SOUND RE-SYNTHESIS FROM RHYTHM PATTERN FEATURES – AUDIBLE INSIGHT INTO A MUSIC FEATURE EXTRACTION PROCESS

Thomas Lidy *Georg Pölzlbauer* *Andreas Rauber*
Vienna University of Technology
Department of Software Technology and Interactive Systems
Vienna, Austria
{lidy, poelzlbauer, rauber}@ifs.tuwien.ac.at

ABSTRACT

For tasks like musical genre identification and similarity searches in audio databases, audio files have to be described by suitable feature sets. Since these feature sets usually try to capture diverse discriminative characteristics, it is interesting and desirable to create an acoustic representation of the feature set to support intuitive evaluation. In this paper, we present an approach for making a specific feature set, namely Rhythm Patterns, instantly human comprehensible by re-assembling sound from the numerical descriptors. The re-synthesized audio chunks represent clearly perceivable rhythmical characteristics on critical frequency bands of the original music.

1. INTRODUCTION

The Music Information Retrieval research domain gained increasing attention in recent years. The sheer amount of music titles available in repositories calls for sophisticated search, retrieval and organization techniques. These, in turn, require representative descriptors for pieces of music in order to measure similarities between music titles. It is, however, unclear, what constitutes the essential "meaning" of music. Most approaches thus focus on perceptually relevant features of music. The descriptors, or features, are derived from the plain audio signal. Although substantial reduction of data is desired, the descriptors have to contain sufficient information from the data to represent some kind of semantics of the music.

Content-based descriptors build the basis for many information retrieval tasks, such as similarity based searches (query-by-example, query-by-humming, etc.), organization and clustering tasks, classification tasks, etc. Numerous different types of descriptors have been proposed. All these descriptors are more or less suitable as a representation of the content of audio, frequently depending on the specific application. The performance of features in similarity retrieval or classification tasks has been evaluated in numerous experiments, some of which showed that a combination of several feature sets improves the results.

In our work we concentrate on Rhythm Patterns as features, describing the loudness amplitude modulation in different frequency bands. The feature set does not merely represent rhythm, or beat, it describes fluctuations in numerous frequency regions covering the complete audible frequency range.

A question frequently raised, particularly for non-standard feature sets, is on the cognitive characteristics of the extracted numbers. What is it, that the features actually represent? For humans it is sometimes difficult to get a notion of the feature set as a whole, as the feature space is often high-dimensional. Relations between the attributes can be elusive, thus the chance for insight into the data is diminished. Since this issue of acoustic interpretation of the feature set has been raised several times since the feature set's inception [10], in this paper we present an acoustical re-synthesis of the feature set. We thus seek to make the numerical descriptors instantly comprehensible to humans allowing to verify characteristics present in the feature set intuitively. Furthermore, the synthesized sound can serve as a control technique for the feature extraction process and provides a notion of the suitability of the feature set for content-based description of musical data. With the audible feature set, one can evaluate the effectiveness of the feature extraction through asking a human for the same task as the computer, working only with the substantially reduced information from the aggregated descriptor, e.g.: Can you discriminate musical genres provided only with the information from the feature set?

In the evaluation of the re-synthesized "audible vectors" we experienced, that the rhythmical structure (i.e. amplitude modulation) on all frequency regions is satisfactorily reassembled. This serves as an indication, that the Rhythm Patterns we chose for content description prove suitable to represent characteristics of a given piece of audio, thus appearing appropriate for classification, organization and retrieval tasks.

The remainder of the paper is structured as follows: In Section 2 we give an overview of related work. Section 3 introduces the feature extraction algorithm that forms the base of the sound synthesis approach. The synthesis process is outlined in Section 4. Section 5 states evaluation results, followed by conclusions in Section 6.

2. RELATED WORK

In recent years, audio analysis received by far more attention than audio synthesis. As stated in the introduction,

this is due to the currently strong interest in music information retrieval tasks. Approaches for deriving content-based audio descriptors are manifold and include the extraction of tempo, beat [2, 5], rhythm [3], pitch [7, 14], and melody [4], to just name a few.

