MODELLING DRUM PATTERNS WITH WEIGHTED FINITE-STATE TRANSDUCERS

Stephan Hübler and Rüdiger Hoffmann

Chair for System Theory and Speech Technology, Technische Universität Dresden

ABSTRACT

In this paper, we present an approach to model bar length drum patterns with weighted finite-state transducers. Motivated by the existing algorithms for speech recognition, we discuss similarities to music, by considering a bar as word and the progression of bars as language. However, in contrast to speech, music has special characteristics, like metrical regularity and multiple notes at one point in time, which have to be taken into account. We use MIDI data to retrieve the drum notes for every bar, which are the input for the training of the WFST models. Once the models are trained, they are used to automatically recognize drum patterns and their time signature in a sequence of unknown drum notes.

We present an experiment on symbolic genre classification to demonstrate the principle of operation, training and recognition. It shows that the sequence of drum notes is representative for the four genres: Rumba, Samba, Tango and Waltz leading to a genre recognition rate of $88.9\% \pm 2.8\%$. Applications that could benefit from this approach include drum loop organization, drum note transcription, music similarity and genre detection.

Index Terms— WFST, music, drums, rhythm pattern, context modeling

1. INTRODUCTION AND RELATION TO PRIOR WORK

Towards the aim of a rhythmic description of music, we demonstrate a statistical approach to model bar length drum patterns with weighted finite state transducers (WFST). With the help of an example, we want to introduce the basic idea. Figure 1 shows the score of two drum patterns (two bars) with the time signature ¾. To get the WFST from the score, the notes at every eighth note step become a transition in the WFST, leading to 6 transitions per bar. The two drum patterns in the score notation begin with the same notes, but have a different note sequence starting from count 2, which leads to two different paths in the WFST. Since bars can have different time signatures, we build one WFST model per time signature calling them the time signature models $\mathcal{T}_{ts}$. In addition to the time signature models, we keep the progression of the time signature of consecutive bars and generate time signature progression model $\mathcal{P}$. For the example, $\mathcal{P}$ would simply hold the information that a bar in ¾ can be followed by a bar with the same time signature ¾.

After the training of those two models, we want to automatically detect drum patterns and their corresponding time signature in a sequence of drum notes, without any knowledge of the bar boundaries and the time signature. We assume that a bar is characterized by the sequence of drum notes, or in other words, we can find the bar boundaries by looking at the drum sequences. In order to recognize bars in an unknown drum note sequence we combine $\mathcal{P}$ and $\mathcal{T}_{ts}$ to a recognition network, which we call the rhythm model $\mathcal{R}$. The rhythm model could be applied for drum transcriptions, similarity retrieval or genre classification.

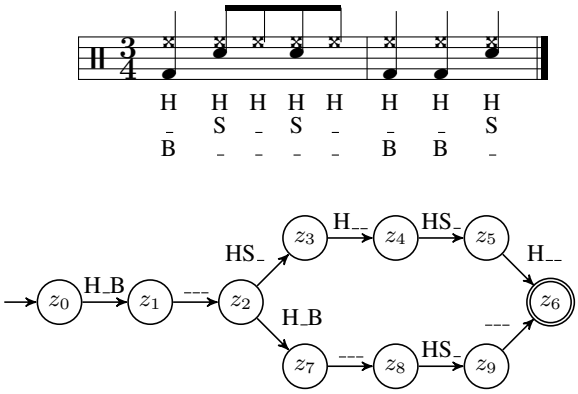

Fig. 1. Drum note example of two bars with time signature ¾ and the resulting time signature model $\mathcal{T}_{3/4}$ as WFST. For clarity, the transitions are only labeled with the input symbol (H: Hi Hat, S: Snare, B: Bass Drum, _ : no note played).

