TIMBRE AND RHYTHMIC TRAP-TANDEM FEATURES FOR MUSIC INFORMATION RETRIEVAL

Nicolas Scaringella
Idiap Research Institute, Martigny, Switzerland
Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
nicolas.scaringella@epfl.ch

ABSTRACT

The enormous growth of digital music databases has led to a comparable growth in the need for methods that help users organize and access such information. One area in particular that has seen much recent research activity is the use of automated techniques to describe audio content and to allow for its identification, browsing and retrieval. Conventional approaches to music content description rely on features characterizing the shape of the signal spectrum in relatively short-term frames. In the context of Automatic Speech Recognition (ASR), Hermansky [7] described an interesting alternative to short-term spectrum features, the TRAP-TANDEM approach which uses long-term band-limited features trained in a supervised fashion. We adapt this idea to the specific case of music signals and propose a generic system for the description of temporal patterns. The same system with different settings is able to extract features describing either timbre or rhythmic content. The quality of the generated features is demonstrated in a set of music retrieval experiments and compared to other state-of-the-art models.

1 INTRODUCTION

As discussed in [2], most state-of-the-art algorithms dedicated to the high-level description of music signals rely on the same basic architecture with different algorithm variants and parameters, i.e. short-term audio features extraction followed by some supervised or unsupervised machine learning algorithm. The most successful approaches rely on features describing the shape of short-term spectral frames of audio signal. Spectral shape is indeed known to be correlated with the perceived timbre of sounds as confirmed by recent experiments in isolated instrument recognition (see notably [8]). As a matter of fact, short-term spectral envelopes have dominated similar research fields for years, like e.g. ASR. These short-term spectral envelopes are typically transformed in accordance with some constraining properties of human hearing such as the nonlinear (critical-band like) frequency resolution (Bark, Mel), the compressive nonlinearity between acoustic stimulus and its percept (logarithm) or the decreasing sensitivity of hearing at lower frequencies (equal-loudness curves). Moreover, these modified spectral frames are typically projected on spectral basis that decorrelate the feature space (cepstrum).

In both speech and music signals, the smallest unit of information, i.e. phonemes on the one hand and notes on the other hand (not only pitched notes but also hits of percussive instruments or whatever that produces sounds), spread on longer time intervals than the usual short-term audio frame. Indeed, typical ASR/MIR (Music Information Retrieval) systems considers slices of audio signals of length 20 to 50 ms (slightly longer when accurate pitch estimates are needed in the lower frequencies). On the contrary, phonemes were demonstrated to spread at least over the interval 200-300ms [26]. As a matter of fact, the minimum discriminable inter onset interval (IOI) is estimated to lie within the range 50-100ms (i.e. two sounds separated by less than the minimum IOI will be perceived as one) so that it is likely that at least 50-100ms of information is needed by a human listener to interpret the incoming sound. Studies on rhythm perception show that the rate of information in music signals is even less. Experiments [16] have indicated that pulse sensation cease to exist outside of the period range of 200-2000ms which is known as the *region of pulse sensation* while the most natural foot-tapping period is approximately 500-600ms. Given these observations, it is reasonable to think that it is probably more perceptually relevant to model audio signals with a longer context than the usual 20-50ms spectrum frames. To preserve information related to the short-term dynamics of sounds and to keep sufficient time-resolution when detecting e.g. musical note onsets, a trade-off consists in building a model of the sequence of short-term feature vectors over a longer time-scale. This process is sometimes referred to as temporal feature integration [14].