In the music information retrieval domain, clearly, there is little work on synthesis of audio. Recently, sound synthesis is applied in computer music and in conjunction with animation and art. A work about additive synthesis using the Inverse Fast Fourier Transformation, both of which is used in our approach, dates back to 1992 [12]. In [9] the authors present a method for preventing artefacts in the re-synthesis of a signal that was previously analysed using the Short Time Fourier Transform with window functions.

A recent method from the music information retrieval domain applies signal synthesis during automatic drum detection [15]. Drum sound is iteratively derived and re-synthesised for progressive detection of further percussive sound in the input signal. Signal analysis and subsequent synthesis is also applied in [1, 8], modelling the timbre of a musical instrument.

3. FEATURE EXTRACTION

The feature set our work is based on is denominated as "Rhythm Patterns". Describing amplitude modulations on various frequency bands covering the complete human audible frequency range, it contains far more than what is commonly considered as rhythm. The Rhythm Patterns features are derived analysing the spectral data of the music signal plus incorporating psycho-acoustic phenomenons. At the final stage they represent fluctuations per modulation frequency on 24 frequency bands according to human perception. The algorithm is described in detail in [11]. In the following, we give a brief outline of the extraction process, depicted in Figure 1.

The algorithm processes audio tracks in standard digital PCM format with 44.1 kHz sampling frequency as input. First, the audio track is segmented into pieces of 6 seconds length. [1] Then, a short time Fast Fourier Transform (STFT) is used to retrieve the energy per frequency bin, i.e. the spectrum, every 11.5 ms, resulting in a spectrogram of the 6 second segment. To reduce the amount of data, the frequency bins of the spectrogram are summed up to 24 so-called critical bands, according to the Bark scale [16]. A further psycho-acoustical phenomenon incorporated is spectral masking, i.e. the occlusion of one sound by another sound. This phenomenon is coped with a spreading function [13]. Successively, the data is transformed into the logarithmic decibel scale, equal-loudness curves are accounted for [16], resulting in a transformation into the unit Phon and afterwards into the unit Sone, reflecting the specific loudness sensation of the human auditory system. At this point, we computed the specific loudness sensation over time on 24 critical frequency bands. Still, we have a time-dependent signal, although reduced to 512 sample values at the time axis due to the window size in the STFT.

[1] The duration of the segment is actually 5.94 seconds, which has an appropriate number of samples (2^{18}) for effective processing through the two Fast Fourier Transforms. Nevertheless, we denote the segment "6 second segment" throughout the paper.

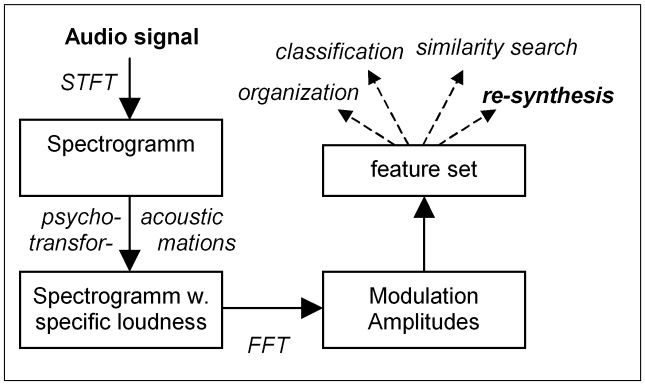

Figure 1. Block diagram of the feature extraction process. Arrows with broken lines do not form part of the feature extraction, but indicate typical post-extraction approaches. Our new approach is the re-synthesis of extracted feature sets.