Previous work on rhythm patterns describe the temporal characteristics of bars, either of the complex musical signal or with a focus on the percussive/drum part of the music. Dixon extracts the signal envelope of the complex musical signal and conducts a genre recognition experiment with 50% accuracy [1]. In contrast, Paulus and Tsunoo separate the drum signal first and then extract low level features to calculate either the rhythmic similarity between songs [2] or to recognize different rhythmic patterns of a song [3]. Ellis calculated basic rhythmic patterns of base drum, snare and hi hat, using principal component analysis [4]. All afore mentioned publications deal with the problems of finding the right meter, tempo and extract meaningful rhythmic characteristics from the audio signal itself. We take symbolic MIDI data as a starting point, leaving the recognition of notes from the audio signal aside, which is discussed in detail in [5, 6]. A recent publication of Mauch puts a focus on possibilities to use MIDI data for statistical methods from ASR by analyzing bar length drum patterns, showing the similarities to a speech corpus and emphasizing the special features of bar length drum patterns [7]. We expand those ideas by demonstrating a data driven approach to model musical context with WFSTs on a MIDI data corpus. Little work has been done to model the context of notes. Paulus used N-Grams in the transcription process, but this can not capture the interactions of drum notes within a longer rhythmic pattern [8]. Correa models the sequence and duration of notes with a bigram [9] and Mauch introduces a statistically language model for chord progressions in the domain of harmony [10].

The remainder of the paper is organized as follows. Section 2 discusses similarities to the domain of ASR. Section 3 explains training of the time signature model, the time signature progression model and their combination to the rhythm model as well as

the recognition process of unknown drum note sequences. Section 4 presents a genre classification experiment on the four genres Rumba, Samba, Tango and Waltz. We discuss advantages, problems and further applications in Section 5.

2. SIMILARITIES TO ASR

Mohri started with ideas from the speech domain and developed a musical phoneme from audio by clustering similar feature vectors [11] and Reed used a similar approach to detect musical phonemes to model temporal information in music tag annotation [12]. In contrast to Mohri and Reed, we consider a group of notes at the same point in time as phoneme. Furthermore, we regard a bar as word and the progression of bars as some kind of language. In the domain of speech, the lexicon model translates from phonemes to words and the language model describes the progression of words [13]. To stress the similarity to music again, the lexicon would be the time signature model $\mathcal{T}_{ts}$ in the music domain, translating from notes to bars with the time signature as output.

3. PROPOSED MODEL

The goal of the paper is to train a rhythmical model $\mathcal{R}$ to recognize drum patterns in unknown drum note sequences. Figure 2 shows an overview of the complete system with the rhythmical model $\mathcal{R}$ at the top. This Section explains the system, which can be divided into three parts: symbol construction, training and recognition.

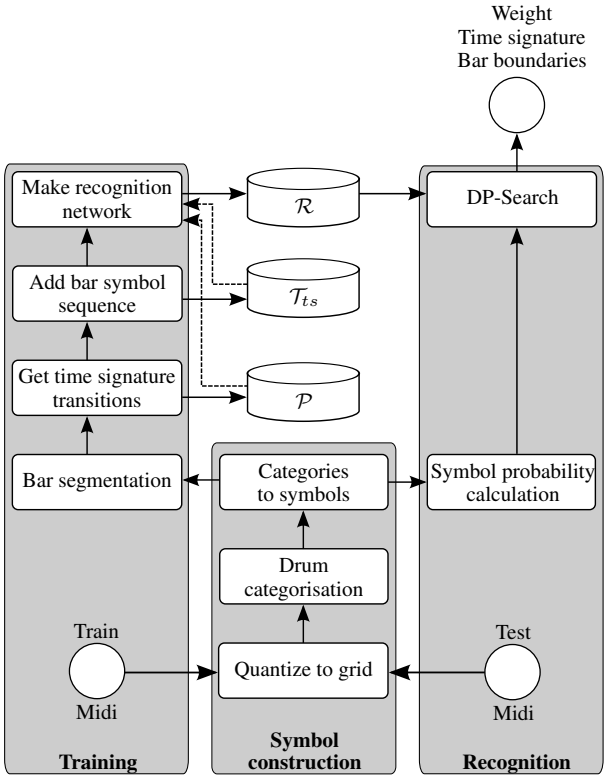

Fig. 2. Training and recognition process of drum patterns leading to the rhythm model $\mathcal{R}$. The rhythm model $\mathcal{R}$ is a combination of the time signature models $\mathcal{T}_{ts}$ and the time signature progression model $\mathcal{P}$. Training and recognition share the symbol construction.

