Timbre-based rhythm analysis

Michael Blaß
Department of Systematic Musicology, University of Hamburg, Germany
michael.blass@uni-hamburg.de - http://www.fbkultur.uni-hamburg.de/de/sm/mitarbeiter/blass.html

In: Jakubowski, K., Farrugia, N., Floridou, G.A., & Gagen, J. (Eds.)
Proceedings of the 7th International Conference of Students of Systematic Musicology (SysMus14)
London, UK, 18-20 September 2014, http://www.musicmindbrain.com/#!sysmus-2014/cfmp

Musical rhythm is a complex experiences, which is structured in time. Furthermore, every musical event has a distinct sound. Thus, it is plausible that investigations in rhythm have to consider its sound, too. Commonly, sounds are discriminated by their timbre. Therefore, rhythm can be described as succession of distinct timbres. We developed a method to model drum patterns in such a manner. Timbre is approximated as a one-dimensional feature consisting of weighted spectral centroid. An onset detection algorithm based on fractal geometry determines the time frames of measurement within the input audio file. The resulting time series is used to train an *m*-state Hidden Markov Model. The model's transition probability matrix serves as a fingerprint of the sample's rhythm. This method can therefore be used to compare music quantitatively and to reveal and cluster musical similarities in sound recording archives.

Listeners of popular music are mostly confronted with a rhythmical base structure, contributed by drums and percussions, over which harmonies are played by different instruments. On top, there is a melody sung by a vocalist. This most typical genre of popular music, the song, is often said to have a certain *groove*. This notion is ambiguous and so is the research in groove. Since rhythm perception is tied to the perception of passing time, a lot of research has been done focusing on the temporal aspect of groove. For example, Frühauf et al. (2013) investigated in the influence of microtiming on groove perception. Madison (2006) conducted an adjective-rating study in order to reveal what is commonly connoted with music having groove. It is widely accepted that groove is a perceptual quality arising from certain rhythmical patterns. These patterns are said to strongly induce movements like tapping one's feet or nodding one's head to the beat of the music.

Besides that, rhythm patterns seem to develop their very own „feeling", when played by an appropriate instrument, e.g. a drum set. This feeling seems to be robust against definite changes in the pattern. For example, replacing the first hi-hat on the second quaver of Figure 2 by two semiquavers does not change the patterns feeling. The same holds when replacing all snares (quavers 3 and 7) by hand claps. However, swapping quaver 6 and 7 alters the feeling of the pattern significantly. From this simple and easily comprehensible example we can conclude, that the length of the inter onset intervals, the microtiming and the other time related parameters are not the only ones crucial for musical rhythm perception. In fact the sound and the order of the single events contribute a lot to how we perceive a rhythm pattern.

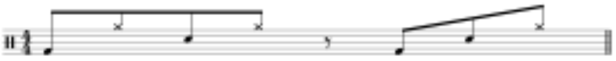

Figure 1. Simple drum pattern. The note F represents the bass drum, C the snare drum and the ghost note G the hi-hat.

Bader and Markuse (1994) found evidence that the perception of meter is influenced by the instruments involved in a pattern. Within a continuous stream of bass drum kicks, each one is rather perceived as on beat, whereas this does not hold for the hi-hat. It is mostly perceived as sounding off-beat. If the feeling of drum patterns can be fundamentally changed by swapping instruments and the choice of instrument influences the perception of meter, we can plausibly assume a mental timbre-related process with strong inducement to rhythm perception.

The matter in hand includes a model to analyze drum patterns in terms of the timbres occurring within them. It comprises of an onset detector, retrieving note onsets in an audio signal, a feature extraction procedure and a Hidden Markov model to investigate time series of quantified timbre representations.

