Rhythm Description for Music and Speech Using the Beat Histogram with Multiple Novelty Functions: First Results

Athanasios Lykartsis[1], Stefan Weinzierl[1]

[1] Audio Communication Group, TU Berlin, 10587 Berlin, Germany, Email: {athanasios.lykartsis, stefan.weinzierl}@tu-berlin.de

Introduction

In the last few years, methods for rhythmic analysis of music signals have become widespread in their use due to their value for diverse tasks of music processing. In the field of Music Information Retrieval (MIR), features describing rhythmic content through the properties of the very low modulation frequencies or periodicities (i.e., between 0.5 and 10 Hz) in the signal have been developed for music transcription, beat tracking, rhythmic similarity calculation and music genre classification. For the latter task in particular, methods have been devised for capturing either specific temporal patterns in order to measure similarities between different tracks and compare against predetermined rhythmic patterns [1]; or for extracting information on the statistical properties of periodicities in the signal, so as to be able to perform supervised classification [2, 3, 4]. In both cases, the basic idea is the same: A novelty function (extracted through onset detection algorithms) of a basic temporal or spectral property of an audio track (e.g., the signal amplitude) is extracted, providing information about salient changes in the signal. This novelty function is then analyzed through an FFT, an Autocorrelation Function (ACF), or resonant filters to provide a representation of the periodicities present in the signal and their relative strength. This form has been dubbed with several names - periodicity/beat histogram, self-similarity-matrix, inter-onset-interval histogram - but the basic goal is the same: the representation provides information concerning the distribution and temporal evolution of signal qualities and therefore describes the rhythmic content of the signal. Up to now, such methods have shown satisfactory results in the rhythm-based genre classification and rhythmic similarity tasks either used alone or in combination with other, non-rhythmic features, inspiring several adaptations and efficient implementations [5, 6]. In related work for speech signals, similar representations based on periodicities detected in the signal amplitude envelope have been used only recently and to a limited extent, in order to analyze their properties and detect differences between languages and speakers [7].

The above mentioned method has, however, some limitations: first, if a strong beat is lacking or the signal periodicities are complex and not distinctive (as is the case, for example, for certain types of jazz music), the extraction leads to noisy and less informative features. Furthermore, if the signals are polyphonic, the features extracted either only express the most prevalent periodicities (which are the ones caused by the instruments or voices having the greatest energy or impact on the signal's waveform and spectrum) amongst others present, or the representation loses its ability to provide meaningful features, essentially blurring information since the rhythms present are interwoven. Finally, the features extracted are not always easy to interpret, since their calculation involves multiple steps which do not allow a clear view of the feature's significance. To tackle those problems many strategies have been followed, such as feature selection (e.g. with mutual information with target data, to identify the most informative features), dimensionality reduction (such as PCA, to increase the feature relevance and independence) and use of more elaborated methods for the representation [8]. However, we wanted to address a basic conceptual problem of this class of methods: Although music and other audio signals mostly comprise of many sources or have properties which change differently in time (e.g., a musical track's harmony does not evolve at the same pace as the drum beat), this information has not been exploited in the past for rhythmic feature extraction. In that sense, two kinds of approaches would be suitable: source separation (for example based on Non-Negative Matrix Factorization - NMF), in order to be able to apply the rhythm extraction on different instruments or voices; or application of the periodicity representation on other signal properties than only amplitude, providing the possibility to analyze several musical properties and extract information pertaining to each of them. This latter approach has the added advantage that it can be adapted for speech signals. In the following, our approach and the first results concerning the application of this method for music and speech are presented.

Method

In order to take account of several signal properties and their periodicities which do not all necessarily evolve in the same way, we extract several features [9] and apply the beat histogram transformation to them [10]. Results have shown that this method provides good performance and can be helpful in determining which exact signal components are responsible for special rhythmic changes - which in this case were the spectral flux, the RMS amplitude and the spectral flatness (concerning the novelty functions), whereas with regards to the statistics on the beat histogram, simple statistics such as the mean and standard deviation but also advanced descriptors such as *tempo* have provided the best results. A similar approach was also used in [11], where we extracted multiple drum components using NMF for rhythm-based genre classification. Being motivated by our results, we decided to adapt and apply this method for speech [12, 13], in order to analyze speech rhythm. So far, only speech

rhythm metrics (analyzing the statistical properties of duration intervals between salient speech elements) have been used up to date (see [14] for a review). In our case, following similar works from [15] and [7] we extracted spectral (spectral flux, centroid and flatness), temporal (RMS) and tonal (F0) measures to check for periodicities and use for automatic language identification (LID). The novelty functions and features on the beat histograms for both music and speech can be seen in Tables 1 and 2 respectively.