The simplest approach consists in computing simple statistics of feature vectors (means and variances) over *texture-windows*. This has been shown [25] to significantly improve music genres classification accuracy when using windows of approximately 1.0 second as opposed to the direct use of short-term frames. However, simple statistics discard dynamical changes of short-term features while the dynamics of sound and notably the attack time and the fluctuations of the spectral envelope over time have proved to be of a great importance in the perception of individual instrument notes (see [11]). Meng [14] modeled the evolution of features over a texture window with an auto-regressive (AR) model and got improved genre classification results than with simple statistics. McKinney and Breebart [12] computed a periodogram for each short-term feature dimension over a frame corresponding to roughly 768ms. Each periodogram was then summarized by its power in 4 predefined frequency bands using a fixed filter bank. This approach was pursued by Arenas-Garcia et al. [1] who trained the filter-bank in a supervised fashion to optimally suit a particular music organization task. Rauber et al. [20] used critical band energies periodograms with a much longer context, i.e. 6 seconds. This longer context was considered to model rhythmic patterns and the range 0-10Hz was considered as higher values are beyond what humans can perceive as rhythm (see again the region of pulse sensation). These approaches to temporal feature integration model the dynamics of each feature independently. Though Meng [14] describes a general multivariate autoregressive model that does take into account correlations between feature trajectories, for the sake of simplicity he experiments in practice with a diagonal multivariate autoregressive model, i.e. an AR model of each feature dimension. Pohle et al. [19] use independent component analysis on short sequences of critical band energies and obtain time-frequency 2D filters that are reminiscent of cortical receptive fields [4]. Though this approach seems more appropriate to take into account correlations between feature trajectories, it is at best of similar quality as short-term features in genre classification experiments.

As a matter of fact, the use of long-term features has been investigated more in depth in the context ASR, notably by Hermansky [7]. These features are extracted in 2 steps. Firstly, rather long-term TempoRAL Patterns (TRAPs) of band-limited (1-3 Bark) spectral energies are considered. Though, a context of 200-300ms seems needed for ASR, an even longer time interval of 1 second is considered so that information about slowly varying noise can be removed from the data (i.e. mean/standard deviation normalization). Hermansky [7] argues that, consistently with color separation in vision, it seems likely that the frequency selectivity of the hearing system is used for separating the reliable (high SNR) part of the signals from the unreliable ones, so that it seems reasonable to use independent frequency localized processors. Consequently, each band-limited TRAP is processed individually by a specifically trained system to build as much knowledge as possible into the feature extraction module to minimize the complexity of the subsequent stochastically trained models. The second step, referred to as the TANDEM part, consists in training a system aiming at the combination of frequency-localized evidence into a set of robust features that can be used in conventional HMM-based ASR systems.

To our knowledge, there's been only one application of these TRAP-TANDEM features to music signals in the context of drum transcription [17]. In this paper, we further investigate some possible applications of the TRAP-TANDEM approach in the context of music information retrieval. More specifically, we describe in section 2 our own implementation of the TRAP-TANDEM feature extraction module, which slightly differs from the original method. As a matter of fact, we propose two different implementations to focus on two different aspects of music signals, namely timbre and rhythm. In section 3, we evaluate the validity of these features for music information retrieval in a set of music clustering experiments. Section 4 reaches conclusion.

2 MUSICAL TRAP-TANDEM FEATURES

The first step of the processing consists in converting the audio signal into some time-frequency representation. In practice, we use the typical short-term Fourier transform (STFT) with Hann windowing of the short-term audio frames. The resulting short-term spectra are projected onto the Mel scale to simulate human critical bands of hearing [13]. The perceptual relevance of this time-frequency representation is further improved by exploiting masking properties of the human ear (see [22]) and frequency response of the outer and middle ear (see [23]). The loudness in Sone of the spectra is finally evaluated according to [3].

These short-term spectra will be later used to build timbre related TRAPs. Rhythmic TRAPs are based on a slightly different representation. More specifically, each critical band of the time-frequency plane is smoothed with a kernel, which width is taken in the range of the minimum IOI so that rhythmically irrelevant details of the band envelopes will be smoothed while the peaks corresponding to two different note onsets will not be merged. The first order derivative of each smoothed critical band is then taken to emphasize sudden changes. Critical bands are finally combined as suggested by Scheirer [21] who demonstrated that the perceived rhythmic content of many types of musical excerpts was mainly contained in 4 to 6 larger critical bands.