Onset Detection

The crucial part of feature extraction is to determine the times of measurement. To this end we developed a onset detection algorithm. In this, we unitized common results of information theory (Shannon, 1998) in order to detect note onsets in digital audio signals. Each note played on a real instrument begins with a transient. A transient is a chaotic event. Chaos means unpredictable fast changes in the signal. Thus, chaos in audio signals is perceived as noise. Therefore, one approach to onset detection is to scan for regions with a high relative noise level. To this end, an input signal is segmented into a number of pieces of equal length *ls*. Each part is then embedded into a two-dimensional pseudo-phase space. For each segment the information entropy H_N can be calculated from its corresponding pseudo-phase space as

$$H_N = -\frac{1}{N}\sum_{i=o}^{N} p(x_i)\ \log\ p(x_i)$$

High levels of noise correspond to high levels of information. Therefore, transient regions in the musical signal represent regions of relative high entropy. A steady tone with little changes has a relative low entropy because the signal does not comprise much information. During a transient, however, the signal changes very fast in an unpredictable way, which leads to a high entropy. Having all the entropy values calculated, the problem of onset detection is reduced to detecting relative maxima in the time series of entropy. The described algorithm applies especially for detecting onsets of unpitched percussion instruments. With simple drum-only samples, the algorithm returns 100 % true positives. However, the performance decreases with musical complexity.

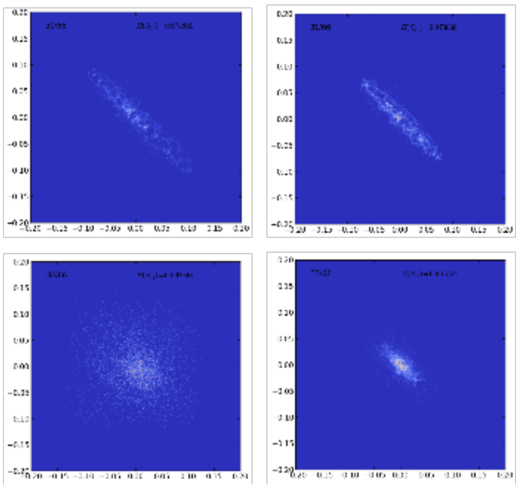

Figure 2. Representation of the pseudo-phase spaces of four successive related segments of an audio input. The lower left picture corresponds to a relative high entropy of H_{32} = *0.992048.* The related audio segment involves an onset.

Timbre features

Different instruments are discriminated by their timbres. Many multi dimensional scaling surveys concerning musical timbre showed that a timbre space of at most three dimensions is adequate to represent the essence of timbre perception. The interpretation of the physical correlates of the perceptual dimensions is similar in most studies (Grey, 1977; Wessel 1979; Inverson & Krumhansl, 1993; Hourdin et al. 1997; Lakatos, 2000). The timbre spaces comprise at least one temporal and one spatial dimension. The spatial dimension was always interpreted as brightness (Bader, 2013). Even experiments using semantic analysis found similar results (von Bismarck, 1974; Zacharakis, et al., 201). Thus, brightness seems to be one of the prominent dimensions of timbre perception. Especially when investigating percussion instruments, brightness seems to be a salient feature (Lakatos, 2000). Brightness strongly correlates with the spectral centroid, which is easily calculated as the „center of gravity" of a power spectrum. The model proposed here should be apt to analyzed drum patterns. In future versions it will be utilized to analyze large data bases of poorly recorded ethnographic audio data. It is thus important to keep its foundation performant but with a maximum precision. Therefore, a one-dimensional feature vector comprising only the spectral centroid is justified.

Model

The goal of this work is to develop a quantitative representation of the perceptional quality of a drum pattern. To this end, a given audio signal is first analyzed with the onset detection algorithm, which returns the audio frame indices at which an onset would be perceived. At these frames feature extraction calculates the spectral centroid and weights it by the maximal amplitude of the segment. The resulting time series is fed to the Hidden-Markov mode procedure. Figure 3 depicts the model's structure. HMMs are stochastic processes commonly used to model discrete as well as continuous valued time series data. They have recently been applied to diverse fields such as speech recognition (Rabiner, 1989), analysis of melody and rhythm (Mavromatis, 2004; Mavromatis, 2005; Mavromatis 2012), chord estimation (Lee & Slaney, 2006), live improvisation and human computer interaction with musical agents (von Nort et al. 2010; Braasch, 2013), segmentation (Aucouturier & Sandler, 2001) and automatic sound classification (Zhang & Jay Kuo, 1998; Zhang & Jay Kuo, 1999). Hidden Markov models are comprised of an unobserved Markov Chain as a state-dependent process and a random distribution representing the parameter process and producing visible observations. The key feature of Markov Chains is the Markov-Property, which links past with present events by loosening the assumption of independence. It states that the probability of a random variable *X* to be in a state *s* at a future time step *t+1* depends only on the present state of *X,* so that