Table 1: Novelty Functions for Beat Histogram Extraction.

Music	Speech
Spectral Flux (SF)	Spectral Flux (SF)
Spectral Flatness (SFL)	Spectral Flatness (SFL)
Spectral Centroid (SCD)	Spectral Centroid (SCD)
RMS Amplitude (RMS)	RMS Amplitude (RMS)
Pitch Chroma Coefficients (1-12)	Fundamental Frequency F0 (HPS)
MFCCs (1-13)	
Tonal Power Ratio (TPR)	

Table 2: Subfeatures extracted from Beat Histograms (both for speech and music).

Distribution	Peak
Mean (ME)	Salience of Strongest Peak (A1)
Standard Deviation (SD)	Salience of 2nd Stronger Peak (A0)
Mean of Derivative (MD)	Period of Strongest Peak (P1)
SD of Derivative (SDD)	Period of 2nd Stronger Peak (P2)
Skewness (SK)	Period of Peak Centroid (P3)
Kurtosis (KU)	Ratio of A0 to A1 (RA)
Entropy (EN)	Sum (SU)
Geometrical Mean (GM)	Sum of Power (SP)
Centroid (CD)	
Flatness (FL)	
High Frequency Content (HFC)	

Table 3: Datasets Used.

Music	Speech
GTZAN	MULTEXT PD
Ballroom	OGI-MLTS
ISMIR2004	
Homburg	
Unique	

Experimental Setup

For the evaluation of the beat histogram features, a baseline feature set was extracted in every case through calculation of a series of non-rhythmic features and the respective novelty functions. The novelty functions for speech and music are shown in Table 1, whereas the features on each novelty function can be seen in the *Distribution* column of Table 2. The reason for this is the need to be able to estimate if the use of rhythmic features provides significantly different results to non-rhythmic ones and consequently, if they present a genuine improvement or degradation in the performance of the associated task.

For supervised classification, we use Support Vector Machines (SVM) [16] in all cases. For the SVM algorithm the Radial Basis Function (RBF) Kernel is used with the parameters C and γ determined through grid search. All experiments take place as multiclass one-vs-one classification problems with 10-fold cross validation and prior standardization of the features (z-score, separately for train and test set). In order to evaluate the classification we use the average *accuracy* (Acc.) as a performance measure.

Concerning the datasets, Table 3 gives an overview of the resources used in both cases. Almost all datasets are unbalanced, which has a negative effect on classification accuracy, but often this problem is circumvented by creating a balanced subset of the dataset. Finally, the quality of the datasets is in both cases, not at an equal level. For the musical ones, signal quality is good, but the ground truth can be challenged. For speech, the MULTEXT dataset has better signal quality, which is important for the outcome of the experiments and the conclusions drawn from them.

Results Comparison - Discussion

Results of genre classification and language identification accuracy can be seen in Fig. 1 and Fig. 2. In Tables 5 and 6, the results of the feature selection for both applications are shown.

Concerning overall classification accuracy, two tendencies can be observed: For music, only for one dataset (Ballroom) the accuracy of the rhythmic features exceeds the one achieved with the baseline feature set. For speech, the accuracy of the rhythmic features is higher in one dataset (MULTEXT PD, almost balanced, read speech, good quality recordings), but lower on the other (OGI-MLTS, unbalanced, spontaneous speech, telephone quality), where results are low at any rate. Comparing music and speech, we can see that for datasets which are balanced, have good sound quality, are rhythmically distinct (for music) or containing less variation (for speech), the performance based on accuracy is good and close to what other studies achieve. This is probably due to noisy end features, resulting from a low quality signal at the beginning of the processing chain and an extraction procedure involving multiple steps.

There are both similarities and differences between the most efficient features in speech and music: For both cases, salient novelty functions denoting spectral change in the signal such as the RMS amplitude, spectral flux and spectral flatness were amongst the most informative features. However, in music, tonal components seem to be as important; their performance, at least for genre classification, is limited across multiple datasets. In speech, however, fundamental frequency appears to be an important feature, particularly in the case where the dataset quality is low. In Fig. 33, feature groups for genre classification shows that those tendencies are also confirmed by the group selection, whereas for speech (Tables 4 and 6), fundamental frequency is a salient feature even in adverse conditions (OGI-MLTS). Those results show that extracting novelty functions which are indicative of salient signal changes provides a good basis for the extraction of informative features. For the subfeatures, no candidate came out as a "winner", stressing the need to extract as much information as possible but also to focus on more meaningful features. On that note, the tempo information provides a good candidate for such a follow-up investigation of its properties.

