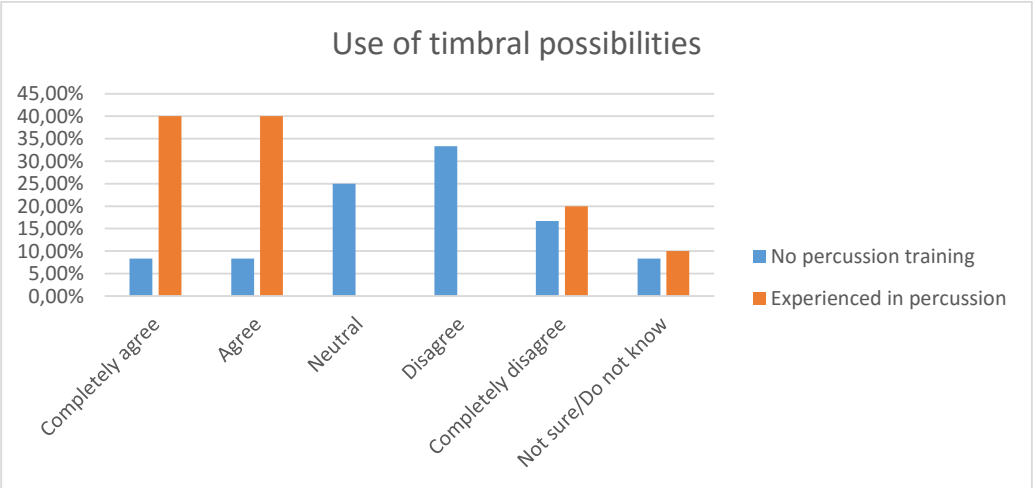

Figure 4.3: Q7: Percentage of responses to the sentence “I took advantage of the timbral possibilities of the box to encode different percussive voices within the rhythm”.

appears to be in disagreement with statement Q10. Figure 4.6 suggest that EP subjects are inspired by general percussion instruments rather than drums when conceiving the finger-tapped pattern. Similarly, around 50% of NEP respondents agree to statement Q11 but rejects Q12. In Figure 4.7 we can see that neither EP nor NEP subgroups tend to overlap strokes when finger-tapping.

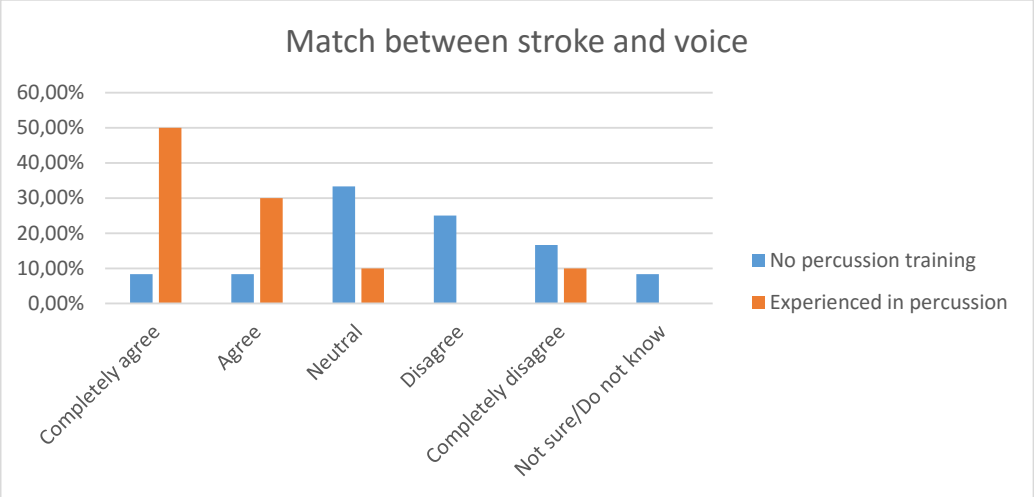

Figure 4.4: Q8: Percentage of responses to the sentence "There was a straight match between the hand or finger used for each stroke and the voice that it represented".

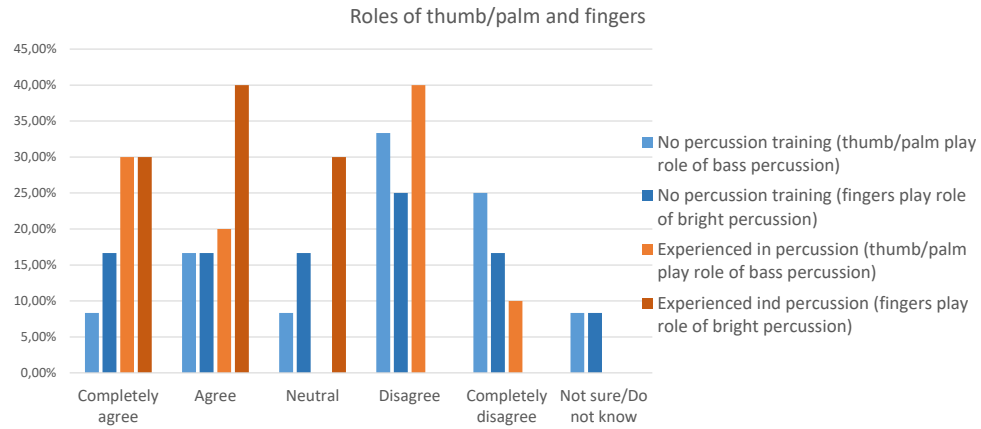

Figure 4.5: Q9: "In general, I used my palm or thumb to play the role of a kick or other bass percussion instrument" and Q10: "In general I used my fingers to play the role of bright percussion instruments such as a snare or hats"'.

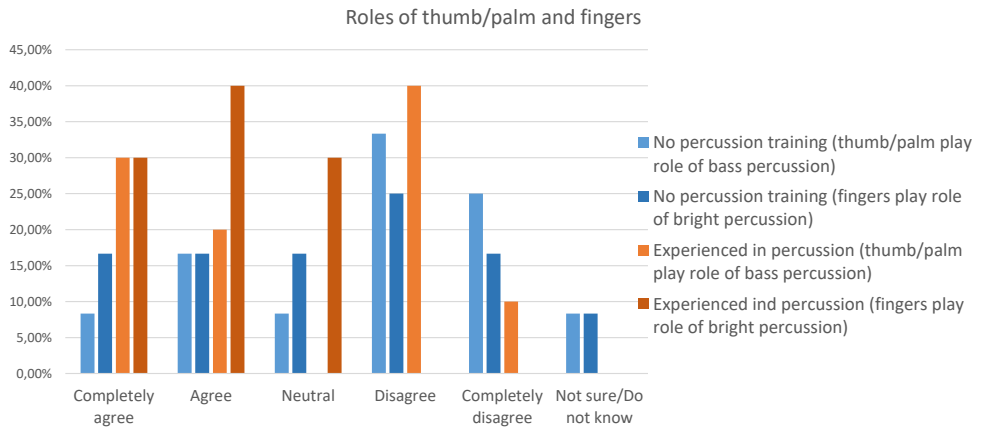

Figure 4.6: Q11: "I was thinking on a percussive instrument when tapping" and Q12: "I was particularly thinking on a real drum instrument when tapping".

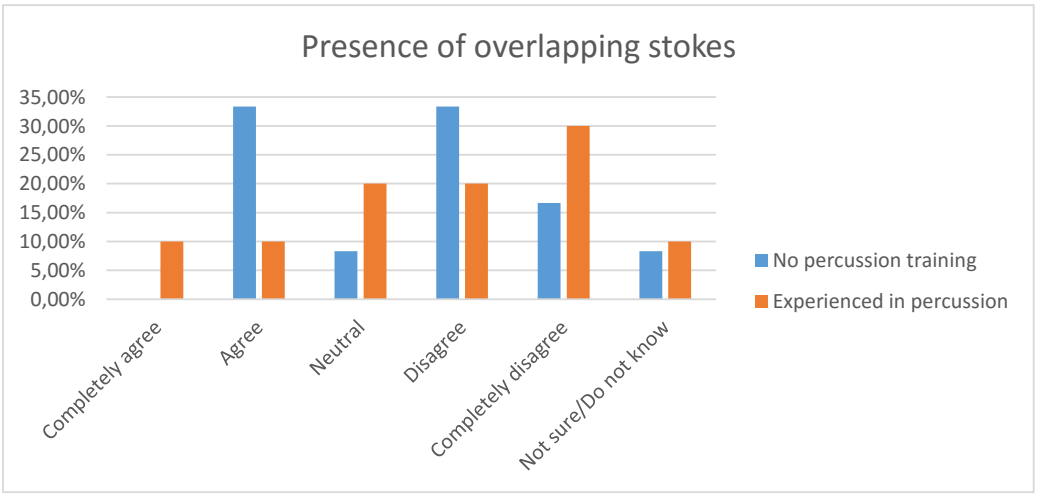

Figure 4.7: Q13: "There are overlapping voices/strokes in my tapping".