In order to obtain a time-independent representation of the data, another Fourier Transform is applied. The idea is to regard the varying energy on a frequency band of the spectrogram as a modulation of the amplitude over time. With the second Fourier Transform, the spectrum of this modulation signal is computed. It is a time-invariant signal that denotes the modulation frequency on the abscissa, and the magnitude of modulation on the ordinate. A high amplitude at the modulation frequency of 2 Hz for example indicates a strong rhythm at 120 bpm (beats per minute = modulation frequency * 60). The abscissa ranges from 0.168 Hz to 43 Hz, with 43 Hz corresponding to 2580 bpm, which is far beyond what any auditory system is able to perceive as rhythm. The notion of rhythm already ends above 15 Hz where the sensation of roughness starts and goes up to 150 Hz, the limit where only three separately audible tones are perceivable. For that reason, the data used for the derived features is cut after a modulation frequency of 10 Hz, which means, that on each of the 24 critical bands, only 60 values are preserved. The final feature vector thus has 24*60 dimensions, containing a time-invariant representation of fluctuation strength between 0.168 Hz and 10 Hz. Subsequently, modulation amplitudes in that range are weighted according to a function of human sensation depending on modulation frequency, accentuating values around 4 Hz.

The 1440-dimensional feature vector represents a descriptor for rhythmical content of the musical signal. We retrieve one feature vector for each 6 second segment. It was shown [11], that averaging all feature vectors retrieved from a musical piece by using the median preserves sufficient characteristics of semantic structure for that given piece.

The resulting Rhythm Patterns feature set proved to be a reliable music descriptor in a number of different applications. Unsupervised learning is applied to the feature vectors to automatically produce a semantic organization on a map upon which intuitive visualization tech-

niques are applied (SOMeJB, the SOM-enhanced Jukebox [11]). The vectors have also successfully been used in music genre classification. A slightly adapted version of the algorithm achieved 82 % accuracy in the ISMIR 2004 Audio Description Contest on rhythm classification [6].

4. SOUND SYNTHESIS FROM FEATURES

Feature sets derived from audio signals represent an aggregation of the original audio data and are usually composed of a number of attributes, that should be mutually independent. In our approach the descriptor is a feature vector with 1440 dimensions. The challenge for classifiers is to find a possible separation between several classes within the feature space. In a high-dimensional space it is nearly impossible for humans to figure out the important attributes in the feature vectors, those best suitable to represent the semantics of the music. Nevertheless, it is desirable to obtain a notion of the content of the features, in order to be able to assess the quality of the feature set and its applicability to a specific task. A way to give insight to the feature set is to make it audible. Apart from giving insight into the features, it provides a possibility for monitoring the feature extraction process, and additionally, it enables the monitoring person to check if he or she as a human would be able to discriminate between classes (such as genres) provided only with feature data.

The process of re-synthesizing sound from the feature vectors seems quite straight-forward. Nevertheless, it faces a number of challenges. The original audio data has a frequency range of 20 to 20000 Hz. The feature vectors represent only 24 critical bands of that data, which is an enormous reduction of information. Moreover, the distribution of energy on frequency bands within a specific critical band is unknown, which means that the only frequency information that we can use as the base for re-synthesis are the centre frequencies of each critical band. During the feature extraction process, the major part of the modulation amplitude information is discarded, only 60 values per band according to modulation frequencies from 0.168 Hz to 10 Hz are retained. From those 60 descriptors of the 24 critical bands we now want to synthesize a sound signal, that resembles the rhythmical structure of the original piece of audio. As the numbers were derived using an FFT, an Inverse Fourier Transform qualifies as a proper method for the synthesis of the modulation signal. As there is no exact information about the original signal, we take the centre frequency f_i of that critical band, i.e.

$$f_i = c_{i-1} + (c_i - c_{i-1})/2 \qquad (1)$$

as the frequency of the base signal of critical band i, where c_i is the upper band limit of critical band i, c_0 being 0. The minimum modulation frequency is 0.168 Hz, so we will have to re-synthesize a sound of 6 seconds length to accommodate 1 period of the lowest modulation frequency. The final signal thus contains 2^{18} samples. In order to apply the Inverse Fourier Transform, we create a "virtual" spectrum with $N = 2^{18}$, with all spectrum bins being zero, except the first 60 values, where the spectrum

