

On the Analysis of Speech Rhythm for Language and Speaker Identification

vorgelegt von

Dipl.-Ing. Athanasios Lykartsis, M. Sc.

an der Fakultät I – Geistes- und Bildungswissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
- Dr. rer. nat. -

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Ulf Schrader

Gutachter: Prof. Dr. Stefan Weinzierl

Gutachterin: Ass. Prof. Dr. Manon Grube

Gutachter: Prof. Dr. Volker Dellwo

Tag der wissenschaftlichen Aussprache:

06. Dezember 2019

Berlin 2020

Acknowledgements

The completion of this thesis would not have been possible without the help and support of a great number of people. In essence, it is a collective work, in which the expertise, wisdom and experience of many individuals and their interactions were incorporated therein; but also the emotional support and coaching provided was paramount for the achievement of the result. To all of these people I want to express my gratitude.

First and foremost, I thank my parents for always being supportive and having made possible that I take on this task and reach this achievement. I hope I have made them proud and hope to continue doing this in the future. One special thanks goes to my cousin, Thanos Lykartsis, who also completed his PhD shortly before me. Knowing we went this road in parallel was a great support and I am proud of him as well - since there will be two people with the same name carrying a title.

Of course, my academic advisors - or my scientific family - are to be thanked. I owe a special thanks to my professor and chief supervisor, Prof. Dr. Stefan Weinzierl, for being a true guide and inspiration on this voyage, both scientifically and personally. To my other advisors and examiners, Prof. Dr. Manon Grube, Prof. Dr. Volker Dellwo and Prof. Dr. Alex Lerch, I owe a great thanks for their input, their remarks and discussions and most of all, for their also acting as role models, which taught me the value of conducting diligent work and always striving for improvement.

In the years at the Audio Communication Group I have had the luck of being surrounded by colleagues who were as amicable as they were competent. For the stimulating discussions and their input, as well as for their emotional support along the way, I would like to thank Dr. Florian Straube, Markus Hädrich, Clemence Büttner, Roman Gebhardt, Henrik von Coler, Paul Luizard, Dr. Melanie Irrgang, Dr. Fabian Brinkmann, Dr. Oliver Schwab-Felisch, David Ackermann, Dr. Steffen Lepa, Dr. Jochen Steffens and Marc Voigt, our IT-Admin for his help with technical matters.

Over the years, there was a number of friends and close people from several backgrounds who have also been working on their PhD and have acquired it in this year. Their support has been more than instrumental, since knowing that other people went through the same challenge helped me carry the load many times. I am really proud for them and I wish to thank Giannis Dafnomilis, Ivana Sidzimovska, Patricio Lopez-Serrano, and Marie Urban.

The list of people who have contributed with their discussions (even on a more general level) to my thought processes and their motivating me is very long. However, I owe an extra thanks to my good friends Christos Tsigaridas, Andreas Pysiewicz and Ferdi Ramadan for being there all along. Same goes for my former flatmates Vaclav Plecity, Omkar Pimple, Andreas Koukias and Helene Battarel, who were kind enough to listen to my pitches and bear me all these years. I would like also like to especially thank Ronja Eggert and Frederike Marmeling-Schulze for their emotional support and their interest, which helped me reflect things from a new perspective.

Last but not least, a great and heartfelt thanks go to my beloved friends from Greece, namely Prokopis Mavridis, Tassos Panagiotopoulos, Giorgos Gravanis, Dimitris Charalampidis, Dimitris Bakyrtsis and Nikos Xydas, for their being there from the beginning and accompanying me with their friendship all this time. They have been following my academic career from the beginning and have always been helpful and supportive.

To all the other people I did not include here but have been supporting me all this time, I extend a thank you to them as well for being there and helping me go forward.

Abstract

In the context of this dissertation, novel methods for rhythm description and extraction originating from the area of Music Information Retrieval (MIR) were adapted and applied to represent speech rhythm and its properties. These methods were then used to extract rhythmic information to be used in two specific classification scenarios relevant to speech technology: language identification (LID) and speaker identification (SID). Specifically, periodicity representations that offer an overview of the prominent “beats” – i.e., the salient, recurring temporal or spectral patterns in the audio signal – were created by using the *Beat Histogram*, an established method for extraction of rhythm information in MIR. The adaptation entailed the analysis of several signal features (e.g., fundamental frequency, energy, spectral change and others) which describe relevant signal properties and directly shape human percepts of, for instance, syllables, phones, accents and prosody. This approach was then thoroughly tested on two multilingual speech datasets with different properties (read vs. spontaneous speech, high vs. low audio signal quality, Indo-European languages only vs. others) using state-of-the-art machine learning algorithms. The results of the experiments for LID show that speech rhythm description based on the proposed methods can be successful, but mostly in the case of read speech with high audio signal quality and for Indo-European languages, pointing towards a potential for improvement of the descriptor robustness. The results are promising, and they surpass the state-of-the-art results of other studies on LID for the used datasets, demonstrating that the proposed features indeed capture a significant part of the variability between languages. Further experiments performed on a dataset of Swiss German for SID showed that rhythmic information is less informative for that task, and that spectral information accounted for much of the variability between speakers. Finally, a feature selection procedure showed descriptors such as tempo (i.e., speech rate), spectral change and fundamental frequency to consistently be among the most useful and informative ones. This finding highlights the fact that it is important to reliably extract salient temporal information, as the descriptors resulting from it are, in many cases, informative as well. Similar results were obtained when the methods were applied to the related task of rhythm-based genre classification on music datasets, suggesting that the findings are not strictly speech specific. Furthermore, listening test experiments for differences in listening to speech vs. listening to singing confirm the findings about the most salient features to be tempo and regularity. Finally, the language rhythm family hypothesis (for example, English and German as the “morse-code” family and Spanish, Italian and French as the “machine-gun” family) could be partially confirmed, but not in its original form. This possibly shows that rhythm classes, which have been difficult to identify with other methods (e.g., other speech rhythm metrics) are also hard to be detected using automatic methods. Alternatively, this might hint at a gap between human-perceived cues and machine-extracted descriptors for speech rhythm. The developed analysis system in the context of the dissertation can be used for rhythm description for various tasks.

Zusammenfassung

Im Rahmen dieser Dissertation wurden neuartige Methoden zur Rhythmusbeschreibung, die aus dem Bereich Music Information Retrieval (MIR) stammen, adaptiert und auf Sprachsignale angewendet, um den Sprachrhythmus und seine Eigenschaften zu erfassen. Diese wurden auf zwei Szenarien aus dem Bereich der Sprachtechnologie angewendet: Die Identifizierung der Landessprache und die Identifizierung einzelner Sprecher*innen. Hierfür wurde eine Analyse von Periodizitäten des Signals mittels *Beat Histogram* durchgeführt. Diese Methode bietet eine Übersicht der prominentesten "Beats", d.h. der salienten, wiederkehrenden temporalen oder spektralen Muster im Audiosignal. Diese ist der etabliertesten Methoden zur Extraktion rhythmischer Information im MIR. Die vorliegende Adaptierung beinhaltet eine Analyse verschiedener Signalmerkmale (z.B. der Grundfrequenz der Stimme, der Energie, der spektralen Änderung und einiger anderer), die relevante Signaleigenschaften beschreiben und mittelbar im Zusammenhang unserer Wahrnehmung sprachlicher Objekte wie z. B. Silben, Phoneme, Akzente und Prosodie beeinflussen. Dieser Ansatz wurde an zwei multilingualen Datensätzen mit unterschiedlichen Eigenschaften getestet (vorgelesene vs. spontane Sprechart, hohe vs. niedrige Signalqualität, nur indoeuropäischen Sprachen vs. andere), wobei verschiedene Algorithmen des maschinellen Lernens zur Anwendung kamen. Die Ergebnisse der Experimente zur Identifizierung der Landessprache konnten zeigen, dass eine Beschreibung des Sprachrhythmus mithilfe der hier benutzten Methoden möglich ist. Sie gelingt allerdings hauptsächlich für vorgelesene Sprache von hoher Signalqualität und für indoeuropäische Sprachen, was auf ein Potential zur Verbesserung der Robustheit der vorgeschlagenen Deskriptoren hinweist. Die Ergebnisse sind insofern vielversprechend, als dass sie schon existierende Werte anderer Studien über Landessprach- und Sprecher*innenidentifizierung für diese Datensätze übertreffen. Das weist darauf hin, dass die Deskriptoren in der Tat einen wichtigen Teil der rhythmischen Variabilität zwischen Sprachen erfassen. Die Experimente zur Identifizierung einzelner Sprecher*innen, die mit einem Datensatz von Aufnahmen in Schweizerdeutsch durchgeführt wurden, haben gezeigt, dass die vorgeschlagenen rhythmischen Deskriptoren für Sprecher*innenidentifizierung weniger informativ für diese Aufgabe sind. In dem Zusammenhang war spektrale Information hilfreicher für die Unterscheidung zwischen Sprecher*innen. Eine Selektionsprozedur für Audiomerkmale hat gezeigt, dass Deskriptoren wie das Tempo, die spektrale Änderung und die Grundfrequenz immer unter den für die Klassifizierung wichtigsten Merkmale waren. Ähnliche Ergebnisse wurden bei Anwendung der Methoden auf Musikdatensätze zur Genreklassifizierung erreicht, was darauf hinweist, dass die Ergebnisse nicht rein sprachspezifisch sind. Außerdem konnten die Ergebnisse bezüglich der wichtigsten Merkmale (die das Tempo und die Regelmäßigkeit waren) durch Hörversuche bestätigt werden, die die rhythmischen Unterschiede zwischen Sprache und Gesang untersuchten. Die Existenz von Rhythmusfamilien, d.h. eine Unterscheidung zwischen akzentzählenden Sprachen wie Englisch oder Deutsch und silbenzählenden Sprachen wie Spanisch oder Italienisch konnte teilweise tendenziell, aber nicht in ihrer ursprünglichen Form bestätigt werden. Das zeigt, dass die Rhythmusklassen, die durch andere Methoden (z. B. Sprachrhythmusmetriken) nicht eindeutig erkannt werden konnten, auch durch automatische Analysemethoden schwer erkennbar sind. Alternativ könnte das auf eine Lücke zwischen von Menschen wahrgenommenen und maschinell extrahierten Merkmalen des Sprachrhythmus hinweisen. Das im Rahmen der Dissertation entwickelte Analysesystem kann für Rhythmusbeschreibung in verschiedenen Kontexten benutzt werden.

List of included publications

This thesis is a cumulative dissertation. It comprises the following eight scientific publications that were authored or co-authored by the doctoral candidate between 2014 and 2019, listed in chronological order of publication.

Peer Reviewed:

1. Lykartsis, Athanasios; Weinzierl, Stefan (2015): "Using the Beat Histogram for Speech Rhythm Description and Language Identification". In: *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.
2. Lykartsis, Athanasios; Wu, Chih-Wei; Lerch, Alexander (2015): "Beat Histogram Features from NMF-based Novelty Functions for Music Classification". In: *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Malaga, Spain.
3. Lykartsis, Athanasios; Lerch, Alexander (2015): "Beat Histogram Features for Rhythm-Based Musical Genre Classification Using Multiple Novelty Functions". In: *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx)*, Trondheim, Norway.
4. Stanev, Madeleine; Redlich, Johannes; Knörzer, Christian; Rosenfeld, Ninette; Lykartsis, Athanasios (2016): "Speech and Music Discrimination: Human Detection of Differences between Music and Speech based on Rhythm". In: *Proceedings of Speech Prosody 2016*, Boston, US.
5. Lykartsis, Athanasios, Weinzierl, Stefan; Dellwo, Volker (2017): "Speaker Identification for Swiss German with Spectral and Rhythm Features". In: *AES Conference on Semantic Audio 2017*, Erlangen, Germany.
6. Lykartsis, Athanasios; Kotti, Margarita; Papangelis, Alexandros; Stylianou, Yannis (2018): "Prediction of Dialogue Success with Spectral and Rhythm Acoustic Features Using SVMs and DNNs". In: *Proceedings of the IEEE Speech and Language Technology Workshop 2018*, Athens, Greece. © 2018 IEEE. Reprinted, with permission, from Margarita Kotti, Alexandros Papangelis, Yannis Stylianou, Prediction of Dialogue Success with Spectral and Rhythm Acoustic Features Using DNNs and SVMs, 2018 IEEE Spoken Language Technology Workshop (SLT), 2018.

Non-peer Reviewed:

7. Lykartsis, Athanasios; Lerch, Alexander; Weinzierl, Stefan (2015): "Analysis of Speech Rhythm for Language Identification Based on Beat Histograms". In: *Fortschritte der Akustik: Tagungsband d. 41. DAGA*, Nürnberg, Germany.
8. Lykartsis, Athanasios; Weinzierl, Stefan (2016): "Rhythm Description for Music and Speech Using the Beat Histogram with Multiple Novelty Functions: First Results ". In: *Fortschritte der Akustik: Tagungsband d. 42. DAGA*, Aachen, Germany.

Content

1 Introduction

1.1 Motivation

1.2 State of the Art

1.2.1 Speech Rhythm Description

1.2.2 Music Rhythm Description

1.3 Methods

1.3.1 Feature Extraction

1.3.2 Machine Learning

1.3.3 Datasets

1.4 Summary of Original Achievements

1.5 Conclusion and Outlook

1.6 References

2 Publications

1 Introduction

The thesis is mainly concerned with the adaptation and evaluation of special audio descriptors for the extraction of *speech rhythm*. These descriptors are also known as *features*, i.e., extractable quantities from the audio data describing various aspects of the acoustic signal, which can be relevant for various analysis tasks¹. The analysis is based on a method from the field of Music Information Retrieval (MIR), called the *Beat Histogram*. The descriptors are then used for the machine learning tasks of language and speaker identification (LID and SID, respectively), with the aim to investigate the role of rhythmic information and the features developed for these and other speech-related classification tasks. Furthermore, related studies were undertaken for musical signals (specifically for rhythm-based genre classification), with the goal of investigating the suitability of those descriptors for music signals, as well as to perform a comparison between the results for music and for speech. Furthermore, a study on speech rhythm perception was performed to determine which descriptors could be important in order to approximate human rhythm perception. Finally, we conducted a study on dialogue success in order to assess the suitability of the rhythmic descriptors for other speech-related tasks. The thesis was conducted at the Audio Communication Group (Fachgebiet Audiokommunikation) of the Technische Universität Berlin and was supervised by Professor Dr. Stefan Weinzierl.

Chapter 1 contains an overview of the papers included in this thesis, with the goal of summarizing their results and providing a common thread between them. In Section 1.1 the basic motivation for analyzing speech rhythm with or without methods from MIR is given. Section 1.2 summarizes the state of the art concerning speech and music rhythm description. Section 1.3 presents a review of the methods and the datasets, whereas Section 1.4 summarizes and discusses the results and achievements attained in the thesis, highlighting their importance for speech rhythm research and related tasks. Section 1.5 contains a conclusion and perspectives for further research, followed by the references used in 1.6. Chapter 2 comprises of the publications included in the thesis, as listed on page III.

1.1 Motivation

Speech rhythm analysis has been both an important as well as a controversial topic in phonetics research and speech technology. On the one hand, speech rhythm analysis could allow to determine if and to what extent there are rhythmic differences between languages. First approaches to tackle this research subject have led to the formulation of the *Rhythm Class Hypothesis* (Pike 1945, Abercrombie 1967). This hypothesis states that some languages have a similar rhythmic character and thus constitute a rhythmic group, which differs from that of other groups. This would place languages such, e.g., English and German in one rhythmic group, and French and Spanish in another. Based on the perception of the interplay between syllables and stresses and the regularity of their appearance, German and English would belong to a group of „stress-timed“ languages. On the other hand, French and Spanish would belong to the group of „syllable-timed“ languages (Pike 1945). Although the hypothesis was first made just on the basis of informal observations, it inspired quite a body of research in the fields of linguistics and phonetics, focusing on the attempt to develop suitable theories and experiments to investigate it. On the other hand, in the field of speech technology applications, better speech rhythm

¹ Further on in the text, the terms “descriptor” and “feature” will be used interchangeably.

description could enhance the performance and quality of speech recognition, translation and speech synthesis systems. Researchers have proposed several definitions of speech rhythm, as well as methods to measure it. The latter is the attempt to quantify its elements, for instance, the statistical properties of durations between linguistic units such as syllables. However, there are different views on what exactly constitutes speech rhythm. The above considerations have sparked scientific research and dialogue on the topic. Subsequent research identified periodicities, abstract phonological patterns and surface timing patterns as important elements for a definition of speech rhythm (see Turk & Shattuck-Hufnagel 2013 for a comprehensive review of the theoretical phonological and phonetical approaches on the topic). Furthermore, there have been attempts to construct measures for the description of speech rhythm, commonly called *rhythm metrics*, such as ΔV , $\%V$, ΔC , *VarcoC* (see Wiget et al. 2010 for a comprehensive list). These were developed to measure the statistical properties of salient speech elements such as vowels and consonants (Ramus et al. 1999, Grabe & Low 2002, Dellwo 2006, Liss et al. 2009). These metrics have produced interesting results but have also been criticized as having several shortcomings (Wiget et al. 2010, Arvaniti 2012). For example, the metrics showed substantial inter-speaker variation, which can reduce the robustness of the metrics when determining if they can be used to measure differences between languages. Other factors, such as elicitation method and syllable complexity, were also found to influence the metrics. Finally, methodological issues, such as rate normalization and boundary definition, also influence the metrics to some extent.

Partly due to these reasons, in the field of speech technology very few attempts were undertaken to describe, extract, and use speech rhythm information in tasks such as LID (Thyme-Gobbel & Hutchins 1996) and SID (Martinez et al. 2012). Most notably, however, some more modern approaches were used for attempting to automatically detect rhythmic differences between languages. One important approach employed the modeling of rhythmic units (dubbed “pseudosyllables”, see Rouas et al. 2005, Rouas 2007). These rhythmic units provided the basis to extract features with which machine learning models could be trained. Another approach involved the direct extraction of signal periodicities and the statistical comparison of their properties (Tilsen & Johnson 2008, Tilsen & Arvaniti 2013). This thesis is to a great extent motivated by these approaches, as they demonstrate two important ideas: Firstly, the perspective of defining speech rhythm through its inherent periodicities. In this view, speech rhythm as it is perceived by humans is the result of the interaction of periodicities between salient speech elements such as the fundamental frequency that can be measured in the speech signal. Secondly, the employment of machine learning methods for the analysis of datasets based on the rhythm analysis data, for tasks such as LID and SID. Another important motivation for this thesis comes from the music science and technology areas. There, theoretical subjects pertaining to rhythm theory (Cooper 1963, Lester 1986) and its perception (Fraisse 1982, Desain 1992, London 2012) as well as description and extraction issues (Bello et al. 2005, Esparza et al. 2015) were thoroughly researched, providing a basis for the analysis of speech rhythm as well.

This research also lead to the development of robust and efficient methods to solve problems such as onset detection, beat tracking and tempo estimation. Since such methods have been tested thoroughly for musical signals, it would be natural to expect that these methods could be also used for speech rhythm analysis. To do that, it is important to pinpoint parallels between music and speech at the signal level, since it would be very improbable to perceive any rhythmic regularities without there being a statistical correlate to measurable signal quantities (Hübler & Hoffmann 2011), as is the case for music. Therefore, methods from the field of music technology (and especially from *Music Information Retrieval*, the

fastest-growing branch of music analysis) should be applicable in a similar way and be able to provide insights into speech rhythm phenomena. Especially methods which were used to extract rhythmic information for genre classification (Tzanetakis & Cook 2002, Burred & Lerch 2003, Gouyon et al. 2004) are taken as the basis for this thesis. Genre classification for music can be seen as the counterpart of LID for speech (Hübler & Hoffmann 2011). This is based on the notion that genre classification attempts to determine the complex, style-specific elements (founded in timbre, rhythm and structure) which make a genre of music be distinguishable from another. This should happen independently from performer, performance situation or the specificities of a given music track. Such an approach is very similar to attempting to define the specific elements of a language (for example regarding its phone inventory, its phonotactics, or its rhythm) differentiating it from other languages independently from speaker, elicitation method, sentence content or speaking situation.

The thesis is therefore also motivated by the above-mentioned music genre classification studies and presents the adaptation of the Beat Histogram method for speech signals, as well as its application on specific speech data for the tasks of LID and SID. In short, the work presented here addresses the following three main questions:

- a) How useful can the proposed speech rhythm analysis be for the associated classification tasks of LID and SID, with an added comparison of results for genre classification?
- b) Which audio features and extraction method parameters are the most relevant for such an analysis?
- c) What do the results reveal regarding the phonetic hypothesis mentioned above, stating that specific rhythm classes can be observed when analyzing rhythm at the signal level and trying to distinguish languages based on this information? What do the results reveal about the human perception of speech rhythm regarding the properties which are important for its perception?

A final aim of the thesis is to provide insights concerning similarities and differences between speech and musical rhythm, as well as perspectives as for further research.

1.2 State of the Art

1.2.1 Speech Rhythm Description

The first mention of a possible description for speech rhythm originates in the phonetics community, with Pike (1945) and Abercrombie (1967) both mentioning that different languages exhibit a specific rhythm, and tentatively suggesting that rhythm language families exist, dubbed the *Speech Rhythm Class Hypothesis*. Inside those hypothesized families, the rhythm of the languages should manifest itself through a perceived *Morse-code* pattern between the syllables (also denoted as stress-timed languages, indicating an isochrony² between the stressed parts of an utterance) or a *machine-gun* pattern (syllable-timed languages), suggesting an isochrony between the syllables themselves³. The first

² The term *isochrony* denotes that for a series of events unfolding in time, there is an equal duration of time occurring between successive events.

³ A third category of *mora*-timed languages is sometimes added to this distinction, from the basic phonological units in Japanese prosody, but the basic distinction in most publications remains between the two other categories.

assumption for the source of this difference was that it lies in the differences between syllable and pattern durations, resulting from phonological differences between languages. Although this hypothesis seems to have a solid basis in the perception of different languages (Ramus et al. 1999). Although Ramus et al. (1999) proposed a first measurement method, there was no clarity whether such methods could capture speech rhythm and differences between languages based on it in a consistent, efficient way. However, the issue sparked a body of research that tried to prove or disprove this claim. To begin with, one study from Lehiste (1977) showed that the concept of isochrony between syllables and stresses cannot be empirically proven for several languages examined in this paper. The use of the metrics on several small corpora showed interesting tendencies, with clusters of languages emerging when analyzed on the basis of two rhythmic metrics (Grabe & Low 2002). Nevertheless, there were two basic shortcomings:

- a) The results were generally seen to have limited validity, due to the rather small amount of data analyzed and the large number of external factors (speaker, sentence, measurer) having a significant effect on the metrics (Wiget et al. 2010).
- b) The results were not entirely consistent between studies, hinting at a limited reliability, which could be attributed both to methodological matters and to the metrics themselves, when performing experiments with a variety of languages, speakers, materials, and elicitation methods; while the metrics also showed limited robustness to speaker, elicitation method and syllable complexity (Arvaniti 2012).

Both of the above observations led to the realization that either those metrics were not pertinent or relevant enough for the distinction between languages to reach an overall valid conclusion. Possible reasons for this observation are that the metrics do not measure only language-specific rhythmic differences, or rhythmic differences at all. It is also possible that the amount of data to be analyzed was too small or too dependent on external factors. Another possibility is that the hypothesis itself refers to a phenomenon taking place primarily at a perceptual level: This would mean that perception of speech rhythm is guided through cues in the signal in a bottom-up way which allow the extraction of rhythmic structure. Apart from that, it could also involve top-down cognitive processes, through which the listener imposing rhythm structure, which cannot be directly found in the signal itself. Another possibility would be that the rhythmic patterns are readily available in the speech signal and can therefore be directly measured, but that more specific metrics should be used to do that.

The observed variability of temporal patterns in speech (perceived or analyzed) can have several sources. Most prominent among them is the character of the specific language under consideration (Pike 1945, Abercrombie 1968, Ramus et al. 1999, Grabe & Low 2002) and the characteristics of the specific speaker involved (Dellwo et al. 2015). Others include the emotional state of the speaker, their gender, or the elicitation type (Wiget et al. 2010). For specific classification tasks or analyzes which focus on differences between languages or between speakers, there were attempts to neutralize other temporal variability sources either by using a large enough sample size or by keeping the other variability sources constant in the respective dataset. However, although the aforementioned literature provides certain substantial evidence on the matter, the main open question is the extent to which each of these factors influences rhythmic variability, how this relates to the metric or feature used to describe it and to the task at hand, and whether this observed variability is a perceptual rather than a signal-related matter.

The considerations mentioned in the previous paragraphs led us to the idea of approaching the speech rhythm question in a different way, using more advanced methods, and thus attempting to redefine the speech rhythm problem (Turk & Shattuck-Hufnagel 2013). Definitions exist which involve the temporal patterns resulting through points of prominence in a phrase (“*Rhythm [is] the systematic patterning of timing, accent, and grouping in sequences of events*”, Patel 2008, p. 96) or which involve aspects of timing of speech elements (“*speech is characterized by timing patterns that, while not necessarily isochronous, are the result of systematic relationships between factors such as segmental identity and context (both segmental and prosodic) on the one hand, and surface duration patterns on the other*”, Turk & Shattuck-Hufnagel 2013, p. 98). However, in earlier studies, it was still not clear which elements (syllables, phonemes or stresses or others) are relevant for timing (Ramus et al. 1999, Ramus 2002). Although the notion has been put forth that speech rhythm implies periodicity of the speech elements (Lehiste 1977), further studies have concluded that this is not necessarily the case (see Turk & Shattuck-Hufnagel 2013). There, it has also been surmised that some kind of periodicity in speech is primarily a perceptual impression or abstraction, which might not necessarily have a direct correlate in the speech signal. One possible mechanism for this phenomenon would be that listeners perceive cues from the speech signal which show certain tendencies of periodicity or temporal structure, but are able to derive a more rigid structure and impose it on the signal by applying well-formedness rules (Lerdahl & Jackendoff 1983), similar to the perceptual phenomenon of meter in music (London 2012). Another possibility is that the correlates are various signal quantities that vary in time according to the behavior of higher-level speech elements. Investigating their periodicities and statistical properties, similarly performed in the analysis from Tilsen & Johnson (2008) and Rouas (2007), could provide information that can be useful for machine learning rhythm classification tasks.

This leads to the idea which was presented and applied in papers [1, 2, 3, 5, 6] of this thesis: to use multiple signal novelty functions in order to track several important signal quantities such as the fundamental frequency, the spectral flux, the spectral centroid, the spectral flatness, and the RMS amplitude. Here we provide a definition: A *novelty function*, also known as *onset function* or *onset strength signal* (Bello et al. 2005), is a signal derived from an audio waveform or spectral representation thereof, denoting time points in the audio where important events such as the onset of a note (for a musical signal), a change in spectral content or the introduction of a new instrument or voice take place. The novelty function takes a larger value whenever a new event emerges and lower values for parts of the audio signal where no changes can be detected. There are several ways to calculate a novelty function, based on spectral changes, phase discontinuities and loudness fluctuations (Bello et al. 2005), or more recent deep learning based methods, in which neural networks can derive the form of an onset by being provided with labeled examples (Schlüter & Böck 2014). The interested readers can refer to the aforementioned articles, the thesis publications [1, 2, 3, 5, 6] or Lerch (2012) for more information on these features and their method of calculation. The novel approach of this thesis lies in the realization that, in principle, every instantaneous feature (a signal-related quantity describing timbre, rhythm or other signal properties) or its first-order difference (the function of its change in time) extracted from the audio signal could serve as a novelty function for further processing and periodicity analysis. Although the relationship of those features to higher-level speech elements such as syllables or phones is an open research topic, the information provided from them should be sufficient to capture a significant amount of the temporal behavior of

speech and music signals, and the approach conforms to models of rhythm perception and cognition for music (Jones 2009). Up to the submission of this thesis, no other works are known to the author that have taken a similarly detailed approach of using multiple novelty functions for rhythm description and assessing their suitability for different tasks.

1.2.2 Music Rhythm Description

Describing and analyzing musical rhythm has been a much more intensively researched topic than that of speech rhythm. This is largely because rhythm in music is much more clearly defined due to the musicological constructs underlying it. It is also much more regular in its structure due to its function, e.g., the facilitation of movement, and the need for reproduction and notation (Patel 2008). Although this thesis focuses primarily on speech rhythm analysis and a thorough discussion of musical rhythm would go beyond the scope of this text, we will provide some information on musical rhythm.