Table 4.3 shows the average, median, mode and standard deviation of the the survey responses. Some of the results appear to be redundant with others previously discussed, thus we will focus on some of the most relevant ones. Regarding the way of tapping, we can see that the average and median are equal to 3 and the deviation is 0.9 among EP subjects. A value of 3 for this particular question corresponds to an approximately equal use of hands and fingers for finger-tapping. The standard deviation is considerably low, which suggests a consensus among the EP subgroup. In the case of NEP subgroup, the standard deviation is slightly higher; the mode raises to 4 and the average lowers down to 2.7. This reveals that these individuals, despite mostly using their hands, may also use fingers or even only their fingers for finger-tapping, but rarely both[1]. In Q6, we can see that NEP subjects are able to play between one and two different voices. EP subgroup differentiated up to three, as previously discussed. In some cases they played two or even less voices since the deviation is 1.1 and the average number of voices played is 2. Due to the small amount of participants in the survey, it is risky to make a generalization of the previous result, but we can expect that a system for finger-tapping transcription will not require to discriminate between more than three voices. Further results show that EP subjects clearly finger-tap using the timbre characteristics of the surface, as opposed to NEP subjects. Regarding Q8, statistics suggest that, in general, neither NEP nor EP subjects tend to overlap strokes. This is an interesting result considering that many percussion instruments, such as a drum set, allow overlapping strokes. This suggests that EP subjects conceive the provided surface (the box) as if it were a single membrane instrument. With regard to the role of the palm and/or the thumb and fingers, EP subjects understand the role of fingers as a high-frequency percussion instrument and disagree, or are neutral, with respect to the role of the thumb or palm as a low-frequency percussion instrument; NEP subjects disagree with both facts. The way in which the pattern is conceived in the mind of the participant appears to be more complex: most of NEP individuals agree or are neutral to think of a percussion instrument different from drums when performing the finger-tapping. A similar trend applies to the results for EP subjects. This suggests that participants understand the provided box as an instrument itself and, rather than emulate a typical drum pattern, conceive the finger-tapping as if they were playing a general percussion instrument such as a djembe[2].

Table 4.2 shows the mean and variance values from tempo estimations manually computed using a regular Digital Audio Workstation (DAW). Results reveal

[1] In fact, during the recordings, it was evident that most of the people that had not taken any percussion training either used their hands (sometimes even only one) or some of their fingers (mainly thumb, middle and index fingers), but never both.

[2] However, the recordings show that participants (even EP), in many cases performed a typical drum pattern like the one notated in Figure 4.8

Table 4.2: Statistics of tempo measurements

Statistics of tempo annotations	No percussion experience		Experienced in percussion	
	Av.	Dev.	Av.	Dev.
BPM	109,5	14,3	110,0	15,0

a clear pattern in the average and variance values of the BPM for both subgroups, reinforcing previous research on human preferred tempo [81]. We can see that most of the estimated tempos range from 95 to 125 bpm.

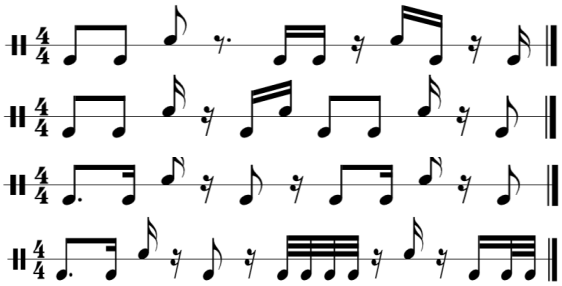

Figure 4.8: Basic notation of the most typical rhythmic pattern that can be listened in the recordings.

Table 4.3: Average, mode and standard deviation of the answers to the second part of the survey.

	Average		Mode		Deviation	
	NEP	EP	NEP	EP	NEP	EP
Q5: Method for finger-tapping	2.7	3.0	4.0	3.0	1.1	0.9
Q6: Number of encoded voices	1.3	2.0	2.0	3.0	1.1	1.2
Q7: Use of timbre	2.5	3.8	2.0	4.0	1.2	1.5
Q8: Match between stroke and voice	2.6	4.1	3.0	5.0	1.1	1.2
Q9: palm/thumb role of bass percussion	2.5	3.2	2.0	2.0	1.3	1.5
Q10: fingers role of bright percussion	2.3	4.0	2.0	4.0	1.4	0.8
Q11: Thinking on percussion instrument	2.9	4.1	4.0	1.0	1.3	1.5
Q12: thinking particularly on drums	2.4	2.7	1.0	1.0	1.3	1.5
Q 13: Escistance of overlapping strokes	2.6	2.4	2.0	1.0	1.1	1.3

1 Completely disagree (* Only fingers)
2 Disagree (* Mainly fingers)
3 Neutral (* Equally fingers and hands)
4 Agree (* Mainly hands)
5 Completely agree (* Only hands)
** values indicate the number of voices

Table 4.4: Cross-correlation of the answers provided by percussion experienced subjects

	Thinking on percussion instrument	Thinking particularly on drums	Existance of overlapping strokes	Number of encoded voices	Match between voice and stroke	Use of timbral possibilities	Palm/thumb role of bass percussion sound	Fingers role of bright percussion sound
Finger-tapping method	0.08	0.15	-0.23	-0.09	0.27	-0.15	-0.38	0.00
Thinking on percussion instrument	-	-0.13	0.47	-0.12	-0.36	0.16	0.33	0.00
Thinking particularly on drums	-	-	-0.30	-0.28	0.29	**0.66**	-0.11	0.43
Existance of overlapping strokes	-	-	-	0.41	0.16	0.02	**0.76**	0.27
Number of encoded voices	-	-	-	-	0.35	0.17	0.35	-0.11
Match between voice and stroke	-	-	-	-	-	0.46	0.38	0.53
Use of timbral possibilities	-	-	-	-	-	-	0.48	0.44
Palm/thumb role of bass percussion sound	-	-	-	-	-	-	-	**0.61**

Table 4.5: Cross-correlation of the answers provided by non-percussion experienced subjects

	Thinking on percussion instrument	Thinking particularly on drums	Existance of overlapping strokes	Number of encoded voices	Match between voice and stroke	Use of timbral possibilities	Palm/thumb role of bass percussion sound	Fingers role of bright percussion sound
Finger-tapping method	-0.32	0.01	-0.08	0.02	**-0.65**	0.10	-0.15	-0.22
Thinking on percussion instrument	-	**0.64**	0.06	-0.44	0.13	0.71	-0.11	0.19
Thinking particularly on drums	-	-	-0.17	-0.55	-0.49	**0.63**	-0.07	0.07
Existance of overlapping strokes	-	-	-	0.28	0.08	0.31	0.12	0.24
Number of encoded voices	-	-	-	-	0.28	-0.05	0.44	0.09
Match between voice and stroke	-	-	-	-	-	-0.19	0.23	0.32
Use of timbral possibilities	-	-	-	-	-	-	0.14	0.09
Palm/thumb role of bass percussion sound	-	-	-	-	-	-	-	**0.63**

Tables 4.4 and 4.5 show the cross-correlation coefficients of the answers provided by both subgroups to each of the questions. This measurement enables to interrelate results, as well as to find behavioral patterns within both types of subjects. We can see, for instance, a moderate correlation (0.66 and 0.63) between the use of the timbre characteristics of the box to play different voices and the fact of thinking particularly on drums among both subgroups. This suggests that those individuals that are inspired by drums when finger-tapping take advantage of the timbre characteristics of the surface to play different voices. On the other hand, those that are not inspired by drums do not consider timbre. In the case of NEP subjects, we find a slight negative correlation between the type of technique used for finger-tapping and the existence of overlapping strokes (-0.65). This indicates that those NEP subjects that mainly use their hands for finger-tapping tend to perform less overlapping strokes than those that mainly use their fingers. Further results show that those NEP subjects that thought about a percussion instrument tend to think particularly in a drum instrument (0.64), as opposed to EP subjects (-0.13). Also, we can see negative correlations between the number of encoded voices in the pattern and the fact of thinking in a percussion and a drum instrument among NEP subgroup (-0.44 and -0.55 respectively). This raises that those NEP subjects that are inspired by these kind of instruments tend to express less instrument layers in the pattern in comparison with those that simply finger-tap without thinking on any particular instrument. Notice the moderately negative correlation (-0.65) between the answers to Q8 and Q5 among NEP subjects. This means that those subjects that mainly finger-tap with their hands do not tend to match each type of different stroke with a particular instrument voice. Conversely, those NEP subjects that mainly use their fingers tend to match each different stroke with a particular voice. Further interrelations among EP subjects show that, for instance, those that take advantage of the timbre of the surface tend to match different strokes with different voices as well as using their thumb/palm and fingers to play bass and bright percussion instruments respectively. Furthermore, those in both subgroups that associate the use of the palm and/or the thumb to play the role of a bass percussion instrument think as well that fingers play the role of a bright percussion instrument, such as snares or high-hats.

Analysis of user needs and potential functionality

The last part of the survey aimed to gather potential capabilities that respondents would expect from a hypothetical software for creating and arranging drum patterns. The goal was to ensure the validity and feasibility of our current ideas and encourage the respondents to provide new ones.

Figures 4.9 through 4.14 show the percentages of answers to Q14 through Q17. In general, we can see that both EP and NEP subgroups appear to be in

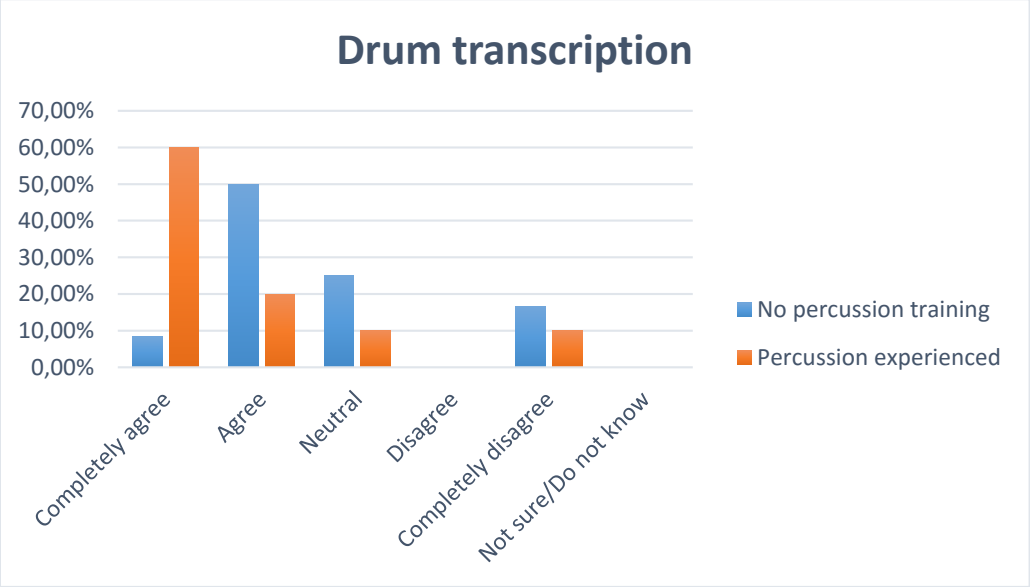

Figure 4.9: Percentage of responses to the sentence “I would like to be able to convert my tapped rhythm into a real drum sound pattern”.