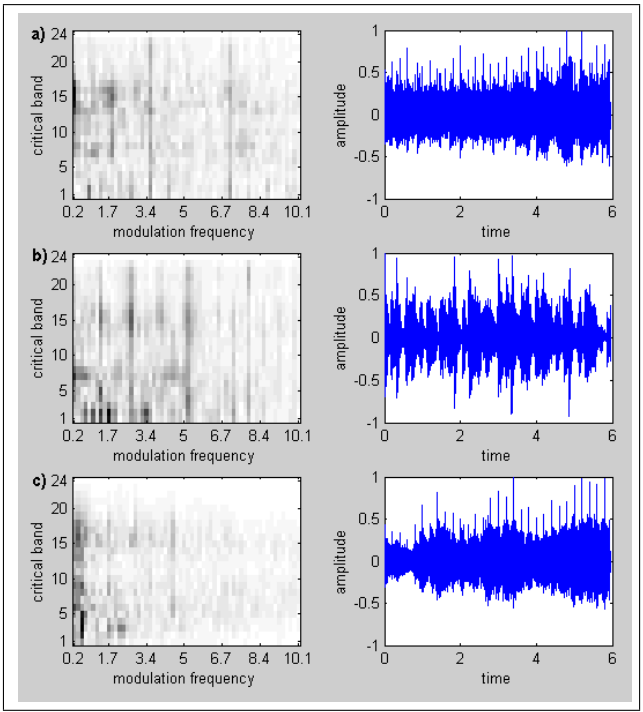

Figure 2. Feature representation: left column: Rhythm Patterns descriptor plot, right column: re-synthesized waveform, for a) pop music, b) reggae, c) classical music.

data of the preserved feature values are used. Additionally, bins $(N - 59)$ to N are generated by mirroring bins 2 through 60 and building the complex conjugates of their values (bin 1 contains the DC component). On that spectrum, we apply the inverse Fourier Transform. The output is the modulation signal of one frequency band, $m_i[t]$, $t \in N$. This signal can now be used to modulate the centre frequency of the critical band, f_i. We, however, do not know about the original amplitude of the signal on that band. We utilize the DC component of the modulation signal, which in general is >0, as the base amplitude A_i of the band signal. Therewith, we can modulate the band centre frequency f_i with the modulation signal. The same process is accomplished on all 24 critical frequency bands and the modulated signals are heterodyned, as in Eq. 2.

$$s[t] = \sum_{i=1}^{24} A_i \times cos(2\pi f_i t) \times m_i[t] \qquad (2)$$

The resulting signal $s[t]$ reflects the structure of the original piece of audio and resembles fluctuations within the critical bands as captured by the feature extraction process. Figure 2 shows plots of three Rhythm Patterns descriptor sets derived from content-based analysis of three music titles from the genres pop, reggae and classics, and the respective waveform plots of the re-synthesized audio. The waveform also reflects the rhythmical structure contained in the descriptor and the audible counterpart establishes the beat of the original piece of music, at least in beat-intensive titles such as pop and reggae songs, as well as modulations of amplitude on all freqof theuency regions. Also, the low modulation of the classical tune is perceivably re-synthesized.

5. EVALUATION

First evaluations showed, that the rhythmic structure of audio pieces can be recognized in the re-synthesized signal. Besides acoustically verifying the sound and comparing the rhythm of the re-synthesized sound with that of the original music through listening, the rhythmic structure can also be seen in the visualization of the waveform (see Figure 2). Music where strong beats do not play an important role (e.g. classical music) can clearly be discriminated from other genres. When a specific rhythm in the form of drums and beats differs by definition from one genre to another genre (e.g. Hip-Hop versus Reggae versus Drum'n'Bass), the genres can easily be distinguished in the acoustical representation of the feature vectors. This result argues in favor of the Rhythm Patterns approach we chose for our feature extraction algorithm and confirms its strong performance in music IR tasks. We emphasize, that (although much of information is cut away) the Rhythm Pattern features contain far more than the conventional definition of rhythm. The features reflect variation (fluctuation) on numerous specific audible frequency regions, up to very high modulation frequencies. This is an explanation why during the evaluation of re-synthesized feature sounds we found, that it is possible to even perceive the voice of a singer in a re-synthesized 6 second sound.