3.1. Symbol construction

Symbol construction is the preprocessing of the MIDI file for the training and recognition process. It involves grid quantization to align the notes to a grid with fixed intervals, drum categorization to reduce the number of drum notes and combining the drum categories to one symbol at a time. The drum notes are quantized to a grid with the fixed interval of a 16th note. The resulting grid is tempo independent, because the actual length of the 16th note differs according to the tempo of the song. Subsequently, we categorize the notes of a drum kit into $N = 5$ categories, dismissing some of the 81 notes of the general drum map:

1. Base Drum [B]: Acoustic Bass Drum (35), Bass Drum (36)
2. Snare [S]: Side Stick (37), Acoustic Snare (38), Electric Snare (40)
3. Hi Hat [H]: Closed Hi Hat (42), Pedal Hi Hat (44), Open Hi Hat (46)
4. Ride [R]: Ride Cymbal 1 (51), Ride Bell (53), Ride Cymbal 2 (59)
5. Cymbals [C]: Crash Cymbal 1 (49), Chinese Cymbal (52), Splash Cymbal (55), Crash Cymbal 2 (57)

We combine the five drum categories to one symbol at a time. Given the number of categories N, one can calculate the number of possible symbols with $M = 2^N$, which leads to 32 symbols for $N = 5$. In other words, we have a set of symbols $\mathcal{S} = \{s_0, ...s_{M-1}\}$ and one song is represented as a sequence $\mathbf{S}$ of symbols $S(k) \in \mathcal{S}$ where $k = 0, \ldots, K-1$. The symbol sequence $\mathbf{S}$ considers two special characteristics of music. First, multiple notes at one point in time are combined to one symbol at a time and second, the metrical regularity is reflected by quantizing the notes to a grid. $\mathbf{S}$ has an empty symbol for grid points with no note played, which preserves the temporal structure within the WFST.

3.2. Training

The training process of the models starts with segmenting $\mathbf{S}$ into bars and continues with updating the time signature progression model $\mathcal{P}$. The model $\mathcal{P}$ is a bigram model and it stores the time signature of the first bar of the song, the progression of the time signature of consecutive bars within the song and the time signature of the last bar. Figure 3 shows an example of $\mathcal{P}$ for 138 Rumba songs with

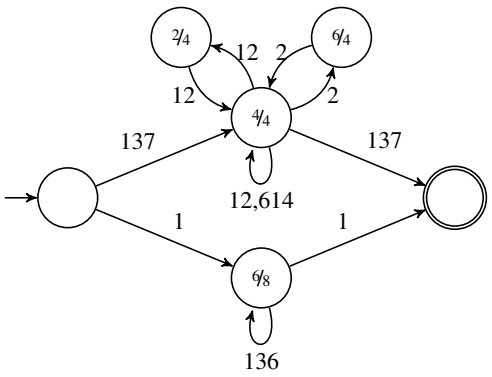

Fig. 3. Example of the time signature progression model $\mathcal{P}$ for 138 Rumba songs. For demonstration purposes the transitions are labeled with the counters instead of the probabilities.