The most influential definitions and descriptions of musical rhythm come from the field of musicology, specifically from Cooper (1960), Fraisse (1982), Lerdahl & Jackendoff (1983), Lester (1986), and, more recently, London (2001). We will focus here on the latter study, which defines rhythm as “*the sequential pattern of durations relatively independent of meter or phrase structure*” (London 2012, pp. 277-309). London’s definition emphasizes the *durational* aspect of rhythm, meaning the fact that rhythmic patterns in music can be precisely defined by the kind and relationships of durations between important sound elements. The latter are frequently defined as *accents*, referring to points in a musical piece which are in some respect different from their neighboring ones or, in general, stand out for perception. These accents can be defined as the onsets of new sounds, as louder – or generally in some way stressed or emphasized – and longer sounds, as well as events which are structurally important in a tonal context (such as the root note in a musical scale). Other, more abstract, rhythmical properties of these important points are considered, such as their relation to the musical bar and their position in a musical phrase, as well as aspects concerning tempo and timing of events. Even more high-level aspects such as the regularity or the complexity of a rhythmic pattern can be defined, from the moment that a pattern has been precisely established. Finally, various definitions of musical rhythm in music analysis exist depending task at hand (for a thorough discussion see Bello et al. 2015).

Rhythm analysis in the field of Music Information Retrieval (MIR) has been an important topic since the late 1990s, primarily because rhythmic elements could be more easily defined for music, at least for western musical styles (Esparza et al. 2015). This also occurred due to the larger range and relevance of commercial applications related to the analysis of musical rhythm. The first efforts were made with the development of algorithms for accurate onset detection and beat tracking (Scheirer 1998). Based on these first results, several other studies used and refined the methods to extract rhythmic information from a signal, whether they related to simple, statistical descriptors or more complex properties such as the tempo of a musical piece. The algorithms developed were based on two assumptions. Firstly, that important, salient points in the signal (for instance based on their high amplitude or energy), such as the ones marking the beginning of notes or chords, are vital to the constitution of rhythmic percepts. This led to the extracting and detecting of onsets and onset functions, and algorithms were developed towards this goal (Bello et al. 2005). Secondly, that an representation of the signal’s inherent periodicities in the “rhythm area” (lower frequencies from 0.5 to 10 Hz) for a musical excerpt can provide information about rhythmic patterns (since periodicity is an important part of most musical rhythms). The

basis for such an assumption concerning musical signals is explained by Gouyon (2006), where it is stated that „Rhythm is about recurring musical events occurring roughly periodically“. If these events are represented by low-level temporal features extracted from the signal, a subsequent analysis of their periodicities should provide us with an acceptable representation of the temporal patterns of instruments⁴. This reasoning led to the development of methods such as the Beat Histogram (Tzanetakis & Cook 2002, Gouyon et al. 2004) and the subsequent extraction of information about rhythm patterns. Based on these a novelty function, it was possible to extract audio descriptors (called *features*) by creating a representation which shows the strength of the prominent periodicities in the signal. These included both so-called low-level features, such as statistics on the distribution, and more complex ones, such as the tempo, as the frequency defined by the most prominent peak in the Beat Histogram. These features could be used as input for machine learning algorithms to solve classification tasks. Finally, specific pattern templates were extracted through self-similarity calculation methods, which could be used for direct similarity tests against another pattern (Foote & Uchihashi 2001).

Similar approaches to the ones described here and their variants have been used in a number of tasks over the last years (Tzanetakis & Cook 2002, Burred & Lerch 2003, Gouyon et al. 2004, Pampalk et al. 2005, Gouyon & Dixon 2005, Peeters 2011). These approaches have also been expanded to include the possibility of performing direct rhythm similarity measurements between patterns (Pohle et al. 2009, Esparza et al. 2015). Finally, more recent approaches use deep learning (which denotes machine learning approaches which use neural networks with a large number of layers and nodes) based methods for directly extracting rhythmic information from the musical signal, such as onsets (Schlüter & Böck 2014). Those approaches have shown satisfactory results in terms of performance when used alone or in connection with other audio features in relevant tasks, such as genre classification, rhythmic similarity calculation, and tempo estimation. However, those representations are still lacking when more complex signals (such as polyphonic music) or tasks (such as downbeat detection) are involved, mostly because the gap between the properties that the features can describe and the complexity of the tasks is too large (also referred to as the semantic gap, see Lerch 2012). Furthermore, the features used are, in most cases, low-level ones and can therefore be noisy and uninformative. Furthermore, the features might relate to the solution of a very specific task without providing information as to which elements are crucial for rhythm analysis. The basic problem in all these cases is the semantic gap between the high-level understanding of rhythm by humans and the low-level representations that are suitable for audio content analysis (Lerch 2012). One way to alleviate this situation is to extensively evaluate which audio features carry the most relevant periodicity information for specific classification tasks. This is the main premise of the publications in this thesis, both for music and for speech signals, providing the basis for further research. This could either be helpful for informing deep learning methods or for creating handcrafted features that are better suited for the tasks at hand.

⁴ For speech signals, a similarly based assumption has been proposed by Turk & Shattuck-Hufnagel (2013), where „surface (near)-periodicity in the signal is imposed by periodicity-based control structures“, with possible congruence of periodicities occurring in the motor control system (for speech production) or in speech perception. Such observations of the important role of periodicity in rhythm-related processes have also formed a basis for its subsequent analysis in this thesis.

1.3 Methods

1.3.1 Feature Extraction

The algorithms developed for the extraction of features describing rhythm – and their adaptations used in this thesis – were based on two assumptions. Firstly, that salient points in the audio track are vital to the constitution of rhythmic percepts. These manifest themselves in the signal due to their high amplitude or energy, a change of melody or key for music, the beginning of a new phrase, note or chord (or any sound after a pause, in essence the appearance of a new sound event), but can also be parts of the audio track which differ greatly in any aspect from their surrounding events, making them perceptually salient (see Lykartsis (2014) for a detailed description of these elements, often named *accents* in the music theory literature). From these salient points and based on principles of grouping, tonal structure, well-formedness rules and the influence of the perception of meter by a listener, patterns of events in the signal can be identified as rhythmic objects, leading to their perception (Lerdahl & Jackendoff 1983, London 2012). The detection and extraction of the correlates of such accents in the signal was seen as a relevant goal. In the MIR community these perceptually salient points are most commonly named *onsets* (whereas their trajectories over time would be named *onset functions*), and several algorithms have been developed towards this goal (Bello et al. 2005). These onset functions have also appeared with other names in several publications, such as onset strength signal or novelty function. There are several ways to capture novelty in the signal, for example, by taking the first difference between consecutive Short-Time-Fourier-Transform (STFT) spectrogram slices, denoting changes in the spectrum (spectral-flux-based methods). Another way involves calculating phase differences between different signal parts, which denote discontinuities in the signal, which can occur at transitions from one sound to another. Finally, one might just look for peaks in the signal amplitude, denoting changes in loudness or energy. All of these methods have been used with success and their suitability has been found to be task- and signal-specific.

In this thesis, we used the nomenclature of novelty function (described briefly in 1.2.1), as we find that it is a general term which indicates change in any possible signal quantity and does not focus on specific note onsets. We calculate novelty in a way specific to this thesis, and here we expand the definition to include any signal (or any instantaneous feature, see Lerch 2012), capturing changes in an audio signal. Rhythmic properties can be then extracted by analyzing the novelty function. The latter can refer to aspects of the signal which involve its spectral content, its energy fluctuations or other quantities⁵. An example of extraction of a novelty function would be calculating the fundamental frequency (F_0) of a voice signal with a pitch tracking algorithm and taking the first difference in order to track changes. This would indicate, for example, a voice tone rising or falling, staying constant or exhibiting certain regularities. For a concrete example, the reader can refer to publication [1] of the thesis, in which two different features of a speech utterance (F_0 and spectral flux) are analyzed and the difference in the resulting novelty function depicted. By tracking several of these signal quantities, it can be assured that different kinds of signal changes (be they spectral-, tonal-, or loudness-based) are captured. Through their different periodicities they can also convey valuable information concerning the overlapping and interacting rhythmic contents of the signal. For example, the fundamental frequency of a voice does not change

⁵ As, for instance, noisiness or the fundamental frequency of a voice audio track.

in time in the same way as the spectrum of the sound in general, which is much more dependent on the kind of phone produced by the speaker (for instance, whether it is a vowel or a consonant). Secondly – and after such novelty functions have been extracted – the methods mentioned here take into account that a representation of the signal's inherent periodicities in the frequency area relevant for rhythm (i.e., lower frequencies, from 0.5 to 10 Hz) for parts of an audio excerpt can provide information about rhythmic patterns - given that periodicity is an important part of most rhythms found in nature or are artificial (e.g., music). This does not mean that every audio signal exhibits just one periodicity, but in every signal there exist predominant periodicities, as well as less articulated, more complex and irregular patterns. Such a periodicity representation is in principle a distribution of these periodicities, whose form could be characteristic for a specific rhythm style, a musical piece, or even to a language or a speaker.

These considerations led to the development of methods such as the Beat Histogram (Tzanetakis & Cook 2002, Burred & Lerch 2003, Gouyon et al. 2004) and the subsequent extraction of (recurring) rhythm patterns. The Beat Histogram can also be found under several names in the literature, e.g., Periodicity hHistogram, Inter-Onset-Interval (IOI)-Histogram or Rhythmogram. The basic principle is to use a periodicity extraction method such as an autocorrelation function, a discrete Fourier transform or a resonant filter bank (Tzanetakis & Cook 2002, Burred & Lerch 2003, Gouyon et al. 2004). In this thesis we used the method proposed by Tzanetakis and Cook (2002) and extract an enhanced autocorrelation function of a relevant novelty function (refer to papers [1, 2, 3, 5, 6] for specifics). In our case however, no special preprocessing or filtering of the signal is undertaken in general, but only in the cases where such processing is justified through the feature extracted. Therefore, this was only performed prior to the extraction of the fundamental frequency F_0 : there, a filtering of frequencies above 800 Hz in the original signal is conducted since these do not play a role for the F_0 , and this way the processing is made easier and more precise. At the end, a periodicity distribution in the desired frequency range is acquired, which can be then treated as a common distribution of values. This is meant in the sense that further statistical properties can be extracted from it, and the specific form of the distribution provides quantifiable rhythmic information for the analyzed excerpt.

1.3.2 Machine Learning

In the thesis the problem of rhythm analysis was framed as one of supervised classification: novel rhythm features, inspired from rhythm analysis methods in the area of MIR were developed and adapted for speech (s. Chapter 1.3.1). Those were then extracted from several recordings categorized per language or per speaker (depending on the dataset used in the respective publication). The features were then used as input to several state-of-the-art machine learning algorithms, so that they could extract rules (i.e., “learn” in the machine learning literature) the specifics of each language/speaker, and be able to identify new, unknown recordings based on their corresponding features.

Such methods are called supervised learning methods, and the most widespread of those are Support Vector Machines (SVMs), Decision Trees and Random Forests (DTs/RFs), k-Nearest-Neighbors (k-NNs) and Neural Networks (NNs) (see Duda et al. (2012) and Bishop (2006) for more information on these algorithms). All of these methods use a variation of a

mathematical optimization method in order to find a separating hypersurface⁶ in a multidimensional space spanned by the vectors containing the features for the samples (called *feature vectors*). This plane should ideally separate samples which have different qualities, expressed through the values in the features. The term “supervised” refers to the fact, that the algorithms take as input the classes (called *labels*) of the samples and learn the special connection of the features to those labels. The learning of this correspondence is called *training*, whereas the evaluation of the algorithm’s performance on unseen samples (which have not been used during training) is called *testing*. For example, for a classification problem with two languages, the labels would be “English” and “French”. A feature used to discriminate between those two languages would be the tempo (i.e., average rate of change) of the fundamental frequency an utterance (one sample) in one of those two languages, calculated through an F_0 novelty function and its corresponding Beat Histogram. The algorithms are usually trained (in learning the patterns) with one part of the samples, creating a model of this rule. This model then predicts the labels of samples which were not used for training and its performance is evaluated. If the performance is not satisfactory, the parameters of the model or the feature extraction are adapted. This procedure of training and testing is usually repeated with several random splits of train and test set (called cross-validation). This is done in order to assure that the results are not biased due to a specific selection of training and test samples. The results are compared to a baseline system (with the same machine learning algorithm but different, well-known features), in order to quantify the improvement of the new development. As with most machine learning systems, the whole procedure is one of trial and error – after deciding on which features to use, the algorithms have to undergo a grid search for their parameter values in order to find the optimal performance. Another optimization that can be performed automatically through machine learning is feature selection. When the parameter optimization does not provide any further gain, an alternative option is to select the best features and discard those which are redundant or irrelevant to the task. This can be done by *filter methods*, which are similar to a correlation, or *wrapper methods*, which use the machine learning methods themselves and retain the features which provide the best performance (Guyon & Elisseeff 2003). This procedure is not only useful for improving the classification performance, but also in order to find the most important features for the task at hand.

The performance of the algorithms is evaluated through the calculation of measures such as accuracy, which is the ratio of the number of correctly classified samples to the number of all classifications made. These are valid for one class only (specifically for one language or speaker) but can be averaged over all classes in a multi-class scenario, where every class is compared against each other, due to the fact that most algorithms allow for binary classification. This classification scheme allows some conclusions: If the rhythm features (used alone) perform better than other, non-rhythm-related features, or should they increase the classification accuracy significantly when used in combination with other features, they will then be thought of as informative enough concerning rhythm properties. If specific rhythm features show more importance for classification (defined through feature selection procedures), the corresponding rhythm properties can be deemed to be generally important for tracking rhythmic differences between languages or speakers. In a classification scenario where only rhythm features are used, the achieved accuracy can allow conclusions concerning the data used. It can also highlight which data are optimal for rhythm-based classification and which are not. The same is valid for pre- and post-processing techniques. Finally, should the classification show that by using the rhythm features only, differentiation

⁶ A hypersurface is a surface in a space which spans four and more dimensions.

between some languages is easier than between others, this would lead to a conclusion: that at least when examined on the basis of these features, some languages share rhythmical properties with others, therefore making discrimination between them difficult. Other languages will have distinct rhythmical properties which allows for their clear classification. Such conclusions are potentially useful for testing linguistic/phonetic hypotheses such as the Rhythm Class hypothesis. Another useful tool for this purpose are the confusion matrices that can be extracted for each classification scenario. These contain in the columns the predicted classes, and in the rows the actual classes, so that the diagonal elements show the correct classifications, and all others that are incorrect. By observing which classes are misclassified for others and knowing the features and parameterizations used, more discriminative features and better parameterizations can be found. In any case, the misclassifications can show which classes are similar to each other, which can be rhythmically similar genres for music or rhythmically close languages.

Recently, a renewed interest in classification and data mining systems employing deep learning, which denotes methods using neural networks consisting of multiple layers and neurons in each one of them. The interest has been rekindled mainly due to the rise of computational power allowing their extended use⁷ but also due to the large amount of data which has become available to companies and research institutions. Deep learning was primarily used in image processing applications (since they offer a large amount of data and potential applications), but then was also expanded to other areas by adapting the methods to further fields. Several variants emerged such as convolutional neural nets (CNN), recurrent neural nets (RNN) and combinations of those, which were then employed for a variety of problems in MIR, such as onset detection and music transcription. In speech technology some attempts have been made, but none for speech rhythm analysis. Therefore, in future work, we want to employ deep learning methods for rhythm-based LID and SID. In order to do so, a substantial amount of data is required, since these methods work with low-level representations of the data (e.g., waveforms or spectrograms). This means that due to the relative simplicity of the input representation these methods require a multiple of the number of instances necessary for basic ML methods (see, e.g., Schlüter & Böck 2014, on onset detection for audio). Therefore, future efforts should focus on acquiring this data and applying deep learning methods on them.

1.3.3 Datasets used in the Thesis

In order to test the usefulness of the rhythm features developed in the context of the thesis, two main datasets were used for experimentation on LID, and one on SID⁸. A short description of each is given in the following:

The MULTEXT Prosodic Database is a dataset of read speech, containing approximately seven hours of speech in five languages, roughly equally distributed (English 150, French 100, German 150, Spanish 100, Italian 100 recordings). For each language, ten speakers (five female and five male) are available, and the passages read are texts describing

⁷ This happens since the computational load rises significantly with the amount of network complexity, requiring GPUs and other high-performance systems in order to process in a plausible time.

⁸ The datasets used for the music related publications [2, 3] are not included in this section for reasons of brevity and because of their lesser relevance to the topic of the thesis. Readers can refer to the citations in the publications themselves.

everyday situations. The average duration of the recordings is 30 seconds. The sample rate is 22050 Hz, with a 16-bit depth. For more information, see Campione & Veronis (1999).

The OGI-MLTS database contains telephone quality speech (16000 Hz sampling rate, 16 Bit depth), which is separated into one word or short phrase recordings (which are not suitable for rhythm feature extraction), and longer monologues (approx. 45 seconds long) that can be analyzed in longer parts. All of the material of longer duration is spontaneously produced. The 11 languages involved are Vietnamese, Tamil, Korean, Japanese, Hindi, French, English, German, Spanish, Mandarin Chinese, and Persian. The dataset is unbalanced (comprising many more utterances for English than for the other languages), so that for most experiments in the studies included here, a balanced version was created to facilitate a more valid classification⁹. For more information, the readers can refer to Muthusamy et al. (1992).

Finally, the TEVOID corpus of Swiss German contains 16 speakers recounting longer phrases of Swiss German, with 256 utterances per speaker in total, resulting in 4096 utterances. This is a balanced dataset with read and spontaneous speech, more information on which can be found in Dellwo et al. (2012).

These datasets were selected due to three reasons: Firstly, their availability, in contrast to other datasets which are too expensive or available only in the specific context of a task contest. Secondly, their use in several other studies for language and/or speaker identification makes comparability with these studies possible. And thirdly, due to the fact that they feature a large variety of languages, speakers and utterances make the results with the methods used in this thesis sufficiently valid.

1.4 Summary of Original Achievements

The main contribution of the thesis is found in paper [1], where the adaptation of the Beat Histogram features for speech, as well as their use for the task of LID are presented in depth. There, it could be shown that features based on Beat Histograms with multiple novelty were helpful for LID, even more than previous features used for the same task on the same dataset (Rouas et al. 2005, Rouas 2007). A marginal improvement for classification accuracy on the two datasets used in this thesis, as well the feature selection performed shows that the features used do indeed capture relevant rhythmic information. The similarity between the features used in the aforementioned studies and the ones used in this thesis hints towards the need to further refine the methods providing less noisy features. The choice of classification algorithm did not seem to play an important role, hinting towards the fact that established algorithms such as the Support Vector Machines are sufficient for the task at hand, at least for datasets of the size used in this thesis. In the future, deep learning algorithms might be able to achieve even better results for LID (e.g., see Lopez-Moreno et al. (2014)), but more data will be necessary to achieve this goal.

These results corroborate and expand the ones in contribution [3], where the method used in [1] is applied to rhythm-based genre classification, on a large and diverse amount of musical data with the multiple novelty function approach. The latter means that several audio features are used as the basis for the calculation of Beat Histograms, capturing various periodicities in the audio signal. In this publication, it could be shown that using multiple

⁹ Most machine learning algorithms tend to favor the largest class in an unbalanced dataset, thus producing erroneous classification results.

novelty functions based on various instantaneous features extracted from the musical audio (Lerch 2012) can provide good classification performance, both when used alone and when used together with other, non-rhythmic features. These results attest to the quality of the multiple novelty based Beat Histogram features and show that using them for further tasks can be promising. Based on those observations, another study was attempted which dealt with more specific musical signals (i.e., percussive ones) so as to determine which factors play a role in the extraction of useful features for rhythm-based genre classification. Contribution [2] applies similar methods as done in [1] and [3] for musical genre classification. A novel Non-negative Matrix Factorization (NMF)-based source separation approach (Wu & Lerch 2015) is used to produce activation patterns¹⁰ of different drums which are then used as novelty functions. The results show that this approach is useful for rhythm-based genre classification for the percussive instruments only. Use of annotated metrical boundaries as a means of extracting information from meaningful audio parts has been shown to improve results, pointing towards the need to apply similar methods for speech, for example using phrase boundaries to segment the signal and extract Beat Histograms from them.

With respect to the datasets used for LID, another important point is the much better classification accuracy achieved using high-quality signals, most probably due to the end features being noisier when the signal itself is noisy. A difference in elicitation method (spontaneous vs. read speech) was also observed, with the latter providing better results, owing to less rhythmic variability in the signal with respect to the speaker, as opposed to the language. Finally, better classification results were achieved when analyzing Indo-European languages. However, since other language families were only part of the lower-quality dataset and less in number, it would be premature to suppose that the features are not applicable in that case. More data is necessary in order to produce clearer results.

Concerning the information used for classification, features such as the speech rate (measured through the tempo, extracted from the Beat Histogram) and basic statistics of the Beat Histogram (mean and standard deviation) have been shown to be among the most informative features. Concerning the novelty functions used to create a Beat Histogram, salient quantities such as the spectral flux, the RMS energy, the fundamental frequency and the MFCCs seem to be most relevant for the extraction of rhythmic information, probably due to their connection to phonetically relevant quantities. However, the relationships of those features or acoustic correlates to perceptual correlates can be expected to be rather complex (Wagner 2008) and are a topic of further research.

After conducting those experiments, a summary and comparison of the results was necessary, especially in order to clarify similarities and differences between rhythm features from speech and musical signals. Both of the included Deutsche Gesellschaft für Akustik (DAGA) publications [7, 8] summarize results up to a certain point of the thesis. Chronologically, [7] is the first one and provides a deeper insight into the results in [1]. Publication [8] presents a comparison between the music related ([3]) and the speech related ([1]) results. These results also led to a need to clarify – on a perceptual level – which cues listeners use in order to differentiate musical and speech rhythm.

Publication [4] provides a first attempt to pinpoint the perceptual phenomena underlying speech rhythm perception and establishes two quantities which are important for

¹⁰ An activation pattern is a temporal pattern showing when a specific drum was active and when not. It closely resembles a novelty function and can be seen as a special form thereof.

discrimination between singing (speech with musical structure) and normal, read speech: the *tempo* and the *regularity*. In order to also test the extent to which the features are speaker-specific, we conducted a further study. The objective was to test whether the speech rhythm features capture mostly language or also speaker variability. Paper [5] applies the same methods as in [1] for speaker identification, showing that, at least for the case of Swiss German, the proposed rhythm features cannot capture variability between speakers as well as regular timbre-describing features do. The analysis has shown that the Beat Histogram-based rhythmic features are less useful for capturing speaker-specific information. This could be owing to the specificity of the language (Swiss German) in the dataset or its size. However, the large discrepancy in performance hints at speaker specificity being mostly based on timbral, and not rhythmic information. Therefore, further research in the SID area could be advised to discard rhythmic information or use simpler metrics that have been shown to be speaker-specific (Dellwo et al. 2015). Nevertheless, using the Beat Histogram features for tasks such as dialogue success prediction could provide satisfactory classification results even for lower signal quality, and even when the language was just one (the English language) and the only variable between the recordings the speaker or their emotional state. This last result shows that the language analyzed as well as the task at hand might play an important role in the suitability of the Beat Histogram features.

On a phonetic analysis level, it was impossible to confirm or reject the Rhythm Class Hypothesis, at least as defined in (Ramus et al. 1999) and adapted to the methodology of the thesis. According to this it is expected that groups of languages would emerge, which would be more difficult to distinguish¹¹ from each other (but less difficult to distinguish between languages belonging in different hypothesized groups), when using machine learning methods with rhythm features as an input. Misclassifications observed for Indo-European languages did not show a clear pattern (which would be a distinct difficulty in differentiating between German and English or between French and Spanish or Italian, which were postulated to be members of the same rhythmic family). Still, tendencies were observed which hinted at rhythmic similarities as predicted by the Rhythm Class hypothesis. These findings show that even when using more complex rhythmic features, a clear proof or rejection of the hypothesis is not possible. In the opinion of the author, the hypothesis should not be discarded, but rather revised. It is possible that rather than fixed families, there exist clusters of languages when observed through their rhythmic qualities – while at the same time, those clusters will vary dependent on other rhythmic variability factors such as speaker specificity, elicitation method, the speaker's emotional state, or even the utterance itself. A revised hypothesis should aim at creating a multifactorial model for rhythm (see Wagner 2008) which, given the specific parameters of the signals under consideration, can result to different clusters for each specific use case.

Finally, it should be noted that the rhythmic differences or similarities between the languages might be the cause of a phenomenon perceptual in nature: Since no marked rhythmic differences could be observed when producing features from the signal itself, it can be surmised that the similarities or differences observed when hearing languages refer to a complex set of rhythmic and other differences which create an overall sense of similarity. Nevertheless, other studies (Ramus et al. 1999, Wagner 2008) and one from this thesis (Stanev et al. 2016) have shown that differentiation between stimuli based on rhythmic information only (in the first case between languages, in the second case between read speech and acapella singing) is possible, especially on the basis of the perceived tempo and

¹¹ Samples of one language in a group would be misclassified as belonging to the another language in a group.

regularity (Stanev et al. 2016). This seems to hint towards the fact that the abstraction of speech rhythm properties is, to a large extent, a perceptual one: Listeners take into account rhythmic qualities that can be extracted from the signal, but their perceptions on rhythmic differences between languages cannot be directly attributed to changes of signal properties, at least not in the way these were extracted in this thesis. Therefore, more research towards the determination of relevant factors in speech rhythm perception is suggested.

Finally, in order to expand the field of research to other possible areas beyond LID and SID or genre classification, the Beat Histogram-based speech rhythm features were used for success prediction of spoken dialogue systems (SDS). The final contribution [6] shows that speech rhythm features, used in combination with other, more general features can be very helpful in categorizing dialogues of spontaneous speech in English as successful or not, hinting at the fact that speech rhythm features can capture mood-related qualities of speakers. These are manifested through features such as speech tempo, which was also shown to be an important factor in LID.

1.5 Conclusion and Outlook

From the preceding summary of original achievements, it becomes clear that there are several research tasks yet to be undertaken. Firstly, although a number of classification experiments were performed in the context of the thesis, the matter of pinpointing the correct parameters for feature extraction and classification is still an open subject. A rudimentary parameter search was already undertaken for the studies included here, but an in-depth analysis of which parameterizations are suitable, for example, for different signal types, can be very helpful. This would help determine how efficient classification can be made in cases where the results of this study were not optimal (in the case of the lower quality, spontaneous speech samples in the OGI-MLTS dataset). Another in-depth analysis would concern the relation of the extracted features and novelty functions to higher-level speech elements such as vowels, consonants and syllables. To that end, it is necessary to acquire reliable manual annotations for the data, which is a difficult and time-consuming task. However, the evaluation of how the novelty functions relate to speech elements might help increase their precision and therefore the quality of the features extracted from them.

Furthermore, the feature extraction methods used in this thesis are handcrafted features that are engineered to provide the machine learning algorithms with the suitable information for classification. One of the drawbacks of such methods is that although theoretically the features are interpretable and capture exactly the information that is deemed to be important for the task at hand, the algorithms may not be able to utilize them for high performance classification. This is because the (often expensive) calculation process might lead to them being noisy or even faulty. Apart from that, the feature design can be a long, tiresome and error-prone process. For that reason, state-of-the-art methods using deep learning can be used. It has been shown that for a multitude of tasks, if enough data is provided, deep learning methods can learn high-level information from low-level input. This can be achieved by learning many small variations of a given pattern, for example a note onset (Schlüter & Böck, 2014). Along these lines, apart from using the latter representations, other rhythmic information can be used, too. These might be onset functions (on the respective level of abstraction similar to a pure audio waveform) or high resolution Beat Histograms, which can be considered as a mid-level representation, similar to a spectrogram but conveying only periodicity information in the lower frequency area. In any case, a larger amount of data is

necessary to train a deep neural network, required to extract the higher-level features. Therefore, one of the crucial next steps would be to acquire more and larger datasets, such as the GLOBALPHONE (Schultz 2002). This comprises many hours of audio, featuring both multiple speakers and 20 languages, making it an important resource for both LID and SID experiments.

Finally, a still-open matter is to which extent the rhythmic differences between languages and speakers is a perceptual, top-down phenomenon, or to which extent there are specific signal cues, which listeners attend directly to in order to detect differences and similarities. Therefore, more listening experiments as in Stanev et al. (2016) are necessary to pinpoint specific adjectives describing the perception of speech rhythm. This is similar to the studies of Gabrielsson (1973a, 1973b, 1973c) who created an inventory of terms for musical rhythm. Specifically, preliminary studies will gather and select a large inventory of terms for general rhythm description (music and speech). A focus group will then create a final list that can be used to design and develop pertinent and relevant audio features describing speech rhythm. These can be expected both to increase classification performance in LID and SID tasks and to shed light on the exact connection between rhythm percepts and their acoustic correlates.