To reduce further processing, we deploy a note onset detection algorithm and we will only compute one TRAP feature vector per onsets instead of using a constant and faster rate of feature extraction. By synchronizing the analysis on detected onsets, an important factor of variability of the data is strongly reduced, i.e. the data is made translation invariant, and consequently, we can expect that the task of learning relevant features will be simplified. The differentiated signals computed for rhythm description are used as a basis to detect note onsets. The signals are first half-wave rectified and peaks are detected by using an adaptive threshold to account for possible loudness changes over the course of the musical piece. Peaks are first combined over the different

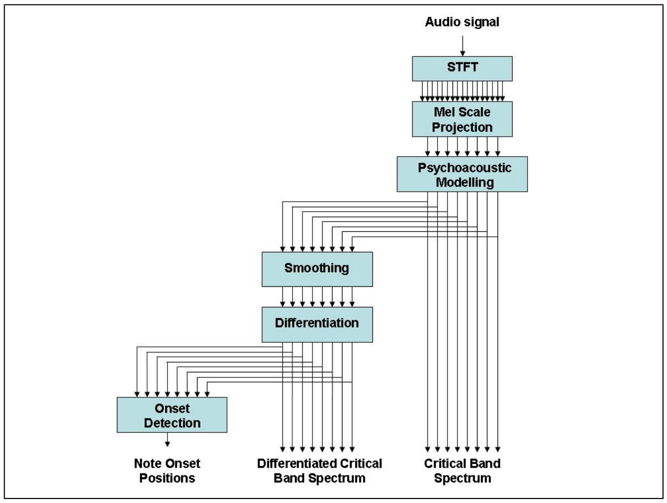

Figure 1. From audio signal to critical band spectra and onset positions.

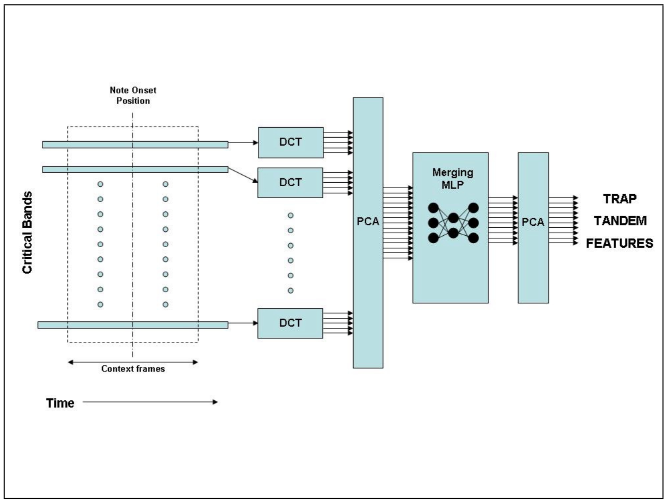

Figure 2. The TRAP-TANDEM feature extraction processing chain.

bands and those closer than the minimum IOI are merged together. Figure 1 illustrates the processing chain that goes from the audio signal to both critical band spectra and onset positions.

The data from each critical band and differentiated critical band surrounding each onset is then parameterized with the cosine transform, which has the good property of producing decorrelated signals. The cosine transform is simple to deploy since it does not need any training phase, plus it has the interesting property that it closely resembles the Principal Component Analysis (PCA) of such critical band signals [7]. The cosine transform also allows for a significant reduction of the dimensionality of the input data. The range of modulation frequencies of the cosines is carefully selected. For timbre description, modulations between 4 and up to 100 Hz can be considered. The lower limit of 4 Hz is set in accordance with the smallest perceivable sound unit discussed in section 1. The higher limit is in the range of modulations contributing to perceptual roughness, which is generally thought to be a primary component of musical dissonance [24]. As a matter of fact, the percept of roughness is considered to correlate with temporal envelope modulations in the range of about 20-150 Hz and maximal at 70 Hz. For rhythm description, lower frequencies are considered. Modulations between 1 and 4 Hz are interesting since they are of the order of typical beat rates, i.e. 60 to 240 Beats Per Minute (BPM).