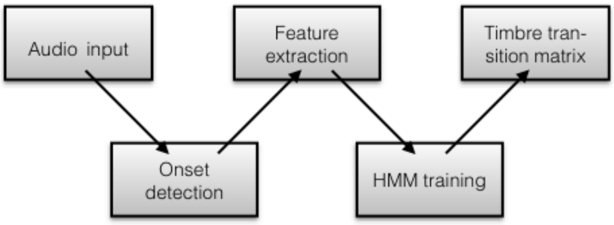

Figure 3. Data flow of the rhythm analysis model.

$$P(X_{t+1}|X_t,\ldots,X_0) = P(X_{t+1}|X_t)$$

holds for every $t \in$ N. Thus, for each possible state of *X* there is a probability of moving to each other state. These values are combined in the *transition probability matrix* **Γ**, where each row and column represents one of *m* given states. To each state, a random distribution is assigned producing observations according to a set of parameters. We combined a Markov Chain with a Poisson Mixture Model as proposed by Zucchini & MacDonald (2009). The latter takes only one parameter, which is the mean λ. Figure 3 depicts a simple 3-state HMM with Poisson mixture. Our basic assumption is that the spectral centroid time series is produced by a mixture of *m* Poisson distribution with means λ_i, i ∊ {1, …, *m*}. The selection of the mean is in turn governed by the hidden Markov Chain with an unknown transition probability matrix. Having in mind that perception supposedly places definite timbres to particular rhythmical positions (bass drum „belongs" to „on beat", etc.) we can model drum patterns as sequences of distinct timbres, whose order is dictated by a stochastic process. The hidden sequence of states then refers to a perceptual process that responds to changes in timbre. The computational task is therefore to estimate a set of parameters, which are in this case the entries of the transition probability matrix **Γ** and the vector of means **λ** from the a given sequence of spectral centroid values. This is done using the Baum-Welch algorithm (Baum, 1970).

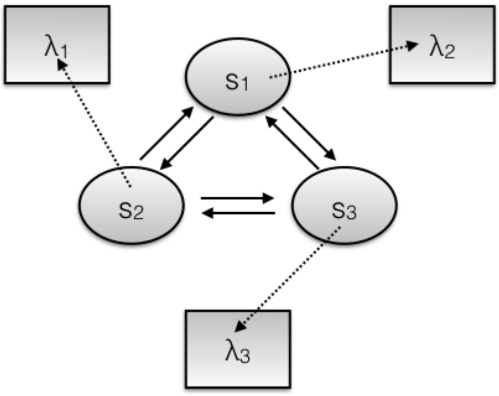

Figure 4. A 3-state Poisson-HMM. The ovals depict the sates of the hidden Markov Chain. The solid arrows represent possible transition between the states. Rectangles represent the Poisson distribution. The dashed arrows illustrate the affiliation of a distribution to a state.

Training data

We trained HMMs for 95 audio samples. These were generated using Apple's Garage Band sequencer. This software provides live recorded drum loops, which can be altered regarding two scales: *soft—loud* and *simple—complex.* Each scale has four discrete steps, so we generated 16 samples for each of five presets. All samples are eight bars long and have a duration between 19 and 24 seconds. They were delivered to the model as 16 bit .wav files of 44.1 kHz sample rate. Additionally, we analyzed two simple drum samples in order to ease the visualization. These samples were synthesized using GarageBand and then converted to audio files using the same specifications as above.