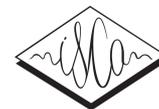
1.6 References

- Abercrombie, D. (ed.) (1967): *Elements of General Phonetics*. Edinburgh University Press.
- Arvaniti, A. (2012): "The usefulness of metrics in the quantification of speech rhythm." In: *Journal of Phonetics*, vol. 40, no. 3, pp. 351-373.
- Bello, J. P.; Daudet, L.; Abdallah, S.; Duxbury, C.; Davies, M.; Sandler, M. B. (2005): "A tutorial on onset detection in music signals." In: *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035-1047.
- Bello, J. P.; Rowe, R.; Guedes, C.; Toussaint, G. (2015): "Five perspectives on musical rhythm." In: *Journal of New Music Research*, vol. 44, no. 1, pp. 1-2.
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.
- Burred, J. J.; Lerch, A. (2003): "A hierarchical approach to automatic musical genre classification." In: *Proceedings of the 6th international conference on digital audio effects*, pp. 8-11.
- Campione, E.; Véronis, J. (1998): "A multilingual prosodic database." In: *Fifth International Conference on Spoken Language Processing*, vol. 98.
- Cooper, G. W.; Meyer, L. B. (1963): *The rhythmic structure of music*. University of Chicago Press.
- Dauer, R. M. (1983): "Stress-timing and syllable-timing reanalyzed." In: *Journal of Phonetics*, vol. 11, pp. 51-62.
- Dellwo, V. (2006): "Rhythm and speech rate: A variation coefficient for ΔC ." In: *Language and language-processing*, pp. 231-241.
- Dellwo, V.; Leemann, A.; Kolly, M. J. (2012): "Speaker idiosyncratic rhythmic features in the speech signal." In: *Thirteenth Annual Conference of the International Speech Communication Association*, Portland, US.
- Dellwo, V.; Leemann, A.; & Kolly, M.-J. (2015): "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors." In: *The Journal of the Acoustical Society of America*, vol. 137, no. 3, pp. 1513-1528.

- Desain, P. (1992): "A (de)composable theory of rhythm perception." In: *Music Perception: An Interdisciplinary Journal*, vol. 9, no. 4, pp. 439-454.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.
- Esparza, T. M.; Bello, J. P.; Humphrey, E. J. (2015): "From genre classification to rhythm similarity: Computational and musicological insights." In: *Journal of New Music Research*, vol. 44, no. 1, pp. 39-57.
- Foote, J.; Uchihashi, S. (2001): "The beat spectrum: A new approach to rhythm analysis." In: *IEEE International Conference On Multimedia and Expo*, pp. 224-228.
- Fraisse, P. (1982): "Rhythm and tempo." In: *The psychology of music*, vol. 1, pp. 149-180.
- Gabrielsson, A. (1973). Similarity ratings and dimension analyses of auditory rhythm patterns. I. In: *Scandinavian Journal of Psychology*, vol 14, no. 1, 138-160.
- Gabrielsson, A. (1973). Similarity ratings and dimension analyses of auditory rhythm patterns. II. In: *Scandinavian Journal of Psychology*, vol 14, no. 1, 161-176.
- Gabrielsson, A. (1973). Adjective ratings and dimension analyses of auditory rhythm patterns. In: *Scandinavian Journal of Psychology*, vol 14, no. 1, 244-260.
- Gouyon, F.; Dixon, S.; Pampalk, E.; Widmer, G. (2004): "Evaluating rhythmic descriptors for musical genre classification." In: *Proceedings of the AES 25th International Conference*, pp. 196-204.
- Gouyon, F.; Dixon, S. (2005): "A review of automatic rhythm description systems." In: *Computer music journal*, vol. 29, no. 1, pp. 34-54.
- Gouyon, F. (2006): *A computational approach to rhythm description - Audio features for the computation of rhythm periodicity functions and their use in tempo induction and music content processing*. Universitat Pompeu Fabra.
- Grabe, E.; Low, E. L. (2002): "Durational variability in speech and the rhythm class hypothesis." In: *Papers in laboratory phonology*, vol. 7, pp. 515-546.
- Guyon, I.; & Elisseeff, A. (2003): "An introduction to variable and feature selection." In: *Journal of machine learning research*, vol. 3, pp. 1157-1182.
- Hübler, S.; Hoffmann, R. (2011): "Comparing the rhythmical characteristics of speech and music—Theoretical and practical issues." In: *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*, pp. 376-386.
- Jones, M. R. (2009): "Musical time." In: *The handbook of music psychology*, pp. 81-92.
- Lehiste, I. (1977): "Isochrony reconsidered." In: *Journal of phonetics*, pp. 253-263.
- Liss, J. M.; White, L.; Mattys, S. L.; Lansford, K.; Lotto, A. J.; Spitzer, S. M.; Caviness, J. N. (2009): "Quantifying speech rhythm abnormalities in the dysarthrias." In: *Journal of speech, language, and hearing research*, vol. 52, no. 5, pp. 1334-1352.
- Lerdahl, F., & Jackendoff, R. (1983): *A generative theory of tonal music* Cambridge, MA: MIT Press.
- Lester, J. (1986): *The rhythms of tonal music*. Pendragon Press.
- London, J. (2001). Rhythm. *The new Grove dictionary of music and musicians*, 21, pp. 277-309.
- London, J. (2012): *Hearing in time: Psychological aspects of musical meter*. Oxford University Press.

- Lopez-Moreno, I.; Gonzalez-Dominguez, J.; Plchot, O.; Martinez, D.; Gonzalez-Rodriguez, J.; Moreno, P. (2014): "Automatic language identification using deep neural networks." In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5337-5341.
- Lykartsis, A. (2014): *Evaluation of Accent-Based Rhythmic Descriptors for Genre Classification of Musical Signals*. Master's Thesis, Technische Universität Berlin, Germany.
- Lykartsis, A.; Weinzierl, S. (2015): "Using the Beat Histogram for Rhythm Description and Language Identification." In: *Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Dresden, Germany.
- Lykartsis, A.; Wu, C.-W.; Lerch, A. (2015): "Beat Histogram features from NMF-based novelty functions for music classification." In: *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*, Malaga, Spain.
- Lykartsis, A.; Lerch, A. (2015): "Beat Histogram features for rhythm-based musical genre classification using multiple novelty functions." In: *Proceedings of the 18th International Conference on Digital Audio Effects (DAFx)*, Trondheim, Norway.
- Lykartsis, A.; Weinzierl, S.; Dellwo, V. (2017): "Speaker Identification for Swiss German with Spectral and Rhythm Features." In: *AES Conference on Semantic Audio 2017*, Erlangen, Germany.
- Lykartsis, A.; Kotti, M.; Papangelis, A.; Stylianou, Y. (2018): "Prediction of Dialogue Success with Spectral and Rhythm Acoustic Features Using SVMs and DNNs." In: *Proceedings of the IEEE Speech and Language Technology Workshop 2018*, Athens, Greece. © 2018 IEEE. Reprinted, with permission, from Margarita Kotti, Alexandros Papangelis, Yannis Stylianou, Prediction of Dialogue Success with Spectral and Rhythm Acoustic Features Using DNNs and SVMs, 2018 IEEE Spoken Language Technology Workshop (SLT), 2018.
- Martinez, D.; Burget, L.; Ferrer, L.; Scheffer, N. (2012): "iVector-based prosodic system for language identification." In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4861-4864.
- Muthusamy, Y. K.; Cole, R. A.; Oshika, B. T. (1992): "The OGI multi-language telephone speech corpus." In: *Second International Conference on Spoken Language Processing*.
- Pampalk, E.; Flexer, A.; Widmer, G. (2005): "Improvements of Audio-Based Music Similarity and Genre Classification." In: *Proceedings of the 6th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 634-637.
- Patel, A. D. (2008): *Music, Language and the Brain*, Oxford University Press, New York.
- 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*, vol. 19, no. 5, pp. 1242-1252.
- Pike, K. L. (1945): *The intonation of American English*.
- Pohle, T.; Schnitzer, D.; Schedl, M.; Knees, P.; Widmer, G. (2009): "On Rhythm and General Music Similarity." In: *Proceedings of the 10th International Society for Music Information Retrieval Conference (ISMIR)*, pp. 525-530.
- Ramus, F.; Nespors, M.; Mehler, J. (1999): "Correlates of linguistic rhythm in the speech signal." In: *Cognition*, vol. 73, no. 3, pp. 265-292.
- Ramus, F. (2002): "Acoustic Correlates of Linguistic Rhythm: Perspectives." In: *Speech Prosody 2002*, pp. 115-120.

- Rouas, J.-L.; Farinas, J.; Pellegrino, F.; André-Obrecht, R. (2005): "Rhythmic unit extraction and modelling for automatic language identification." In: *Speech Communication* vol. 47, no. 4, pp. 436-456.
- Rouas, J.-L. (2007): "Automatic prosodic variations modeling for language and dialect discrimination." In: *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1904-1911.
- Scheirer, E. D. (1998): "Tempo and beat analysis of acoustic musical signals." In: *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588-601.
- Schlüter, J.; Böck, S. (2014): "Improved musical onset detection with Convolutional Neural Networks." In: *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6979-6983.
- Schultz, T. (2002): "GlobalPhone: a multilingual speech and text database developed at Karlsruhe University." In: *Seventh International Conference on Spoken Language Processing*.
- Stanev, M.; Redlich, J.; Knörzer, C.; Rosenfeld, N.; Lykartsis, A. (2016): "Speech and music discrimination: Human detection of differences between music and speech based on rhythm." In: *Speech Prosody 2016*, pp. 222-226.
- Thymé-Gobbel, A. E.; Hutchins, S. E. (1996): "On using prosodic cues in automatic language identification." In: *Fourth International Conference on Spoken Language Processing*.
- Tilsen, S.; Johnson, K. (2008): "Low-frequency Fourier analysis of speech rhythm." In: *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. 34-39.
- 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*, vol. 134, no. 1, pp. 628-639.
- Turk, A.; Shattuck-Hufnagel, S. (2013): "What is speech rhythm? A commentary on Arvaniti and Rodriguez, Krivokapić, and Goswami and Leong." In: *Laboratory Phonology* vol. 4, no. 1, pp. 93-118.
- Tzanetakis, G.; Cook, P. (2002): "Musical genre classification of audio signals." In: *IEEE Transactions on Speech and audio processing*, vol. 10, no. 5, pp. 293-302.
- Wagner, P. (2008): *The rhythm of language and speech: Constraints, models, metrics and applications*. Habilitation-Dissertation, University Bonn.
- Wiget, L.; White, L.; Schuppler, B.; Grenon, I.; Rauch, O.; Mattys, S. L. (2010): "How stable are acoustic metrics of contrastive speech rhythm?" In: *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1559-1569.
- Wu, C. W.; Lerch, A. (2015): "Drum transcription using partially fixed non-negative matrix factorization." In: *23rd European Signal Processing Conference (EUSIPCO)*, pp. 1281-1285.



Using the Beat Histogram for Speech Rhythm Description and Language Identification

Athanasios Lykartsis¹, Stefan Weinzierl¹

¹Audio Communication Group, Technische Universität Berlin, Germany

athanasios.lykartsis@tu-berlin.de, stefan.weinzierl@tu-berlin.de

Abstract

In this paper we present a novel approach for the description of speech rhythm and the extraction of rhythm-related features for automatic language identification (LID). Previous methods have extracted speech rhythm through the calculation of features based on salient elements of speech such as consonants, vowels and syllables. We present how an automatic rhythm extraction method borrowed from music information retrieval, the beat histogram, can be adapted for the analysis of speech rhythm by defining the most relevant novelty functions in the speech signal and extracting features describing their periodicities. We have evaluated those features in a rhythm-based LID task for two multilingual speech corpora using support vector machines, including feature selection methods to identify the most informative descriptors. Results suggest that the method is successful in describing speech rhythm and provides LID classification accuracy comparable to or better than that of other approaches, without the need for a preceding segmentation or annotation of the speech signal. Concerning rhythm typology, the rhythm class hypothesis in its original form seems to be only partly confirmed by our results.

Index Terms: speech rhythm, beat histogram, language identification, novelty functions, rhythm typology

1. Introduction

Speech rhythm and its quantification has been an interesting and controversial matter of research, with implications for language rhythm typology and the possible existence of two or more rhythmic classes of languages (stress- and syllable-timed) [1, 2], dubbed the *Rhythm Class Hypothesis* [3]. In the last years, analysis of speech rhythm has focused on the attempt to obtain metrics of acoustic correlates of speech rhythm which could provide information about the rhythmic patterns of speech, generally by manually annotating vowels, consonants and stresses in the speech signal and consequently calculating statistics of the durations between intervals of those language prominence units, resulting in measures such as the ΔC , $\%V$, $nPVI$ and $VarcoC$ [3, 4, 5, 6]. Those metrics have been proven useful as first attempts to design descriptors of speech rhythm and were very often used to investigate language rhythm typology, by testing for significant differences between languages and attempting to position the languages in a rhythm continuum between stress- and syllable-timed [4, 5]. For various small speech corpora, they have provided evidence that supports the rhythm class hypothesis and have therefore been seen as adequate measures of speech rhythm [4, 5, 6]. However, the scientific discussion about speech rhythm and its measurement continues up to the present day [7, 8, 9, 10]. In this context, the aforementioned metrics have also been criticized [9, 11, 12, 13] as not

being robust with respect to the information they hold about speech rhythm, since differences between languages have not been consistent or significant across all studies, presumably due to the existence of many non-language specific factors affecting speech rhythm [9]. Other shortcomings are the manual annotation necessary for the procedure (which is tedious and can be subjective or erroneous), their derivation on basis of abstract language elements (i.e., syllables) as opposed to quantities physically manifest in the speech signal (e.g., its amplitude envelope or other measures) and, finally, their variability with respect to other, non-rhythm-related speech parameters such as speaker or elicitation method [6, 9]. However, several promising studies on the description of speech rhythm have taken a different direction, attempting to extract rhythmic quantities directly from the acoustic signal, specifically by extracting salient periodicities and their characteristics from its amplitude envelope [14, 15]. Furthermore, studies from the field of language discrimination [16, 17] have used measures derived from fundamental frequency and amplitude to discriminate between pairs of languages with relative success. Other studies from the field of rhythm- or prosody-based automatic language identification (LID) [18, 19, 20, 21, 22] have conducted rhythm modeling by using schemes such as automatic segmentation of the speech signal in pseudosyllables and extracted statistical features describing energy and fundamental frequency which produced good results (in the area of 60 – 80%) in LID tasks, showing that those features can indeed be useful for describing speech rhythm. The crux of those approaches is that the focus is shifted on quantities in the speech signal rather than on the regularities of more linguistically defined speech elements. This paper follows in that rationale, introducing a novel method for speech rhythm description, inspired from similar rhythm analysis methods from the field of Audio Content Analysis [23], which have been used for tasks such as musical genre classification [24] with success. We assume that the rhythmic content of a sound can be captured through the signal-inherent periodicities and their properties. This definition does not differentiate between musical and speech signals, providing a unified concept for rhythm which has been called for [25, 26]. In the following chapters, the rhythm features are described, after determining the signal properties whose periodicities are relevant for rhythm. The features are evaluated in an automatic LID task for two established multilingual speech corpora in order to draw conclusions about their suitability for rhythm-based LID and on rhythm language typology. Results are encouraging regarding the feature capacity, but with certain caveats which are discussed. Moreover, findings concerning language rhythm classes are ambivalent. Finally, advantages and disadvantages of the proposed method and the most informative features are discussed and perspectives for further research are given.

2. Method

Various approaches for rhythm description and quantification have been developed in the field of Music Information Retrieval (MIR) [27]. In the context of musical genre classification, the focus lay on the extraction of signal periodicities from a musical excerpt. Beginning with the work of Scheirer [28], a representation for periodicities of the signal amplitude envelope in the lower frequency area was introduced for beat tracking. Tzanetakis and Cook [29] modified and used this representation, called the *beat histogram*, for extracting rhythmic content features. Similar approaches followed also by Burred and Lerch [30] and Gouyon et al. [31]. The fundamental assumption is that those features are representative of the regularities in the temporal structure of an acoustic signal, describing multiple aspects of the signal’s inherent periodicities. For both music and speech, the beat histogram captures periodicities related to strong, recurring ‘beats’, in effect salient onsets of the signal’s constituent elements.

The beat histogram calculation can take place on basis of the trajectory of various relevant signal quantities over time [32]. As such, the representation will then express periodicities related to this quantity, which might have different statistical and other properties than those which are amplitude or energy related. Those temporal trajectories are called *novelty functions* [33]. A careful consultation of the most important works in phonetics, MIR and rhythm-based LID (mentioned in Section 1), as well as a study of the important rhythm definition approaches in music theory and cognition [34, 25, 35] reveals that there are three essential quantities whose temporal evolution must be taken into account for the extraction of speech rhythm: The **amplitude** of the signal envelope is an acoustic correlate of perceived loudness. This makes it the basis for the detection of rhythm which results from the changing energy of the signal due to the application of *stresses* on specific parts of speech in comparison to others. As such, it denotes *intonation*. The **pitch** or value of a salient (for speech, the fundamental *F0*) frequency in the signal and the temporal trajectory thereof is the most important tonal rhythm carrier in the signal and expresses *speech prosody*. Features derived from its beat histogram can describe changes in voice melody trajectories, regularities in rising or falling voice pitch or related changes. **Spectral** changes are an acoustic correlate of change in sound texture and timbre, which essentially characterize different categories of sounds (such as tonal or noisy) or changes in spectral content (e.g. high or low-frequency content). Features from a beat histogram based on spectrum novelty can serve as descriptors for change of speech elements, such as consonants and vowels, or even different formants. In our study, amplitude novelty is extracted through the calculation of the **RMS amplitude** of the signal. The **fundamental frequency (F0)** is extracted through the use of a spectral harmonic product algorithm on a filtered version of the speech signal (using a 4th-order Butterworth lowpass with a 800 Hz cutoff-frequency), so as to ensure tracking of the fundamental frequency alone. Three standard features are extracted to track spectral changes [23]: the **spectral flux (SF)** (indicating general spectral change), the **spectral flatness (SFL)** (as a measure of signal tonalness or noisiness) and the **spectral centroid (SCD)** (a measure of the spectral centre-of-weight), the latter also on a filtered version of the signal (using a 4th-order Butterworth bandpass filter between 300 Hz and 3200 Hz) to ensure that only formant area frequencies are considered. More information on those features can be found in [23]. Experiments in musical genre classifica-

tion using features based on similar amplitude, tonal and spectral shape novelty functions have shown promising results for a wide range of datasets [32], suggesting their suitability for LID, a task analogue to genre classification [26].

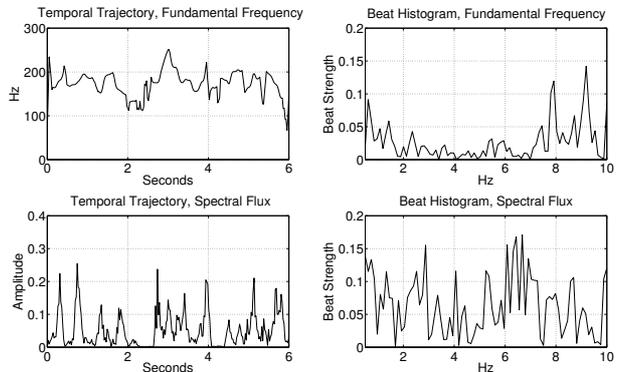


Figure 1: Novelty functions and corresponding beat histograms.

Distribution	Peak
Mean (ME)	Saliency of Strongest Peak (A1)
Standard Deviation (SD)	Saliency of 2nd Stronger Peak (A2)
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 1: Subfeatures extracted from beat histograms.

Fig. 1 gives an overview of two novelty functions (F0 and Spectral Flux) and their corresponding beat histograms for the utterance *Gestern war ich in einem Selbsterfahrungskurs. Ich bin mir nicht wirklich sicher, ob es mir gefallen hat* (from the german subset of the MULTEXT corpus, signal 11). It is clear that the two novelty functions show different periodicities and therefore carry valuable information on multiple speech rhythm levels. More specifically, the fundamental frequency follows the prosody of the given utterance, whereas the spectral flux measure is expected to track general changes of spectrum in the signal, i.e. phoneme or stress changes.

The beat histogram computation follows the computation in [29, 32]. All spectral-based features are calculated through a Short-Time-Fourier-Transform (STFT), whereas the fundamental frequency and the RMS measure on basis of the time-domain signal, both of which with the same temporal resolution parameters from the time domain signal. The complete procedure for the generation of a feature vector representing each utterance includes the following steps: the audio signal is down-mixed to mono, resampled to 22.5 kHz, DC-freed and normalized. Afterwards, the signal is separated in texture windows with a length of 3 s and 50% overlap, on which the beat histograms are extracted. The STFT is performed with a frame-length of 46.4 ms, a Hann window and an overlap of 75%, whereas for the time-domain features the same parameters are used. The novelty function is computed through the calculation of the temporal trajectory of the features and half-wave rectification. The beat histogram is extracted through an Autocorrelation Function (ACF) for each texture window, retaining the

area between 0.5 Hz and 10 Hz, as representative for the relevant periodicities in speech [25]. Finally, the beat histograms extracted from all 3 s frames for an utterance are averaged. For each beat histogram, two categories of features can be extracted (Table 2). Similar features on beat histograms have been used in [29, 30, 31], providing valuable statistical information on the temporal features of each language, similar to the work on LID in [15]. In total, 5 novelty functions are used for the production of as many beat histograms, from each of which 19 subfeatures are extracted, producing in total 95 features.

3. Experimental Setup and Evaluation

For evaluation, extraction of a series of non-rhythmic features was undertaken, by calculating their values over all texture windows (keeping the average value inside an analysis window) on a speech file. Those non-rhythmic features serve as a baseline for the comparison, since acoustic feature-based LID-approaches are among those providing very high performance [36, 37]. Acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs) and Shifted Delta Cepstral (SDCs) features have been used widely for non-rhythmic LID with good results [38, 39]. However, for the sake of comparability with the novel rhythm features, we used a baseline set which comprises all five novelty functions which were also utilized for the calculation of the beat histograms. From their temporal trajectories, the distribution features listed in Table 2 were extracted. In total, the baseline feature set comprises 5 novelties times 11 subfeatures = 55 features. For supervised classification, the Support Vector Machines (SVM) [40] algorithm under MATLAB with a Radial Basis Function (RBF) kernel in a multiclass setting was used. A grid search procedure (i.e. a search for the optimal parameter values) was applied to determine the hyperparameters for this kernel (C , γ). All experiments took place with a 10-fold cross-validation, with results averaged over the folds. Z-score standardization was conducted prior to classification, separately for the train and test set. Classification performance was evaluated through *accuracy*, defined as the proportion of correctly classified samples to all samples.

As speech material, two established multilingual speech corpora for automatic LID were used: the MULTEXT PD [41] and the OGI-MLTS corpus [42]. The first is a corpus of read, high quality speech which contains five indoeuropean languages which are assumed belong to the two basic rhythm groups (english and german to the stress-timed, french, italian and spanish to the syllable-timed), making it useful to test the rhythm class hypothesis when using the proposed novel features. The corpus 10 speakers per language (5 male and 5 female with an average of 15 passages per speaker) and an average length of 20 s for each utterance. The OGI-MLTS corpus contains spontaneous, telephone quality speech from eleven languages (featuring apart from indoeuropean also tonal languages such as mandarin chinese, or even others, such as hindi or vietnamese), multiple speakers per language (male and female) and an average length of 45 s for each utterance. For the experiments in this paper, we retained only the four languages which are common with the MULTEXT PD corpus and which can be used for rhythm typology research. The two selected datasets represent two cases of speech material with very different properties.

In order to identify the best performing descriptors and novelty functions we conducted feature selection following two approaches: First, we apply a filter method (Mutual Information with Target Data [43], using the maximum relevance CMIM metric [44] from the MI-Toolbox [45]). From the feature rank-

ing, we retain the N best features which gave comparable accuracy to the full rhythmic feature set. Second, we evaluate each of the five novelty functions separately, by retaining only the 19 subfeatures resulting from the corresponding beat histogram.

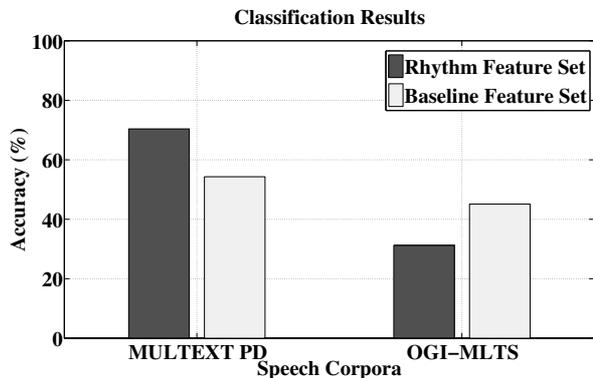


Figure 2: Classification results, both datasets and feature sets

True/Predicted	EN	FR	GE	IT	SP
EN	110	22	5	5	8
FR	4	76	5	9	6
GE	15	22	121	23	19
IT	3	21	5	114	7
SP	7	25	5	6	107
Acc. (%)	73.3	76	60.5	76	71.3
Prior (%)	20	13.3	26.7	20	20

Table 2: Confusion matrix for MULTEXT PD corpus languages, rhythmic features, average accuracy: 70.4%.

True/Predicted	EN	FR	GE	SP
EN	58	56	33	39
FR	38	76	27	45
GE	44	55	39	48
SP	44	56	27	49
Acc. (%)	31.2	40.9	21.0	26.3
Prior (%)	25	25	25	25

Table 3: Confusion matrix for OGI-MLTS corpus languages, rhythmic features, average accuracy: 31.2%.

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

Table 4: Best features after filter feature selection. Abbreviation left of point denotes subfeature, otherwise novelty function.

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: Group feature selection, results given in percentages.

4. Results

Results of the classification procedure are presented in Fig. 2. For the MULTEXT PD corpus, the rhythmic feature set per-

forms better than the baseline set. The performance of the baseline set (54.3%) lies close to that of the rhythmic feature set (70.4%). For the OGI-MLTS dataset, results show the exact opposite tendency: The baseline set with an accuracy of (37.5%) outperforms the rhythmic set (31.5%). With regards to the performance of the corpora, a great difference in accuracies can be observed: Whereas the MULTEXT PD corpus shows a satisfactory performance which lies well above the average prior (20%), the OGI-MLTS corpus accuracy stays at relatively low levels (which are, however, comparable to those in other rhythm modeling LID studies [22]). In the case of the MULTEXT PD corpus, high accuracy can be achieved for all languages. In the case of the OGI-MLTS corpus, only french shows better performance, whereas the accuracy for other languages is only moderately above the prior. The confusion matrices for both cases are given in Tables 2 and 3. It can be seen that in the case of the MULTEXT PD corpus, the rhythm class hypothesis is confirmed only partly: The hypothesized stress-timed languages english and german are not confused with each other more than with others outside this group. In the syllable-timed group, italian and spanish are confused with french, but not with each other. However, the tendency towards misclassifications toward french can be observed for all languages. For the OGI-MLTS corpus, specific misclassifications between languages in the hypothesized same rhythm class, such as english-german or french-spanish, cannot be observed in this case as well. Finally, concerning feature selection for the rhythmic feature set, results show that the same accuracy can be achieved with the first 19 (MULTEXT PD) or 21 (OGI-MLTS) features of the CMIM ranking. In Table 4, a list of the best features for both datasets is given. It is noted that between the novelty functions, such based on spectral flux, spectral flatness and pitch are most commonly among the best ones. Finally, selection based on novelty feature groups (Table 5) shows that all novelty functions are almost equally important for accuracy, a result which is true for both corpora. In both cases, the F0 feature subgroup seems to perform marginally better than the others.