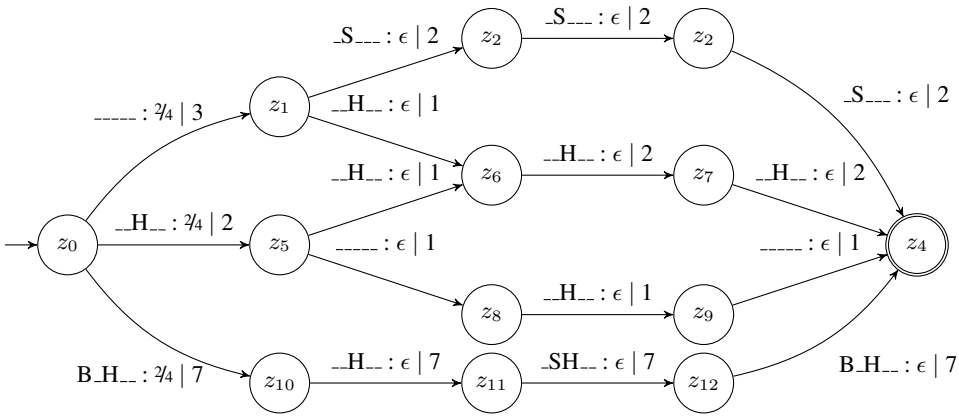

Fig. 4. Time signature model $\mathcal{T}_{2/4}$ of 12 bars from 121 Rumba songs. For demonstration purposes the transitions are labeled with the counter instead of the weight and an 8th-note grid is used instead of an 16th-note grid.

counters of the progressions. There are bars with four different time signatures in the training material, where 137 songs start with a bar in 4/4 and one song starts with a bar in 6/8. The song starting with 6/8 at the bottom of the Figure has 137 bars and all of them are in 6/8, which makes this Rumba song an exception. For the rest of the songs, most of the time a 4/4 bar is followed by a 4/4 (12,614 times) and 12 times a bar in 2/4 is inserted. Two times a bar in 6/4 occurs, which might be interpreted as a 4/4 + 2/4.

The time signature models $\mathcal{T}_{ts}$ hold the actual information of the bar length note sequences. Every transition of the WFST $\mathcal{T}_{ts}$ holds an input symbol, an output symbol and a weight. The input symbol is the combination of drum categories and can be an empty symbol if no note is played at the current grid point. The output symbol is the time signature at the transitions starting from an initial state and an empty output symbol ϵ for the rest of the transitions. The weight is the probability of every transition within the model. Every bar is added to the model with the appropriate time signature, leading to a path within the model. Different training bars could use the same transition, when sharing the same input symbols. The counters of the associated transitions are incremented, causing a higher path probability for common rhythmic patterns. Only complete bars are processed and bars with no note events are removed from the training, since they do not hold any information about possible drum note sequences. After adding all training bars, the probabilities of the different transitions for each state are estimated using the counters and converted to negative logarithmic probabilities. Finally every model $\mathcal{T}_{ts}$ is minimized.

Figure 4 shows the time signature model $\mathcal{T}_{2/4}$ of 12 bars from 138 Rumba songs. Seven identical 2/4 bars appear in the training material starting with `B_H__` (lowermost path). Five different drum patterns could be observed, leading to 5 different paths within the WFST. After adding all bars of the training data, every state of the bigram $\mathcal{P}$ is replaced by the correspondent WFST model $\mathcal{T}_{ts}$ to form the rhythmical model $\mathcal{R}$.

3.3. Recognition

For the recognition process the rhythmical model $\mathcal{R}$ is used as recognition network. With the help of dynamic programming the best path for a given symbol sequence $\mathbf{S}$ within $\mathcal{R}$ is determined. The output of the recognition process is a sequence of time signature labels at the beginning of the bars and a path weight, which reflects the suitability of $\mathbf{S}$ to the rhythmic model $\mathcal{R}$. To recognize drum sequences, which were not part of the training, it is possible to change symbols in the sequence. Every symbol s_i of $\mathcal{S}$ is likely to occur with a certain probability $p(k, s_i)$ at every point in time k.

$$p(k, s_i) = \begin{cases} a & \text{for } s_i = S(k) \\ \frac{1-a}{M-1} & \text{for } s_i \neq S(k) \end{cases} \quad (1)$$

The overall probability is shared between all possible symbols, giving a high value to the actual symbol $S(k)$ and equally sharing the rest between the other $M-1$ symbols. When the recognition process is done on symbolic music, the symbol $S(k)$ is known, which is why we set $a = 0.9999$. Nevertheless, this procedure is needed to recognize unknown drum note sequences in symbolic music. The change of a symbol leads to a low probability of the symbol, which increases the path weight.