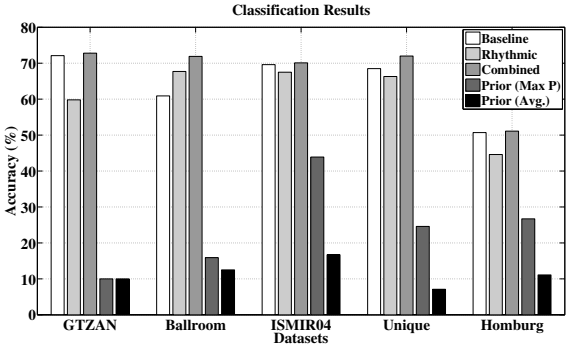

Figure 1: Classification results, comparison between datasets (music). Figure from [10].

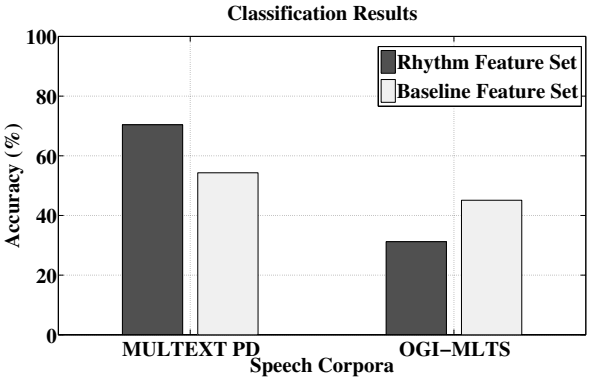

Figure 2: Classification results, comparison between datasets (speech). Figure from [13].

Table 4: *Feature group comparison (speech).*

Rhythmic feature subset	MULTEXT PD	OGI-MLTS
All features	70.4 %	31.2 %
RMS Amplitude	67.5 %	25.4 %
Fundamental Pitch	70.4 %	27.4 %
Spectral Flux	67.5 %	25.5 %
Spectral Flatness	66.8 %	24.7 %
Spectral Centroid	64.9 %	24.9 %

Table 5: *Best features after feature selection (music). Left: subfeature, right: novelty function. Table from [10].*

Rank	GTZAN	Ballroom	ISMIR04	Unique	Homburg
1	MD.RMS	P1.SF	MD.MFC2	SD.MFC1	SD.RMS
2	FL.RMS	A0.SFL	CD.MFC1	GM.SFL	SD.SPC3
3	GM.SFL	SD.SPC3	A0.SF	MD.MFC2	FL.SFL

Table 6: *Best features after feature selection (speech). Left: subfeature, right: novelty function. Table from [13].*

Rank	MULTEXT PD	OGI-MLTS
1	FL.SFL	SP.HPS
2	GM.SF	P3.HPS
3	A2.SF	A2.HPS

Conclusions

In this paper we present first results on the use of novel features for rhythm analysis and rhythm-based LID. The expansion of the use of periodicity representation methods from the field of MIR such as the beat histogram for speech rhythm analysis has provided promising results. For the rhythm descriptors, not only the signal amplitude but also other rhythm-relevant signal quantities were used as basis for creating the beat histogram and were found to be relevant. Furthermore, a comprehensive array of subfeatures was extracted from the periodicity representation, which provides ample information about the periodicities in the signal and their patterns. We could show that classification performance for one multilingual speech corpus using the SVM algorithm is comparable to the results of similar studies and close to those using other basic, non-rhythmic features. Similar results can be observed for music, where for two out of five datasets, performance is acceptable and in one case even better than when using more general features. In general, concerning the datasets, rhythmic features provide good or at least acceptable performance for balanced, high-quality sound datasets, both for music and for speech. Furthermore, the proposed method has the advantage that it takes into account the rhythmic properties on the signal (signal properties and features) and not on the speech element level (syllables), providing a new perspective for the analysis of speech rhythm and the related signal properties (such as fundamental frequency for speech). Another important advantage of the proposed method for speech rhythm analysis is that it is fully automatic and can be extended to larger datasets.