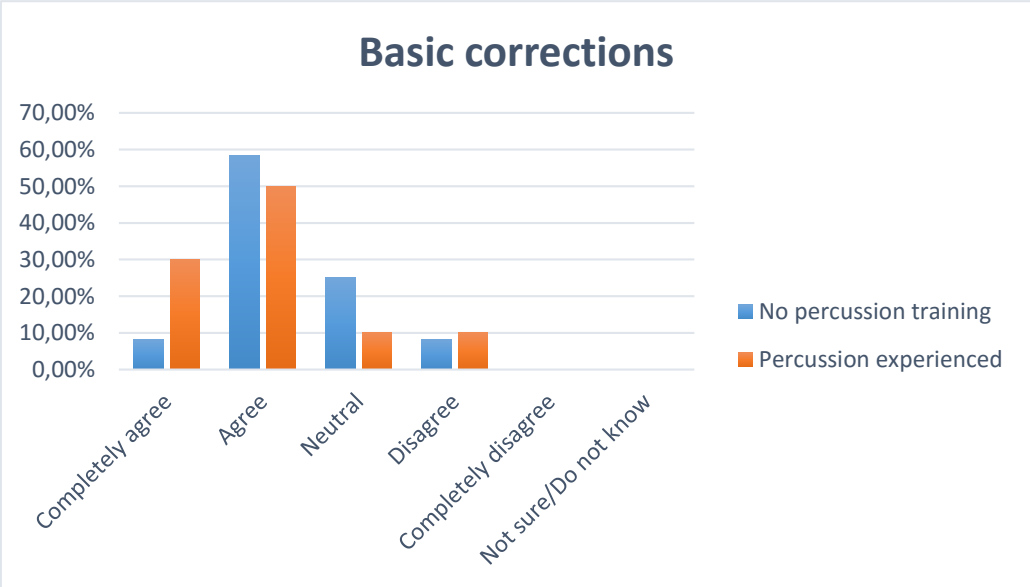

Figure 4.10: Percentage of responses to the sentence “I would like the software to improve my rhythm by correcting events out of tempo or other clear mistakes”.

agreement with all the proposals. Nevertheless, as depicted in Table 4.6, the least rated option among EP are the enrichment of the pattern by adding new instrument layers, while the NEP do not consider necessary the recommendation of drum sounds. This suggests that EP do not find so useful to enrich the number of percussion layers in a transcribed version of the finger-tapped pattern. Also, NEP appear to be less concerned about the timbre aspects in the pattern (as it was already noticed in previously discussed results). In addition, many respondents took the time to suggest valuable ideas (but out of the scope of this project) such as to enable the use of the surface as a simple controller, following the idea of the Tabledrum mentioned in 2.1.3 and also to enable the variation of the measure or even apply the groove of a given drummer.

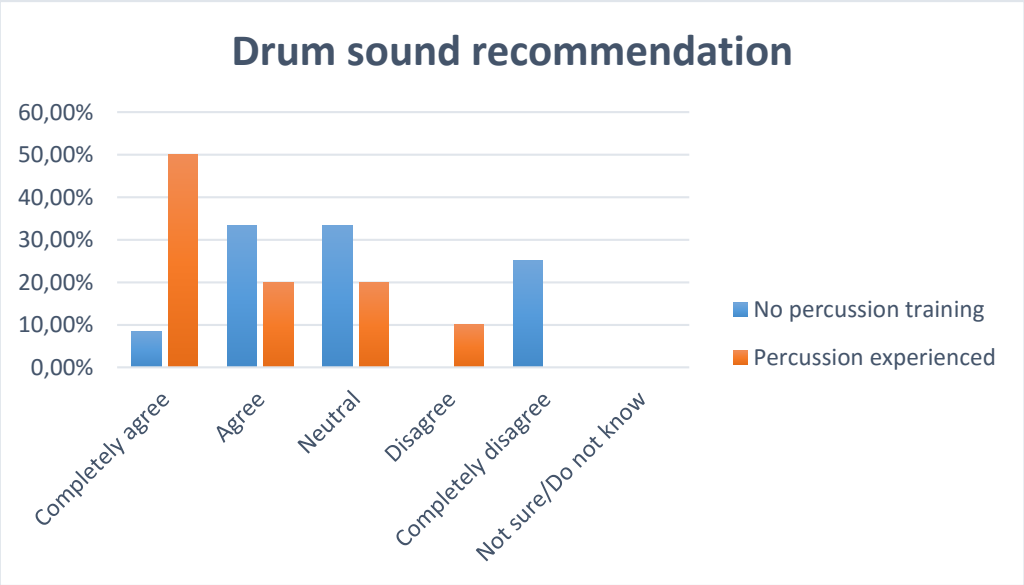

Figure 4.11: Q14: "I would like the software to recommend drum or percussive sounds for each of the voices in my rhythm".

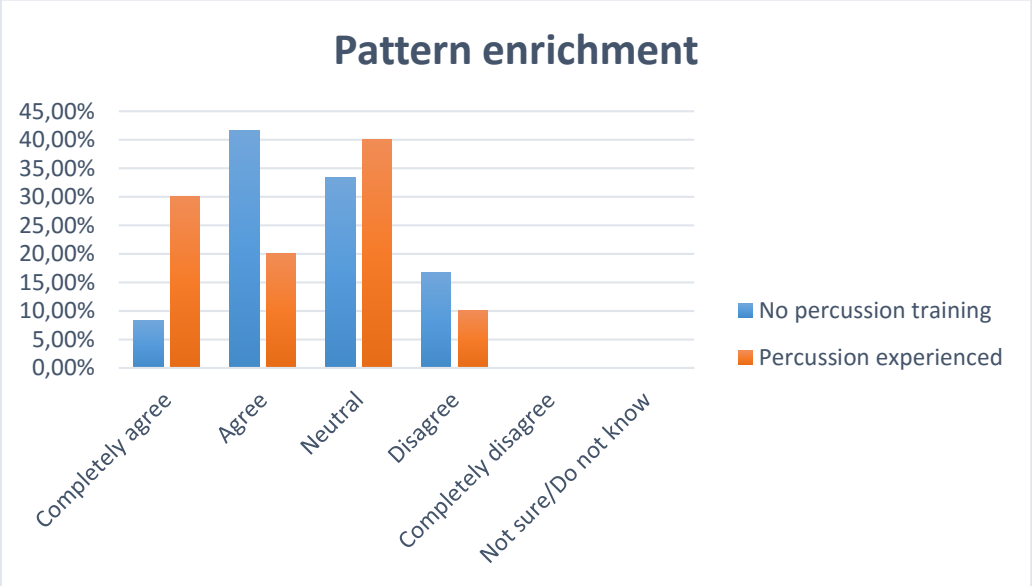

Figure 4.12: Q15: "I would like the software to automatically enrich the rhythm by adding new drum lines to the basic rhythm I played".

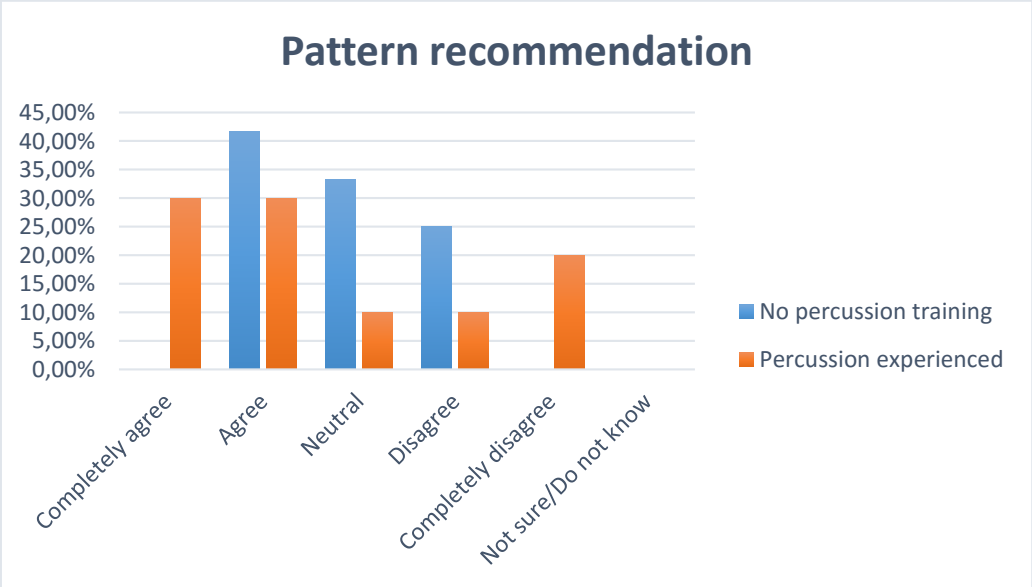

Figure 4.13: Q16: "I would like the software to recommend new patterns as variations of the rhythm I played".