5. Discussion

The results presented in Section 4 suggest that the application of the beat histogram features for automatic LID is indeed valuable, since it provides comparable performance to that of other rhythm-based LID approaches [18, 19, 21, 22], although latest i-vector-based methods provide even higher results [46, 47]. The differences observed between the rhythmic and baseline feature sets are telling with respect to the robustness and quality of the proposed features. In the case of the MULTEXT PD corpus, which is a prosodic database, rhythmic features seem better suited to capture differences between languages than more general acoustic features. On the other hand, for the more generic OGI-MLTS corpus, non-rhythmic features perform better, showing that in that case rhythm features are informative enough. Other reasons which could explain the difference in performance between the two datasets are the signal quality, which in case of the OGI-MLTS corpus might impair the extraction of rhythm features or features in general significantly; and the difference in speech elicitation method, showing that spontaneous speech not only makes the extraction of robust features much more difficult, but also does not allow rhythmic features to achieve acceptable performance. Those observations are useful in determining the scope of use of the suggested rhythmic features, suggesting that they could be more suitable for read speech with good signal quality, but their robustness could be

further improved. With regards to the best features, the fact that novelty functions of pitch and spectral change features produce the most salient beat histograms is a hint for their eligibility for speech rhythm analysis. It is interesting that features such as P1 and P2 (showing periodicities of prominent beats in speech) are not among the best ones. This hints towards the fact that either speech periodicities cannot help differentiate between languages (as they could be noisy because of variability due to other factors) or that they cannot be reliably extracted from the beat histograms through the subfeatures presented here. Concerning language typology on basis of the beat histogram features, the rhythm class hypothesis does not seem to be corroborated in its pure form from our results on the MULTEXT PD corpus: on the one hand it is clear that languages supposed to be rhythmically close to each other, such as english and german are not confused with each other more than with languages from different supposed rhythm classes. On the other hand, spanish and italian are more confused with french than with english or german (which would hint towards a rhythmic similarity in this group), however this can be an artifact of the specific dataset, since french seems to act as an attractor for all other languages, hinting that its rhythmic features are somehow representative of other languages as well. In the case of the OGI-MLTS corpus, results also do not confirm the rhythm class hypothesis directly. Those results can indicate that the novel features in their present form are better suited for specific languages. However, they might also be the consequence of our features not capturing speech rhythm in the same form as the rhythm class hypothesis first posited. More experiments are needed in order to determine of those results are dependent on dataset or the feature extraction and classification methods.

6. Conclusions

The presented beat histogram features are shown to be good descriptors of speech rhythm since they have been shown to provide good accuracy in an automatic LID task. Furthermore, the features achieved accuracies comparable to those of other speech rhythm feature approaches [21, 22] for the same datasets, further attesting to the merit of the method. Amongst the advantages of the presented rhythm description scheme is that it does not require any preprocessing such as syllable annotation or even automatic segmentation which is time-consuming or could potentially insert erroneous assumptions. Furthermore, the method allows the automatic processing of greater datasets and provides a novel perspective on the description of speech rhythm through solely signal-based measures. However, more experiments with greater corpora (such as GLOB-ALPHONE [48]), extraction parameters (to test, e.g., for effects concerning the texture window size) and other classification methods (such as artificial neural nets, as well as unsupervised methods) will be conducted, so as to be able to check for result consistency and improve robustness. Furthermore, the relationship between the features and more abstract speech elements is not entirely clear, prompting future research to establish concrete connections. Further future work on feature selection will attempt to find out which novelty functions and features are the most informative across many datasets and experimental setups, in order to compare the results with those from phonetics or human speech rhythm perception research. Concerning language typology, the presented rhythm-based LID does not seem to corroborate the rhythm class hypothesis in its pure form, but gives incentives to attempt and reformulate the hypothesis in a new version so as to account for the empirical evidence.

7. References

- [1] K. L. Pike, *The Intonation of American English*. Ann Arbor: University of Michigan Press, 1945.
- [2] D. Abercrombie, *Elements of general phonetics*. Edinburgh University Press Edinburgh, 1967, vol. 203.
- [3] R. M. Dauer, "Stress-timing and syllable-timing reanalyzed," *Journal of phonetics*, 1983.
- [4] F. Ramus, M. Nespore, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.
- [5] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515-546, 2002.
- [6] V. Dellwo, A. Fourcin, and E. Abberton, "Rhythmical classification of languages based on voice parameters," in *ICPhS '07*, 2007, pp. 1129–1132.
- [7] P. Wagner, "The rhythm of language and speech: Constraining factors, models, metrics and applications," *Germany: Habilitationsschrift, University of Bonn*, 2008.
- [8] L. Wiget, L. White, B. Schuppler, I. Grenon, O. Rauch, and S. L. Mattys, "How stable are acoustic metrics of contrastive speech rhythm?" *The Journal of the Acoustical Society of America*, vol. 127, no. 3, pp. 1559–1569, 2010.
- [9] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, vol. 40, no. 3, pp. 351–373, 2012.
- [10] A. Turk and S. Shattuck-Hufnagel, "What is speech rhythm? a commentary on arvaniti and rodriguez, krivokapic, and goswami and leong," *Laboratory Phonology*, vol. 4, no. 1, pp. 93–118, 2013.
- [11] P. Roach, "On the distinction between stress-timed and syllable-timed languages," *Linguistic controversies*, pp. 73–79, 1982.
- [12] W. J. Barry, B. Andreeva, M. Russo, S. Dimitrova, T. Kostadinova et al., "Do rhythm measures tell us anything about language type," in *ICPhS '03*, 2003, pp. 2693–2696.
- [13] A. Arvaniti, "Rhythm, timing and the timing of rhythm," *Phonetica*, vol. 66, no. 1-2, pp. 46–63, 2009.
- [14] S. Tilsen and K. Johnson, "Low-frequency fourier analysis of speech rhythm," *The Journal of the Acoustical Society of America*, vol. 124, no. 2, pp. EL34–EL39, 2008.
- [15] S. Tilsen and A. Arvaniti, "Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages," *The Journal of the Acoustical Society of America*, vol. 134, no. 1, pp. 628–639, 2013.
- [16] F. Cummins, F. Gers, J. Schmidhuber, and C. Elvezia, "Automatic discrimination among languages based on prosody alone," *Speech Communication*, 1999.
- [17] A. Thymé-Gobbel and S. E. Hutchins, "Prosodic features in automatic language identification reflect language typology," in *ICPhS '99*, 1999.
- [18] J. Farinas, F. Pellegrino, J.-L. Rouas, and R. André-Obrecht, "Merging segmental and rhythmic features for automatic language identification," in *ICASSP '02*, vol. 1, 2002, pp. 1–753.
- [19] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Modeling prosody for language identification on read and spontaneous speech," in *ICASSP '03*, vol. 6, 2003, pp. 1–40.
- [20] J.-L. Rouas, J. Farinas, and F. Pellegrino, "Automatic modelling of rhythm and intonation for language identification," in *ICPhS '03*, 2003, pp. 567–570.
- [21] J.-L. Rouas, J. Farinas, F. Pellegrino, and R. André-Obrecht, "Rhythmic unit extraction and modelling for automatic language identification," *Speech Communication*, vol. 47, no. 4, pp. 436–456, 2005.
- [22] J.-L. Rouas, "Automatic prosodic variations modeling for language and dialect discrimination," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 6, pp. 1904–1911, 2007.
- [23] A. Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*. Wiley & Sons, 2012.
- [24] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, March 2006.
- [25] A. D. Patel, *Music, language, and the brain*. Oxford university press, 2008.
- [26] S. Hübler and R. Hoffmann, "Comparing the rhythmical characteristics of speech and music—theoretical and practical issues," in *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues*. Springer, 2011, pp. 376–386.
- [27] F. Gouyon and S. Dixon, "A review of automatic rhythm description systems," *Computer music journal*, vol. 29, no. 1, pp. 34–35, 2005.
- [28] E. D. Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.
- [29] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [30] J. J. Burred and A. Lerch, "A hierarchical approach to automatic musical genre classification," in *DAFX '03*, 2003.
- [31] F. Gouyon, S. Dixon, E. Pampalk, and G. Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *AES '04*, 2004, pp. 196–204.
- [32] A. Lykartsis, "Evaluation of accent-based rhythmic descriptors for genre classification of musical signals," Master's thesis, Audio Communication Group, Technische Universität Berlin, 2014.
- [33] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies, and M. B. Sandler, "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.
- [34] F. Lerdahl and R. S. Jackendoff, *A generative theory of tonal music*. MIT press, 1983.
- [35] J. London, *Hearing in time*. Oxford University Press, 2012.
- [36] Y. K. Muthusamy, E. Barnard, and R. A. Cole, "Reviewing automatic language identification," *Signal Processing Magazine, IEEE*, vol. 11, no. 4, pp. 33–41, 1994.
- [37] M. A. Zissman and K. M. Berkling, "Automatic language identification," *Speech Communication*, vol. 35, no. 1, pp. 115–124, 2001.
- [38] E. Singer, P. A. Torres-Carrasquillo, T. P. Gleason, W. M. Campbell, and D. A. Reynolds, "Acoustic, phonetic, and discriminative approaches to automatic language identification," in *INTER-SPEECH*, 2003.
- [39] W. M. Campbell, E. Singer, P. A. Torres-Carrasquillo, and D. A. Reynolds, "Language recognition with support vector machines," in *ODYSSEY04*, 2004.
- [40] V. Vapnik, *The nature of statistical learning theory*. Springer, 2000.
- [41] E. Campione and J. Véronis, "A multilingual prosodic database," in *ICSLP*, vol. 98, 1998, pp. 3163–3166.
- [42] Y. K. Muthusamy, R. A. Cole, B. T. Oshika, L. D. Consortium et al., "The ogi multi-language telephone speech corpus," in *ICSLP*, vol. 92, 1992, pp. 895–898.
- [43] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [44] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.
- [45] G. Brown, A. Pocock, M.-J. Zhao, and M. Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.
- [46] H. Li, B. Ma, and C.-H. Lee, "A vector space modeling approach to spoken language identification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 1, pp. 271–284, 2007.
- [47] M. Li and W. Liu, "Speaker verification and spoken language identification using a generalized i-vector framework with phonetic tokenizations and tandem features," *submitted to INTER-SPEECH*, 2014.
- [48] T. Schultz, "Globalphone: a multilingual speech and text database developed at karlsruhe university," in *INTER-SPEECH*, 2002.

BEAT HISTOGRAM FEATURES FROM NMF-BASED NOVELTY FUNCTIONS FOR MUSIC CLASSIFICATION

Athanasios Lykartsis

Technische Universität Berlin
Audio Communication Group
alykartsis@mail.tu-berlin.de

Chih-Wei Wu

Georgia Institute of Technology
Center for Music Technology
cwu307@gatech.edu

Alexander Lerch

Georgia Institute of Technology
Center for Music Technology
alexander.lerch@gatech.edu

ABSTRACT

In this paper we present novel rhythm features derived from drum tracks extracted from polyphonic music and evaluate them in a genre classification task. Musical excerpts are analyzed using an optimized, partially fixed Non-Negative Matrix Factorization (NMF) method and beat histogram features are calculated on basis of the resulting activation functions for each one out of three drum tracks extracted (Hi-Hat, Snare Drum and Bass Drum). The features are evaluated on two widely used genre datasets (GTZAN and Ballroom) using standard classification methods, concerning the achieved overall classification accuracy. Furthermore, their suitability in distinguishing between rhythmically similar genres and the performance of the features resulting from individual activation functions is discussed. Results show that the presented NMF-based beat histogram features can provide comparable performance to other classification systems, while considering strictly drum patterns.

1. INTRODUCTION

The description of musical rhythm remains an important and challenging topic in Music Information Retrieval (MIR) with applications in several areas [12, 16]. The difficulty of rhythm extraction lies in its multifaceted character, which involves periodicity and structural patterning in the signal as well as perceptual components such as musical meter [19]. An approach which has achieved some popularity over the last years is based on the creation of a periodicity representation — commonly called the beat histogram (BH) — and the subsequent extraction of features from this histogram to be used, e.g., in genre classification [4, 13, 33]. A common first processing step of all approaches is the extraction of a so-called novelty function [2] or its derivatives as the starting point for further analysis. Since a complete rhythm representation of a musical track results from the superposition of the temporal progressions of different instruments or voices [12, 16], it makes sense to include features taking into account individual temporal and spectral properties.

In western popular music (which is the focus of this paper), rhythm is most often carried from the drum section, providing the temporal grid on which other instruments can unfold their melodic or harmonic patterns. This makes the analysis of the drum track appealing for the description of rhythmic character. In order to obtain the rhythmic properties of the drum section, the extraction of temporal novelty functions per instrument is necessary. Although such methods for the extraction of specific voices or instruments have been commonly used in the area of source separation or automatic instrument transcription (the most notable being non-negative matrix factorization (NMF) [31]), their application to rhythm extraction problems is, to the best of our knowledge, sparse. We therefore propose to use a technique for source separation and drum transcription based on partially fixed NMF using the resulting activation functions as a source material for the extraction of rhythmic features based on beat histograms. This paper investigates the suitability of the proposed features in the context of rhythm-based genre classification for dance music and other styles.

The paper is structured as follows. In the second section, an overview of previous work and the goals of the current paper are presented. In section 3, the drum transcription procedure and the feature extraction are described. In the fourth section, the evaluation of the proposed features and the results are given. After discussing the results in section 5, we close by giving conclusions and suggestions for future work (sect. 6).

2. PREVIOUS WORK AND GOALS

Beat histograms have been used for a long time as rhythmic descriptions. Initially introduced in studies on beat tracking and analysis [11, 29] as a useful very low frequency periodicity representation, they were only later referred to as the *beat histogram* [33] or *periodicity histogram* [13]. The histogram is useful as an intermediate representation that can be used to extract musical parameters such as tempo as well as low-level features (e.g., statistical properties of the histogram). Traditionally, a measure of the signal amplitude envelope or its change over time is utilized as the novelty function for the extraction of a beat histogram [4, 13, 33]. However, in the field of onset detection, the proposed novelty functions take into account spectral content changes [3, 10, 15, 27]. Genre classification systems based on such representations have generally shown



© Athanasios Lykartsis, Chih-Wei Wu, Alexander Lerch.
Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** Athanasios Lykartsis, Chih-Wei Wu, Alexander Lerch. "Beat histogram features from NMF-based novelty functions for music classification", 16th International Society for Music Information Retrieval Conference, 2015.

promising results, although rhythm features do usually not perform as well as features from other domains such as timbre descriptors [4,28,33]. However, studies have shown that for highly rhythmical music, beat histogram features can achieve very high performance [13], a fact which has been confirmed in current studies investigating the role of using multiple novelty functions as a basis for beat histogram features [20].

Since drum tracks convey essential information about tempo, rhythm and possibly genre, they could potentially provide better representation for extracting rhythm features. To extract drum tracks from complete mixtures of music, a drum transcription system for polyphonic music would be necessary. Gillet and Richard divide systems for the drum transcription from mixtures into three categories [9]: (i) segment and classify, (ii) separate and detect, and (iii) match and adapt. Here, we focus on the second type of approaches (separate and detect). Based on the assumption that the music signal is a superposition of different sound sources, the music content could be transcribed by first decomposing the signal into source templates with corresponding activation functions, and then detecting the activities of each template. Different methods such as Independent Subspace Analysis [7], Prior Subspace Analysis [6], and Non-negative Matrix Factorization [1, 21] fall into this category. These approaches are usually easy to interpret since most of the decompositions result in spectrum-like representations. Furthermore, these approaches do not require additional classes for simultaneous events, which could potentially reduce the model complexity.

In the context of NMF for music transcription, the following issues have to be taken into consideration: First, the number of sound sources and notes within a music recording is usually unknown. It is therefore difficult to determine a suitable rank r in order to obtain a clear differentiation of the decomposed components in the dictionary matrix. Second, after the unsupervised NMF decomposition process, it is difficult to identify the associated instrument of each component in the dictionary matrix W when rank is too high or too low. Third, when multiple similar entries exist in the dictionary matrix, the corresponding activation matrix could be activated at these entries simultaneously, which in turn increases the difficulty of intuitively interpreting the results.

To address the above issues, Yoo et al. proposed a co-factorization algorithm [35] to simultaneously factorize a prior drum track and a target signal, and use the basis matrix from the drum track to identify the drum components in the target signal. This method ensures that the drum components in both dictionary matrices remain percussion only over the iterations, and thus proper isolation of the harmonic components from the drum components. Since they focus on drum separation rather than drum transcription, their selection of ranks can be higher, but the approach is not directly applicable to the transcription problem because of the probable lack of interpretability of the dictionary matrix. Wu and Lerch proposed a variant of the co-factorization algorithm using partially fixed NMF (PFNMF)

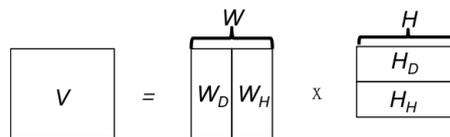


Figure 1. Illustration of the factorization process. W: dictionary matrix. H: activation matrix. Subscript D: drum components, Subscript H: harmonic components.

for drum transcription in polyphonic signals [34]. Instead of co-factorization, this method uses a pre-determined drum dictionary matrix during the decomposition process, and extracts one activation function for each of the three drums (Hi-Hat, Snare Drum, and Bass Drum).

In this paper, we apply PFNMF to transcribe drum events in polyphonic signals, and use the activation functions as the basis for the extraction of beat histogram features. The idea of using NMF with prior knowledge of targeting source within the mixture has been applied in source separation tasks [32], multi-pitch analysis [26] and drum transcription [34]. Furthermore, the use of multiple novelty functions for the extraction of beat histograms has been proposed in [20]. Here, we combine both approaches for the generation of rhythmic features which are descriptive of the percussive rhythmic content of polyphonic tracks and therefore of their general rhythmic character. We focus on two tasks: the investigation of their overall performance, in order to determine the salience of the features for genre classification; and their performance for each percussive component (drum track) separately, attempting to extract conclusions regarding the importance of drum based rhythm features and the salience of NMF activation functions.

3. METHOD

The basic concept of NMF is to approximate a matrix V with matrices W and H as $V \approx WH$ with non-negativity constraints. Given a $m \times n$ matrix V , NMF will decompose the matrix into the product of a $m \times r$ dictionary (or basis) matrix W and an $r \times n$ activation matrix H , with r being the rank of the NMF decomposition. In most audio applications, V is the spectrogram to be decomposed, W contains the magnitude spectra of the salient components, and H indicates the activation of these components with respect to time [31]. The matrices W and H are estimated through an iterative process that minimizes a distance measure between the target spectrogram V and its approximation [30].

To effectively extract drum activation functions from the polyphonic signals, PFNMF is used in this study. Figure 1 visualizes the basic concept from the work of Yoo et al.: the matrices W and H are split into the matrices W_D and W_H , and H_D and H_H , respectively. Instead of using co-factorization, PFNMF initializes the matrix W_D with drum components and to not modify it during the factorization process. Matrices W_H , H_H , and H_D are initialized with random numbers. The distance measure used in this paper is the generalized KL-divergence (or I-divergence), in which



Figure 2. Flowchart of NMF and beat histogram feature extraction and classification system.

$D_{\text{KL}}(x | y) = x \cdot \log(x/y) + (y - x)$. The cost function as shown in (1) is minimized by applying gradient descent and multiplicative update rules, the matrices W_{H} , H_{H} , and H_{D} will be updated according to Eqs. (2)–(4).

$$J = D_{\text{KL}}(V | W_{\text{D}}H_{\text{D}} + W_{\text{H}}H_{\text{H}}) \quad (1)$$

$$H_{\text{D}} \leftarrow H_{\text{D}} \frac{W_{\text{D}}^T (V / (W_{\text{D}}H_{\text{D}} + W_{\text{H}}H_{\text{H}}))}{W_{\text{D}}^T} \quad (2)$$

$$W_{\text{H}} \leftarrow W_{\text{H}} \frac{(V / (W_{\text{D}}H_{\text{D}} + W_{\text{H}}H_{\text{H}})) H_{\text{H}}^T}{H_{\text{H}}^T} \quad (3)$$

$$H_{\text{H}} \leftarrow H_{\text{H}} \frac{W_{\text{H}}^T (V / (W_{\text{D}}H_{\text{D}} + W_{\text{H}}H_{\text{H}}))}{W_{\text{H}}^T} \quad (4)$$

PFNMF can be summarized in following steps:

1. Construct an $m \times r_{\text{D}}$ dictionary matrix W_{D} , with r_{D} being the number of drum components to be detected.
2. Given a pre-defined rank r_{H} , initialize an $m \times r_{\text{H}}$ matrix W_{H} , an $r_{\text{D}} \times n$ matrix H_{D} and an $r_{\text{H}} \times n$ matrix H_{H} .
3. Normalize W_{D} and W_{H} .
4. Update H_{D} , W_{H} , and H_{H} using (2)–(4).
5. Calculate the cost of the current iteration using (1).
6. Repeat step 3 to step 5 until convergence.

In our current setup, the STFT of the signals is calculated using a window size and a hop size of 2048 and 512, respectively. A pre-trained dictionary matrix is constructed from the training set, consisting of isolated drum sounds. The templates are extracted for the three classes Hi-Hat (HH), Bass Drum (BD) and Snare Drum (SD) as the median spectra of all individual events of one drum class in the training set. Next, the PFNMF will be performed with rank $r_{\text{H}} = 10$ on the test files. More details of the training process and the selection of rank r_{H} can be found in [34]. Finally, the activation Matrix H_{D} can be extracted from the audio signals through the decomposition process.

Once the activation functions of the three drum tracks have been extracted as described above, they are used as novelty functions for the calculations of beat histograms, similar to [20]. The complete procedure for the generation of a feature vector representing each track includes the following steps: For each activation function, the beat histogram is extracted through the calculation of an Auto-correlation Function (ACF) and the retaining of the area between 30 and 240 BPM. For each beat histogram, the sub-features listed in Table 1 are extracted. The concatenation

Distribution	Peak
Mean (ME)	Saliency of Strongest Peak (A1)
Standard Deviation (SD)	Saliency 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 1. Subfeatures extracted from beat histograms.

of all subfeature groups for each novelty function produces the final feature vector for an audio excerpt. Similar sub-features as listed in Table 1 can be found in the literature, e.g., in [33] (Peak), and [4, 13] (Distribution). In total, 3 novelty functions are used for the production of as many beat histograms, from each of which 19 subfeatures are extracted, resulting in a total count of 57 features.

4. EVALUATION

4.1 Dataset Description

In order to evaluate the features for multiple track kinds possessing different rhythmic qualities, two datasets were considered: the Tzanetakis Dataset (**GTZAN**) [33], as an example of a dataset which is widely used, comprising 100 30 sec excerpts for each of 10 diverse musical genres; and the Ballroom Dataset [5, 13] (**Ballroom**), comprising 698 very rhythm/dance-oriented tracks of length 10 sec and therefore suitable for the evaluation of our NMF-based beat histogram features. Both datasets contain tracks with a drum section and others with only non-percussive instruments. This does not only allow to investigate if the extracted features are also suitable for music where a drum section is present and if they can generalize to other music styles, but also allows conclusions as to what genres in particular are represented satisfactorily or insufficiently by the features.

4.2 Evaluation Procedure

The features were tested using the Support Vector Machine (SVM) algorithm for supervised classification. For our multiclass setting, an RBF kernel was used and the optimal parameters (C, γ) were determined through grid search. We chose the SVM classifier since it has been frequently used in similar genre classification experiments, shows generally good results (see [8]) and allows for comparability with those studies. Since the focus here lay on the features and not the classification algorithms, we refrained from using more state-of-the-art approaches such as deep learning algorithms. All experiments took place with a 10-fold cross-validation (using 90% of the data for training and 10% for testing over 10 randomly selected folds, taking the average accuracy over the folds for each dataset) and standardization (z-score) of the training and testing data. After the full

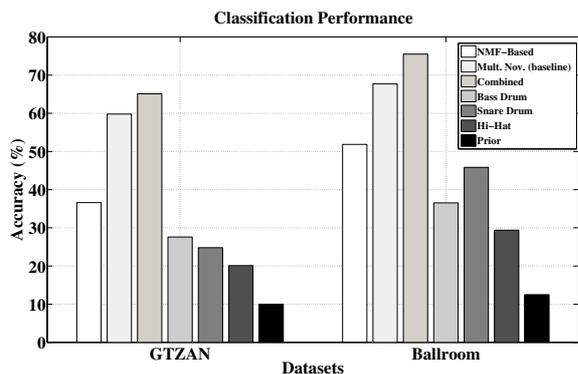


Figure 3. Classification results for both datasets.

NMF-based feature set (i.e., the features originating from all three drum activation functions) was tested, the features from each individual activation function were evaluated in turn in order to study the importance of each drum track separately. Finally, the NMF-based features are combined with other beat histogram features from a current study [20], extracted from novelty functions of amplitude (RMS), spectral shape (spectral flux, centroid, flatness and the first 13 MFCCs) and tonal components (pitch chroma coefficients and tonal power ratio) on 3 second-long frames. Those features resulted from a similar procedure as the one used here, where 30 different novelty functions were extracted and their beat histograms computed through the calculation of an ACF. A subsequent two-stage feature selection scheme (mutual information with target data [14] using the CMIM metric [25], followed by a sequential forward selection with an SVM wrapper [17]) was applied to retain the best-performing features, resulting in a total of 20 features in each case.

4.3 Results

The results are shown in Figure 3. On both datasets, the full NMF feature set (comprising features from all three drum activation functions) performs better than the individual ones (BD, SD, HH), with an attained accuracy of 36.6% and 51.9% for GTZAN and Ballroom, respectively. Those values lie considerably above the average priors of both datasets. The differences between the accuracies of the feature sets are not large (especially between the individual drum based feature sets) but are significant at the 0.05 level in all cases (based on a comparison test of the Cohen’s Kappa extracted from the confusion matrices). Due to their small values (ranging from 0.2% to 0.6%), standard deviations between accuracies of the folds for each feature set are not presented in Figure 3.

The multiple novelty feature set (from [20]) outperforms the NMF-based features, reaching an accuracy of 59.8% for the GTZAN and 67.7% for the Ballroom dataset, whereas the combined set (NMF and multiple novelty) demonstrates the best performance (accuracy of 65.1% (GTZAN) and 75.5% (Ballroom)). The individual feature sets from each drum track provide performance inferior to that of the

	Ch.	Ji.	Qu.	Ru.	Sa.	Ta.	Vw.	Wa.
Ch.	54	10	10	11	14	3	1	8
Ji.	17	13	5	6	10	2	2	5
Qu.	10	2	44	5	3	8	7	3
Ru.	8	3	2	53	4	7	2	19
Sa.	15	4	8	2	50	3	1	3
Ta.	2	1	6	6	2	55	7	7
Vw.	5	3	9	6	0	6	17	19
Wa.	5	0	4	16	1	4	4	76
Acc.	49	22	54	54	58	64	26	69
Pr.	15.9	8.6	11.7	14.0	12.3	12.3	9.3	15.8

Table 2. Confusion matrix for Ballroom dataset, average accuracy: 51.9%. Accuracy and Prior are given in %.

	Bl.	Cl.	Co.	Di.	Hi.	Ja.	Me.	Po.	Re.	Ro.
Bl.	15	11	16	15	4	7	9	9	11	3
Cl.	4	63	3	1	1	14	5	3	1	5
Co.	6	6	38	12	4	5	6	6	11	6
Di.	13	1	6	43	6	1	8	5	12	5
Hi.	8	4	5	4	21	8	10	20	13	7
Ja.	8	17	5	0	7	38	9	7	7	2
Me.	3	11	7	7	2	6	51	2	2	9
Po.	7	6	6	5	14	5	5	33	12	7
Re.	6	3	6	6	6	4	1	11	53	4
Ro.	6	4	10	10	17	10	11	11	10	11
Acc.	15	63	38	43	21	38	51	33	53	11
Pr.	10	10	10	10	10	10	10	10	10	10

Table 3. Confusion matrix for GTZAN dataset, average accuracy: 36.6%. Accuracy and Prior are given in %.

full NMF-based set, but still considerably higher than the prior. The best individual drums are the BD and SD for the GTZAN and Ballroom datasets, respectively. The worst individual percussion instrument is in both cases the HH. For the full NMF-based feature set, confusion matrices resulting from the classification can be seen in Tables 2 and 3. In general, features achieved better average performance on the Ballroom dataset than on the GTZAN. In order to evaluate the misclassifications and the performance of the individual genres, a closer observation of the confusion matrices of each dataset should be taken.

For the **Ballroom** dataset, confusions between genres appear to be plausible based on what one would expect when extracting rhythm features only from drums tracks: genres with strongly pronounced, stable rhythm played from a drum section such as samba and *chachacha* (*Ch.*) are confused with each other, whereas the *waltz* (*Wa.*) and *tango* (*Ta.*) genres, having no drum section (but still a succinct rhythm) are not confused much with other genres. The latter are the two genres which also achieve the best individual performance, followed by *chachacha*, *quickstep* (*Qu.*), *rumba* (*Ru.*) and *samba* (*Sa.*). *Jive* (*Ji.*) and *viennese waltz* (*Vw.*) display the worse performance, and are confused with *chachacha* and *waltz* respectively, a result which is also expected when one considers the rhythmic proximity of those genres, whether they possess a drum section or not.

For the **GTZAN** dataset, misclassifications present a more mixed picture: On the one hand, genres which possess tracks featuring a well articulated, distinct rhythmic performed by a drum section (such as *reggae* (*Re.*), *metal*

(*Me.*) and *disco* (*Di.*) as well as the only genre without drums (*classical* (*Cl.*)) achieve satisfactory performance and are confused with genres which are rhythmically relatively close (*classical* with *jazz* (*Ja.*), *metal* with *rock* (*Ro.*), *disco* with *reggae*, and *reggae* with *pop* (*Po.*)). On the other hand, genres possessing tracks with a more “generic” rhythm (such as *country* (*Co.*) and *pop*) are confused with multiple other genres. Finally, *hiphop* (*Hi.*), *blues* (*Bl.*) and *rock* attain the last places in individual performance and are confused with multiple other genres.