In the original TRAP approach, for each critical band some algorithm, typically a non-linear feedforward multilayer perceptron (MLP), is trained to estimate posterior probabilities of the classes under consideration. In ASR, the targets are phonemes and there exist plenty of annotated datasets. Ideally, we would like to have instrument annotations to translate the TRAP idea to the case of music signals. Unfortunately, no such dataset exists for real-world polyphonic music and as a matter of fact, the annotation problem would become even more complex since we're considering mixture of instruments. For the time being, we left aside the use of critical band MLPs.

The TANDEM part of the system is in charge of combining evidence from the different frequency bands into a single estimate. A MLP is typically trained to combined these band limited features into a set of class posterior probabilities. Again, we lack appropriate annotated datasets. As an alternative, we use music genres annotations that are much cheaper to obtain, using e.g. some online music guide such as the AllMusic guide[1]. The merging MLP is fed with the concatenation of all band limited features, which are whitened with PCA to make them decorrelated. The MLP is trained to associate acoustic evidence with 50-dimensional binary vectors for which each dimension corresponds to a particular music genres. Notice that one song may be characterised by multiple genres. Once properly trained, the outputs of the MLP are decorrelated by PCA and can be used as a feature vector describing the different genres in which the acoustic data has been observed. Figure 2 illustrates the processing chain that goes from the critical band or differentiated critical band spectra to the final TRAP-TANDEM like features.

3 EVALUATIONS

The TRAP-TANDEM features describe the timbre and rhythmic context of a note onset. They may need to be aggregated into a song-level model in e.g. song retrieval applications. One application scenario consists in retrieving sets of songs

[1] http://www.allmusic.com

similar to some query song according to some similarity measure between song-level models. Though listening tests have proved to be a valid evaluation method of such music retrieval systems due to the consistency of judgements observed over different listeners [15], they are very demanding in terms of time and human resources. Previous works have shown that evaluations based on genre data correspond to evaluations based on similarity ratings gathered in listening tests [15]. Consequently, we will base our evaluation of the descriptive quality of our features on some genre annotated data. However, since we are more interested in music retrieval than automatic genre labelling, we will use measures of the ranking quality of the system rather than classification accuracy.

3.1 Timbre TRAP-TANDEM evaluation

To evaluate the timbre TRAP-TANDEM features, we have gathered a set of 210 songs annotated with genre and styles from the AllMusic guide. The quality of the labels was cross-checked by systematic listening tests. Each selected song is performed by a different artist to avoid the *album effect* [9] in the evaluation. Moreover, the songs selected were not previously used while training the TRAP-TANDEM feature extractor. Six main genre clusters of songs are considered and each genre cluster is composed of a set of smaller style clusters with 10 songs per style. The **Rock** cluster is composed of the *Grunge*, *British-Invasion*, *Punk-Blues*, *Glam-Rock*, *New-Wave*, *Folk-Rock* sub-clusters. The **Jazz** cluster is composed of the *Soul-Jazz*, *Swing*, *Hard-Bop*, *Free-Jazz* sub-clusters. The **Hip-Hop** cluster is composed of the *West-Coast*, *East-Coast*, *Turntablism*, *Old-School* sub-clusters. The **Electronica** cluster is composed of the *House*, *Trip-Hop*, *Drum'n'Bass* sub-clusters. The last two clusters, **Soul** and **Adult Contemporary**, are both divided into two sub-clusters according to the gender of the lead singer.