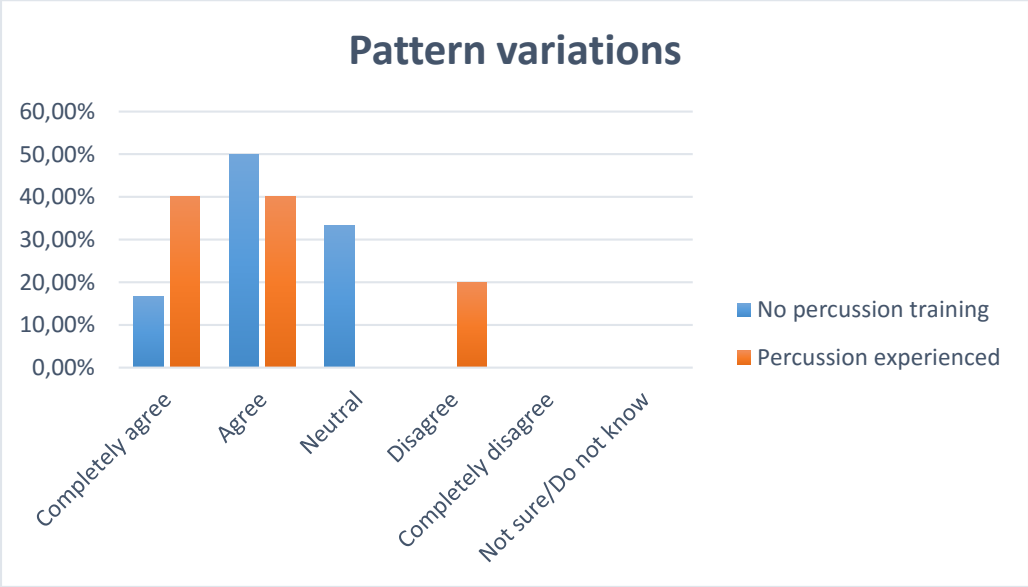

Figure 4.14: Q17: “I would like the software to recommend an entirely new drum pattern approaching the style I played”.

Table 4.6: Statistical measurements of the preferences of respondents regarding the capabilities of a hypothetical software for the arrangement of drum patterns

	Average		Mode		St. Deviation	
Respondent profile	NEP	EP	NEP	EP	NEP	EP
Q14: Drum transcription	3.3	4.2	4.0	5.0	1.2	1.2
Q15: Drum sound recommendation	3.0	4.1	3.0	5.0	1.6	1.0
Q16: Enrich rhythm	3.4	3.7	4.0	3.0	0.9	1.0
Q17: Basic corrections	3.7	4.0	4.0	4.0	0.7	0.9
Q18: Recommend variations	3.8	4.0	4.0	4.0	0.7	1.1
Q19: Recommend new patterns	3.2	3.4	4.0	5.0	0.8	1.5

1 Completely disagree (* Only fingers)
2 Disagree (* Mainly fingers)
3 Neutral (* Equally fingers and hands)
4 Agree (* Mainly hands)
5 Completely agree (* Only hands)
** values indicate the number of voices

In conclusion, the results so far suggest the existence of two general strategies for human finger-tapping: the NEP and EP profiles. The former subgroup is characterized by mainly using the hands for tapping, which constraints to two the number of percussion voices that they are able to play in the pattern. Also, they appear not to be concerned about timbre aspects when it comes to hit different spots of the given surface and that way confer timbre differences to each stroke. Moreover, they do not think of each different stroke as a different voice but actually do play different voices, as evidenced from the recordings. In general, NEP do not think of a drum or a percussion instrument for conceiving the pattern. The latter profile is characterized by using equally hands and fingers for finger-tapping. In spite of using a more elaborated technique, EP subgroup states not to play more than three voices in the pattern. Moreover, they claim to take advantage of timbre to provide differences to each stroke as well as to associate different strokes with a particular voice. Even though they do not think of a drum to conceive the rhythm, they may think of a general percussion instrument. We can conclude that there are two considerably different ways of finger-tapping depending on the musical training of the performer. These differences are not only evident in the method used for tapping but can also be perceived by hearing. This suggests that a potential software for transcribing finger-tapping could have different implementation approaches depending on the type of final user. Particularly in the case of timbre, we conclude that the system should not be so sensitive to this music aspect. Not only because of the gathered results on this matter, but also because we expect that timbre will depend mostly on the characteristics of the object being tapped, rather than the tapping technique. This issue and others related with the acoustic properties of finger-tapping are better understood by studying the audio content, as follows.

4.1.2 Audio content analysis

This Section presents the results from the analysis of the finger-tapping recordings. The goal is to find relationships between the audio content and previously discussed results, as well as to infer useful information for the implementation of the desired system.

Previous discussion on the survey answers revealed the existence of two general finger-tapping profiles: one that includes people with no experience at all in percussion and another composed of people with at least two years of experience in some percussion facet. Some of the mentioned differences between both groups can also be seen in low-level features derived from the audio content. In the previous section it was stated that EP subjects used the timbre characteristics of the surface to confer different sound to each stroke, contrary to NEP participants. Conversely, NEP subjects use loudness variations to provide different

sound to each stroke, as it can be perceived by listening the audio. Thus, our hypothetical software should not rely so much on timbre variations to discriminate between different voices. Figures 4.15 and 4.16 show the temporal evolution of the spectral centroid, spectral kurtosis, spectral flux and RMS for an excerpt of a finger-tapping recording of a NEP and an EP subject respectively. The audio features were computed using MIR.EDU open source vamp plug-in in Sonic Visualizer (Hanning window with size equal to 2048 samples and hop-size of 1024 samples). We can see that RMS clearly identifies the audio onsets while spectral features appear to be more noisy for both excerpts. It can be derived from the evolution of the kurtosis and spectral centroid that these magnitudes are negatively correlated. In general, during the attacks of the finger-taps, the spectral centroid is high, while kurtosis remains low. This means that the spectrum during the attack has higher frequency content and tends to flatten while, during the release, high frequency content weakens and the energy spectrum tends to concentrate symmetrically around a fundamental frequency, so the kurtosis grows. However, we can see some strange inter-onset behavior for both magnitudes which could be attributed to noise. Spectral flux looks more reliable for the task of onset identification, although it still appears to be slightly noisy in some parts. Below we compare these measurements with those derived from real drums.

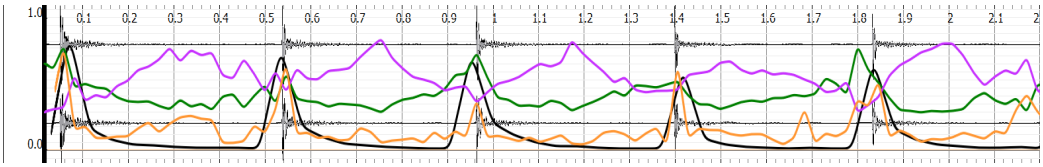

Figure 4.15: Example of NEP Finger-tapping recording. Green: evolution of the spectral centroid; purple: spectral kurtosis; orange: spectral flux; black: evolution of the RMS.

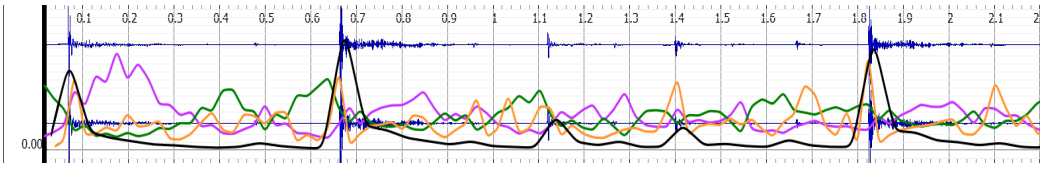

Figure 4.16: Example of EP Finger-tapping recording. Green: evolution of the spectral centroid; purple: spectral kurtosis; orange: spectral flux; black: evolution of the RMS.

Figure 4.17 shows the same measurements as in Figures 4.15 and 4.16 but in the case of real drum loops. We can see that RMS has a similar behavior to previous examples, while spectral features clearly have a smoother shape. This

small example, together with a precise and repeated listening of the recordings, suggest that regular spectral features which have been used in previous research on the classification of percussion instruments will not be so accurate in the task of finger-tapping transcription. Deeper analysis over the computed descriptors and classification of onsets, in terms of the performer (by EP or NEP), will provide a better insight of the characteristics of finger-tapping.

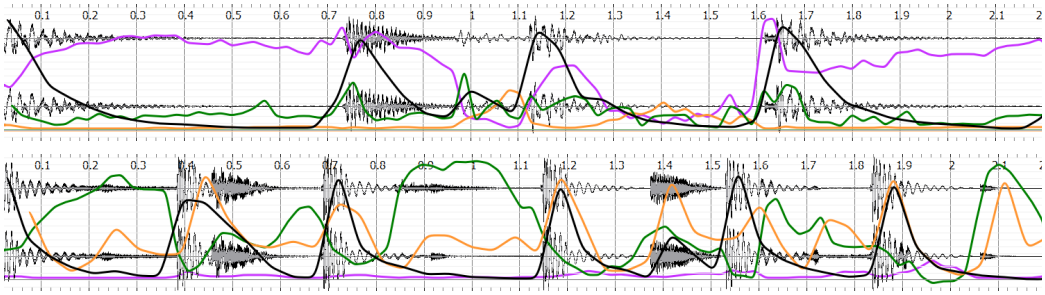

Figure 4.17: Excerpts of a drum loop. Green: evolution of the spectral centroid; purple: spectral kurtosis; orange: spectral flux; black: evolution of the RMS.

Table 4.7 compares the inter-onset mean and variance measurements of the computed descriptors for both EP and NEP subgroups. Features were computed with Essentia library, using a Hanning window with a length of 2048 samples and a hop-size of 1024. Notice the differences between both subgroups in the variance of the Bark-bands kurtosis (5.111E04 in EP against 1.337E02 in NEP subjects), the Bark-bands skewness, the spectral roll-off and the inharmonicity. In contrast, the spectral centroid appears to be distributed in a similar range of values for both EP and NEP subgroups (in the range of 2200-2300 Hz) as depicted in Figure 4.18. We can expect from this results that, except for spectral centroid, the above mentioned descriptors contain information to be considered in a finger-tapping classification task as the one presented in the following pages.

4.1.3 Finger-tapping method classification

A database containing 343 finger-tapped strokes from 14 recordings was used in this study. Even though we recorded a total of 47 finger-tapping excerpts, many had to be discarded due to the level of noise or because the performer did not answer the test, critical for the correct annotation of the categories being classified. Both NEP and EP classes were derived from the demographic information presented in section 4.1 and manually annotated. Figure 4.19 details the distribution of sounds across both classes.