4. EXPERIMENT ON GENRE AND TIME SIGNATURE CLASSIFICATION

Various publications address the problem of automatic genre classification [14]. Especially Tzanetakis included rhythmic descriptors as part of the system [15] and Gouyon introduced a dataset of 8 ball room dance styles, to explicitly investigate rhyhtmic features [16]. This data set has also been used by various other researchers to evaluate rhythmic features and patterns for genre recognition [1, 17, 18, 19]. In contrast to those works on audio genre classification, we classify symbolic data into genres, which has been also done by McKay and Lidy [20, 21]. In our experiment, the classification of the genre is solely based on the drum notes played and their sequences within a bar.

4.1. Database

The experiment uses midi files of four different ball room dance styles: Rumba (Rum), Samba (Sam), Tango (Tan) and Waltz (Wal). The midi files are part of the midiart database[1]. Table 1 shows the number of files per genre and the number of bars per time signature of the experiment database. Remarkably, 137 bars of the genre Waltz have the time signature 4/4, which are two songs, having parts

[1] www.midi.de

Genre	Songs	Bars per time signature						
		2/4	3/4	4/4	6/4	3/8	6/8	9/8
Rumba	138	12	-	12.765	2	-	137	-
Samba	161	102	-	16.125	8	-	98	-
Tango	47	19	-	4.673	-	-	-	-
Waltz	158	141	22.872	137	-	9	3.747	3
Total	504	274	22.872	33.700	10	9	3.982	3

Table 1. Number of songs per genre and the number of bars per time signature.

in 3/4 and 4/4. Those two Waltz songs have segments from two different genres: Waltz and Rock. Furthermore, both Rumba and Samba have bars in 6/8. For Rumba, it is one complete song in 6/8 and for Samba one song having 87 bars in 4/4 and 98 bars in 6/8. We decided to leave all of those uncommon songs in the database to investigate the resulting effects.

4.2. Settings

To classify drum note sequences **S** with respect to genre, one rhythm model $\mathcal{R}_g$ per genre is trained. With the help of dynamic programming, we determine the best path for a sequence **S** of the test set within every genre rhythm model $\mathcal{R}_g$. The rhythmical model $\mathcal{R}_g$ with the minimal path weight determines the genre of the test song. Additionally, the output symbols of $\mathcal{R}_g$ are time signature labels at the beginning of every bar, which can be evaluated to determine the bar boundaries and the time signature of a song. The song-wise time signature is the one most bars belong to. A 10-fold cross validation is performed to achieve the recognition results.

4.3. Results

The genre classification of the 504 files with 10-fold cross validation leads to a correctness of $88.9\% \pm 2.8\%$. Figure 5 shows the confusion matrix. We obtain a fairly high recognition rate, which could be explained with the use of symbolic MIDI data and the restriction on only four distinctive rhythmic genres. A closer look on the results reveals the influence of the uncommon songs within the training database. For example, the two Waltz songs that are classified as Rumba fit to the drum pattern of the Rumba song in 6/8 and the drum patterns of the two Samba songs classified as Waltz match the bars in 4/4 of the uncommon Waltz songs. We expected, that those errors would be suppressed by the time signature progression model $\mathcal{P}$, but it seems that the ratio of the weights between $\mathcal{P}$ and the time signature model $\mathcal{T}_{ts}$ have to be adjusted. In the domain of ASR, this is done by the language model factor. Other confusions are caused simply by the fact that the drum note sequence **S** of a song matches the sequence of a different genre. For example, some basic drum patterns are used in more than one genre.