We will measure how well the timbre TRAP-TANDEM features are able to recover this organisation in terms of genre/style clusters. In practice, we summarize the distribution of TRAP-TANDEM feature vectors over each song by a simple average. Though more complex models, like e.g. HMM, could be deployed, our goal is to demonstrate the descriptive quality of the TRAP-TANDEM features so that we leave more complex song-level modelling strategies to future work. Two songs can then be compared by evaluating the cosine similarity of their average song-level TRAP-TANDEM vectors. The similarity between each pair of songs of the dataset is computed and the quality of the system is assessed by comparing the labels of a song and its nearest neighbours. More specifically, the precision at 1, 2 and 3 are computed for each query and averaged over the dataset. The precision at *n* is a quality measure commonly used to evaluate information retrieval systems. It accounts for the quality of ranking, i.e. it is high if the most relevant hits are in the top documents returned for a query. It is measured by computing the precision at different cut-off points (for example, if the top 10 documents are all relevant to the query and the next ten are all nonrelevant, we have 100% precision at a cut off of 10 documents but a 50% precision at a cut off of 20 documents).

Table 1. Timbre TRAP-TANDEM features.

	Prec. at 1	Prec. at 2	Prec. at 3
Genre	72.86	71.90	69.84
Style	50.00	43.10	39.37

Table 2. Full Gaussian of MFCCs.

	Prec. at 1	Prec. at 2	Prec. at 3
Genre	65.24	57.62	54.13
Style	35.71	30.00	25.87

To ease the comparison of our approach with the state-of-the-art, we have implemented one of the most popular timbre similarity measure based on spectral shape features. It simply consists of a Mel-Frequency Cepstrum Coefficients (MFCCs) frame-based parameterization of the audio signal (20 coefficients including the energy coefficient). These features are aggregated over the song as a Gaussian with full covariance matrix and compared using the symmetric version of the Kullback-Leibler (KL) divergence. This approach has been originally introduced by Mandel and Ellis [10] and was used in the winning algorithm of the 1st Annual Music Information Retrieval Evaluation exchange (MIREX 2005) [2] artist identification contest and ranked 3rd on 13 at the MIREX 2005 music genres classification contest while being almost 3 times faster than the first two winning algorithms. It can be considered as a simplified, yet competitive, implementation of Aucouturier's timbre model [2].

Tables 1 and 2 summarize the average precision at 1, 2 and 3 for both models. Results are given for both genres and styles targets, i.e. in the first case precision increases if the nearest neighbours are of the same genre, and in the second case precision increases if the nearest neighbours are of the same style.

It is clear from this experiment that the TRAP-TANDEM features are more reliable than the short-term spectral shape features in a music retrieval context. It is worth noticing that MFCCs with a simple average and cosine similarity leads to poorer results than MFCCs with mean, full covariance matrix and KL divergence, while for the TRAP-TANDEM features, the use of a KL-based distance function leads to slightly inferior results than the simpler cosine similarity.

[1] 1

[2] http://www.music-ir.org/mirex2005/index.php/Main_Page

Table 3. Rhythm TRAP-TANDEM features.

	Prec. at 1	Prec. at 2	Prec. at 3
Style	79.37	77.01	76.36

Table 4. 10-fold 1-NN classification accuracy.

	Accuracy
Rhythm TRAP-TANDEM features (without annotated tempo)	79.49
Gouyon & Dixon (without annotated tempo	67.60
Gouyon & Dixon (with annotated tempo)	82.10
Peeters (with annotated tempo)	90.40
Dixon et al. (with annotated tempo and semi-automated beat tracking)	96.00

This suggests that it is possible to have better retrieval results with a simpler—and especially faster—similarity function. Indeed, even if we're considering here 50-dimensional TRAP-TANDEM vectors against 20-dimensional MFCC vectors, the cosine similarity remains much faster than the symmetric KL of two full Gaussians since the later requires some matrix multiplications. This advantage becomes crucial when dealing with industrial databases with millions of songs.