Table 4.7: Inter-onset mean and variance measurements of spectral and temporal low-level features.

	NEP		EP	
Descriptor	Inter-onset mean	Inter-onset variance	Inter-onset mean	Inter-onset variance
Bark_bands Kurtosis	1.718 e01	2.991 e01	1.337 e02	5.111 e04
Bark_bands Skewness	2.901 e00	2.153 e01	5.524 e00	1.287 e01
Bark_bands Spread	1.094 e01	2.112 e01	9.024 e00	1.846 e01
HFC	5.250 e-01	1.689 e-01	1.450 e-01	2.235 e-02
Spectral_centroid	2.196 e03	7.086 e00	2.293 e03	1.023 e01
Spectral_complexity	1.411 e00	9.199 e-01	4.841 e-01	3.952 e-01
Spectral_crest	1.353 e01	9.384 e00	1.548 e01	7.564 e00
Spectral_decrease	-3.968 e-10	1.077 e-19	-1.722 e-10	1.415 e-20
Spectral_energy	1.545 e-03	1.586 e-06	6.631 e-04	2.076 e-07
Spectral_energy low	3.742 e-04	7.899 e-08	1.724 e-04	2.104 e-08
Spectral_energy mid-low	9.253 e-04	9.093 e-07	3.066 e-04	1.256 e-07
Spectral_energy mid-high	1.119 e-04	1.277 e-08	2.740 e–05	1.197 e-09
SPectral_energy high	1.404 e-05	3.570 e-10	2.224 e-05	8.213 e-12
Spectral_flatness	1.836 e-01	1.809 e-03	2.185 e-01	2.427 e-03
Spectral_flux	2.167 e-02	1.044 e-04	1.444 e-02	2.919 e-05
Spectral_RMS	8.006 e-04	1.919 e-07	5.042 e-04	4.554 e-08
Spectral_rolloff	4.492 e02	3.491 e04	3.642 e02	3.665 e04
Spectral strong_peak	2.970 e-02	4.025 e-04	1.435 e-02	1.837 e-04
Zero_crossing_rate	2.196 e-02	6.211 e-05	1.852 e-02	4.730 e-05
Inharmonicity	9.456 e-02	8.620 e-04	7.730 e-02	7.755 e-04

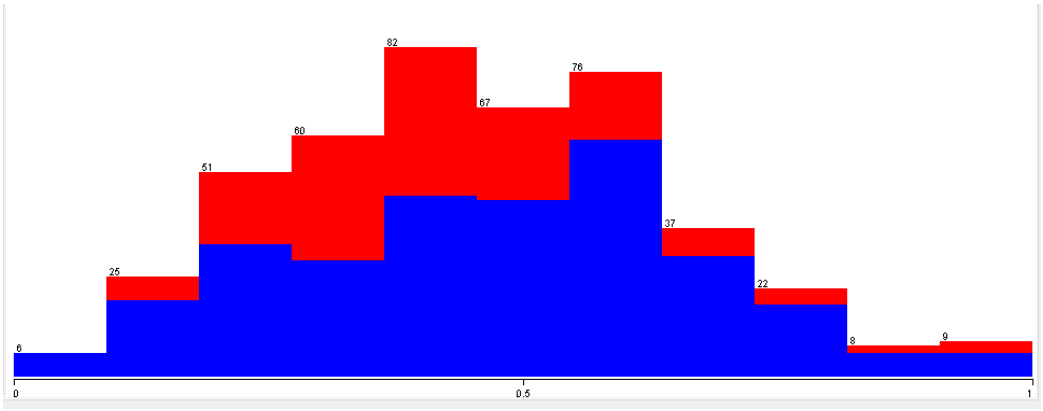

Figure 4.18: Distribution of the spectral centroid (blue: EP subjects; red: NEP subjects).

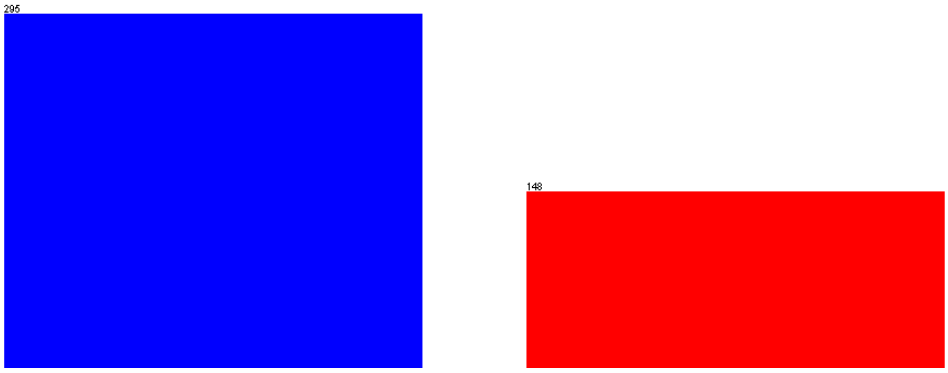

Figure 4.19: Onset class distribution. In blue: 295 EP onsets; in red: 148 NEP onsets.

Table 4.8: Hit rates for different learning algorithms (rows) and different feature selection strategies (columns)

	All features	BestFirst + CFS (5)	Greedy Step-Wise + CFS	Ranker + GainRation attribute	Ranker + InfoGain attribute
NaiveBayes	83.74 %	83.5 %	84.0 % (20)	84.0 % (20)	77.6 % (20)
C4.5	98.64 %	98.2 %	98.2 %	98.9 % (10)	**99.1 % (20)**
SVM	99.09 %	82.8 %	82.2 % (20)	**99.5 % (30)**	**99.1 % (30)**
AdaBoost	**99.32 %**	-	-	-	-

Table 4.9: Confusion matrix; Support Vector Machine with Information Gain Ranking filter (30 best features)

a	b	<– classified
295	0	a = Onset-EP
2	146	b = Onset-NEP

Table 4.8 summarizes the main results. We have first tested a set of 67 descriptors including those that were also used in previous studies on percussive sound classification. Regarding the feature selection technique, we compared several attribute evaluators and search methods. Except for Naive Bayes, we mostly obtained very high accuracies for all classification strategies. Correlation Feature Selection (CFS) drives down the accuracy regardless of using Best First or Greedy Step-wise as searching method, and also regardless of the classification method. Only using Information Gain or Gain Ratio filters with 10 to 30 attributes and a Ranker search yielded best hits, with an overall maximum of 99.5% using Support Vector Machine (SVM). Only two NEP-onsets where confused as depicted in Table 4.9, achieving an relative absolute error of 1% using 10-fold cross validation procedure (4.10). A very high accuracy is also obtained using C4.5 decision tree classifier with only 20 selected features using Information Gain Ranker filter. In Figure 4.21 we can see the visualization of the former decision tree, which only considers the energy at Bark-band 9 and the inharmonicity. Moreover, in Figure 4.22 it is depicted the decision tree only using the best features selected by Information Gain Evaluation algorithm. The best ten features ranked by the filter are shown in Figure 4.20.

Information Gain Ranking evaluation measures the change in the information entropy from a prior state to a state that takes the evaluated feature for attempting the classification of a given attribute class. Then, features are ranked in decreasing order of entropy. As depicted in Figure 4.20, all the ranked attributes belong to the spectral domain: three Bark bands, second, third and fourth moments of the Bark

Table 4.10: Stratified cross-validation. Support Vector Machine and feature selection

Correctly Classified Instances	441	99.5485 %
Incorrectly Classified Instances	2	0.4515 %
Kappa statistic	0.9898	
Mean absolute error	0.0045	
Root mean squared error	0.0672	
Relative absolute error	1.014 %	
Root relative squared error	14.2448 %	
Total Number of Instances	443	

```
Attribute Evaluator (supervised, Class (nominal): 68 class):
        Information Gain Ranking Filter

Ranked attributes:
 0.903    35 Bark_band_Spread.mean
 0.898    33 Bark_band_Kurtosis.mean
 0.898     7 Bark_band3.mean
 0.887    64 Spectral_Rolloff.mean
 0.882    34 Bark_band_Skewness.mean
 0.881    67 inharmonicity.mean
 0.876    29 Bark_band25.mean
 0.876    46 MFCC_9.mean
 0.873    18 Bark_band14.mean
 0.873    65 Spectral_Strong_Peak.mean
```

Figure 4.20: Ranking of the ten best features according to Information Gain algorithm.

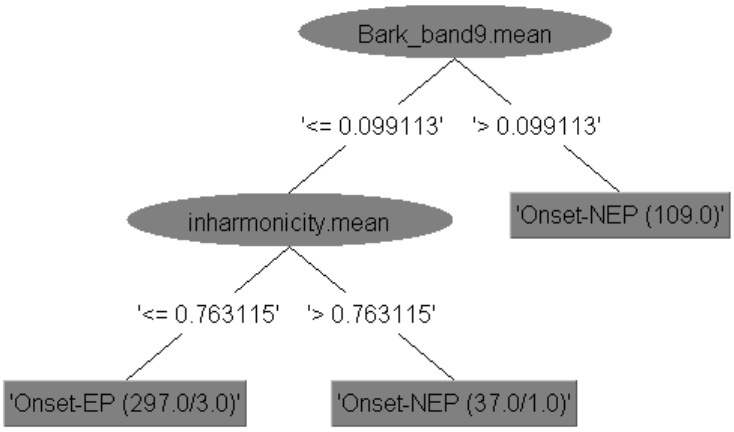

Figure 4.21: Visualization of the decision tree using C4.5 classifier with InfoGain Evaluation feature selection.