6. SUMMARY AND FUTURE WORK

Content-based descriptors of music are the core of every music information retrieval task. Many of the established descriptors, or features, typically differ heavily regarding both their dimensionality and the kind of semantics they try to capture. We present an unconventional way for the evaluation of content-based descriptors used in music information retrieval tasks. Making the feature sets audible enables a person to instantly get a notion of the content of the feature set. We perform a re-synthesis of the Rhythm Patterns feature set. Re-synthesis faces a number of challenges as a direct reversal of the feature extraction process is not possible. Nevertheless, we show that it is possible to synthesize sound from the feature set, that still contains sufficient information in order to enable recognition of typical characteristics in the audio.

Future work will include investigation on how to cope with the non-linear transformations performed throughout the feature extraction. Some frequency bands currently are too dominant in the signal, producing a metallic sound. Further investigation in how to calculate the appropriate relation between magnitudes of the modulation signal and the centre frequency signal should improve the perceived quality of the re-synthesized signal.

7. REFERENCES

[1] H. De Paula, M. Loureiro, and H. Yehia. Timbre representation of a single musical instrument. In *Proc. Intl. Computer Music Conf.*, Miami, USA, 2004.

[2] S. Dixon. An interactive beat tracking and visualisation system. In *Proc. Intl. Computer Music Conf.*, Havana, Cuba, 2001.

[3] S. Dixon, E. Pampalk, and G. Widmer. Classification of dance music by periodicity patterns. In *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, Baltimore, USA, 2003.

[4] E. Gomez, A. Klapuri, and B. Meudic. Melody description and extraction in the context of music content processing. *J. New Music Research*, 32(1), 2003.

[5] M. Goto and Y. Muraoka. A real-time beat tracking system for audio signals. In *Proc. Intl. Computer Music Conf.*, Banff, Canada, 1995.

[6] ISMIR 2004 Audio Description Contest. Website, 2004. `http://ismir2004.ismir.net/ISMIR_Contest.html`.

[7] P. Lepain. Polyphonic pitch extraction from musical signals. *J. New Music Research*, 28(4), 1999.

[8] M. Loureiro, H. De Paula, and H. Yehia. Timbre classification of a single musical instrument. In *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004.

[9] P. Masri and A. Bateman. Improved modeling of attack transients in music analysis-resynthesis. In *Proc. Intl. Computer Music Conf.*, Hong Kong, 1996.

[10] A. Rauber and M. Frühwirth. Automatically analyzing and organizing music archives. In *Proc. European Conf. on Digital Libraries (ECDL)*, Darmstadt, Germany, 2001.

[11] A. Rauber, E. Pampalk, and D. Merkl. The SOM-enhanced JukeBox: Organization and visualization of music collections based on perceptual models. *J. New Music Research*, 32(2):193–210, 2003.

[12] X. Rodet and Ph. Depalle. Spectral envelopes and inverse FFT synthesis. In *93rd Convention of the Audio Engineering Society*, New York, USA, 1992.

[13] M.R. Schröder, B.S. Atal, and J.L. Hall. Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66:1647–1652, 1979.

[14] G. Tzanetakis, A. Ermolinskyi, and P. Cook. Pitch histograms in audio and symbolic music information retrieval. In *Proc. Intl. Conf. on Music Information Retrieval (ISMIR)*, Paris, France, 2002.

[15] A. Zils, F. Pachet, O. Delerue, and F. Gouyon. Automatic extraction of drum tracks from polyphonic music signals. In *Proc. Intl. Conf. on WEB Delivering of Music*, Darmstadt, Germany, 2002.

[16] E. Zwicker and H. Fastl. *Psychoacoustics, Facts and Models*, Springer, Berlin, 1999.