3.2 Rhythm TRAP-TANDEM evaluation

We will evaluate the descriptive power of the rhythm TRAP-TANDEM features in a similar fashion. Rhythm TRAP-TANDEM features are averaged over each song and two songs are compared with the cosine similarity. We used here a well known dataset that contains 698 pieces of ballroom dance music divided into 8 sub-styles having different rhythmic characteristics, namely **Cha Cha Cha**, **Jive**, **Quickstep**, **Rumba**, **Samba**, **Tango**, **Viennese Waltz** and **Waltz**. It is interesting to notice that the TRAP-TANDEM system for rhythm was trained with the same 50 genres targets as the system for timbre, and that these genres are far from being as restricted as **Jive** or **Cha Cha Cha**. Table 3 summarizes the results obtained.

To ease the comparison with state-of-the-art algorithms, we also computed the classification accuracy on a 10-fold cross validation experiment with a 1-Nearest Neighbour classifier since various authors have reported results on this dataset using this particular evaluation procedure (see table 4). Gouyon and Dixon [6] reports up to 82.10% accuracy using a set of rhythmic features including the manually annotated tempo. The accuracy drops to 67.60% when using the tempo automatically extracted by their algorithm. Peeters [18] reaches up to 90.40% using *spectral rhythm patterns* normalised by the manually annotated tempo. While these two algorithms extract features from some periodicity representation of the audio signal, Dixon et al. [5] characterise the amplitude envelope of musical patterns, synchronised on musical bar positions and normalised to a reference tempo. They obtain an impressive 96.00% accuracy, but they also make use of the annotated tempo and a semi-automated beat tracking algorithm. On the contrary, our approach is fully automatic and reaches 79.49% classification accuracy. Moreover the results obtained by Dixon et al. with tempo normalised/bar synchronised temporal patterns suggest that the TRAP-TANDEM rhythmic features could become even more effective if synchronised on higher level musical events (musical bar positions instead of note onsets) and if made independent of the tempo. However, though on this particular dataset, a tempo normalisation may be needed since the clusters exhibit clearly defined rhythmic patterns with variable tempi, a tempo normalisation may not be so crucial for a general purpose music similarity engine since *slow*/*fast* songs should probably be similar to other *slow*/*fast* songs independently from the rhythmical pattern they're built on, i.e. the percept of speed would be more important than the perception of a particular rhythmical pattern.

4 CONCLUSION

We have presented a new set of features for music content description based on the work by Hermansky [7] in the context of ASR. The original approach has been adapted to the specific case of music signals and two different implementations based on the same architecture have been proposed to describe two apparently dissimilar dimensions of music, namely timbre and rhythmic patterns. Instead of using a relatively simple low-level characterization of the audio signal (like e.g. MFCCs), the TRAP-TANDEM approach is a complex feature extraction module that encodes temporal patterns and as much prior knowledge as possible. The distribution of TRAP-TANDEM features over a song can be described with simple models that can be used together with fast similarity measures. First experimental results confirm that the TRAP-TANDEM approach is competitive against state-of-the-art algorithms. Future work will focus on experimenting at a larger scale to confirm—or infirm—the descriptive quality of the TRAP-TANDEM features.

5 REFERENCES

[1] J. Arenas-Garcia, J. Larsen, L. Kai Hansen, A. Meng, “Optimal filtering of dynamics in short-time features for music organization”, in Proc. ISMIR 2006, pp. 290-295, Victoria, Canada, 2006.

[2] J.J. Aucouturier, “Dix Expériences sur la Modélisation

du Timbre Polyphonique", Ph.D. Dissertation, Université Paris 6, France, 2006.

[3] R.A.W. Bladon, B. Lindblom, "Modeling the judgment of vowel quality differences", in J. Acoustical Society of America, vol. 69, no.5, pp. 1414-1422, 1981.

[4] D.D. Depireux, J.Z. Simon, D.J. Klein, S.S. Shamma, "Spectro-temporal response fields characterization with dynamic ripples in ferret primary auditory cortex", in J. Neurophysiology, vol. 85, pp. 1220-1234, 2001.

[5] S. Dixon, F. Gouyon, G. Widmer, "Towards Characterisation of Music via Rhythmic Patterns", in Proc. ISMIR 2004, pp. 509-516, Barcelona, Spain, 2004.