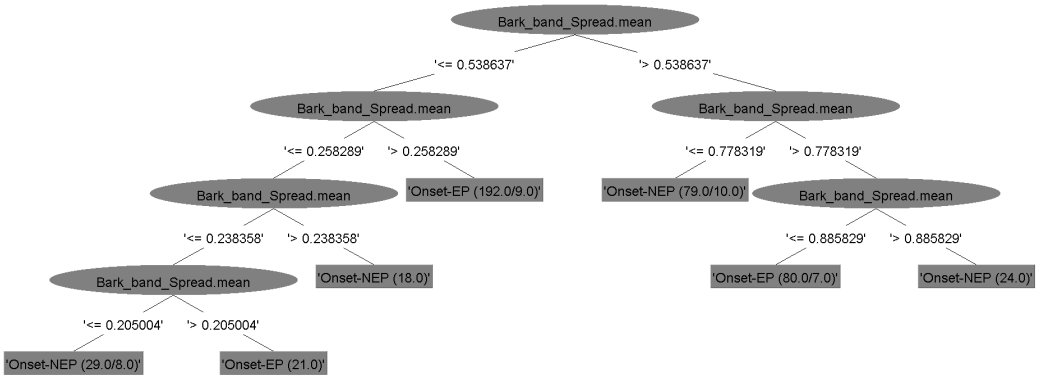

Figure 4.22: Visualization of the decission tree using C4.5 classifier with just 1 feature selected by InfoGain Evaluation algorithm.

scale spectrum, the spectral strong peak, inharmonicity, spectral roll off, spectral spread and the ninth MFCC. Notice that many of these descriptors coincide with the ones pointed out in section 4.1.2. These descriptors are closely linked with the timbre characteristics of the sound. In this regard, the ranking suggests that timbre descriptors underlie the main differences between both finger-tapping methods. It can, thus, be concluded that the finger-tapping technique over the provided box, confers some timbre quality that characterizes unequivocally the profile of the performer. Also, it can be definitely stated that NEP subjects do not consider timbre in the same extent as those belonging to EP subgroup when finger-tapping.

4.2 Finger-tapping onset transcription

The aim of this system is to transcribe human finger-tapping signals into a higher level of representation for improving human-computer interaction in drum composition applications. The information automatically extracted from the signal includes the voice category played on each stroke (differentiate between types of stroke), the onset time of each event and several spectral and temporal features. The system architecture is thus based on three major parts:

1. a segmentation module

2. a feature extraction module

3. and a clustering module for which different approaches were tested

The database used for this study consists of 27 finger-tap recordings from which 372 strokes were extracted. To segment each finger stroke, we used an onset detection algorithm based on Super Flux [82] implemented in Essentia Standard library. Since finger-tap signals consist of localized events with abrupt onsets, this algorithm obtains very satisfying results.

To select an appropriate features set, we experimented with several group combinations of those descriptors mentioned in 3.3.1. A simple clustering classifier (k-Means and I-Nearest Neighbor) was then used for differentiating between types of stroke. Experimental results on the different feature sets have, for a large part, confirmed previously discussed hypothesis on the use of time-depending energy descriptors rather than timbre descriptors for the transcription of finger-tapping signals. Therefore, the best transcription result is achieved using as input feature for a given onset the 10-step quantized version of the normalized energy of the current, previous and next onsets. The main disadvantage of this approach is that it constraints real-time applicability.

In Figures 4.23 through 4.25 we can see the visualization of the onset clustering into three different classes using exclusively time-dependent descriptors. Each color represents a different class.

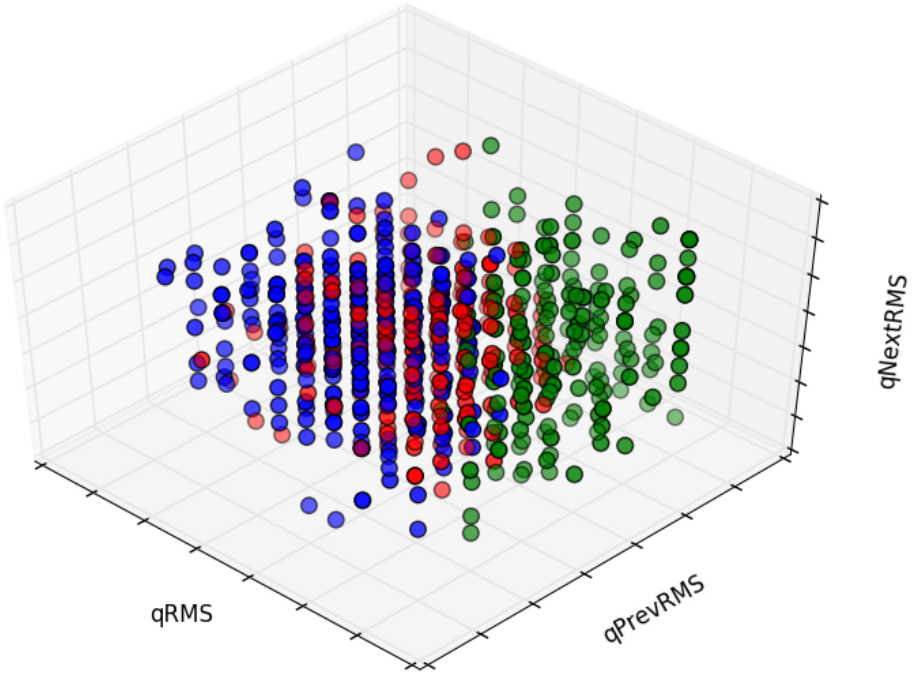

Figure 4.23: Visualization of the clusters obtained using k-Means algorithm with 3 classes (a).

4.3 Discussion

The results in 4.1 have proven the existence of two overall finger-tapping strategies: one addressed by people with no percussion training and another by people with at least two year of experience in some percussion facet, as stated in 3.2. By analyzing the gathered data and the recorded audio content, we have deeply described the behavior and characteristics of finger-tapping for each of them. We suggested that a potential software for the arrangement of drums through finger-tapping would be approached in different ways depending on the type of user that would make use of it. Our first approach focuses in the individuals from the former group. In a condensed form, this finger-tapping strategy is characterized by not considering timbre to confer differences to each stroke when finger-tapping, we can normally distinguish 2 different voices in the pattern and there are no

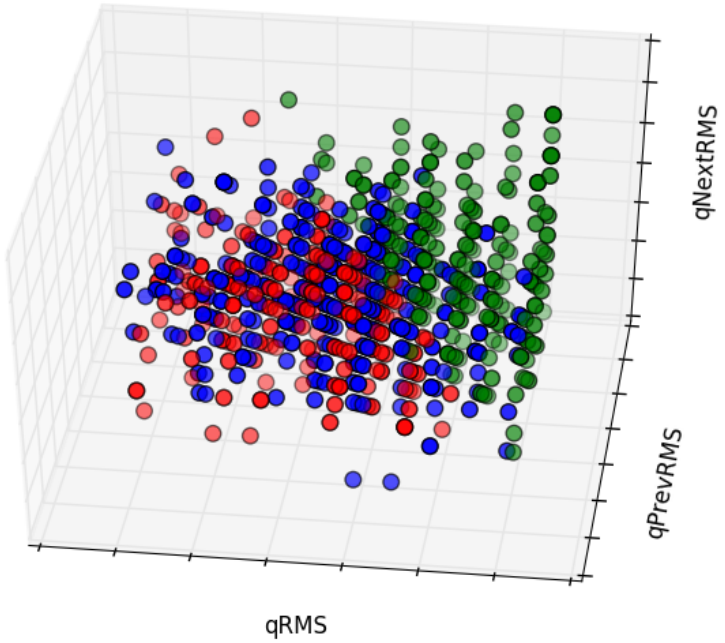

Figure 4.24: Visualization of the clusters obtained using k-Means algorithm with 3 classes (b).

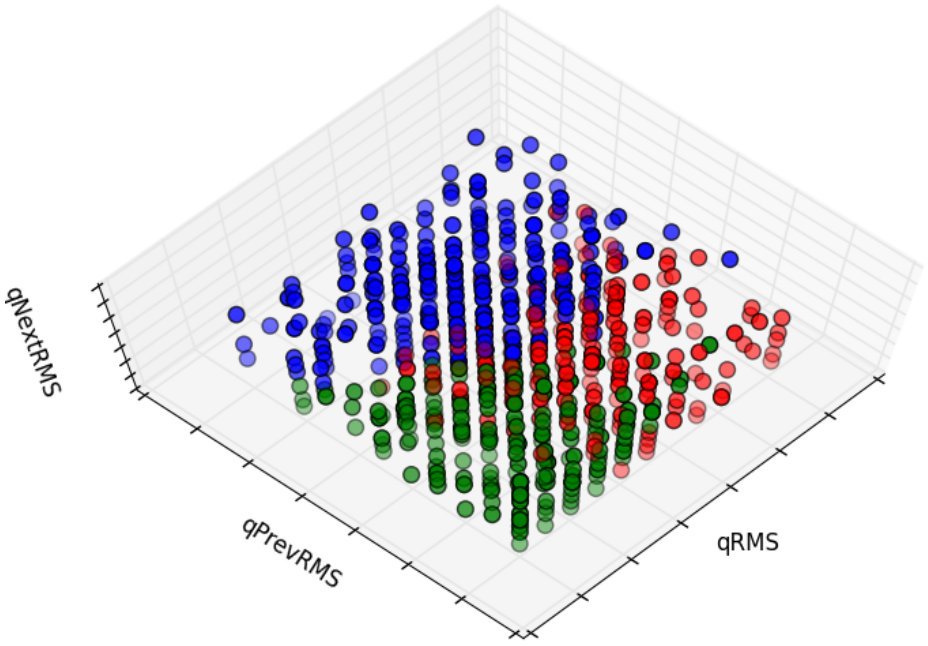

Figure 4.25: Visualization of the clusters obtained using k-Means algorithm with 3 classes (c).

overlapping strokes. For this reason, time-dependent energy descriptors have resulted to provide more accurate transcription than with timbre descriptors. Timbre descriptors have resulted to work well for differentiating between finger-tapping strategies but not between type of onsets within a particular pattern.