After choosing the best genre model $\mathcal{R}_g$, the output of time signature labels at every bar boundary is evaluated. Figure 6 shows the song-wise time signature confusion matrix. For $98.6\% \pm 1.0\%$ songs the time signature is correctly detected. Within the 504 midi files 60,850 bars occur. The correctness of the label at the bar boundary is $95.8\% \pm 0.2\%$. Error types are division of a bar in 4/4 into two 2/4 bars or the mapping of three bars in 4/4 to four bars in 3/4.

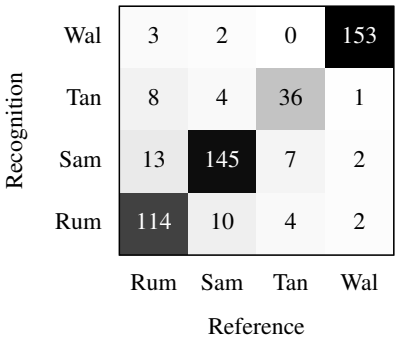

Fig. 5. Confusion matrix of the four genres Rumba, Samba, Tango and Waltz.

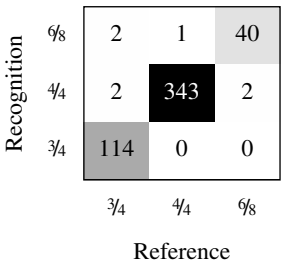

Fig. 6. Confusion matrix of the song-wise time signature.

5. CONCLUSIONS

We showed how to model bar-length drum note patterns with WFSTs using MIDI files for the training and evaluation. For every bar of the training material the sequence of drum notes is modeled with the time signature model $\mathcal{T}_{ts}$. Combined with the time signature progression model $\mathcal{P}$ it forms a recognition network, called the rhythm model $\mathcal{R}$. The rhythm model $\mathcal{R}$ can be used to recognize drum patterns in a sequence of unknown drum notes. The suitability of the model to describe rhythm could be shown with the help of a symbolic genre classification experiment, leading to $88.9\% \pm 2.8\%$ correctness for the four genres Rumba, Samba, Tango and Waltz.

Future work will concentrate on extending the approach to other instruments and using the model with real audio. Challenges involve the grid with fixed intervals, which has to be calculated from the audio material. This issue is addressed by work on the metrical analysis and the fastest pulse of music [22, 23]. However, the grid must not necessarily belong to one specific note type, which might be difficult to determine automatically, but could be defined as time interval range. For example, with the time intervals between 100 - 200ms, a song in 120 bpm would lead to a 16th note grid with $125ms$ grid interval and a song in 70 bpm to a 32th note grid with an interval of $107ms$.

Besides genre classification the rhythm model $\mathcal{R}$ could be used for similarity retrieval, by modeling exactly one song and getting a weight for drum note sequences of other songs within $\mathcal{R}$. Within the transcription process of drum notes from audio, $\mathcal{R}$ could serve as musical model [6]. The presented approach is a contribution towards a better modeling of the temporal features of music.

6. REFERENCES

[1] Simon Dixon, Fabien Gouyon, and Gerhard Widmer, "Towards characterization of music via rhythmic patterns," in *ISMIR, 5th International Society for Music Information Retrieval Conference*, Barcelona, Spain, 2004, pp. 509–516.

[2] Jouni Paulus and Anssi Klapuri, "Measuring the similarity of rhythmic patterns," in *ISMIR, 3rd International Society for Music Information Retrieval Conference*, Paris, France, 2002, pp. 150–156.

[3] Emiru Tsunoo, Nobutaka Ono, and Shigeki Sagayama, "Rhythm map: Extraction of unit rhythmic patterns and analysis of rhythmic structure from music acoustic signals," in *ICASSP, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Taipei, Taiwan, 2009, pp. 185–188.