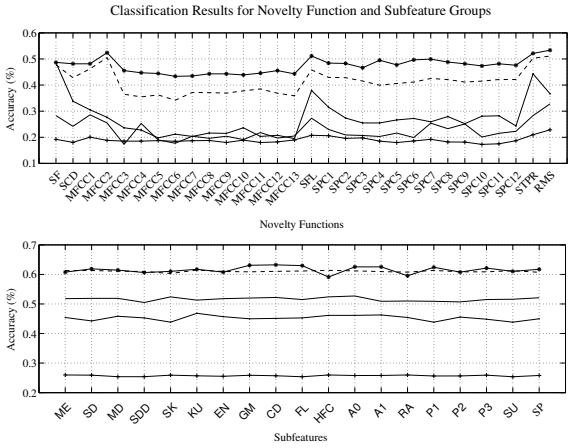

Figure 3: Feature group comparison (music). Table from [13].

These conclusions provide several objectives for further research, such as the application of the method to more diverse and comprehensive speech corpora (such as the GLOBALPHONE [17]). At this point, the relation of the rhythm features to other speech rhythm metrics and language elements such as syllables and consonant-vowel clusters is unclear, suggesting another direction for future work. Another promising direction is focusing on specific salient features (such as the tempo, which has been shown to be easier to extract and understand where music is concerned, but which has these properties in speech as well) over different languages and/or genres, in order to study their behavior and draw conclusions about whether they can serve as a discriminatory feature. The

use of rhythmic similarity measures as complementary methods to the beat histogram is also a possible goal, so as to capture language specific rhythm patters instead of features describing periodicities. Future goals include the investigation of optimal parameter settings for feature extraction, as well as the utilization of unsupervised classification methods and novel classifiers, such as Deep Neural Networks (DNNs).

References

[1] Pohle, T.; Schnitzer, D.; Schedl, M.; Knees, P.; Widmer, G. (2009): "On rhythm and general music similarity." In: *ISMIR*.

[2] Tzanetakis, G.; Cook, P. (2002): "Musical genre classification of audio signals." In: *Speech and Audio Processing, IEEE transactions on*, **10**(5):293–302.

[3] Burred, J.J.; Lerch, A. (2003): "A hierarchical approach to automatic musical genre classification." In: *DAFx*.

[4] Gouyon, F.; Dixon, S.; Pampalk, E.; Widmer, G. (2004): "Evaluating rhythmic descriptors for musical genre classification." In: *Proceedings of the AES 25th International Conference*, 196–204, Citeseer.

[5] Peeters, G. (2011): "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal." In: *IEEE Transactions on Audio, Speech and Language Processing*, **19**(5):1242–1252.

[6] Holzapfel, A.; Flexer, A.; Widmer, G. (2011): "Improving tempo-sensitive and tempo-robust descriptors for rhythmic similarity." In: *Proceedings of the 8th Sound and Music Computing Conference*.

[7] Tilsen, S.; Arvaniti, A. (2013): "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages." In: *The Journal of the Acoustical Society of America*, **134**(1):628–639.

[8] Marchand, U.; Peeters, G. (2014): "The modulation scale spectrum and its application to rhythm-content description." In: *DAFx*.

[9] Lerch, A. (2012): *An introduction to audio content analysis: Applications in signal processing and music informatics*. John Wiley & Sons.

[10] Lykartsis, A.; Lerch, A. (2015): "Beat histogram features for rhythm-based musical genre classification using multiple novelty functions." In: *DAFx*.

[11] Lykartsis, A.; Wu, C.W.; Lerch, A. (2015): "Beat histogram features from nmf-based novelty functions for music classification." In: *ISMIR*.

[12] Lykartsis, A.; Weinzierl, S. (2015): "Analysis of speech rhythm for language identification based on beat histograms." In: *Fortschritte der Akustik: Tagungsband d. 41. DAGA*.

[13] Lykartsis, A.; Weinzierl, S. (2015): "Using the beat histogram for speech rhythm description and language identification." In: *Sixteenth Annual Conference of the International Speech Communication Association*.

[14] Wagner, P. (2008): "The rhythm of language and speech: Constraining factors, models, metrics and applications." In: *Germany: Habilitationsschrift, University of Bonn*.

[15] Rouas, J.L.; Farinas, J.; Pellegrino, F.; André-Obrecht, R. (2005): "Rhythmic unit extraction and modelling for automatic language identification." In: *Speech Communication*, **47**(4):436–456.

[16] Vapnik, V. (2000): *The nature of statistical learning theory*. springer.

[17] Schultz, T. (2002): "Globalphone: a multilingual speech and text database developed at karlsruhe university." In: *INTERSPEECH*.