5. DISCUSSION

The results show that beat histogram features based on NMF activation functions of specific drums can be helpful in rhythm-based genre classification, as their accuracy for the used datasets is comparable to that achieved by other rhythmic feature sets used up to date (59.8% [20] and 28% [33] for the GTZAN, 67.7% [20] and 56.7% [13] for the Ballroom dataset). When taking into account that the features are solely based on drum novelty functions, their performance, especially for the Ballroom dataset, can be seen as satisfactory. It is clear, though, that for this reason, our results cannot achieve as high accuracy as other studies which use very sophisticated methods [8, 18, 22–24]. Our results are somewhat lower than the state of the art using rhythm [22, 24] or combined features [8, 23], however staying in the same range. For the sake of comparison, we report here the highest performances reached when using advanced rhythmic features: on the GTZAN dataset an accuracy of 92.4% [22] has been achieved, for the Ballroom dataset one of 96.1% [24]. The advantage of our proposed methods and features lies in the ability to pinpoint the importance of the rhythm patterns from specific drums for specific genres.

The misclassifications (reported in Tables 2 and 3) show that genres which do not feature genre-specific rhythm patterns, even if those are clearly articulated by the drum section (e.g., a 4/4 BD and SD alternating beat), tend to be confused with other similar genres (especially when drum tracks are present, such as in *rock*). Genres containing non-percussive tracks (such as *classical* and *waltz*) or very specific rhythmic patterns (*reggae*) are more easily distinguished from others. Those results indicate that the NMF-based beat histogram features indeed capture rhythmic properties related to the drum section and the regularities of their periodicities, pointing towards the suitability of those features for the extraction of drum-based rhythmic properties and the use in the discrimination of musical tracks which contain drums from ones which do not.

With regards to the feature sets, the satisfactory accuracy of the NMF-based feature set is a hint towards the appropriateness of the features for the analysis of the rhythmic character of a musical track. However, it is clear that those features, being derived only from drum tracks, cannot represent as much information as features resulting from the use of multiple novelty functions covering many aspects of the signal temporal progress. The improved performance of the combined set (NMF and multiple novelty based)

is a consequence of incorporating specific, drum-related rhythm information in the feature base, showing that the NMF-based rhythm feature set can contribute information not provided by more general rhythm features and lead to significant improvement for the two evaluated datasets. The analysis of the features derived from the activation function of a specific drum track showed that mainly the snare drum and to a lesser extent the kick drum are the most important components. The tendency is strong for the Ballroom dataset, where the SD outperforms the BD, whereas for the GTZAN dataset the result is reversed but with a smaller difference. In all cases (also between the individual drum sets), the differences in accuracies between the feature sets are significant at the 5% level. Those results can be due to the very pronounced sound texture and greater power of those drums which leads to a salient activation function, as well as their role in providing the basic metric positions in most of western popular music. However, the accuracy of each subset lies below that of their combination, leading to the conclusion that the activation functions of all three percussion instruments contribute valuable information to the feature description of musical genre.

Concerning the datasets, the poorer classification performance observed for the GTZAN dataset is a sign of the more diverse character of tracks and genres in this set, containing music styles which lack a specific rhythmic character and can therefore not be distinguished effectively through beat histogram features derived from drum activation functions. Results were still better than the ones reported in [33], but their inferiority compared to the ones in other studies [13, 20] shows that when considering a multitude of different genres, solely drum based activation functions can not provide a complete rhythmic characterization. This, however, points towards the possible goal of using NMF in order to transcribe not only drums but also other instruments in order to use their activation functions as a basis for beat histogram features. The Ballroom dataset shows better performance, which was to be expected since the tracks therein are selected for belonging to different dance styles, requiring a special rhythmic pattern which is mostly conveyed by the drum section. The results are in the same range as those provided in [13] (56.7%) when using only periodicity histogram features. Furthermore, in the same study it was shown that using the tempo of the given tracks as a feature they could achieve very high results using a simple 1-NN classifier (51.7% for the “naive tempo” derived from the periodicity histogram and 82.3% for the ground-truth tempo provided with the recordings), reaching as much as 90% when combining the correct tempo with other descriptors (MFCCs) from the periodicity histogram. This shows that beat histograms (from which the tempo can be extracted) are a good tool for rhythmic analysis in datasets containing dance music such as the Ballroom.

Regarding specific genres, it is clear from the results that the NMF-based features have a twofold use: first, in representing genres which are characterized by distinct patterns in their drum sections (e.g., *reggae* or *samba*) and second, in characterizing genres which lack a drum section

at all (*waltz*, *classical*) in contrast to genres which do; the activation functions transcribed in this case are maximally different, leading to beat histogram features which can be easily discriminated by a classifier. Such a finding shows that drum-based rhythm features can be very helpful for rhythmic characterization of specific genres, which could be an argument for their further application when a specific kind of music is involved. As a general remark, it can be seen that genres possessing a stable rhythm articulated by a drum section such as *reggae* and *samba* or genres lacking drums in general (*waltz* and *classical*) perform better, whereas genres which have a very uncharacteristic rhythm (such as *rock* or *blues*) get more easily confused.

6. CONCLUSIONS

The work presented in this paper focuses on the creation of novel, NMF-based beat histogram features for rhythm-based musical genre classification and rhythmic similarity. The difference in comparison to other well-known studies for rhythm features based on beat histograms [4, 13, 24, 33] is the use of the activity functions of specific drums provided through NMF as a basis for the calculation of the beat histogram. We showed that the classification accuracy using these beat histogram features is comparable to that of other rhythm features, whereas our proposed features are better especially for characterizing tracks with specific rhythmic patterns or for distinguishing between songs with and without a drum section. It was observed that the most important percussion patterns for dance music classification were generated by the snare and the kick drum, which underlines the importance of its activation function for further tasks.

One future goal is the expansion of the use of NMF to identify more instruments or voices and use them as possible novelty functions. The goal would be to therefore capture the rhythmic patterns of every instrument, essentially joining source transcription and rhythm feature extraction into one module. Another possibility is the use of our proposed features for larger and more specific datasets, in order to further investigate their suitability for specific genres, as well as the strengths and weaknesses of the patterns extracted from individual drums in discriminating between musical genres. As an expansion of the feature selection procedure, a further idea would be to profit from the combination of NMF-based features and other acoustic features using a classifier that is capable of learning feature importance (e.g. random forest) to quantitatively investigate the importance of NMF-derived features. While NMF-based beat histogram features have been evaluated only in the context of rhythmic genre classification, we believe that they can prove useful in other tasks. Future research will focus on adjusting and using the proposed features for MIR tasks such as rhythmic similarity computation and structural analysis.

7. REFERENCES

- [1] David S Alves, Jouni Paulus, and José Fonseca. Drum transcription from multichannel recordings with non-negative matrix factorization. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Glasgow, Scotland, UK, 2009.
- [2] Juan P. Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*, 13(5):1035–1047, 2005.
- [3] Juan P. Bello, Chris Duxbury, Mike Davies, and Mark Sandler. On the use of phase and energy for musical onset detection in the complex domain. *IEEE Signal Processing Letters*, 11(6):553–556, 2004.
- [4] Juan José Burred and Alexander Lerch. A hierarchical approach to automatic musical genre classification. In *Proceedings of the 6th international conference on digital audio effects*, pages 8–11, 2003.
- [5] Simon Dixon, Elias Pampalk, and Gerhard Widmer. Classification of dance music by periodicity patterns. In *Proceedings of the 4th International Conference on Music Information Retrieval (ISMIR)*, 2003.
- [6] Derry FitzGerald, Bob Lawlor, and Eugene Coyle. Drum transcription in the presence of pitched instruments using prior subspace analysis. In *Proceedings of the Irish Signals and Systems Conference (ISSC)*, 2003.
- [7] Derry FitzGerald, Robert Lawlor, and Eugene Coyle. Sub-band independent subspace analysis for drum transcription. In *Proceedings of the Digital Audio Effects Conference (DAFX)*, pages 65–59, 2002.
- [8] Zhouyu Fu, Guojun Lu, Kai Ming Ting, and Dengsheng Zhang. A survey of audio-based music classification and annotation. *IEEE Transactions on Multimedia*, 13(2):303–319, 2011.
- [9] Olivier Gillet and Gaël Richard. Transcription and separation of drum signals from polyphonic music. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(3):529–540, 2008.
- [10] Masataka Goto and Yoichi Muraoka. Music understanding at the beat level – real-time beat tracking for audio signals. In *Computational auditory scene analysis*, pages 157–176, August 1995.
- [11] Masataka Goto and Yoichi Muraoka. A real-time beat tracking system for audio signals. In *Proceedings of the International Computer Music Conference (ICMC)*, pages 171–174, 1995.
- [12] Fabien Gouyon and Simon Dixon. A review of automatic rhythm description systems. *Computer Music Journal*, 29(1):34–35, 2005.

- [13] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer. Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, pages 196–204, 2004.
- [14] Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [15] Stephen Hainsworth and Malcolm Macleod. Onset detection in musical audio signals. In *Proceedings of the International Computer Music Conference (ICMC)*, 2003.
- [16] Enric Guaus i Termens. New approaches for rhythmic description of audio signals. Technical report, Music Technology Group, Universitat Pompeu Fabra, 2004.
- [17] Ron Kohavi and George H John. Wrappers for feature subset selection. *Artificial intelligence*, 97(1):273–324, 1997.
- [18] Chang-Hsing Lee, Jau-Ling Shih, Kun-Ming Yu, and Hwai-San Lin. Automatic music genre classification based on modulation spectral analysis of spectral and cepstral features. *Multimedia, IEEE Transactions on*, 11(4):670–682, 2009.
- [19] Justin London. *Hearing in time*. Oxford University Press, 2012.
- [20] Athanasios Lykartsis. Evaluation of accent-based rhythmic descriptors for genre classification of musical signals. Master’s thesis, Audio Communication Group, Technische Universität Berlin, (www.ak.tu-berlin.de/menue/abschlussarbeiten/), 2014.
- [21] Arnaud Moreau and Arthur Flexer. Drum transcription in polyphonic music using non-negative matrix factorization. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, pages 353–354, 2007.
- [22] Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R Arce. Music genre classification using locality preserving non-negative tensor factorization and sparse representations. In *Proceedings of the 10th International Conference on Music Information Retrieval (ISMIR)*, 2009.
- [23] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce. Music genre classification via joint sparse low-rank representation of audio features. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 22(12):1905–1917, 2014.
- [24] Geoffroy Peeters. Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal. *IEEE Transactions on Audio, Speech and Language Processing*, 19(5):1242–1252, 2011.
- [25] Hanchuan Peng, Fulmi Long, and Chris Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.
- [26] Stanisław A Raczynski, Nobutaka Ono, and Shigeki Sagayama. Multipitch analysis with harmonic nonnegative matrix approximation. In *Proceedings of the 8th International Conference on Music Information Retrieval (ISMIR)*, 2007.
- [27] Axel Roebel. Onset detection in polyphonic signals by means of transient peak classification. In *Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, 2005.
- [28] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek. Automatic genre classification of music content: a survey. *IEEE Signal Processing Magazine*, 23(2):133–141, 2006.
- [29] Eric D Scheirer. Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America*, 103(1):588–601, 1998.
- [30] D Seung and L Lee. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.
- [31] Paris Smaragdis and Judith C Brown. Non-negative matrix factorization for polyphonic music transcription. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, pages 177–180, 2003.
- [32] Paris Smaragdis, Bhiksha Raj, and Madhusudana Shashanka. Supervised and semi-supervised separation of sounds from single-channel mixtures. In *Independent Component Analysis and Signal Separation*, pages 414–421. Springer, 2007.
- [33] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE transactions on Speech and Audio Processing*, 10(5):293–302, 2002.
- [34] Chih-Wei Wu and Alexander Lerch. Drum transcription using partially fixed non-negative matrix factorization. In *Proceedings of the European Signal Processing Conference (EUSIPCO)*, 2015.
- [35] Jiho Yoo, Minje Kim, Kyeongok Kang, and Seungjin Choi. Nonnegative matrix partial co-factorization for drum source separation. In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 1942–1945, 2010.

BEAT HISTOGRAM FEATURES FOR RHYTHM-BASED MUSICAL GENRE CLASSIFICATION USING MULTIPLE NOVELTY FUNCTIONS

Athanasios Lykartsis

Audio Communication Group
Technische Universität Berlin
Berlin, Germany

athanasios.lykartsis@tu-berlin.de

Alexander Lerch

Center for Music Technology
Georgia Institute of Technology
Atlanta, Georgia, US

alexander.lerch@gatech.edu

ABSTRACT

In this paper we present beat histogram features for multiple level rhythm description and evaluate them in a musical genre classification task. Audio features pertaining to various musical content categories and their related novelty functions are extracted as a basis for the creation of beat histograms. The proposed features capture not only amplitude, but also tonal and general spectral changes in the signal, aiming to represent as much rhythmic information as possible. The most and least informative features are identified through feature selection methods and are then tested using Support Vector Machines on five genre datasets concerning classification accuracy against a baseline feature set. Results show that the presented features provide comparable classification accuracy with respect to other genre classification approaches using periodicity histograms and display a performance close to that of much more elaborate up-to-date approaches for rhythm description. The use of bar boundary annotations for the texture frames has provided an improvement for the dance-oriented Ballroom dataset. The comparably small number of descriptors and the possibility of evaluating the influence of specific signal components to the general rhythmic content encourage the further use of the method in rhythm description tasks.

1. INTRODUCTION

The extraction of features describing musical rhythm is an important and a challenging topic in Music Information Retrieval (MIR) with applications in many areas [1, 2, 3, 4]. Rhythm features can be of importance, for example, in musical genre classification, where they have been shown to improve classification accuracy and, in specific cases (such as for special, dance-music oriented datasets) allow very successful rhythm-based classification or similarity computation [2, 5, 6, 7, 4]. Since the concept of 'genre' can be related to various musical parameters [8, 9, 5], it is generally attempted to include features from all possibly relevant musical categories (timbre, pitch, loudness, and rhythm) in order to achieve good classification results. In this paper, we chose to focus on rhythm features for musical genre classification since computational rhythm description remains a challenging subject, even when considerable advances have been made lately in this field [7, 10, 11, 4]. A common approach to the description of the rhythmic content of a musical track focuses on the extraction of features based on changes in the signal's amplitude envelope, possibly applying frequency-band filtering or attempting to track energy changes from the signal's short time spectrum. The basic assumption behind this approach is that rhythm is strongly related to

long-term amplitude periodicities and their statistical properties [2, 5, 6]. This assumption is, in our view, well-founded, since many music theoretical works on rhythm have stressed the multidimensional nature of rhythm, establishing the need to consider different musical properties and their temporal evolution when attempting to represent the rhythmic content of music [12, 13, 14, 15].

Several approaches for the automatic extraction of features describing rhythmic content for genre classification through a periodicity distribution representation have been proposed [16, 17, 18]. They are based on earlier studies on beat analysis [19], which aimed at creating a very low frequency periodicity representation (later commonly dubbed the *beat histogram* [16] or *periodicity histogram* [18]), used to extract musical parameters such as tempo as well as low-level features (e.g., statistical properties of the histogram such as mean or standard deviation). Traditionally, a measure of the signal amplitude envelope or its change over time is utilized as the novelty function for the extraction of a beat histogram [18, 16, 17]. Genre classification systems based on such representations have generally shown promising results, although the rhythm features did not perform as well as other subsets of the feature set such as timbre features [5, 16, 17]. There still exist only relatively few studies [20, 21, 18, 22, 3] where rhythm features alone were so efficient as to achieve high accuracy in a genre classification or similarity task. The dataset used in most of those cases (Ballroom [23]) includes only tracks from dance genres with presumably clearly distinguishable rhythmic properties, leading to very high performance but providing little information as to the suitability of the features for music with less distinctive rhythm.

Additionally, there exist many approaches for rhythm description which do not only base themselves on direct feature extraction from a periodicity representation, but rely on direct similarity measurements, computing distances between rhythm representations also being based on periodicities present in the signal [24, 25, 26, 7]. Finally, latest works have attempted to model rhythm using probabilistic models [10, 3, 27] or deep neural networks [11, 28]. Such methods have shown excellent results in rhythmic classification and similarity computation tasks (with accuracies ranging up to 95% [3, 27, 11]), attesting to the plausibility of addressing rhythm-based genre classification tasks with rhythmic similarity methods or elaborate rhythm modeling [4]. However, we identify two main drawbacks with this category of approaches: first, the interpretability of the features describing rhythm is limited, since the method of their generation is either too complex or based on purely technical and not music-theoretical considerations. This is not a problem, of course, if the goal is to develop a method providing high accuracy; it does, however, limit the possible conclusions which can be drawn from the classification task in itself and the

features involved in it. Secondly, the complexity of the methods makes them prone to errors and random influences. Furthermore, the very high results achieved by some studies, although proving the suitability of such classifier systems for e.g. commercial use, should in our view be seen in a critical light for two reasons: on the one hand, it is questionable whether systems based only on rhythmic features could theoretically achieve such performance, since listening experiments show that even human subjects cannot distinguish genres perfectly [29]; on the other hand, it has been demonstrated very often in the intrinsically fuzzily defined genre classification task that complex systems relying on elaborate transformations and trained with small amounts of data might provide very high results, but in effect do not generalize to real-world data (a case of *overfitting*) or do not describe the quantities they are supposed to, their performance being an artifact of an erroneous ground truth or highly dataset-specific features [30, 31, 32]. We chose to focus on the periodicity representation methods and the features which can be extracted on their basis, conducting a detailed examination of the features which can be used in their context and their behavior for many different datasets. This approach allows to investigate the merits of those methods in depth and to identify which signal-based features bear the most importance for rhythmic description.

A common element of previous studies using periodicity representations allowing rhythmic content feature extraction for genre classification is that the features are derived from a beat histogram created on the basis of the signal amplitude and energy changes, as for example in [16, 17, 18]. This approach might seem intuitive at first, since it is based on the assumption that rhythmic properties are conveyed through amplitude or energy changes in the musical signal, which is a common consequence of rhythm perception modalities [33, 34]. However, relevant literature in music theory [12, 15, 14] and cognition [35, 36] indicates that rhythm arises as the combination or interplay of periodicities associated with different sound properties such as *amplitude* (e.g., accents), *spectral* (e.g., instrumentation changes) and *tonal* changes (e.g., chord changes). In the field of onset detection, novelty functions have been proposed which take into account spectral content changes [37, 38, 39]. The goal of this work is to capitalize on this observation by extracting novelty functions from different signal properties (which will allow to take not only amplitude-based but also spectral and tonal changes and their related periodicities into account), using the beat histogram method to create a basis for features describing the rhythmic content of the signal. We investigate the impact of using various musical qualities as novelty functions for beat histogram calculation on classification accuracy, identify the most and least descriptive features and feature groups, and compare the results to those of a traditional timbre-related feature set. Furthermore, the effect of bar boundary annotations from manually annotated data is investigated, since beat histograms based on musically more meaningful data are expected to produce more qualitative features. Finally, we examine the misclassifications in the confusion matrices to draw conclusions on the suitability of the features for rhythm-based classification and possible shortcomings.

The paper is structured as follows. In Section 2, the feature extraction procedure is described. In the third Section, the evaluation of the proposed features is given, along with information about feature selection and the datasets. In the fourth Section, results are presented and discussed in Section 5. Finally, we give conclusions and suggestions for future work (Section 6).

2. METHOD

Novelty functions are generally defined as temporal curves designating prominent changes of a signal [40, 41, 42, 43], resulting from a reduced or filtered version of the original signal from which the first difference is computed, in order to accentuate substantial changes in the monitored quantity. In this paper, we expand this definition somewhat and consider every temporal trajectory of a signal feature to effectively be a novelty function in order to use it as a basis for the beat histogram calculation. This assumption is justified from a practical point of view, since prominent changes in the magnitude of a signal feature (such as, e.g., spectral flux) are still represented (but not as accentuated), their inherent periodicities detectable by methods such as a Discrete Fourier Transform (DFT), an Auto-Correlation Function (ACF) or a resonant filter bank. Furthermore, the avoidance of taking the first difference function has the added advantage of reducing noise in the feature temporal trajectory. For rhythm analysis, signal characteristics beyond amplitude are included here since their periodicities also contribute to the overall rhythm. Examples include changes in instrumentation, which cause a universal, broadband change in the spectral content of a signal, or chord changes, which can be tracked by changes in the instantaneous pitch content.

Fig. 1 shows the novelty function (top) and the resulting beat histogram (bottom) for two features representing spectral change (Spectral Flux) and spectral shape (Spectral Flatness), respectively. The audio track is an excerpt from the *disco* class of the GTZAN dataset [16] and the features are extracted over the whole length of a single texture frame (3 s). In the excerpt, the drum section plays a straight 4/4 measure with a beat each 0.48 s, clearly visible in the spectral flux temporal curve (tracking general spectral change) and as the prevalent periodicity (126 BPM) in the corresponding beat histogram. The changes tracked by the spectral flatness measure, however, more strongly reflect tonalness changes, taking place every 0.96 s, with a main periodicity of 63 BPM. Both the general form (distribution) of the histogram and the strength and exact BPM value of the most prevalent periodicities can be seen to differ significantly in the two examples. Based on this example, it is obvious that the changes tracked by different novelty functions can lead to different beat histograms for the same audio excerpt, with each histogram potentially providing a rhythmically meaningful description.

The novelty functions used in this paper cover envelope, spectral shape, and tonal content changes. The selected features are the following (for details on their computation, see [44]):

- **Spectral Shape** features include Spectral Flux (SF), Spectral Centroid (SCD), Mel Frequency Cepstral Coefficients (MFCC 1-13) and Spectral Flatness (SFL).
- **Tonal** novelty features comprise Spectral Tonal Power Ratio (STPR) and the Pitch Chroma Coefficients (SPC 1-12).
- **Envelope** novelty is tracked through the Root Mean Square (RMS) measure of the signal amplitude.

The beat histogram computation is similar to the one proposed by Tzanetakis [16] without implementation of the wavelet filtering. All features are calculated through a Short-Time-Fourier-Transform (STFT), except for the RMS which is extracted with the same temporal resolution parameters from the time domain signal. The complete procedure for the generation of a feature vector representing each track includes the following steps:

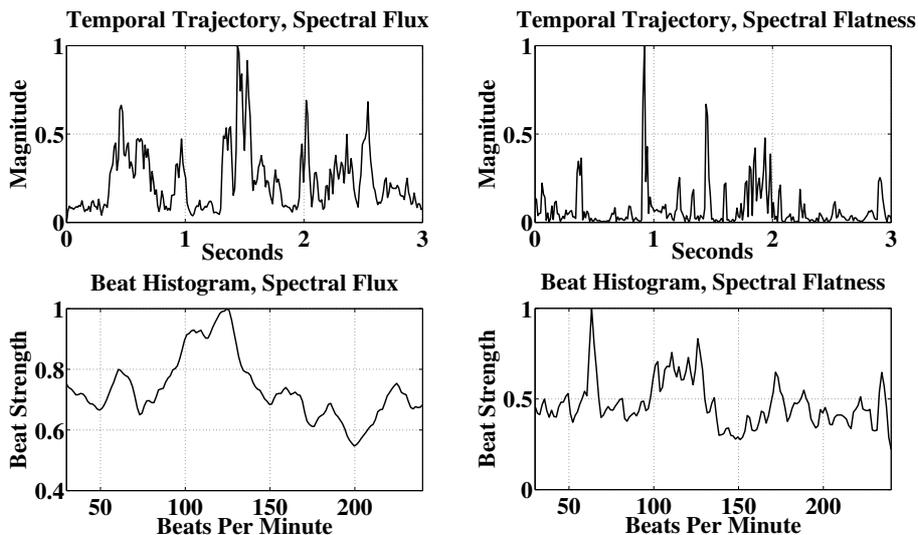


Figure 1: Temporal Trajectory of Spectral Flux and Spectral Flatness (upper row) and corresponding Beat Histograms from 30 – 240 BPM (lower row).

- Preprocessing:** the audio signal is down-mixed to mono, resampled to 22.5 kHz, lowpass-filtered to remove DC components and normalized.
- STFT:** the transform is computed with a frame-length of approximately 46.4 ms, windowed by a Hann window and with 75% overlap between consecutive frames. The resulting frequency resolution is $\Delta f = 10.77$ Hz.
- Novelty function:** the features listed earlier are extracted from the STFT frames, and the novelty function is computed through the calculation of the temporal trajectory and half-wave rectification, as is commonly practiced in novelty function extraction [43].
- Beat histogram:** the beat histogram is extracted through the calculation of an Autocorrelation Function (ACF) for each texture window of length 3 s. The overlap of the texture windows is 75%.
- Subfeature computation:** for each beat histogram, the subfeatures listed in Table 1 are extracted. The concatenation of all subfeature groups for each novelty function produces the final feature vector for an audio excerpt.

Similar subfeatures can be found e.g. in [16] (Peak), and [17, 18] (Distribution). In total, 30 novelty functions are used for the production of as many beat histograms, from each of which 19 subfeatures are extracted, resulting in a total count of 570 features. This effectively means that from the temporal trajectory of every MFCC or chroma coefficient, as well as the other features mentioned in the previous paragraph, a beat histogram is extracted, ensuring that all relevant periodicities and their properties — in different frequency bands and describing various audio aspects — are accounted for.

Table 1: Subfeatures extracted from Beat Histograms.

Distribution	Peak
Mean (ME)	Saliency of Strongest Peak (A1)
Standard Deviation (SD)	Saliency 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)	

3. EXPERIMENTAL SETUP AND EVALUATION

In order to be able to compare the results for the rhythm content features we extracted a *baseline feature set* by calculating the feature value over all texture windows of an excerpt (keeping the average value inside a window) without extracting periodicities. Those features are considered as a baseline since they represent a standard set of features used in genre classification. They include all features listed in Sect. 2 except the *Pitch Chroma Coefficients* and additionally include the features *Spectral Spread*, *Peak Amplitude Value*, and *Zero Crossing Rate*. The subfeatures on each of these features’ temporal trajectory throughout each track are given in the *Distribution* column of Table 1. In total, the baseline feature set comprises 21 features times 11 subfeatures = 231 features. We chose not to include other, more state-of-the-art non-rhythmic features in the baseline mainly because we aimed at investigating rhythmic features rather than non-rhythmic features. Second, since the main goal of the paper was to evaluate the beat histogram fea-

tures derived from all relevant novelty functions of the signal, it seemed plausible to include as a baseline the same novelty functions. This, however, was performed without applying the beat histogram transformation, so as to assess the added value of this certain processing step and the resulting features.

3.1. Classification

For the classification part, we apply the Support Vector Machines (SVM) [45] algorithm under MATLAB with a Radial Basis Function (RBF) kernel. The two hyperparameters for this kernel C and γ were determined with a grid search procedure. For all the experiments presented here, a 10-fold cross-validation took place, and results are averaged over the folds. The goal of the given classification setup is to compare the performance of the rhythm content feature set to a standard set of features, while also testing the combined set in order to assess the improvement when using the rhythm feature set in addition to the baseline set. All features are subjected to standardization (z-score) prior to classification (train and test set separately). The performance measure used to evaluate the classification is *accuracy*, defined as the proportion of correctly classified samples to all samples classified.

3.2. Datasets

In order to ensure comparability of the results with other publications and to evaluate the rhythm features for different genre hierarchies and tracks, five datasets were evaluated:

- GTZAN [16]
- Ballroom [18, 23]
- ISMIR04 [46]
- Unique [47]
- Homburg [48]

Although none of the datasets raises claims to being either complete or error-free (since the definition of genre itself is a debatable subject), their previous extensive use makes them suitable as benchmarks for the musical genre classification task. In order to avoid having results which could be artifacts of one specific dataset (e.g., for the GTZAN which has been criticized for its content [31], or the Ballroom which has been seen to be easily classifiable by the tempo feature [18]), we chose to include five very diverse datasets here. An interesting task which came under consideration would be training with one dataset and testing with another. However, the different class/genre structure across them did not allow for such a use in this context.

3.3. Feature Selection

The large number of resulting features mentioned in Sect. 2 makes direct feature evaluation a tedious task. In order to identify the best and worst performing descriptors we conducted a two stage feature selection: First, we apply a filter method (Mutual Information with Target Data [49], using the maximum relevance CMIM metric [50] from the MI-Toolbox [51]). Second, we run a sequential forward feature selection using the SVM as a wrapper method [52]. We retain the N best features (the feature number N is dataset-dependent, but ranges between 10 and 20) which gave comparable to or better accuracy than the full feature sets. This procedure is applied to the baseline and rhythm feature sets separately, and the

best features from both are then pooled to produce the combined feature set. Finally, we apply feature selection by separating feature subsets in the rhythm set in order to determine the effectiveness of different novelty functions and subfeature groups.