[4] Daniel P.W. Ellis and John Arroyo, "Eigenrhythms: Drum pattern basis sets for classification and generation," in *ISMIR, 5th International Society for Music Information Retrieval Conference*, Barcelona, Spain, 2004, pp. 554–559.

[5] Olivier Gillet and Gaël Richard, "Automatic transcription of drum loops," in *ICASSP, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, may 2004, vol. 4, pp. 269–272.

[6] Jouni Paulus, *Signal Processing Methods for Drum Transcription and Music Structure Analysis*, Ph.D. thesis, Tampere University of Technology, 2009.

[7] Matthias Mauch and Simon Dixon, "A corpus-based study of rhythm patterns," in *ISMIR, 13th International Society of Music Information Retrieval Conference*, Porto, Portugal, 2012, pp. 163–168.

[8] Jouni K. Paulus and Anssi P. Klapuri, "Conventional and periodic n-grams in the transcription of drum sequences," in *Inproceedings of 2003 International Conference on Multimedia and Expo*, Baltimore, USA, 2003.

[9] Debora C. Correa, Jose H. Saito, and Luciano da F. Costa, "Musical genres: Beating to the rhythms of different drums," *New Journal of Physics*, vol. 12, no. 5, pp. 053030, 2010.

[10] Matthias Mauch, Daniel Müllensiefen, Simon Dixon, and Geraint Wiggins, "Can statistical language models be used for the analysis of harmonic progressions?," in *Proceedings of the 10th International Conference on Music Perception and Cognition*, Sapporo, Japan, 2008.

[11] Mehryar Mohri, Pedro J. Moreno, and Weinstein Eugene, "Efficient and robust music identification with weighted finite-state transducers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 18, no. 1, pp. 197 – 207, Jan. 2010.

[12] Jeremy Reed and Chin-Hui Lee, "On the importance of modeling temporal information in music tag annotation," in *ICASSP, Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, 2009, pp. 1873–1876.

[13] Mehryar Mohri, Fernando Pereira, and Michael Riley, *Springer Handbook of Speech Prosessing*, chapter Speech Recognition with Weighted Finite-State Transducers, pp. 559–583, Springer, 2008.

[14] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang, "A survey of audio-based music classification and annotation," *IEEE Transactions on Multimedia*, vol. 13, no. 2, pp. 303 – 319, 2011.

[15] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," in *IEEE Transactions on Speech and Audio Processing*, July 2002, vol. 10.

[16] Fabien Gouyon, Anssi P. Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano, "An experimental comparison of audio tempo induction algorithms," in *IEEE Transactions on Speech and Audio Processing*, 2006, vol. 14, pp. 1832–1844.

[17] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *AES, 25th International Conference*, London, UK, June 2004.

[18] Geoffroy Peeters, "Spectral and temporal periodicity representations of rhythm for the automatic classifcation of music audio signal," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, pp. 1242 –1252, 2011.

[19] Stephan Hübler and Rüdiger Hoffman, "A study on the metrical structure of music with similarity experiments," in *Speech Prosody, 6th International Conference*, Shanghai, China, 2012.

[20] Cori McKay, "Automatic genre classification of midi recordings," Master thesis, McGill University, Montreal, 2004.

[21] Thomas Lidy, Andreas Rauber, Antonio Pertusa, and Jose Manuel Inesta, "Improving genre classification by combination of audio and symbolic descriptors using a transcription system," in *ISMIR, 8th International Society for Music Information Retrieval Conference*, Vienna, Austria, 2007, pp. 61–66.

[22] Jarno Seppänen, Antti Eronen, and Jarmo Hiipakka, "Joint beat & tatum tracking from music signals," in *ISMIR, 7th International Society for Music Information Retrieval Conference*, Victoria, Canada, 2006, pp. 23–28.

[23] Anssi P. Klapuri, Antti J. Eronen, and Jaako T. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 342–355, Jan 2006.