Results

Results will be presented using the two synthesized example files *beat.wav* and *edgy.wav*. In the first file the drum computer plays the rhythm pattern given in Figure 1. To visualize the model's performance we plotted the waveform of the input and coloured the areas where the spectral centroid was measured. Each colour represents a hidden state of the Markov Chain. The areas of measurement are much larger then the segments in which the onsets were found. Evaluation of the method revealed that the quality of the results strongly depends on the area of measurement. The detector localizes an onset during the transient, in an area of 800 sample points length. This parameter is adaptable and influences the precision. *During* the onset spectral centroid has another value than during steady state. Hence, we expanded the region of measurement to five times the length of one segment, which yielded good results while experimenting.

Local decoding of an HMM means calculating the most probable sequence of hidden states given a sequence of observations. That implies for our model to find for every measured spectral centroid value the poisson that most probably produced it. For each mean there is a connected and yet un-interpreted hidden state. Matching the sequence of the most probable states with observations made, reveals how the states are to be understood. Figure 5 shows a local decoding of a 3-state HMM, which was trained on *beat.wav*. Obviously the states can be interpreted as *bass drum* (blue), *snare drum* (yellow) and *hi-hat* (purple). This shows that the model can distinguish the instruments of a drum set in human-like way by only learning spectral centroids data. Furthermore, the model managed to assign every observation to the right state. The second result is not really surprising since the model was trained from the same data.

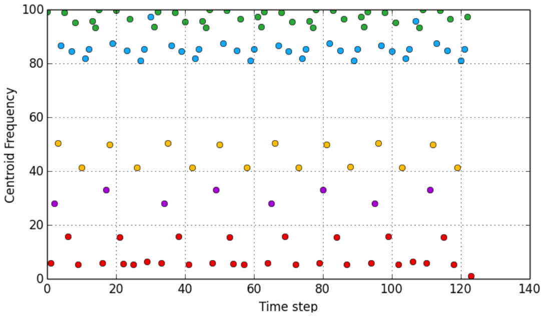

Figure 5. Local decoding of *beat.wav*. X-axis represents the frame index of the audio signal, y-axis the amplitude. The colors represent a hidden state of the Markov model: blue corresponds to the bass drum, yellow to the snare and purple to the hi-hat.

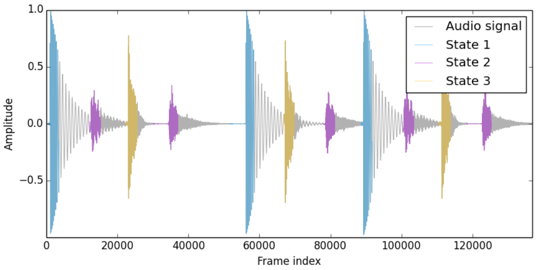

Figure 6: Centroid frequencies measured in *edgy.wav* plotted according their order of measurement. Clearly visible are the streams and characteristic patterns.

Interesting results can be delivered by plotting the values of the spectral centroids against the number of measurements, as in Figure 6 using a 5-state HMM of *edgy.wav*. As above, each colour refers to one hidden state.Again, there are several interesting things to mention. First, it is seems obvious that the model chooses the means in a plausible way. One can clearly recognize five separate streams. Second, within these streams there are apparently patterns, which refer to small but steady changes of the spectral centroid. This possibly represents slight changes of articulation. At this time we are unfortunately not able to explain this phenomenon. The third thing to mention is that model could derive 5 different states, but the samples it was trained on does only involve three different instruments: snare, bass drum and hi-hat. These extra states refer to mixture-timbres, which arise from two instruments played at the same time.

Conclusion

We proposed a timbre-based system in order to model and analyze rhythm patterns. The above results show that the model can recognize and distinguish different instruments according to their spectral centroid. Small changes in timbre, probably resulting from patterns of accents are assigned to the right mean. On the other hand, mixtures of timbres are assigned to an extra state. Given that mixtures arise only at certain positions on the pattern, the model learned hidden, timbre-related structures of the rhythmical pattern. In this light the proposed model is a promising candidate for rhythm fingerprinting. Still, much evaluation has to be done. At this time we work with covariance matrices of the resulting transition probabilities as well as on a self-organizing map to illustrate similarity between different models.