Chapter 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

We have presented a conducted study on the behavioral and acoustic properties of human finger-tapping. This work has been drawn towards the implementation of an audio-driven software for the automatic arrangement of drums, capable of using this human behavior as a mean for interaction. The general goal was to study the feasibility of finger-tapping for improving human-computer interaction in the context of computer-music. To this end we have accomplished different sub-tasks. First, we have collected a data-set of 43 finger-tapping recordings performed by individuals with different music backgrounds. Secondly, we have gathered demographic details as well as information about the finger-tapping performance by submitting an online survey to 23 subjects from the total. In this survey we also collected user-context information by requesting people's preferences and needs with regard to a hypothetical interactive software for the arrangement of drumsets. Thirdly, we detected onsets and computed several low-level descriptors from the recordings. Analysis over the survey results and the extracted audio features suggested the existence of two general strategies for finger-tapping: one that includes people with no experience at all in percussion (NEP) and another composed of people with at least two years of experience in some percussion facet (EP). These two profiles presented considerable differences in the answers to the survey as well as in many timbre descriptors. Further experiments allowed us to prove previous hypothesis through the classification of both strtegies, achieving an encouraging accuracy of 99%. Finally, we implemented a simple finger-tapping transcription system as a first step on the overall software for the arrangement of drums. The system, differently from other transcription methods, implements

as feature the relative 10-step quantized energy of a given onset with respect to the next and previous onsets within the pattern. Although this method does not allow real-time performance, it was proven to achieve encouraging results in the transcription of finger-tapping excerpts without the use of timbre descriptors.

5.2 Future work

Following the idea of understanding finger-tapping behavior, directions for future work may be in different subjects. First, we propose to approach experiments from a functional perspective (i.e. for the purpose of implementing an audio-driven drum arrangement software) rather than from a cognitive or more general perspective. For example, we propose to submit a group of volunteers to listen a certain drum pattern and, after some time, ask them to perform it by finger-tapping. Through this experiment it would be clear the types of instrument within a drum-set that match with each stroke in the finger-tapping pattern. Results have shown that neither subjects with experience in percussion nor inexperienced tend to think in a drum or percussion instrument in the conditions of the undertaken tests. In this regard, contextualizing the experiments may be useful for gathering more significant data for the application under consideration. Secondly, we propose to increase the finger-tapping collection with new recordings. Unfortunately, collecting the data has been one of the most arduous tasks in this thesis since most of the people felt overwhelmed when asked to be recorded. For this reason and due to time-line limitations we have not been able to acquire the expected data-set. From the implementation perspective, we propose the following points that may improve the performance of our finger-tapping transcription function:

- Add some pre-processing based on timbre that keeps frequencies where finger-tapping percussion is more likely to happen and discards information from other bands.
- Experiment with features derived from rhythmic pattern representations such as IOIH, ACF, DFT and other spectral and temporal representations (see 2.2).
- Experiment with time-dependent modeling algorithms such as Hidden-Markov Models or Dynamic Bayesian Networks.
- Carry out the perceptual evaluation of the generated finger-tapping transcriptions.

Assuming that these suggestions would enhance the transcription of finger-tapping transcription, the next direction to work on is on the development of high

level features based on the sequence of MIDI events for music information retrieval. Concretely, and following the general prospect of the project, we propose to include automatic arrangement capabilities such as onset time correction, addition of new percussion layers, audio-driven drum-kit recommendation, etc. We also hope, after attempting the first finger-tapping transcription system, that this extended behavior will be considered in the future for improving human-computer interaction in the field of interfaces for rhythm expression.

Appendix A

CODE AND DATA DEVELOPED/GENERATED

Below we provide a link to the full finger-tapping survey[1] as well as to a git-hub repository[2] containing the following contributions:

- Finger-tapping recordings collection: a set of 40 spontaneous finger-tapping excerpts recorded by western people of many different music backgrounds.
- Python script for extracting onset-based descriptors and generating a feature-set in JSON, CSV and ARFF formats. This script needs certain external python modules in order to work: ARFF[3], Essentia[4] and Scikit-learn[5]. This script receives as input the path to the folder containing the finger-tapping audio files and outputs an onset-based feature-set in the above mentioned formats.
- Python script for finger-tapping transcription. This script receives as input the path to a finger-tapping audio recording, the name and path of the output file and generates a MIDI file (.mid) with the symbolic transcription of the input pattern.
- Already computed feature-sets in csv, json and arff format.

[1] https://docs.google.com/forms/d/1LpuaNHymt2gEgjqjIBup$_0 o2wkFsjpbn5Om4N1nTYB8/viewform$

[2] https://github.com/cukinhou/Rhythmic-arrangement-from-finger-tapped-audio-recordings

[3] https://pypi.python.org/pypi/liac-arff

[4] http://essentia.upf.edu/

[5] http://scikit-learn.org/

Bibliography

[1] S Lakatos. A common perceptual space for harmonic and percussive timbres. *Perception & psychophysics*, 62(7):1426–1439, 2000.

[2] Wikipedia. Tapping.

[3] The Editors of Encyclopædia Britannica. Music Arrangement.

[4] Richard Cook. *Richard Cook's Jazz Encyclopedia*. London, penguin bo edition, 2005.

[5] Wallace Berry. *Structural Functions in Music*. New York, second edition, 1987.

[6] Mark Feezell. Music Theory Fundamentals. *Music Theory Fundamentals*, pages 1–46, 2011.

[7] Margaret C. Tiffin-Richards, Marcus Hasselhorn, Michael L. Richards, Tobias Banaschewski, and Aribert Rothenberger. Time reproduction in finger tapping tasks by children with attention-deficit hyperactivity disorder and/or dyslexia. *Dyslexia*, 10(4):299–315, 2004.

[8] Pablo Arias, Verónica Robles-García, Nelson Espinosa, Yoanna Corral, and Javier Cudeiro. Validity of the finger tapping test in Parkinson's disease, elderly and young healthy subjects: Is there a role for central fatigue? *Clinical Neurophysiology*, 123(10):2034–2041, 2012.

[9] Simon Dixon and Werner Goebl. Pinpointing the beat: Tapping to expressive performances. *To appear in: 7th International Conference on Music ...*, (July):2000–2003, 2002.

[10] M. Franek, J. Mates, T. Radil, K. Beck, and E. Pppel. Finger tapping in musicians and nonmusicians. *International Journal of Psychophysiology*, 11(3):277–279, 1991.

[11] Wikipedia. Timbre.

[12] Perfecto Herrera, Alexandre Yeterian, and Fabien Gouyon. Automatic classification of drum sounds : a comparison of feature selection methods and classification techniques. *Music and Artificial Intelligence*, pages 69–80, 2002.

[13] Perfecto Herrera, Amaury Dehamel, and Fabien Gouyon. Automatic Labelling of Unpitched Percussion Sounds. *Aes 114Th Convention*, 2003.

[14] Sofia Cavaco and Hugo Almeida. Automatic Cymbal Classification Using Non-Negative Matrix Factorization. *Systems, Signals and Image Processing (IWSSIP), 2012 19th International Conference on. IEEE*, pages 468–471, 2012.

[15] Gustavo E A P A Batista and Nilson E Souza-filho. Automatic Classification of Drum Sounds with Indefinite Pitch. pages 1–8, 2015.

[16] E Pampalk, P Hlavac, and P Herrera. Hierarchical Organization and Visualization of Drum Sample Libraries. *Proc Intl Conf Digital Audio Effects*, pages 3–8, 2004.

[17] Perfecto Herrera, Vegard Sandvold, and Fabien Gouyon. Percussion-related semantic descriptors of music audio files. *Proceedings of 25th International AES . . .* , pages 1–5, 2004.

[18] O. Gillet and G. Richard. Automatic transcription of drum loops. *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 4:2–5, 2004.

[19] Olivier Gillet and Gal Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech and Language Processing*, 16(3):529–540, 2008.

[20] Derry Fitzgerald and Jouni Paulus. Unpitched Percussion Transcription. *Signal Processing Methods for Music Transcription*, II:131–162, 2006.

[21] Emmanouil Benetos, Sebastian Ewert, and Tillman Weyde. Automatic transcription of pitched and unpitched sounds from polyphonic music. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, (May):3107–3111, 2014.

[22] Christian Dittmar, I Fraunhofer, and D Gärtner. Real-time Transcription and Separation of Drum Recording Based on NMF Decomposition. *Proceedings of the International Conference on Digital Audio Effects (DAFX)*, pages 1–8, 2014.

[23] Marius Miron, Matthew E P Davies, and Fabien Gouyon. An open-source drum transcription system for Pure Data and Max MSP. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, (October):221–225, 2013.

[24] George Tzanetakis, Ajay Kapur, and Richard I. McWalter. Subband-based drum transcription for audio signals. *2005 IEEE 7th Workshop on Multimedia Signal Processing*, pages 9–12, 2006.

[25] J. K. Paulus and A. P. Klapuri. Conventional and periodic N-grams in the transcription of drum sequences. *Proceedings - IEEE International Conference on Multimedia and Expo*, 2(Icme):II737–II740, 2003.

[26] Lucas Thompson, Simon Dixon, and Matthias Mauch. Drum Transcription via Classification of Bar-Level Rhythmic Patterns. *International Society for Music Information Retrieval Conference*, (Ismir):187–192, 2014.

[27] Koen Tanghe, Sven Degroeve, and Bernard De Baets. An algorithm for detecting and labeling drum events in polyphonic music. ... *of the 1st Annual Music Information* ..., 2005.

[28] Vegard Sandvold and Fabien Gouyon. Percussion Classification in Polyphonic Audio. *Proc International Conference on Music Information Retrieval*, pages 2–5, 2004.

[29] Kazuyoshi Yoshii, Masataka Goto, and Hiroshi G Okuno. Automatic Drum Sound Description for Real-World Music. *Science And Technology*, pages 184–191, 2014.

[30] J Paulus and A Klapuri. Drum Sound Detection in PolyphonicMusic with Hidden {Markov} Models. *EURASIP Journal on Audio, Speech, and Music Processing*, 2009.