3.4. Bar Annotation

The parameters for the rhythm feature extraction algorithm were given in Sect. 2. It is a common problem of many such algorithms that the overlapping texture windows used in the block-wise processing of the audio file are of pre-defined length that does not necessarily represent "meaningful" parts of the music, such as, e.g., a bar. Placing the frames exactly at the boundaries of a bar or a musical phrase could increase the precision of the beat histogram representation, since the periodicities extracted from the segment would be musically meaningful, without onsets added or being left out because of random framing. In order to adapt the texture window boundaries to the bar boundaries, annotations of the audio files are necessary. In the case of the Ballroom dataset, such a manual annotation is available [53] and will be used here.

4. RESULTS

Results of the classification after feature selection for all datasets are presented in Fig. 2 and Table 2. The *priors* (percentage of a class/genre samples in a dataset) as average (Avg) and of the greatest class (Max P) are also provided. Sample results of the feature selection process concerning feature ranking (the three best and three worst features per dataset) are given in Table 3. Apart from that, in Fig. 3, accuracy results for all novelty function and subfeature groups for each dataset can be seen. Finally, confusion matrices are provided for the GTZAN and Ballroom datasets (Tables 4 and 5) for the rhythmic feature set in order to examine the misclassifications for those very well-known datasets.

Some tendencies can be clearly identified: The baseline feature set performs always better than the rhythm feature set alone except for the Ballroom dataset. The difference is mostly small but also significant at the 0.05 level in all cases except for the ISMIR04 dataset (based on a comparison test of the Cohen's Kappa extracted from the confusion matrices [54]); it ranges from 1.9% for the ISMIR04 dataset to 12.3% for the GTZAN dataset. Only for the Ballroom dataset the accuracy using the rhythm feature set is 6.8% above that observed when using the baseline set. The combined feature set outperforms the individual sets in all cases, the achieved accuracy being very close to that of the baseline feature set. This difference in accuracies is significant at the 0.05 level only in two cases (Ballroom and Unique). With regard to the datasets, results show accuracies in the area of 44.6 (rhythm feature set, Homburg) to 72.8% (combined feature set, GTZAN). The best performance of the rhythm features can be observed for the Ballroom dataset (67.7%), whereas the poorest performance can be found for the Homburg dataset. It should be noted that for the unbalanced datasets (i.e., all except GTZAN) it was observed from inspecting the confusion matrices that the achieved accuracy is mostly influenced by the most prominent class being classified correctly, whereas for the other classes, the performance is inferior but still in most cases above the prior of the respective class. An important result is the performance of the Ballroom dataset when using the annotated bars as texture window length (Table 2): It shows a clear improvement for the rhythm feature set alone and reaches 88.4% for the combined set.

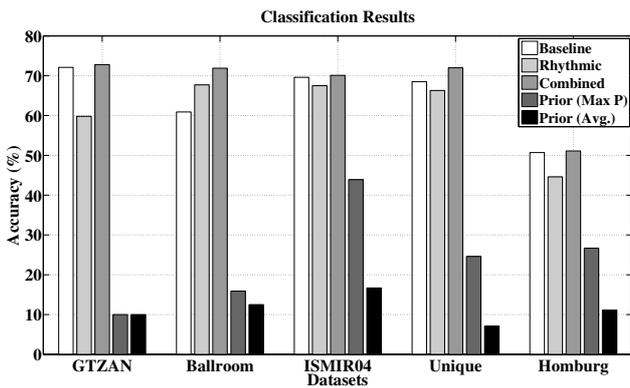


Figure 2: Classification results after feature selection.

Table 2: Classification results (accuracy in %) for various settings and best classification results. For the Ballroom dataset, a second accuracy value is reported for the case of using the manual annotation for frame boundaries.

Setting	GTZAN	Ballroom	ISMIR04	Unique	Homburg
Baseline	72.1	60.9	69.6	68.5	50.7
Rhythmic	59.3	67.3/76.7	67.5	66.3	44.6
Combined	72.8	71.9/88.4	70.1	72	51.1
P-MaxP	10	15.9	43.9	24.6	26.7
P-Avg	10	12.5	16.7	7.1	11.1

Concerning feature selection, it can be seen in Table 3 that the best features resulting from the information-theoretical feature selection procedure comprise proportionally more features based on spectral (SF, SFL) or amplitude (RMS) novelty, whereas with respect to the subfeatures the image is not so clear — only the SD feature appears more consistently (4 times) in the first ranks. The worst features can be seen to be based on tonal (STPR, SPC) novelty functions, with subfeatures giving a clearer image, with Peak features such as the A1 being frequent in the ranking.

Those tendencies can also be partially observed in Fig. 3, where the feature groups were tested individually: the only novelty functions showing to provide slightly better results on its own for all datasets are the SFL, MFCC2, STPR and RMS, whereas other MFCC and SPC features show relatively lower performance on their own. However, the performance of all novelty functions appears to be relatively similar, except for the SFL, STPR and RMS novelty functions which seem to provide higher performance for all datasets. In the case of the subfeature groups, no specific feature seems to be standing out.

Table 3: Best and worst features after feature selection. Abbreviation left of point denotes subfeature, otherwise novelty function.

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
568	A0.STPR	A0.STPR	SP.MFC3	A0.STPR	A1.MFC2
569	SP.MFC1	A1.MFC1	A1.RMS	A1.MFC3	A0.MFC2
570	A1.RMS	EN.MFC1	A0.STPR	A0.MFC3	A1.MFC1

Table 4: Confusion matrix for Ballroom dataset, average accuracy: 67.3%. Accuracy and Prior are given in %.

	Ch.	Ji.	Qu.	Ru.	Sa.	Ta.	Vw.	Wa.
Ch.	87	4	3	4	8	3	2	0
Ji.	9	40	1	1	6	2	1	0
Qu.	2	5	50	5	10	6	3	1
Ru.	10	0	3	62	0	5	2	16
Sa.	9	6	7	3	55	4	2	0
Ta.	3	0	9	2	2	58	7	5
Vw.	0	0	11	9	0	5	25	15
Wa.	0	0	2	12	0	1	2	93
Acc.	78	67	61	63	64	67	38	85
Pr.	15.9	8.6	11.7	14.0	12.3	12.3	9.3	15.8

Table 5: Confusion matrix for GTZAN dataset, average accuracy: 59.3%. Accuracy and Prior are given in %.

	Bl.	Cl.	Co.	Di.	Hi.	Ja.	Me.	Po.	Re.	Ro.
Bl.	59	2	4	4	5	5	8	1	7	6
Cl.	2	79	1	1	0	11	2	0	1	5
Co.	11	3	56	4	2	7	3	1	1	13
Di.	3	0	4	56	6	3	7	8	6	10
Hi.	3	0	0	8	64	2	2	5	13	3
Ja.	7	13	4	1	0	64	4	1	3	3
Me.	2	2	1	8	1	1	72	2	0	12
Po.	7	2	3	9	5	4	1	42	10	16
Re.	6	1	3	6	14	4	0	4	59	3
Ro.	7	4	8	4	4	4	15	9	4	42
Acc.	59	79	56	56	64	64	72	42	59	42
Pr.	10	10	10	10	10	10	10	10	10	10

5. DISCUSSION

The results given in Table 2 provide strong support for the view that the presented rhythm features can perform in the same range as non-rhythmic baseline features. The overall better performance of the combined feature set is a consequence of using information related to both timbre and rhythm features, with significant improvement for two out of five datasets.

Concerning the datasets, the poor classification performance observed for the Homburg dataset could be an indication that the latter has a special genre or track selection, which cannot be predicted efficiently using our features. The very similar results observed for all feature sets for the ISMIR04 dataset is most probably due to its largely unbalanced character. The Ballroom dataset stands out as a good example where the rhythm features alone could offer a satisfactory performance. The results reported here are comparable to or better than those reported in studies using similar methods [16, 17, 18], but lie below those of newer studies employing more sophisticated features which depart from the beat histogram [55, 7, 10, 3, 56, 11, 57]. However, the simplicity of the beat histogram calculation and feature extraction and the possibility to assess features and feature groups individually are, in our view, advantages which have to be taken into account.

The use of bar boundaries as texture window boundaries for the Ballroom dataset gave encouraging results: A maximum accuracy of 76.7% was achieved, which is a notable improvement to the 67.7% with fixed-length segmentation. The result is close to the one reported by Gouyon et al. [18] using features extracted

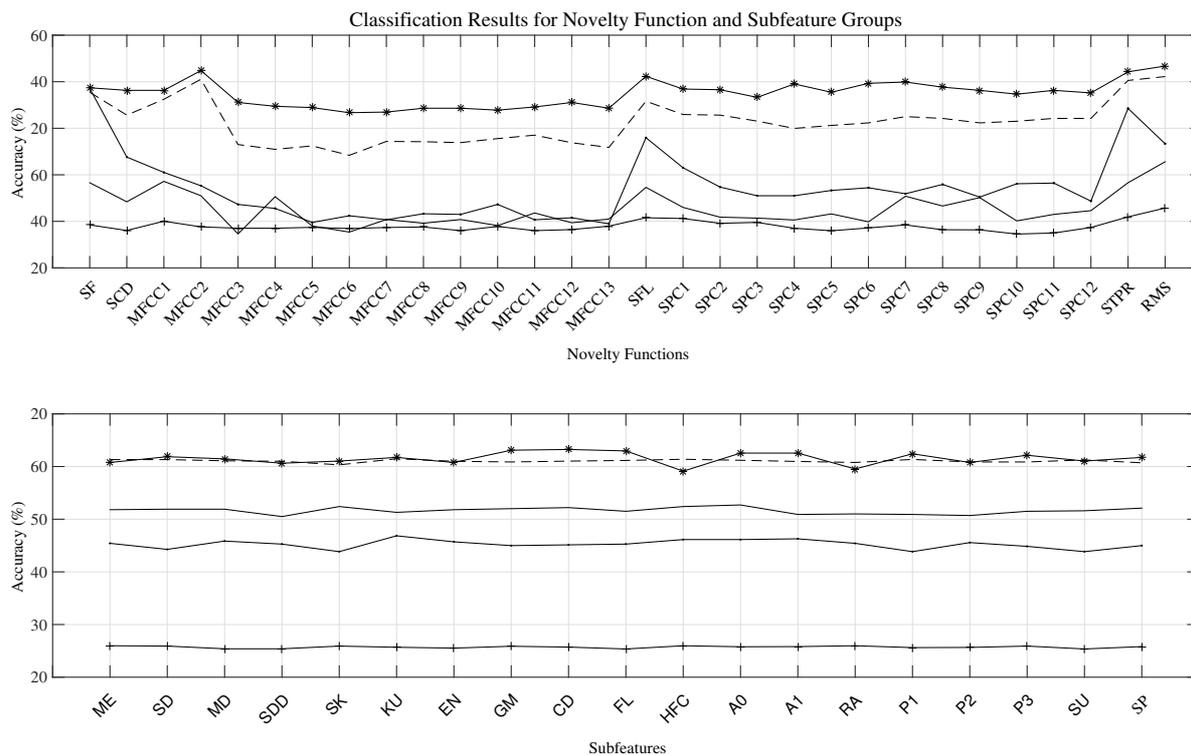


Figure 3: Classification results for feature groups. Upper graphic shows Novelty Function, lower Subfeature groups. Datasets are denoted by the following line styles: "-" (GTZAN), "-." (Ballroom), "*-" (ISMIR04), "-" (Unique), "+-" (Homburg).

from the beat histogram. This suggests that the application of prior knowledge regarding the “real” boundaries of the musical surface can help to considerably improve the accuracy of rhythm-based genre classification.

The misclassifications (reported in Tables 4 and 5) show that genre confusions are those that could be expected when using features capturing the rhythmic character of the pieces, thus making it difficult to distinguish between genres containing similar patterns: for the Ballroom dataset, the most prominent misclassifications take place between Rumba/Waltz, Quickstep/Samba and Waltz/Viennese Waltz. For the GTZAN dataset, Classical is confused with Jazz, Country with Rock, Hiphop with Reggae and Pop with Rock and Metal. Those results indicate that the multiple novelty-based beat histogram features indeed capture the specific rhythmic properties of the genres and the regularities of their periodicities, pointing towards the suitability of those features for the extraction of general rhythmic properties.

With regards to the feature selection, it was shown that mainly novelty functions which are based on amplitude or spectral shape changes in the signal give the best results, a result which could be confirmed both from the standard feature selection procedure, as well as from the feature subset-based selection. Those results can be due to important turning points of a track in most popular western music being mostly mediated through loudness, energy or spectral form (for example, in the case of many instruments playing together in a bar change or a new instrument entering the

scene) which leads to those features possessing more salient novelty functions out of which more qualitative beat histogram features can be extracted. Another reason could be the role of those novelty functions in stressing the basic metric positions in most of western popular music. Concerning the subfeatures and the results of the information-theoretical feature selection (Table 3), the higher performance of simple statistics such as the SD of the beat histogram attest to them having more discriminative power, possibly because they express very basic statistical tendencies of the periodicity distribution. Furthermore, higher-level features (such as the P1, which is a good estimate of the excerpts’ tempo, already shown to be important in rhythm-based genre classification [18]), also possess discriminative power since their value is an important indicator of a genre character (e.g. dance vs. classical music). However, since no such tendencies are observed in the feature group selection (Fig. 3), it can be deduced that no specific subfeature on the beat histogram bears special discriminative power, but it is rather the combination with a salient novelty function which allows for better results.

6. CONCLUSIONS

The work presented in this paper focuses on the creation of novel features for rhythm-based musical genre classification. The difference in comparison to previous studies in the field [16, 17, 18], using the signal amplitude envelope only, is the use of the tempo-

ral trajectory of other signal quantities such as SF as the novelty function for the calculation of the beat histogram. We showed that performance using these beat histogram features is higher or in a similar range than related work using periodicity histograms. It has also been shown that specific novelty functions relating to amplitude or spectral shape are among the most informative when analyzed with a periodicity representation method. Finally, we showed the positive impact of manual bar-boundary annotation for the extraction of rhythm features on classification performance.

There are many more features which can be considered as novelty functions [38, 43], as well as possible subfeatures on the beat histogram (such as MFCCs, presented in [18]) and other methods for the periodicity representation calculation [19, 17, 18]. Future goals include an even more extensive and detailed feature selection, identifying the features or feature groups which are informative with respect to specific genres (in order to associate specific novelty functions with relevance to specific genres) and a test of the feature robustness against signal degradations. Those subfeatures and novelty functions could be then suitable for future use in more specific rhythm description tasks. Furthermore, the investigation of optimal parameter settings for feature extraction and classification, the utilization of other classification methods and performance evaluation measures, as well as the usage of other methods for feature aggregation are other possible research directions. The high accuracy achieved especially for the Ballroom dataset indicates the suitability of the descriptors for further application and points to the importance of bar-boundary annotation (which can also be performed automatically) for rhythm features.

While the features for beat histogram calculation have been evaluated only in the context of genre classification, we believe that they will prove useful in other tasks as well. Future research will concentrate on adjusting and using rhythm content features for MIR tasks such as audio similarity and mood recognition. Preliminary research has also been undertaken concerning the use of novelty functions of specific instruments (e.g. Drums) extracted through Non-Negative Matrix Factorization (NMF) [58], or the application of the features to other signals, such as speech [59]. Their application in a task of automatic spoken language identification based on the rhythmic elements of speech has shown promising results and points to further research directions.

7. REFERENCES

- [1] Enric Guaus i Termens, "New approaches for rhythmic description of audio signals," Tech. Rep., Universitat Pompeu Fabra, Music Technology Group, 2004.
- [2] Fabien Gouyon and Simon Dixon, "A review of automatic rhythm description systems," *Computer Music Journal*, vol. 29, no. 1, pp. 34–35, 2005.
- [3] Geoffroy Peeters, "Spectral and temporal periodicity representations of rhythm for the automatic classification of music audio signal," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 19, no. 5, pp. 1242–1252, 2011.
- [4] Tlacaél Miguel Esparza, Juan Pablo Bello, and Eric J Humphrey, "From genre classification to rhythm similarity: Computational and musicological insights," *Journal of New Music Research*, no. ahead-of-print, pp. 1–19, 2014.
- [5] Nicolas Scaringella, Giorgio Zoia, and Daniel Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, March 2006.
- [6] Enric Guaus i Termens, *Audio content processing for automatic music genre classification: descriptors, databases, and classifiers.*, Ph.D. thesis, Universitat Pompeu Fabra, Barcelona, 2009.
- [7] Tim Pohle, Dominik Schnitzer, Markus Schedl, Peter Knees, and Gerhard Widmer, "On rhythm and general music similarity," in *ISMIR*, 2009, pp. 525–530.
- [8] François Pachet, Daniel Cazaly, et al., "A taxonomy of musical genres," in *Proceedings of the Conference on Content-Based Multimedia Information Access*, 2000.
- [9] Jean-Julien Aucouturier and François Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.
- [10] Andre Holzapfel, Arthur Flexer, and Gerhard Widmer, "Improving tempo-sensitive and tempo-robust descriptors for rhythmic similarity," in *Proceedings of the 8th Sound and Music Computing Conference*, 2011.
- [11] Aggelos Pikrakis, "A deep learning approach to rhythm modelling with applications," in *6th International Workshop on Machine Learning and Music (MML13)*, 2013.
- [12] Grosvenor Cooper, *The rhythmic structure of music*, vol. 118, University of Chicago Press, 1963.
- [13] Paul Fraisse, "Rhythm and tempo," in *The psychology of music*, Diana Deutsch, Ed., Series in Cognition and Perception, chapter 6. Academic Press, 1982.
- [14] Fred Lerdahl and Ray S Jackendoff, *A generative theory of tonal music*, MIT press, 1983.
- [15] Joel Lester, *The rhythms of tonal music*, Pendragon Press, 1986.
- [16] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Trans. on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.
- [17] Juan José Burred and Alexander Lerch, "A hierarchical approach to automatic musical genre classification," in *DAFX*, 2003.
- [18] Fabien Gouyon, Simon Dixon, Elias Pampalk, and Gerhard Widmer, "Evaluating rhythmic descriptors for musical genre classification," in *AES*, 2004.
- [19] Eric D Scheirer, "Tempo and beat analysis of acoustic musical signals," *The Journal of the Acoustical Society of America*, vol. 103, no. 1, pp. 588–601, 1998.
- [20] Elias Pampalk, Simon Dixon, and Gerhard Widmer, "Exploring music collections by browsing different views," *Computer Music Journal*, vol. 28, no. 2, pp. 49–62, 2004.
- [21] Simon Dixon, Fabien Gouyon, Gerhard Widmer, et al., "Towards characterisation of music via rhythmic patterns," in *ISMIR*, 2004.
- [22] Andre Holzapfel and Yannis Stylianou, "Rhythmic similarity of music based on dynamic periodicity warping," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*. IEEE, 2008, pp. 2217–2220.
- [23] Simon Dixon, Elias Pampalk, and Gerhard Widmer, "Classification of dance music by periodicity patterns," in *ISMIR*, 2003.
- [24] Jonathan Foote and Shingo Uchihashi, "The beat spectrum: A new approach to rhythm analysis," in *ICME*, 2001.

- [25] Kristopher West and Stephen Cox, "Features and classifiers for the automatic classification of musical audio signals," in *ISMIR*, 2004.
- [26] Elias Pampalk, Arthur Flexer, Gerhard Widmer, et al., "Improvements of audio-based music similarity and genre classification," in *ISMIR*. London, UK, 2005, vol. 5, pp. 634–637.
- [27] Geoffroy Peeters and Helene Papadopoulos, "Simultaneous beat and downbeat-tracking using a probabilistic framework: theory and large-scale evaluation," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 6, pp. 1754–1769, 2011.
- [28] Bob L Sturm, Corey Kereliuk, and Aggelos Pikrakis, "A closer look at deep learning neural networks with low-level spectral periodicity features," in *Cognitive Information Processing (CIP), 2014 4th International Workshop on*. IEEE, 2014, pp. 1–6.
- [29] George Tzanetakis, Georg Essl, and Perry Cook, "Human perception and computer extraction of musical beat strength," in *Proc. DAFx*, 2002, vol. 2.
- [30] Bob L Sturm, "An analysis of the gtzan music genre dataset," in *Proceedings of the second international ACM workshop on Music information retrieval with user-centered and multimodal strategies*. ACM, 2012, pp. 7–12.
- [31] Bob L Sturm, "The gtzan dataset: Its contents, its faults, their affects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.
- [32] B Sturm, "A simple method to determine if a music information retrieval system is a "horse"," *IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 16, NO. 6, OCTOBER 2014*, 2014.
- [33] H Christopher Longuet-Higgins and Christopher S Lee, "The perception of musical rhythms," *Perception*, vol. 11, no. 2, pp. 115–128, 1982.
- [34] Richard Parncutt, "A perceptual model of pulse salience and metrical accent in musical rhythms," *Music Perception*, pp. 409–464, 1994.
- [35] David Temperley, *The cognition of basic musical structures*, MIT press, 2004.
- [36] Justin London, *Hearing in time*, Oxford University Press, 2012.
- [37] Stephen Hainsworth and Malcolm Macleod, "Onset detection in musical audio signals," in *ICMC*, 2003.
- [38] Juan P. Bello, Chris Duxbury, Mike Davies, and Mark Sandler, "On the use of phase and energy for musical onset detection in the complex domain," *IEEE Signal Processing Letters*, vol. 11, no. 6, pp. 553–556, 2004.
- [39] Axel Roebel, "Onset detection in polyphonic signals by means of transient peak classification," in *ISMIR*, 2005.
- [40] Anssi Klapuri, "Sound onset detection by applying psychoacoustic knowledge," in *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing, 1999*. IEEE, 1999, vol. 6, IEEE.
- [41] Chris Duxbury, Mark Sandler, and Mike Davies, "A hybrid approach to musical note onset detection," in *Proc. Digital Audio Effects Conf.(DAFX,i02)*, 2002, pp. 33–38.
- [42] Chris Duxbury, Juan Pablo Bello, Mike Davies, Mark Sandler, et al., "Complex domain onset detection for musical signals," in *Proc. Digital Audio Effects Workshop (DAFx)*, 2003, pp. 6–9.
- [43] Juan P. Bello, Laurent Daudet, Samer Abdallah, Chris Duxbury, Mike Davies, and Mark B Sandler, "A tutorial on onset detection in music signals," *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 5, pp. 1035–1047, 2005.
- [44] Alexander Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*, John Wiley & Sons, 2012.
- [45] Vladimir Vapnik, *The nature of statistical learning theory*, Springer, 2000.
- [46] Adam Berenzweig, Beth Logan, Daniel PW Ellis, and Brian Whitman, "A large-scale evaluation of acoustic and subjective music-similarity measures," *Computer Music Journal*, vol. 28, no. 2, pp. 63–76, 2004.
- [47] Klaus Seyerlehner, Gerhard Widmer, and Dominik Schnitzer, "From rhythm patterns to perceived tempo," in *ISMIR*, 2007.
- [48] Helge Homburg, Ingo Mierswa, Bülent Möller, Katharina Morik, and Michael Wurst, "A benchmark dataset for audio classification and clustering," in *ISMIR*, 2005.
- [49] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.
- [50] Hanchuan Peng, Fulmi Long, and Chris Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2005.
- [51] Gavin Brown, Adam Pocock, Ming-Jie Zhao, and Mikel Luján, "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," *The Journal of Machine Learning Research*, vol. 13, pp. 27–66, 2012.
- [52] Ron Kohavi and George H John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1, pp. 273–324, 1997.
- [53] Florian Krebs, Sebastian Böck, and Gerhard Widmer, "Rhythmic pattern modeling for beat and downbeat tracking in musical audio.," in *ISMIR*, 2013.
- [54] Giles M Foody, "Thematic map comparison," *Photogrammetric Engineering & Remote Sensing*, vol. 70, no. 5, pp. 627–633, 2004.
- [55] Klaus Seyerlehner, Markus Schedl, Tim Pohle, and Peter Knees, "Using block-level features for genre classification, tag classification and music similarity estimation," *MIREX 2010*, 2010.
- [56] Yannis Panagakis, Constantine Kotropoulos, and Gonzalo R Arce, "Music genre classification using locality preserving non-negative tensor factorization and sparse representations.," in *ISMIR*, 2009.
- [57] Yannis Panagakis, Constantine L Kotropoulos, and Gonzalo R Arce, "Music genre classification via joint sparse low-rank representation of audio features," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1905–1917, 2014.
- [58] Athanasios Lykartsis, Chih-Wei Wu, and Alexander Lerch, "Beat histogram features from nmf-based novelty functions for music classification," in *ISMIR*, 2015.
- [59] Athanasios Lykartsis and Stefan Weinzierl, "Using the beat histogram for speech rhythm description and language identification," in *INTERSPEECH*, 2015.



Speech and Music Discrimination: Human Detection of Differences between Music and Speech based on Rhythm

Madeleine Stanev, Johannes Redlich, Christian Knörzer, Ninett Rosenfeld, Athanasios Lykartsis

Audio Communication Group, Technische Universität Berlin, Germany

madeleine.stanev@mailbox.tu-berlin.de, johannes.redlich@mailbox.tu-berlin.de,
christian.knoerzer@mailbox.tu-berlin.de, ronin@mailbox.tu-berlin.de,
athanasios.lykartsis@tu-berlin.de

Abstract

Rhythm in speech and singing forms one of its basic acoustic components. Therefore, it is interesting to investigate the capability of subjects to distinguish between speech and singing when only the rhythm remains as an acoustic cue. For this study we developed a method to eliminate all linguistic components but rhythm from the speech and singing signals. The study was conducted online and participants could listen to the stimuli via loudspeakers or headphones. The analysis of the survey shows that people are able to significantly discriminate between speech and singing after they have been altered. Furthermore, our results reveal specific features, which supported participants in their decision, such as differences in regularity and tempo between singing and speech samples. The hypothesis that music trained people perform more successfully on the task was not proved. The results of the study are important for the understanding of the structure of and differences between speech and singing, for the use in further studies and for future application in the field of speech recognition.

Index Terms: speech-music discrimination, speech perception, speech rhythm, computational paralinguistics

1. Introduction

The identification of rhythm in language as it can be detected in music has been a growing field in many disciplines such as linguistics, music psychology or neuroscience. The nature and understanding of rhythm is vital in order to locate similar patterns in speech. Even though periodicity has been widely linked to the description of rhythm, it is important to treat this term carefully, especially with regard to speech rhythm [1]. Also according to Patel [1], rhythm can be best described as “the systematic pattern of sound in terms of timing, accent and grouping” since “both speech and music are characterized by systematic temporal, accentual, and phrasal patterning”. Rhythm in speech, on the other hand, refers to the way languages are organized in time. Most linguists argue that certain languages differ in their rhythmic structure. The basis of the exact distinction is yet hard to define. It is assumed that the differences depend on whether sequences of syllables or stresses are isochronous. English and Dutch, for example, manifest highly varied syllable patterns and therefore are referred to as “stress-timed”, whereas languages like French and Spanish are “syllable-timed” due to their low varied syllable structure [2]. Nevertheless, critics argue that these distinctions are nonexistent [3]. Lehiste [4], for example,

mentioned that isochrony is considered as a perceptive phenomenon. In recent years, studies have turned away from the idea of isochrony and, to a greater extent, have focused on vocalic and intervocalic intervals in speech [3, 5, 6]. This shift away from classes and towards rhythmic differences has evoked the idea of a rhythmic continuum in which the clusters become blurry if more languages are added [2, 7].

The identification of durational vowel patterns in speech requires the application of specific methods. The development of the *Pairwise Variability Index (nPVI)* by Patel & Daniele [8] allowed the researchers to quantify prosody, which was directly comparable to music. With the nPVI the durational contrast between successive elements in a sequence can be detected and measured. The method has been widely applied to exploit more accurate differences between “stress-timed” and “syllable-timed” languages. Vowel-based measures such as the nPVI are the most plausible way to compare speech to music since musical notes can be roughly compared to syllables, with vowels forming the core of syllables [9].

Ramus and Mehler [10] examined the ability of infants that have been raised in a bi- or multilingual environment to discriminate between languages. They used *speech resynthesis* to further explore the components of prosody by measuring all relevant acoustic parts of a speech signal and subsequently resynthesizing the speech material with an appropriate algorithm. This enabled them to freely select or dispose of the components they wanted to use, such as phonemes, rhythm, and intonation. The results showed that when preserving intonation, rhythm, and broad phonetic categories alternately, that the use of these suprasegmental cues is indeed a sufficient way to allow the discrimination of languages.