[6] F. Gouyon, S. Dixon, "Dance Music Classification: a Tempo-Based Approach", in Proc. ISMIR 2004, pp. 501-504, Barcelona, Spain, 2004.

[7] H. Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech", in Proc. IEEE. ASRU-2003, no. 50, St. Thomas, Virgin Islands, 2003.

[8] P. Herrera-Boyer, G. Peeters, S. Dubnov, "Automatic classification of musical instrument sounds", in J. of New Music Research, no. 32 (2), pp. 3-21, 2003.

[9] Y.E. Kim, D.S. Williamson, S. Philli, "Towards quantifying the Album Effect in artist identification", in Proc. ISMIR 2006, pp. 393-394, Victoria, Canada, 2006.

[10] M. Mandel, D. Ellis, "Song-level Features and Support Vector Machines for Music Classification", in Proc. ISMIR 2005, pp. 594-599, London, UK, 2005.

[11] S. McAdams, S. Winsberg, S. Donnadieu, G. De Soete, J. Krimphoff, "Perceptual scaling of synthesized musical timbres: common dimensions, specificities and latent subject classes", in Psychological Research, no. 58, pp.177-192, 1995.

[12] M.F. McKinney, J. Breebart, "Features for audio and music classification", in Proc. ISMIR 2003, Baltimore, Maryland, USA, 2003.

[13] S. Stevens, J. Volkman, E. Newman, "A scale for the measurement of the psychological magnitude of pitch", in J. of the Acoustical Society of America, vol. 8(3) pp. 185-190, 1937.

[14] A. Meng, "Temporal feature integration for music organisation", Ph.D. Dissertation, IMM Technical University of Denmark, Denmark, 2006.

[15] E. Pampalk, "Computational models of music similarity and their application in music information retrieval", Ph.D. Dissertation, Vienna Institute of Technology, Austria, 2006.

[16] R. Parncutt, "The perception of pulse in musical rhythm", in Action and Perception in Rhythm and Music, pp. 127-138, Stockholm, Sweden, 1987.

[17] J. Paulus, A. Klapuri, "Combining Temporal and Spectral Features in HMM-Based Drum Transcription", in Proc. ISMIR 2007, Vienna, Austria, 2007.

[18] G. Peeters, "Rhythm Classification using Spectral Rhythm Patterns", in Proc. ISMIR 2005, pp. 644-647, London, UK, 2005.

[19] T. Pohle, P. Knees, M. Schedl, G. Widmer, "Independent Component Analysis for Music Similarity Computation", in Proc. ISMIR 2006, pp. 228-233, Victoria, Canada, 2006.

[20] A. Rauber, E. Pampalk, D. Merkl, "Using psychoacoustic models and self-organizing maps to create a hierarchical structuring of music by sound similarity", in Proc. ISMIR 2002, Paris, France, 2002.

[21] E. Scheirer, "Tempo and beat analysis of acoustic musical signals", in J. of the Acoustical Society of America, 103(1): 588-601, 1998.

[22] M.R. Schroeder, B.S. Atal, J.L. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear", in J. of the Acoustical Society of America, vol. 66, issue 6, pp. 1647-1652, December 1979.

[23] E. Terhardt, "Calculating virtual pitch", in Hearing Research, vol. 1, pp. 155-182, 1979.

[24] E. Terhardt, "On the perception of periodic sound fluctuations (roughness)", in Acustica, 30, pp. 201-213, 1974

[25] G. Tzanetakis, P. Cook, "Musical genre classification of audio signals", in IEEE Trans. Speech Audio Processing, vol.10, no.5, pp. 293-302, July 2002.

[26] H.H. Yang, S. Sharma, S. van Vuuren, H. Hermansky, "Relevance of Time Frequency Features for Phonetic and Speaker/Channel Classification", in Speech Communication, vol. 31, issue 1, pp. 35-50, August, 2000.