[31] Jouni Paulus. Acoustic Modelling of Drum Sounds with Hidden Markov Models for Music Transcription. *Technology*, pages 241–244, 2006.

[32] Hazan and Amaury. Towards automatic transcription of expressive oral percussive performances. *IUI '05: Proceedings of the 10th international conference on Intelligent user interfaces*, pages 296–298, 2005.

[33] E Sinyor, C M Rebecca, D Mcennis, and I Fujinaga. Beatbox classification using ACE. *Proceedings of the International Conference on Music Information Retrieval*, 2005.

[34] Dan Stowell and Mark D Plumbley. Characteristics of the beatboxing vocal style. *Electronic Engineering*, pages 1–4, 2008.

[35] Michael Proctor, Erik Bresch, Dani Byrd, Krishna Nayak, and Shrikanth Narayanan. Paralinguistic mechanisms of production in human "beatboxing": a real-time magnetic resonance imaging study. *The Journal of the Acoustical Society of America*, 133(2):1043–54, 2013.

[36] Kyle Hipke, Michael Toomim, Rebecca Fiebrink, and James Fogarty. BeatBox. *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces - AVI '14*, pages 121–124, 2014.

[37] Geoffroy Peeters. Spectral and Temporal Periodicity Representations of Rhythm for the Automatic Classification of Music Audio Signal. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5):1242–1252, 2011.

[38] Jonathan Foote and Shingo Uchihashi. The beat spectrum: A new approach to rhythm analysis. *Proceedings - IEEE International Conference on Multimedia and Expo*, 00(C):881–884, 2001.

[39] Jonathan Foote, Matthew Cooper, and U. Nam. Audio retrieval by rhythmic similarity. *Proceedings of the International Conference on Music Information Retrieval*, 3:265–266, 2002.

[40] Iasonas Antonopoulos, Aggelos Pikrakis, Olmo Cornelis, Dirk Moelants, Marc Leman, and Sergios Theodoridis. Music Retrieval by Rhythmic Similarity Applied on Greek and African Traditional Music. *Austrian Computer Society (OCG)*, pages 6–9, 2007.

[41] G. Tzanetakis. Factors in automatic musical genre classification of audio signals. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2003-Janua:143–146, 2003.

[42] Elias Pampalk, Simon Dixon, and Gerhard Widmer. Exploring Music Collections by Browsing Different Views. *Computer Music Journal*, 28(2):49–62, 2004.

[43] Chih-wei Wu and Alexander Lerch. Beat Histogram Features from NMF-based Novelty Functions for Music Classification. pages 434–440, 2015.

[44] Athanasios Lykartsis, Chih-Wei Wu, and Alexander Lerch. Beat Histogram Features for Rhythm-based Musical Genre Classification using Multiple Novelty Functions. *Proceedings of the 16th ISMIR Conference*, (JANUARY):434–440, 2015.

[45] Jouni Paulus and Anssi Klapuri. Measuring the Similarity of Rhythmic Patterns. *Signal Processing*, 1:44, 2002.

[46] A Pikrakis. A deep learning approach to rhythm modelling with applications. ... *Workshop on Machine Learning and Music (MML13)*, (SEPTEMBER 2013), 2013.

[47] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer. Evaluating rhythmic descriptors for musical genre classification. *AES 25th International Conference*, pages 1–9, 2004.

[48] Fabien Gouyon and Simon Dixon. Dance music classification: A tempo-based approach. *Proc International Conference on Music Information Retrieval*, pages 501–504, 2004.

[49] Simon Dixon, Fabien Gouyon, and Gerhard Widmer. Towards characterisation of music via rhythmic patterns. *Proceedings of the 5th International Conference on Music Information Retrieval (ISMIR 2004)*, 5(April):509–516, 2004.

[50] Simon Dixon. An Interactive Beat Tracking and Visualisation System The Audio-Graphical User Interface. *Proceedings of the International Computer Music Conference ICMC*, pages 215–218, 2001.

[51] Matthew Wright, W Andrew Schloss, and George Tzanetakis. Analyzing Afro-Cuban Rhythm Using Rotation-Aware Clave Template Matching With Dynamic Programming. pages 647–652, 2008.

[52] Holzapfel Andre and Stylianou Yannis. Rhythmic Similarity of Music based on Dynamic Periodicity Warping. *Transform*, pages 317–320, 2009.

[53] Andf Holzapfel and Yannis Stylianou. A Scale Transform Based Method for Rhythmic Similarity of Music. *Transform*, pages 317–320, 2009.

[54] Jesper Hjvang Jensen, Mads Groesbll Christensen, and Sren Holdt Jensen. A TEMPO-insensitive representation of rhythmic patterns. *European Signal Processing Conference*, (Eusipco):1509–1512, 2009.

[55] Matthias Gruhne, Christian Dittmar, and Daniel Gaertner. Improving Rhythmic Similarity Computation by Beat Histogram Transformations. *Ismir*, (Ismir):177–182, 2009.

[56] Chunta Chen and Jyh-shing Roger Jang. A Shifted Alignment Algorithm For Query By Tapping. pages 701–705, 2014.

[57] JS Jang, HR Lee, and Chia-hui Yeh. Query by tapping: A new paradigm for content-based music retrieval from acoustic input. *Advances in Multimedia Information Processing - PCM*, pages 590–597, 2001.

[58] Gunnar Eisenberg, Jm Batke, and Thomas Sikora. BeatBank - An MPEG-7 compliant Query by Tapping System. *Proc. of the 116th AES Conv*, 7, 2004.

[59] Gunnar Eisenberg, Jan-mark Batke, and Thomas Sikora. Efficiently Computable Similarity Measures. (April):189–192, 2004.

[60] Geoffrey Peters, Caroline Anthony, and Michael Schwartz. Song search and retrieval by tapping. *Proceedings of the National Conference on Artificial Intelligence*, page 1696, 2005.

[61] A Kapur, M Benning, and G Tzanetakis. Query-by-beat-boxing: Music retrieval for the DJ. *Proceedings of the International . . .* , 2004.

[62] Dan Stowell and Mark D. Plumbley. Delayed Decision-making in Real-time Beatbox Percussion Classification. *Journal of New Music Research*, 39(3):203–213, 2010.

[63] A Hazan. Billaboop real-time voice-driven drum generator. In *118th Audio Engineering Society Convention*, 2005.

[64] a Hazan. Performing Expressive Rhythms With Billaboop Voice-Driven Drum Generator. *Citeseer*, pages 20–23, 2005.

[65] Tiphaine de Torcy, Agnès Clouet, Claire Pillot-Loiseau, Jacqueline Vaissière, Daniel Brasnu, and Lise Crevier-Buchman. A video-fiberscopic study of laryngopharyngeal behaviour in the human beatbox. *Logopedics, Phoniatrics, Vocology*, 39(1):38–48, 2014.

[66] Polytechnique Fpms and Mons Umons. Analysis and automatic recognition of human Beat-box: a compartive study. pages 4255–4259, 2015.

[67] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53, 2004.

[68] Jonathan L Herlocker, Joseph a Konstan, and John Riedl. Explaining collaborative filtering recommendations. *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250, 2000.

[69] Robin Van Meteren and Maarten Van Someren. Using Content-Based Filtering for Recommendation. *ECML/MLNET Workshop on Machine Learning and the New Information Age*, pages 47–56, 2000.

[70] E Gstrein, F Kleedorfer, and R Mayer. Adaptive personalization: A multidimensional approach to boosting a large scale mobile music portal. *... : Integration of Music in ...*, 2005.

[71] Vito Claudio Ostuni, Sergio Oramas, Tommaso Di Noia, Xavier Serra, and Eugenio Di Sciascio. A Semantic Hybrid Approach for Sound Recommendation. A knowledge graph for sounds. *WWW 2015 Companion: Proceedings of the 24th International Conference on World Wide Web*, pages 85–86, 2015.

[72] Peter Hlavac, Brigitte Krenn, and Erich Gstrein. SOUNDSCOUT : A song recommender based on sound similarity for huge commercial music archives. . 2007.

[73] E Aoki and K Maruyama. Automatic musical arrangement apparatus generating harmonic tones, 1995.

[74] Tomomasa Nagashima and Jun Kawashima. Experimental study on arranging music by chaotic neural network. *International Journal of Intelligent Systems*, 12(4):323–339, 1997.

[75] Axel Berndt, Knut Hartmann, Niklas Röber, and Maic Masuch. Composition and Arrangement Techniques for Music in Interactive Immersive Environments. *Proceedings of the Audio Mostly Conference - a Conference on Sound in Games*, (April 2016):53–59, 2006.

[76] Jae-woo Chung and G Scott Vercoe. The affective remixer: personalized music arranging. *Proceedings of ACM CHI 2006 Conference on Human Factors in Computing Systems*, 2:393–398, 2006.

[77] Jeffrey M A K Ka-Hing, Clifford Choy Sze-Tsan, Clifford So Kwok-Fung, and Henry Ma Chi-Fai. Emotion-driven automatic music arrangement. {*ACM*} {*SIGGRAPH*} *2006 Research posters*, (April 2016):108, 2006.

[78] D.R. Tuohy and W.D. Potter. GA-based Music Arranging for Guitar. *2006 IEEE International Conference on Evolutionary Computation*, (April):1065–1070, 2006.

[79] Daniel R Tuohy and W D Potter. An Evolved Neural Network / HC Hybrid for Tablature Creation in GA-based Guitar Arranging.

[80] Jiun-Long Huang, Shih-Chuan Chiu, and Man-Kwan Shan. Towards an automatic music arrangement framework using score reduction. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 8(1):1–23, 2012.

[81] Dirk Moelants. Preferred tempo reconsidered. *Proceedings of the 7th international conference on music perception and cognition*, pages 580–583, 2002.

[82] S. Böck and G. Widmer. Maximum Filter Vibrato Suppression For Onset Detection. *16th Int. Conference on Digital Audio Effects (DAFx-13)*, pages 1–7, 2013.