Ohala and Gilbert [11] tried a different approach by providing participants with conversations instead of single sentences. The use of long passages instead of short ones had no effect on participants’ performances. They furthermore converted the speech signal with the help of a voltage-controlled-frequency and voltage-controlled-amplitude signal generator into a ‘buzz’ while still maintaining the same frequency, timing, and amplitude of the original signal. The results were highly significant statistically ($p < 0.0001$) but the scores were still low enough to indicate that the signal lacks crucial prosodic information after the conversion. Syllables and word boundaries, for example, are not recognizable anymore. Since syllables may be the closest comparable cue to a music tone due to their salience for perception, their preservation might lead to a better performance.

Several studies [11, 12, 13, 14] have succeeded in demonstrating that the rhythmic pattern in speech and language is able to act as clue even after the extraction of different prosodic components. In this paper we aim to further investigate this subject, also with regard to the rhythmic structure to music. Instead of solely comparing music with spoken sentences, we make use of singing sequences without any instrumental background (a capella singing). Thus, the signals are quite similar with regard to their prosodic structure and are barely left with any musical features. This gives us the opportunity to explore the rhythmic patterns more specifically without the distraction of other cues or the dominance of instrumental rhythm. However, discoveries have been already made regarding the phenomenon of musicians composing their music on the basis of language-specific-rhythmic patterns [15]. By overlaying the envelope curve of the speech and singing sequences with a carrier signal we are able to eliminate any other linguistic information apart from rhythm, allowing us to further explore rhythmical characteristics and differences of speech and music.

2. Method

The question is: "Are people able to distinguish singing from spoken words only by the rhythm?". Based on personal experience, the current status of research and the experiments of Ramus et al. [14], we find the investigation of this question worthwhile. According to Ramus et al., speech has a significant time structure. We will use this time structure to test if people are able to distinguish singing from spoken words. Therefore, the H_0 hypothesis is that singing and speech are distinguishable based only on rhythmic cues, whereas H_1 states that it is not possible to distinguish them better than chance on a significant level. To that purpose we used recordings from a multilingual prosodic corpus (MULTEXT PD) [16]. The corpus derived from the EUROM 1 speech database that was developed within the Esprit SAM (Multilingual Speech Input/Output Assessment, Methodology and Standardisation) project [17]. The collection contains 40 different thematically connected passages, each comprising five sentences. 50 different speakers read in five languages (French, English, Italian, German and Spanish). We used a mere 2 out of the 40 passages, spoken by 10 different speakers. Each speaker spoke one of these two passages so that we got 10 passages. With this selection we want to get a pool of different ways of speaking, to represent different rhythms of speaking. Furthermore, we only made use of the German passages since most of our participants were of German nationality, so that we are primarily researching rhythm perception for native speakers. The acoustic quality of the recording is high with sampling rate at 20 kHz, 16 bits, and the recordings having been performed in an anechoic room. An example of a recorded message can be seen below:

I have a problem with my water softener. The water level is too high, and the overflow keeps dripping. Could you arrange to send an engineer on Tuesday morning, please? It's the only day I can manage this week. I'd be grateful if you could confirm the arrangement in writing.

We downloaded the singing files from youtube. Any videos uploaded on youtube are publicly available and can therefore be legally downloaded and processed for research purposes.

Furthermore, the files are altered in an unrecognizable way. We used five a capella songs, three choir sequences, and two rap songs. They were all in German language in order to ensure a better comparison basis with the spoken sentences.

To create the stimuli we followed a self-developed method, because of transparency and simplicity in the algorithm and the accuracy of the envelope. It is also possible to produce the envelope by using the Hilbert transform, but in our case it was not so well fitted to the signal. We created the stimuli by overlaying a carrier signal (noise or sine tone) with the envelope curve of the speech and singing signals. The noise is in our case a band-pass filtered noise between 200 Hz and 400 Hz. Figure 1 shows the detection of the peak points. The detection by itself was accomplished with a MATLAB script. Connecting the peak points of the useful signal formed the envelope curve of a signal. The distance between the peak points was defined by the desired maximum frequency of 2000 Hz of the envelope curve.

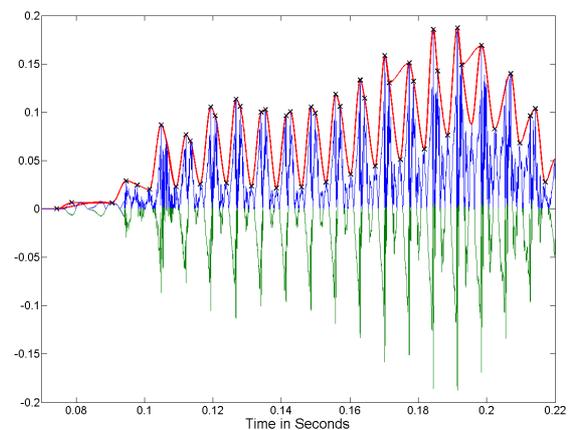


Figure 1: *Generation of the envelope curve for a stereo signal. The green curve is the speech-signal, the blue curve is the absolute value of the speech-signal, black crosses represent signal peaks and the red curve is the interpolated envelope curve.*

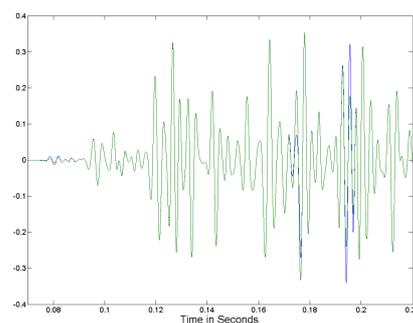


Figure 2: *Final stimuli with sine wave as carrier.*

The connection of the points was generated with a piece-by-piece cubic interpolation creating a high frequency useful signal on a low frequency carrier signal. This method can have as a result the appearance of unintended artifacts, like "bubbling" or "scratching" sounds, which we will accept due to its otherwise high effect of eliminating every linguistic aspect but the rhythm and the outcome of a reduced signal.

The results did not show any evidence of participants feeling disturbed by these artifacts. We are convinced that other methods could bring fewer artifacts, but then the extraction of the rhythm could be less effective. During the survey participants are asked to comment on their perception of the stimuli, to gain a deeper understanding of the applicability. We decided to limit the lengths of the stimuli to ten seconds, which still ensured the detectability of the rhythm by simultaneously preserving the participants from overstimulation or fatigue. We decided to conduct our research via the online survey platform LimeSurvey to reach more potential participants. A recent study by Pysiewicz [18] suggested that results produced in online settings hold as much validity as the ones within laboratory settings. Since this research was conducted in the field of auditory perception research, it can be applied here. The survey was also designed to query about further personal information such as the degree of musical education or the valuation of musical understanding. This information may help to further specify any influential aspects on the ability to discriminate. Since the study was distributed via the Internet, the participants could choose their own test setting, which gave us no influence on the environment. They used their own equipment, either loudspeakers or headphones, to listen to the sound samples. This is not a disadvantage, since the perception of the rhythm as a high-level element is relatively disconnected from the quality of reproduction. However, a test phase with two stimuli, pure sine tone and pure noise, allowed the participants to adjust the volume accordingly and to get a general idea of the transformed sequences. In the next step, 20 out of the 40 possible testing samples were presented randomly with regard to the alteration pattern (sine or noise), alternating between spoken sentences and singing. The randomization of the stimuli counteracted the appearance of any potential systematic sequential effects or response biases. To get a better understanding of the clues that had helped participants with their decision, they were asked for further explanations in the end of the survey. Thirty-two volunteers participated in the survey. 84.4% of the contestants completed the survey in German and 15.6 % in English. 78.1 % were German native speakers and the remaining 21.9 % were almost equally distributed between Spanish, Russian, Chinese, Japanese and Greek. Participants were mostly male (65.6%) than female (34.4%), with a mean age of 28 years.

3. Results

Table 1 presents a first overview of our results including hit rates, false alarm rates, discrimination scores A' [19] and the response bias measures B''_D [20]. Hit rates serve as the percentage of correctly identified singing sequences and false alarm rates include the percentage of speech passages that were selected as singing.

Table 1. Hit rates, false alarm rates, discrimination scores, and response bias measures for each stimuli condition. A' was compared to 0.5 (chance level).

Signal Carrier	Hit rates	False Alarms	A'	B''_D
Sine	0.78	0.27	0.84	-0.135
Noise	0.66	0.18	0.83	0.402

To perform an analysis of discrimination, we first applied a Kolmogorov-Smirnov test for normality to the discrimination scores. The results showed that the distribution of the score could be considered as normal with all p values < 0.134 . Afterwards, a t-Test was computed to compare the discrimination scores to the chance level of 0.5. The results showed that the discrimination scores were significantly above chance level in the sine ($t(16) = 4.39, p \leq 0.001$) and noise ($t(16) = 4.523, p \leq 0.001$) conditions. Hence, participants were able to distinguish between singing and speech regardless of its transformation, confirming the H_0 . According to the results, we assume that their decisions were not random but based on the detection of meaningful cues.

To gain a deeper knowledge on any potential effects of music education on the ability to discriminate, we performed an ANOVA. Test results show with $p > 0.462$ no significance. Therefore, music education or training does not correlate with a better performance on the test. Results with the regard to age, sex and nationality showed no significance either. After the discrimination task we provided participants with five possible cues that could have helped with their decision and asked them to determine the ones that had contributed to their choice. Most participants opted for the sound-pause-ratio (81.25%) followed by regularity (62.5%), accentuation (56.25%), timing (46.9%), and grouping (18.75%). When asked to further define the aspects that lead participants to a certain conclusion, several stated that they focused on regularity or periodic sequences. Irregular patterns and a faster tempo were assumed to belong to speech passages, whereas more regular patterns and the impression of fluency were associated with singing sequences. A few participants referred to perceived rhythm as a main characteristic.

4. Discussion

The results presented in Table 1 indicate that participants were able to discriminate between singing and speech in both conditions. Ramus et. al [5] noted that for single frequencies, intonation-stressed pitches and syllables were still detectable. In addition, signals without any intonation cues showed that “syllabic rhythm was a robust cue for discrimination” [5]. The B''_D scores show a slight difference of response bias between the two conditions. Participants were conservative with the noise stimuli and liberal with the sine stimuli, meaning that the tendency to recognize the noise stimuli was slightly higher with the sine stimuli. This could relate to the clearer acoustic nature of the sine stimuli in comparison to the noise ones. The test results of the ANOVA implied that music education in any form has ultimately no impact on the discrimination of the stimuli. The right perception of stimuli seems to be independent from musical preparatory training or talent. A reason for this could be the different significance of rhythm in speech and singing. Human beings each learn at least a language in their lifetime and are native speakers of it. This circumstance allows them to develop a special consciousness for speech rhythm. According to Patel [1], speech rhythm is determined by its irregularity and musical patterns are defined by its periodicity. Therefore, musical patterns seem to be structured in a more elementary way than speech patterns. We suppose that simpler patterns are in general also easier to detect. Assuming that people have the same knowledge with regard to complex speech patterns, no difference exists between musically trained people and non-musicians. Hence, participants mostly relied on the sound-pause-ratio while listening to the stimuli. Since rhythm correlates with temporal

aspects, the explicit focus on that ratio seems to be meaningful. Regularity plays another important role and is often associated with rhythmic structures in music. Recurring periodic patterns served as a hint to the existence of some musical structure.

5. Conclusions

The aim of this paper was to test the capacity of subjects to discriminate between speech and singing with rhythm serving as the only indicator. While most studies have been focusing on the discrimination of different languages, especially with infants [21, 22], or the connection between language and corresponding rhythms in compositions [13], researchers are still at odds with the exact definition of rhythm in speech. Therefore, we examined the musical rhythmic structure as well as the linguistic ones. Even though our results resemble the previous ones, we could discover that rhythmic structures vary between language and music, as participants stated these variations as a crucial criterion for the discrimination of presented signals. As mentioned above, participants rated signals regarding to their regularity and periodicity. Participants found irregularity in rhythmic patterns and faster pace to be indicators for speech whereas they characterized singing sequences by regular patterns and a medium pace. The rhythm of speech is defined by the use of syllables and stresses depending on the context, systemizing the sequences of sounds in the sense of time, accentuation and grouping [1]. As a consequence, language is never just a periodic repetition of linguistic units [1]. In contrast, Western music has a regular rhythmic structure, which allows a high complexity on other musical and more important dimensions [1]. Flexibility of pace is often used as a way of expression [1]. We also offer a new method that allows the complete transformation of a signal without losing desired valuable components. By overlaying the signals' envelope curves with sine and noise carrier signals we ensured that all linguistic cues were eliminated besides rhythm. The discoveries in this study are applicable to various fields. An explicit attribution of singing stimuli as such may be, for example, useful for algorithms in applications that aim to match a singing sequence to the appropriate song. We suggest further research to explicitly identify the specific rhythmic patterns that contribute to a correct perception of singing and speech.

A better comprehension of rhythm in speech and singing can be meaningful in the development of speech therapies. Patients' entry into the therapy could be simplified by using the periodic rhythm patterns of singing to give them a better understanding of the rhythmic nature of language. A specific linguistic rhythm can be better suited for a specific, more efficient use of the language in music [23]. It is probably no coincidence that most operatic arias were composed in Italian. However, this assumption contradicts with the categorization of languages into "stress-timed" and "syllable-timed" since the English language is commonly regarded as "most singable". Future studies could draw a comparison between several different languages to further exploit our results. The high performance in our study could derive from specific characteristics of the German language that may simplify discrimination. One should also take into account that we used different singing styles (rap, ballad, etc.). Some of these are more similar to speech (like rap) than others. Therefore, they are not clearly recognizable as singing sequences because the rhythmical features are less typical for musical rhythm structures. It would be interesting to investigate if a more

homogenous selection or a selection with a larger variance, consisting of greater number of sequences than used here might offer even clearer results. Furthermore, it would be interesting to see if the test subjects perform differently when only comparing rap (or other more specific vocal singing types) sequences to speech, in contrast to the comparison of a capella singing, so as to determine if the detection of rhythm differences is dependent on the kind of vocal song involved. Likewise, this could also be performed for spontaneous speech rather than read as in our study. Finally, future work could also explore the significance of our results when applied to training with stutterers, as they are often more capable in singing, which could point to a better understanding of rhythm in singing than normal speech.

6. References

- [1] Patel, Aniruddh D.: "Rhythm." In: Aniruddh D. Patel (Ed.) *Music, Language and the Brain*. Oxford, The Oxford University Press, pp. 96 – 177, 2007.
- [2] Grabe, Esther, and Ee Ling Low: "Durational variability in speech and the rhythm class hypothesis." In: *Papers in laboratory phonology*, 7, pp. 515-546, 2002.
- [3] Roach, P.: "On the distinction between 'stress-timed' and 'syllable-timed' languages," In: D. Crystal (Ed.) *Linguistic Controversies: Essays in Linguistic Theory and Practice in Honour of F.R. Palmer*, Edward Arnold, London, pp. 73–79, 1982.
- [4] Lehiste, Ilse: "Isochrony reconsidered." In: *Journal of phonetics*, 5 (3), pp. 253 – 263.
- [5] Low, Ee Ling, Grabe, Esther, and Francis Nolan: "Quantitative Characterizations of Speech Rhythm: Syllable-Timing in Singapore English." In: *Language and speech*, 43 (4), pp. 377-401, 2000.
- [6] Ramus, Franck, Nespor, Marina, and Jacques Mehler: "Correlates of linguistic rhythm in the speech signal." In: *Cognition*, 73 (3), pp. 265-292, 1999.
- [7] Farinas, Jérôme, and François Pellegrino: "Automatic rhythm modeling for language identification." In: *INTERSPEECH*, pp. 2539–2542, 2001.
- [8] Patel, Aniruddh D., and Joseph R. Daniele: "An empirical comparison of rhythm in language and music." In: *Cognition*, 87 (1), pp. 35-45, 2003.
- [9] Patel, Aniruddh D.: „Rhythm in Language and Music – Parallels and Differences.“ In: *Annals of the New York Academy of Science*, 999 (1), pp. 140-143, 2003.
- [10] Ramus, Franck, and Jacques Mehler: "Language identification with suprasegmental cues: A study based on speech resynthesis." In: *The Journal of the Acoustical Society of America*, 105 (1), pp. 512-521, 1999.
- [11] Ohala, J. J., and J. B. Gilbert: "Listeners' ability to identify languages by their prosody." In: *Report of the Phonology Laboratory Berkeley*, 2, pp. 126-132, 1978.
- [12] Patel, Aniruddh D., Peretz, Isabelle, Tramo, Mark, and Raymonde Labreque: „Processing Prosodic and Musical Patterns: A Neuropsychological Investigation.“ In: *Brain and Language*, 61, pp. 123-144, 1998.
- [13] Patel, Aniruddh D., Iversen, John R., and Jason C. Rosenberg: „Comparing the rhythm and melody in speech and music: The case of British English and French.“ In: *The Journal of Acoustical Society of America*, 119 (5), pp. 3034-3047, 2006.
- [14] Ramus, Franck, Dupoux, Emmanuele, Zangl, Renate, and Jacques Mehler: „An empirical study of the perception of language rhythm.“, 2000.
- [15] Hannon, Erin E.: "Perceiving speech rhythm in music: Listeners classify instrumental songs according to language of origin" In: *Cognition*, 111, pp. 403-409, 2009.
- [16] Campione, Estelle, and Jean Véronis: "A Multilingual Prosodic Database." In: *ICSLP*, 98, pp. 3163-3166, 1998.
- [17] Chan, D., Fourcin, A., Gibbon, D., Grandström, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno,

- A., Mouropoulos, J., Senia, F., Transcoso, I., Velt, C. and Zeiliger, J.: "EUROM – A Spoken Language Resource for the EU." In: *Proceedings of Eurospeech '95*, Madrid, 1995.
- [18] Pysewicz, Andreas: *Validity and Reliability of Internet-based Auditory Perception Experiments*. MA-Arb. Berlin: Technische Universität, 2014.
- [19] Snodgrass, J. G., G. Levy-Berger, and M. Haydon: *Human Experimental Psychology*. New York: Oxford University Press, 1985.
- [20] Donaldson, W.: „Accuracy of d' and A' as estimates of sensitivity.“ In: *Bulletin of the Psychonomic Society*, 31, pp.271-274, 1993.
- [21] Nazzi, Thierry, Josiane Bertoncini, and Jacques Mehler: "Language discrimination by newborns: toward an understanding of the role of rhythm." In: *Journal of Experimental Psychology: Human perception and performance*, 24 (3), pp. 756-766, 1998.
- [22] Ramus, Franck: "Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues." In: *Annual Review of Language Acquisition*, 2, pp. 85-115, 2002.
- [23] Schafroth, Elmar: "Sprache und Musik. Zur Analyse gesungener Sprache anhand von Opernarien." (Erweiterte Fassung [S. 33] In: <http://docserv.uni-duesseldorf.de/servlets/DocumentServlet?id=23678>.



Audio Engineering Society Conference Paper

Presented at the Conference on
Semantic Audio
2017 June 22 – 24, Erlangen, Germany

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Speaker Identification for Swiss German with Spectral and Rhythm Features

Athanasios Lykartsis¹, Stefan Weinzierl¹, and Volker Dellwo²

¹Audio Communication Group, Technische Universität Berlin, Germany

²Phonetics Laboratory, Universität Zürich, Switzerland

Correspondence should be addressed to Athanasios Lykartsis (athanasios.lykartsis@tu-berlin.de)

ABSTRACT

We present results of speech rhythm analysis for automatic speaker identification. We expand previous experiments using similar methods for language identification. Features describing the rhythmic properties of salient changes in signal components are extracted and used in an speaker identification task to determine to which extent they are descriptive of speaker variability. We also test the performance of state-of-the-art but simple-to-extract frame-based features. The paper focus is the evaluation on one corpus (swiss german, TEVOID) using support vector machines. Results suggest that the general spectral features can provide very good performance on this dataset, whereas the rhythm features are not as successful in the task, indicating either the lack of suitability for this task or the dataset specificity.

1 Introduction

The efficient description of speech rhythm is a challenging task which has been solved with limited success so far. The reason for this is the difficulty to define, measure and quantize what exactly constitutes speech rhythm. However, many studies up to now have shown that the rhythmic characteristics or even the general temporal evolution of speech, together with other factors, play an important role in the perception of language, especially for tasks such as speaker identification (SID) and language identification (LID), or even speech intelligibility [1, 2, 3, 4, 5, 6]. Therefore, further research on the subject could serve determining the important constituent elements of speech rhythm which contribute to language and speaker variability; and the creation of better features for speech processing.

Concerning speech rhythm feature extraction, the most influential studies have been performed in linguistics and phonetics. The basic assumption of those approaches is that rhythm-related speech phenomena take place on the level of the duration of intervals, phonemes, syllables, words and phrases. Therefore, metrics such as ΔC , $\%V$, $nPVI$ and $VarcoC$ [7, 1, 2, 8, 3] have been developed to capture the variability in the duration of syllables or consonant-vowel cluster intervals. However, recent observations [9, 10, 11] also criticize that those metrics are not necessarily characteristic of (solely) language variability. One novel approach for speech rhythm description are the attempts to describe speech rhythm related periodicities inherent in the signal. Such approaches for rhythm-based LID have been introduced based on automatic segmentation and feature extraction [12, 13, 14, 15, 16], low-frequency

periodicity analysis [17, 18, 6] and lately with methods borrowed from the field of Music Information Retrieval (MIR), e.g. with the beat histogram [19]. When looking specifically at the task of speaker ID, such approaches have been applied only to a lesser extent. However, recent studies [4, 20, 17] on speaker idiosyncratic speech rhythm features point toward the need to experiment with novel rhythm description methods. Standard SID approaches using machine learning methods with the help of basic features [21] and i-vectors [22, 23, 24, 25, 26] have provided good performance results in speaker recognition. Especially the i-vector approach in combination with Deep Learning has shown very high performance [27, 28, 29, 30]. These methods, however, are computationally complex and expensive and require a large amount of data for the building of the Universal Background Model (UBM), as well as for the training of the Deep Neural Nets (DNNs). Furthermore, it is largely unclear which features function well and why, as well as how they relate to specific qualities of speech (e.g. rhythm), with rhythm related features almost totally absent. Finally, the methods are applied to datasets which are not widely accessible since they are very expensive to obtain or only available in a challenge context (e.g. the NIST datasets), making the reproducibility of results difficult.

In this paper, we have therefore applied a novel method to extract speech rhythm related features for SID using the data of the swiss language TEVOID corpus [17] in order to determine if the proposed rhythmic features can be as successful for SID as they have been for LID [19]. Those features were selected, since speech rhythm metrics have been shown to provide interesting results for speaker identification. It is therefore interesting to evaluate our approach to rhythm features on the same dataset in order to check for consistencies or differences and draw conclusions about the features. At the same time we will test standard features in audio content analysis [31] as well as from speech processing - Shifted Delta Cepstral Coefficients (SDCs) and Mel Frequency Cepstral Coefficients (MFCCs) - as a baseline. We chose this dataset since it was accessible and it has been analyzed using the speech rhythm metrics [17], to which we wanted to compare our approach.

The paper is structured as follows: The feature extraction method is shortly described. The steps of the experimental setup feature evaluation for the TEVOID corpus are presented and discussed. Finally, conclusions and perspectives for further research are given.

2 Methods

2.1 Feature Extraction

For the extraction of rhythmic features for the SID task, we utilize the method proposed in [19], where five different novelty functions, i.e. temporal trajectories of different signal properties or their derivatives [32], are calculated and used as the basis for the creation of beat histograms, similar to the periodicity representations in [33, 34, 35]. We extract five such novelty functions:

- **Spectral Flux (SF)**, following strong changes in (wideband) spectral properties.
- **Spectral Flatness (SFL)**, indicating whether the signal is strongly tonal or noisy.
- **Spectral Centroid (SCD)**, giving information about the spectral center of weight.
- **RMS Amplitude (RMS)**, the standard amplitude/level information of the signal.
- **Fundamental Frequency (F0)**, following the basic F0 information in the speech signal (extracted using the harmonic product sum method, see [31]).

The interested reader can refer to [31] for more information on the mathematical definition and the properties of those audio features. A beat histogram from the temporal trajectories of those features (in a given texture frame, i.e., a smaller window of the whole audio file) is extracted by computing the scaled autocorrelation function for frequencies from 0.5 to 10 Hz. From the beat histograms, the following statistical and other features (subfeatures) are extracted in turn (95 in total, resulting from 5 novelty functions and 19 subfeatures):

- **Distribution statistics:** Mean (ME), Standard Deviation (SD), Mean of the Derivative (MD), Standard Deviation of the Derivative (SDD), Skewness (SK), Kurtosis (KU), Entropy (EN), Beat Histogram Centroid (CD) and High Frequency Content (HFC).
- **Peak related:** Strength and Position of the First and Second Strongest Peak (P1, A1, P2, A2), Ratio (RA) of the Strength of the first Peak (A1) to that of the Second one (A2), Peak Centroid (P3), Sum (SU) and Sum of Beat Histogram Power (SP).

Almost the same parameterization as in [19] was used here; all files were resampled to 22050 Hz, and a window of 512 samples with an overlap of 75% was applied. A texture window of 4 seconds with a 50% overlap was used for creating several beat histograms, which were then averaged across the whole audio file. Other values for those parameters were considered, but those provided the best results. Apart from the rhythmic features, spectral ones were extracted by calculating the feature value over analysis frames of a Short-Time-Fourier-Transform (STFT) with the same parameterization as above for the whole audio file. We included the following 34 features (for more information, see [31]):

- **Spectral Shape and Change:** Spectral Flux (SF), Spectral Centroid (SCD), SDCs derived from the MFCCs (1 – 13, resulting in 13 SDCs in the 7 – 1 – 1 – 1 setting, see [36] for more details), the MFCCs themselves, Spectral Flatness (SFL) and Spectral Spread (SSP).
- **Tonal:** Spectral Tonal Power Ratio (STPR) and Zero Crossing Rate (ZCR).
- **Envelope:** Root Mean Square Amplitude (RMS) and Envelope Max (EMX).

2.2 Classification

In order to perform supervised classification we have used the Support Vector Machines (SVM) [37] algorithm in a MATLAB implementation with a Radial Basis Function (RBF) kernel for a multi-class setting. The hyperparameters for the RBF kernel (C , γ) were determined through a grid search procedure. For all experiments, a 10-fold cross-validation took place. This means that the dataset was randomly separated in 10 equally large subsets (folds), out of which 9 were used for training and 1 for testing (validation). This procedure was repeated 10 times (corresponding to the number of the folds) and the average accuracy over all trials was computed. This represents a common way to perform machine learning experiments (e.g. in the MIR community) and assures that no skewed results are produced because of a single random advantageous or disadvantageous partitioning of the dataset. When the dataset is small, this could lead to problems with insufficient training material, which is why we chose

a partitioning with relatively many folds (10). Z-score standardization was conducted prior to classification, separately for the training and test set. The *accuracy*, as the number of correct classifications for one class, to the number of overall classifications, was used to evaluate classification performance. We are primarily interested in this measure, as we are performing a 1-vs-1 multiclass supervised classification setting - that is, for each speaker pair, classifications are performed (in each fold), as we wish to know how well the algorithm can distinguish one speaker in comparison to another, and not to all others together (as in a 1-vs-all setting), since we can then interpret misclassifications in an easier way. The final result is calculated by summing the individual results for each class. This way we can also detect effects misclassified classes, which would point at speakers having similar properties (as measured through our features) or some speakers having not enough variance to stand out in comparison to any other class.

2.3 Datasets

For the speaker ID task, the TEVOID corpus was used [17]. It contains sixteen spontaneous utterances from sixteen male and female (50% for each category) Swiss German speakers (i.e. 256 utterances in total) transcribed and read by all speakers, resulting in 4096 sentences. The audio signal quality is high, and the corpus has already been analyzed [17] using many established speech rhythm metrics ($\%V$, $\Delta V(ln)$, $\Delta C(ln)$, $\Delta Peak(ln)$) and was found to contain considerable between-speaker variability, even when strong within-speaker variability was introduced. In this sense, it is expected that the speakers could be identified from a supervised learning algorithm successfully. It must be mentioned, however, that the database is relatively small, which could make the generation of reliable results difficult. Furthermore, the fact that the dataset deals with only variety of the german language (swiss german) could lead to results of the SID experiment might be specific for this language.