References

Aucouturier, J.J., & Sandler, M. (2001). Segmentation of musical signals using Hidden Markov models. *10th convention of the audio engineering society*.

Bader, R., Markuse, B. (1994). Perception and analyzing groove in popular music. *Proceedings of the 3rd international conference for music perception and cognition*, 401 – 402.

Bader, R. (2013). *Nonlinearities and Synchronization in Musical Acoustics and Music Psychology*.

Baum, L. Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique of probabilistic functions of Markov Chains. *The Annals of Mathematical Statistics*, 41(1), 164 – 1771.

Braasch, J. (2013). The μ cosm project: an introspective platform to study intelligent agents in the context of musical ensemble improvisation. Bader, R. *Sound - Perception - Performance*, 257 – 270.

Frühauf, J., Kopiez, R. & Platz, T. (2013). Music on the timing grid: The influence of microtiming on perceived groove quality of a simple drum pattern performance. *Musicae Scientiae 17: 246.*

Grey, J. M. (1977). Multidimensional perceptual scalings of musical timbres. *Journal of the Acoustical Society of America*, 61(5), 1270 – 1277.

Hourdin, C., G., Charbonneau, & Moussa, T. (1997). A Multidimensional Scaling Analysis of Musical Instruments' Time-Varying Spectra. *Computer Music Journal*, 21(2), 40 – 55.

Inverson, P., & Krumhansl, C.L. (1993). Isolatingthe dynamicattributesof musical-timbre. *Journal of the Acoustical Society of America*, 94(5), 2595 – 2603.

Lakatos, S. (2000). A Common Perceptual Space for Harmonic and Percussice Timbres. *Perception & Psychophysics*, 62(7), 1426 – 1439.

Lee, K., & Slaney, M. (2006). Automatic Chord Recognition from Audio Using an

HMM with Supervised Learning. *ISMIR Proceedings.*

Madison, G. (2006). Experiencing Groove Induced by Music: Consistency and Phenomenology. *Music Perception*, 24(2), 201-208.

Mavromatis, P. (2004). A Hidden Markov model of melody in greek church chant. *Proceedings of the 8th international conference on music perception and cognition*.

Mavromatis, P. (2005). A Hidden Markov model of melody production in greek church chant. *Computing in Musicology,* 14, 93 – 112.

Mavromatis, P. (2012). Exploring the rhythm of the palestrina style. *Journal of music theory*, 2(56), 169 – 223.

Rabiner, L. (1989). A tutorial to Hidden Markov models d selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257 – 286.

Shannon, C. E. (1998). Communication in the presence of noise. *Proceedings of the IEEE*, 86, 447-457.

van Nort, D. Braasch, J. & Oliveros, P. (2009). A system for musical improvisation combining sonic gestures recognition and genetic algorithms. *Proceedings of the 6th sound and music computing conference*, 131 – 136.

van Nort, D. Braasch, J. & Oliveros, P. (2010). Developing systems for improvisation based on listening. *Proceedings of the international computer music conference*, 108 – 115.

von Bismarck, G. (1974). Timbre of Steady Sounds: A Factorial Investigation of its Verbal Attributes. *Acustica*, 3(3), 146 – 159.

Wessel, D. L. (1979). Timbre Space as a Musical Control Structure. *Computer Music Journal*, 3(2), 45 – 52.

Zacharakis, A., Pastiadis, K., Papadelis, G., & Reiss, J.D. (2011). An Investigation of Musical Timbre: Uncovering Salient Semantic Descriptors and Perceptual Dimensions. Proceedings of the 11th International Society for Music Information Retrieval Conference.

Zhang, T. & Jay Kuo, C.C. (1998). Hierarchical system for content-based audio classification and retrieval. *SPIES's conference on multi- media storage and archiving systems III*, 398 – 409.

Zhang, T. & Jay Kuo, C.C. (1999). Heuristic approach for generic audio data segmentation and annotation. *ACM multimedia conference*, 67 – 76.

Zucchini, W., & MacDonald, I.L. (2009). *Hidden Markov Models for Time Series*.