3 Results

Using the rhythm feature set (see the confusion matrix, Fig. 1), it was observed that all speakers are identified with an accuracy above chance level (Accuracy > Prior = 6.25%) while speakers S4, S7, S8, S10, S12 and S16 are identified with relatively low accuracy, below

20%. On the other side, two out of sixteen speakers are identified with relatively high absolute accuracy (S2 with 53.9% and S3 with 54.7%), three others with moderately good accuracy (S1, 36.3%, S9, 30.9% and S14, 31.6%) and for the rest of the speakers an accuracy of 20 to 30% is achieved. The average accuracy is 26.95%, which is more than four times greater than chance classification accuracy, but still in absolute values not entirely satisfactory. Using the spectral feature set (Fig. 2), the results were unambiguous: the overall performance was 87.6%, without much variation between speakers (around 10%). Speakers S3, S6, S10, S14, S15 are identified with an accuracy above 90%. This points towards the fact that simple, spectral features capture very speaker-specific characteristics such as voice timbre or F0. This confirms findings from other SID studies [22, 21, 23]. When combining both feature sets (Fig. 3), an 82.3% average accuracy is reached, which does not show much variation between speakers. This shows two interesting effects: First, accuracy actually decreases when using spectral features together with the rhythm related ones, hinting towards the fact that when using all the features with the same SVM classifier, the determination of a good class separation becomes more difficult. A similar effect was observed when using the linear SVM and the kNN classifiers, however with lower accuracy. Secondly, the variation pattern follows that of the spectral features, showing that they dominate in the task.

4 Discussion

The results presented in the previous section give a mixed picture. Using the rhythm features, it can be seen that the overall performance (as measured by accuracy) on the TEVOID corpus is relatively low (26.95%). This points towards the fact that the features do not necessarily capture speech rhythm in the same way as the rhythm metrics do, since when using the latter, it could be shown that between-speaker rhythmic variability in this corpus is robust and even with respect to certain kinds of within-speaker variability [17, 38]. However, the fact that recognition stays well above the prior in most cases is encouraging with respect to the features capturing some speaker related rhythmic variability. The spectral features have achieved a very high overall performance (87.6%) showing that SID with good results is possible even with the use of an uncomplicated, fast feature extraction scheme, opting for their use in future experiments and applications.

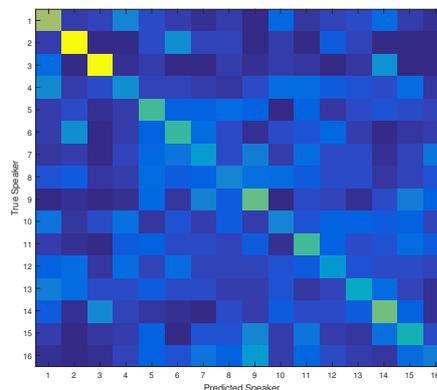


Fig. 1: Confusion Matrix for the TEVOID corpus, rhythm features (16 speakers).

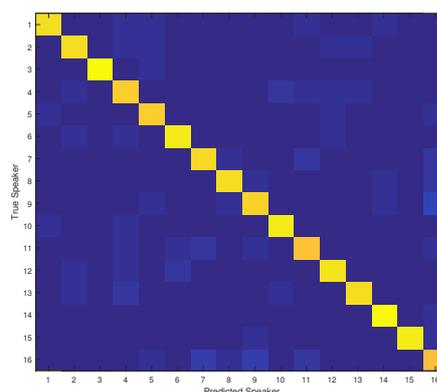


Fig. 2: Confusion Matrix for the TEVOID corpus, spectral features (16 speakers).

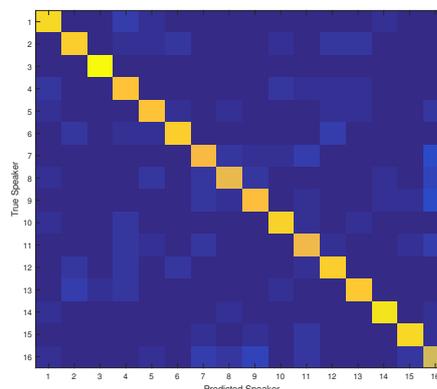


Fig. 3: Confusion Matrix for the TEVOID corpus, combined features (16 speakers).

The reasons for the unsatisfactory performance of the rhythm features might lie in the specific variety of swiss german in the corpus, which might be a special, difficult case to analyze in terms of rhythmic variability. Also interesting is the fact that specific users (two in particular) are identified with relatively high accuracy. This is a hint towards the assumption that our rhythm features capture very specific rhythmic patterns of certain individuals, which might have to do with the specific dialect of german, rate of speech or elicitation method (as the rhythm features did not perform well on spontaneous speech for LID either, see [19]) although those parameters have to be investigated further. A listening probe into the speaker characteristics of the best and worst cases did not reveal any rhythm-specific reasons for them performing better or worse, other than the fact that speakers S2 and S3 speak relatively slowly and somewhat more clearly. In this context, the investigation of perceptual similarities in speech rhythm between speakers through a listening experiment could also be helpful. To summarize, the fact that the results are significantly above the chance shows that rhythm features can indeed be helpful for SID but need to be further refined for use in such tasks. Possible problems could include the temporal resolution of the rhythm features (which could be adjusted to, e.g., fit the speaker rate) or the elicitation method. All of the above imply that SID (in contrast to LID) is much better served by just using spectral features, as they apparently capture a great part of speaker specificity. This might be a result of different speakers of the same language having very different voice timbre characteristics, which are readily captured through features such as the MFCCs, the SDCs and similar ones. In general, the high performance of the spectral features is similar to results shown elsewhere (e.g., the studies that use i-vector methods derived from MFCCs and SDCs [22, 23, 24, 25, 26, 27, 28, 29, 30]), which achieve error rates of 5% or lower on various datasets, ranking just a bit higher than our spectral features, but with a much more effortful analysis. On the other hand, speaker specific rhythm characteristics might either be absent (in general or for the dataset and language used here), very confounded with other sources of rhythm variability (such as elicitation method, emotional speech) or might just not be captured through our rhythm extraction method. Since using those rhythm features has shown good classification results both in MIR tasks [33, 35, 39, 40]) and in LID [19], we surmise that they are not as suitable for SID.

5 Summary

The analysis presented reveals tendencies concerning the application of multiple novelty beat histogram-based rhythm descriptors for SID. It has been shown that at least on one dataset of swiss german, the rhythmic features are not very helpful to achieve high accuracy in SID, although it has been shown that other rhythm metrics can capture the idiosyncrasy present in the corpus [20, 17, 41]. The reasons for that are not clear yet, but possible candidates are the specificity of the corpus language, the size of the dataset or that the features do not capture speech rhythm characteristics in a way that is speaker-specific. The latter might well be the case, as we were able to show in a previous study [19] that the same features are indeed descriptive of speech rhythm when it comes to the task of LID.

Another clue pointing to this direction is that the features achieved good accuracy for a few speakers, showing that they could partly capture characteristics of specific speakers, but not in every case. However, further tests with other datasets are necessary to confirm this tendency. From a theoretical perspective they are nevertheless very useful, as they give clues to the importance of speech rhythm for the corresponding task. The simple spectral features have shown very high performance with a low computational cost and should therefore be further applied.

Future work will include the following tasks: The use of larger datasets as the GLOBALPHONE [42] in order to be able to draw conclusions across languages and to test for rhythmic variability both between speakers and between languages at the same time. Further feature analysis is also scheduled, so as to investigate if the tendencies observed in the present study are robust across datasets and other settings (speaker, elicitation methods), as well as further investigating which aspect of the speech data (language, dataset size, feature parameterization etc) is the most important in generating better results. Specifically with respect to the speech tempo, an automatic tempo extraction scheme similar to the one used here, such as the tempogram [43], will be used in combination with manually obtained ground truth data in order to investigate the validity of the automatic tempo extraction procedure. Finally, further rhythm feature extraction algorithms, e.g. the modulation scale spectrum [44] or similarity detection schemes [45] will be adapted so as to be used for speech rhythm description.

References

- [1] Ramus, F., Nespors, M., and Mehler, J., “Correlates of linguistic rhythm in the speech signal,” *Cognition*, 73(3), pp. 265–292, 1999.
- [2] Grabe, E. and Low, E. L., “Durational variability in speech and the rhythm class hypothesis,” *Papers in laboratory phonology*, 7(515-546), 2002.
- [3] Dellwo, V., Fourcin, A., and Abberton, E., “Rhythmical classification of languages based on voice parameters,” in *ICPhS '07*, pp. 1129–1132, 2007.
- [4] Dellwo, V. and Koreman, J., “How speaker idiosyncratic is measurable speech rhythm,” in *Abstract presented at the annual IAFPA meeting*, 2008.
- [5] Arvaniti, A. and Ross, T., “Rhythm classes and speech perception,” *Understanding Prosody: The Role of Context, Function and Communication*, 13, p. 75, 2012.
- [6] Tilsen, S. and Arvaniti, A., “Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages,” *The Journal of the Acoustical Society of America*, 134(1), pp. 628–639, 2013.
- [7] Dauer, R. M., “Stress-timing and syllable-timing reanalyzed,” *Journal of phonetics*, 1983.
- [8] Dellwo, V., “Rhythm and speech rate: A variation coefficient for DeltaC,” *Language and language-processing*, pp. 231–241, 2006.
- [9] Arvaniti, A., “The usefulness of metrics in the quantification of speech rhythm,” *Journal of Phonetics*, 40(3), pp. 351–373, 2012.
- [10] Arvaniti, A. and Rodriguez, T., “The role of rhythm class, speaking rate, and F0 in language discrimination,” *Laboratory Phonology*, 4(1), pp. 7–38, 2013.
- [11] Turk, A. and Shattuck-Hufnagel, S., “What is speech rhythm? A commentary on Arvaniti and Rodriguez, Krivokapić, and Goswami and Leong,” *Laboratory Phonology*, 4(1), pp. 93–118, 2013.
- [12] Farinas, J., Pellegrino, F., Rouas, J.-L., and André-Obrecht, R., “Merging segmental and rhythmic features for automatic language identification,” in *Audio, Speech and Signal Processing, 2002. ICASSP 2002. International Conference on*, volume 1, pp. I–753, 2002.
- [13] Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R., “Modeling prosody for language identification on read and spontaneous speech,” in *Acoustics, Speech and Signal Processing, 2003. ICASSP 2003. IEEE International Conference on*, volume 6, pp. I–40, 2003.
- [14] Rouas, J.-L., Farinas, J., and Pellegrino, F., “Automatic modelling of rhythm and intonation for language identification,” in *International Conference on Phonetic Sciences*, pp. 567–570, 2003.
- [15] Rouas, J.-L., Farinas, J., Pellegrino, F., and André-Obrecht, R., “Rhythmic unit extraction and modelling for automatic language identification,” *Speech Communication*, 47(4), pp. 436–456, 2005.
- [16] Rouas, J.-L., “Automatic prosodic variations modeling for language and dialect discrimination,” *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(6), pp. 1904–1911, 2007.
- [17] Dellwo, V., Leemann, A., and Kolly, M.-J., “Speaker idiosyncratic rhythmic features in the speech signal,” in *INTERSPEECH*, 2012.
- [18] Tilsen, S. and Johnson, K., “Low-frequency Fourier analysis of speech rhythm,” *The Journal of the Acoustical Society of America*, 124(2), pp. EL34–EL39, 2008.
- [19] Lykartsis, A. and Weinzierl, S., “Using the beat histogram for speech rhythm description and language identification,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [20] Dellwo, V., Kolly, M.-J., and Leemann, A., “Speaker identification based on temporal information: a forensic phonetic study of speech rhythm and timing in the Zurich variety of Swiss German, International Association for Forensic Phonetics and Acoustics Conference,” *Santander, Spain*, 2012.

- [21] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., and Torres-Carrasquillo, P. A., "Support vector machines for speaker and language recognition," *Computer Speech & Language*, 20(2), pp. 210–229, 2006.
- [22] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Jones, D. A., and Leek, T. R., "Phonetic speaker recognition with support vector machines," in *Advances in neural information processing systems*, p. None, 2003.
- [23] Campbell, W. M., Sturim, D. E., and Reynolds, D. A., "Support vector machines using GMM supervectors for speaker verification," *Signal Processing Letters, IEEE*, 13(5), pp. 308–311, 2006.
- [24] Campbell, W. M., Campbell, J. P., Reynolds, D. A., Singer, E., and Torres-Carrasquillo, P. A., "Support vector machines for speaker and language recognition," *Computer Speech & Language*, 20(2), pp. 210–229, 2006.
- [25] Mandasari, M. I., McLaren, M., and van Leeuwen, D. A., "Evaluation of i-vector speaker recognition systems for forensic application." in *INTER-SPEECH*, pp. 21–24, Citeseer, 2011.
- [26] Matějka, P., Glembek, O., Castaldo, F., Alam, M. J., Plhot, O., Kenny, P., Burget, L., and Černocký, J., "Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification," in *Acoustics, Speech and Signal Processing, 2011. ICASSP 2011. IEEE International Conference on*, pp. 4828–4831, IEEE, 2011.
- [27] Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., and Vaquero, C., "Unsupervised domain adaptation for i-vector speaker recognition," in *Proceedings of Odyssey: The Speaker and Language Recognition Workshop*, 2014.
- [28] Garcia-Romero, D. and McCree, A., "Supervised domain adaptation for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 4047–4051, IEEE, 2014.
- [29] Greenberg, C. S., Bansé, D., Doddington, G. R., Garcia-Romero, D., Godfrey, J. J., Kinnunen, T., Martin, A. F., McCree, A., Przybocki, M., and Reynolds, D. A., "The NIST 2014 speaker recognition i-vector machine learning challenge," in *Odyssey: The Speaker and Language Recognition Workshop*, pp. 224–230, 2014.
- [30] Senior, A. and Lopez-Moreno, I., "Improving DNN speaker independence with i-vector inputs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pp. 225–229, IEEE, 2014.
- [31] Lerch, A., *An introduction to audio content analysis: Applications in signal processing and music informatics*, Wiley & Sons, 2012.
- [32] Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., and Sandler, M. B., "A tutorial on onset detection in music signals," *IEEE Transactions on Speech and Audio Processing*, 13(5), pp. 1035–1047, 2005.
- [33] Tzanetakis, G. and Cook, P., "Musical genre classification of audio signals," *IEEE transactions on Speech and Audio Processing*, 10(5), pp. 293–302, 2002.
- [34] Burred, J. J. and Lerch, A., "A hierarchical approach to automatic musical genre classification," in *DAFx*, 2003.
- [35] Gouyon, F., Dixon, S., Pampalk, E., and Widmer, G., "Evaluating rhythmic descriptors for musical genre classification," in *AES '04*, pp. 196–204, 2004.
- [36] Campbell, W. M., Singer, E., Torres-Carrasquillo, P. A., and Reynolds, D. A., "Language recognition with support vector machines," in *ODYSSEY04 – The Speaker and Language Recognition Workshop*, 2004.
- [37] Vapnik, V., *The nature of statistical learning theory*, Springer, 2000.
- [38] Dellwo, V., Leemann, A., and Kolly, M.-J., "Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors," *The Journal of the Acoustical Society of America*, 137(3), pp. 1513–1528, 2015.
- [39] Lykartsis, A., Wu, C.-W., and Lerch, A., "Beat histogram features from NMF-based novelty functions for music classification," in *ISMIR*, 2015.

-
- [40] Lykartsis, A. and Lerch, A., “Beat histogram features for rhythm-based musical genre classification using multiple novelty functions,” in *DAFx*, 2015.
- [41] Dellwo, V., Leemann, A., and Kolly, M.-J., “The recognition of read and spontaneous speech in local vernacular: The case of Zurich German,” *Journal of Phonetics*, 48, pp. 13–28, 2015.
- [42] Schultz, T., “Globalphone: a multilingual speech and text database developed at karlsruhe university,” in *INTERSPEECH*, 2002.
- [43] Grosche, P. and Müller, M., “Extracting predominant local pulse information from music recordings,” *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(6), pp. 1688–1701, 2011.
- [44] Marchand, U. and Peeters, G., “The Modulation Scale Spectrum and its Application to Rhythm-Content Description.” in *DAFx*, pp. 167–172, 2014.
- [45] Pohle, T., Schnitzer, D., Schedl, M., Knees, P., and Widmer, G., “On Rhythm and General Music Similarity.” in *ISMIR*, pp. 525–530, 2009.

PREDICTION OF DIALOGUE SUCCESS WITH SPECTRAL AND RHYTHM ACOUSTIC FEATURES USING DNNS AND SVMs

Athanasios Lykartsis

Technische Universität Berlin
Audio Communication Group
Berlin, Germany

*M. Kotti, A. Papangelis, Y. Stylianou**

Toshiba LTD
Cambridge Research Laboratory
Cambridge, UK

ABSTRACT

In this paper we investigate the novel use of exclusively audio to predict whether a spoken dialogue will be successful or not, both in a subjective and in an objective manner. To achieve that, multiple spectral and rhythmic features are inputted to support vector machines and deep neural networks. We report results on data from 3267 spoken dialogues, using both the full user response as well as parts of it. Experiments show an average accuracy of 74% can be achieved using just 5 acoustic features, when analysing merely 1 user turn, which allows both a real-time but also a fairly accurate prediction of a dialogue successfulness only after one short interaction unit. From the features tested, those related to speech rate, signal energy and cepstrum are amongst the most informative. Results presented here outperform the state of the art in spoken dialogue success prediction through solely acoustic features.

Index Terms dialogue success, acoustic features, deep neural networks, support vector machines

1. INTRODUCTION

In the last years, there has been much research into Statistical Spoken Dialogue Systems (SDS), as they increasingly find more and more applications such as assisted living [1], phone banking [2], intelligent virtual agents [3] and health care [4], to name just a few. Although many of the aforementioned systems for these applications use speech to communicate with the user, the audio channel is typically not directly used to enhance the user experience, for example by predicting dialogue success, as is the case in this paper. However, this acoustic information can be very helpful, especially at the early dialogue stages (e.g., the first few turns), as it can allow the system to engage in a different strategy, such as not repeating the same question if the dialogue is predicted unsuccessful, adapting the way the answer is presented to the user, adjusting the synthesised speech parameters, or even forwarding the user to a human agent. An automatic way to measure task success

would also be useful for evaluating conversations among humans, e.g., for evaluating agents in a call center as the authors of [5] underline.

In the dialogue success prediction area, various features such as linguistic and belief state features [6], automatic speech recognition features [7], and acoustic features [8] can be used to obtain some information concerning the current state of the interaction. In human-computer dialogues, predicting the task success after just a first few turns of the conversation could avoid disappointment if the conversation is predicted unsuccessful. However, in the dialogue successfulness prediction area, most of the efforts use exclusively the linguistic channel [5] or the dialogue logs [9], disregarding the audio part altogether. In the work done previously by various authors [5, 7, 8, 10], it has been shown that employing all of these feature categories, separately or together, can provide satisfactory classification accuracy when analysing a dataset of interactions, where users communicate with such an automatic system. In a limited number of works the speech channel is used to measure user satisfaction. However, most efforts done so far take into account the systems voice, instead of the users. For example, in [11] and in [12] it was found that automatic speech recognition (ASR) quality is a key component of user satisfaction.

However, there are certain possible shortcomings when ignoring the audio channel. For example, using belief state features to predict dialogue success implies that we have a robust mechanism to track the users' goal as it evolves over time, for each domain, i.e. subject, the system operates in. However, the user goal is most of the time hidden. Therefore, our aim is to infer if the goal is met based only on readily available attributes which can be extracted from speech. We observe that the exclusive use of the speech signal per se has numerous advantages: Firstly, the speech signal is often already analysed for ASR purposes (in the sense of feature extraction). Hence, by utilizing this information for predicting dialogue successfulness, we are receiving an extra gain at a very low (or even none, depending on the extracted features) computational cost. Secondly, speech features are domain-independent, in the sense that their values are not affected by

*Y. Stylianou is also with the Department of Computer Science, University of Crete, Heraklion, Greece.

the subject the conversation is about. If the user is dissatisfied with the system, the quality of his or her voice is affected in the same way, whether the subject of the conversation is booking a hotel room, purchasing a product or receiving information. Thirdly, speech signal features are not affected by the ontology size of an SDS, in the sense that the number of slots and values of an ontology is not connected with the user's cognitive state while interacting with the system. For all the reasons mentioned above, we choose to take a closer look at the gains that acoustic features alone, extracted solely from the human user voice, can provide when predicting dialogue success. Such features have been used with great success in many other audio classification tasks [13].

In the next section, related work is given. Contributions are presented in Section 3 and the method is detailed in Section 4. We continue with the presentation of the results in Section 5, their discussion in Section 6 and sum up with conclusions and directions for future work.

2. RELATED WORK

The number of research effort that take into account acoustic features to improve the dialogue experience is limited, since, as already mentioned, most works disregard the audio channel. For example, in [14] features derived from the ASR output and the preceding system action are used, disregarding the audio channel, in order to predict the user satisfaction. A limited number of systems, like [15] have used multiple features and sparingly some acoustic descriptors for predicting problematic dialogues. More current attempts have been made, e.g., in [16], [17] and [18] to detect irregularities in dialogues using acoustic features. In other works, [19, 20] prosodic features were also used to detect quotations and dialogue acts, respectively. Prediction of dialogue success has also been carried out, using energy, cepstrum, and pitch-related features [8, 10]. Also, [21] have implemented a multimodal system and extracted a very broad range of acoustic features to predict the user's cognitive state. However, predicting the cognitive state is out of the scope of this paper since the target values provided with the presented database refer to success.

To the best of our knowledge, there are no works using a wider range of acoustic features to predict the success of the dialogue. To that end, we aim to use simple, thus easy to compute at low cost and without much data, acoustic features, suitable for a real-time SDS. Such features have shown promising results for alternative applications [22, 23, 24, 13].

3. CONTRIBUTIONS

Contributions of this paper include:

- the extraction of a broader pool of acoustic features, consisting a substantially larger dataset than that already found in related literature [8, 10], and the iden-

- tification of the acoustic features most informative and useful for the task,

- determining the minimum necessary number of dialogue turns (and, accordingly, length) to achieve a satisfactory success prediction accuracy,

- whether it is more feasible to predict user satisfaction (i.e., if the user reported that they were content with course and outcome of the dialogue and interaction) or the dialogue's objective success (if the goal set for the interaction was actually met).

4. METHOD

The method consists of the pipeline shown in Fig. 1. More information about the individual steps will be given in the following subsections. All experiments reported were performed on a Linux CentOS system using python 2.7.5.

4.1. Audio Data

The dataset consists of a series of human-machine interactions. Specifically, the interactions are calls between Amazon mechanical turk workers concerned with finding appropriate laptops using an SDS. Each person was given a set of preferences prior to the session and was asked to retrieve a laptop that fulfils those preferences. Regarding the laptop ontology, it comprises 9 slots, i.e., database attributes, 6 of which are informable, i.e., information the user provides with a mean of 3 values per slot, and 3 of which are requestable, i.e., information the system requests. The person then interacted with the dialogue system until the item was retrieved or the person hung up. We used a dataset of a total length of approximately one hour, comprising of 3267 dialogues. The audio files comprise light background noise and some artifacts, while being of telephone speech quality (with a sampling rate of 16000 Hz). The files are segmented per dialogue turn, i.e., every part of the utterance, in which the user interacts with the system, is available, as well as the full-length interaction. Therefore, using only some initial dialogue turns for classification would potentially speed up the classification procedure and lead an earlier recognition of dialogue success. We experimented with the whole dialogue files, as well as with their parts, i.e., turns, up to a maximum of 5 dialogue turns. Transcripts and details about the nature of the dataset can be found in [10]. An example of a dialogue is given in Table 1.

The labels for classification comprise two groups: *subjective*, denoting the success of the dialogue as expressed in the feedback of the user to the question "Did you find all the information you were looking for?"; and *objective* ones, resulting as the outcome of the system response offering the user a laptop adhering to the requirements provided by the user in the beginning. For example, if the goal is to find a low-price, light-weight laptop, if the final item offered to the

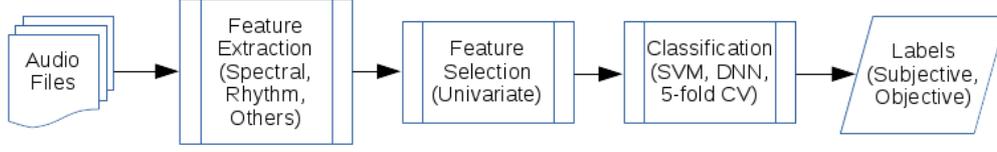


Fig. 1. acoustic feature extraction, selection and classification pipeline.

user did in fact fulfil those two requirements, the dialogue was labelled as objectively successful. The two label sets were largely non-identical, i.e., the subjective satisfaction was not a consequence of objective success, or vice-versa.

Since the initial 3267 dialogues were highly unbalanced both for the case of objective (with a prior of 64% for the unsuccessful case) and for that of subjective (with a prior of 73% for the successful case) labels, we create balanced versions to avoid spurious high accuracy by randomly under-sampling the larger class to the size of the smaller. This results to a sample size of 1786 for the subjective label data and 2404 for the objective labelled ones.

Criteria: You want to find a laptop in the large hard drive range that is in the mid weight range and has a battery in the exceptional rating category. Make sure you get the name and family the laptop belongs to.

System: Hello, I am here to help you find a suitable laptop. Now please tell me your requirements.

User: With large hard drive, ehh, mid weight, and exceptional battery.

System: Would you like your battery rating to be standard, good, or exceptional?

User: Exceptional!

System: Regarding weight, would you like something mid weight or heavy?

User: Mid.

System: The tecra-z50-a-106 is in the mid weight range, is in the moderate price range and has an exceptional battery rating.

User: What's family?

System: This laptop is called tecra-z50-a-106.

User: Thank you, good bye.

Table 1. The transcript of a dialogue between an Amazon mechanical turk worker and the statistical SDS for the case of a dialogue that was objectively unsuccessful, but subjectively successful. User is still satisfied, although family info was not provided by the system, hence not meeting the given requirements.

4.2. Feature Extraction

Since we wanted to primarily investigate the effectiveness of low-level acoustic information (which is computationally

easy and cheap to obtain, this being an advantage for a possible real-time application), we extracted three categories of features, using the librosa library [25].

The first category comprises features providing spectral information. In all cases, X refers to the signal magnitude spectrum (as a result of a DFT), K refers to the total number of frequency bins, for which k is a running index, whereas N refers to the total number of time frames, for which n is a running index. Those are:

The three first Mel-frequency cepstral coefficients (MFCCs) (excluding the zeroth coefficient), as a compact spectral representation (here, X' , K' and k' refer to the mel-warped magnitude spectrum):

$$MFCC^{1-3}(n) = \sum_{k'=1}^{K'} \log(|X'(k', n)|) \cdot \cos(j(k - 0.5) \frac{\pi}{K'}) \quad (1)$$

Spectral flux, indicating spectral shape change:

$$SF(n) = \frac{\sqrt{\sum_{k=0}^{K/2-1} (|X(k, n)| - |X(k, n-1)|)^2}}{K/2} \quad (2)$$

Spectral flatness, indicating noisiness or tonalness:

$$SFL(n) = \frac{\sqrt{\prod_{k=0}^{K/2-1} |X(k, n)|}}{\frac{2}{K} \sum_{k=0}^{K/2-1} |X(k, n)|} \quad (3)$$

Spectral contrast, indicating difference between high and low frequency content, (refer to [26] for more information on this feature's extraction).

Spectral centroid, the spectral center of weight,

$$SC(n) = \frac{\sum_{k=0}^{K/2-1} k \cdot |X(k, n)|^2}{\sum_{k=0}^{K/2-1} |X(k, n)|^2} \quad (4)$$

Spectral rolloff, indicating the limit frequency, from which on only very little energy is existent in the spectrum (κ is here equal to 0.85, i the frequency index of a specific bin):

$$SR(n) = i \quad \text{with} \quad \sum_{k=0}^i |X(k, n)| = \kappa \cdot \sum_{k=0}^{K/2-1} |X(k, n)| \quad (5)$$

Spectral bandwidth, describing the signal effective bandwidth.

$$SB(n) = \sum_{k=0}^{K/2-1} |X(k, n)| * (i - SC) \quad (6)$$

Also, temporal features are computed, such as:

Root mean square (RMS) energy value, an indicator for the energy of the signal (the equation here refers to one frame of samples; i_s is the sample start index, i_e end index inside the block, n refers to the frame index, K is the sample amount in the frame):

$$RMS(n) = \sqrt{\frac{1}{K} \sum_{i=i_s(n)}^{i=i_e(n)} x(i)^2} \quad (7)$$

Zero crossing rate, showing tonalness or noisiness (the indexes are as for the previous temporal feature, the RMS, and f_s refers to the sample frequency):

$$ZCR(n) = \frac{2 \cdot (i_e(n) - i_s(n))}{f_s \cdot \sum_{i=i_s(n)}^{i=i_e(n)} |sign[x(i)] - sign[x(i-1)]|} \quad (8)$$

Pitch, i.e., for voice, the fundamental frequency, as estimated in librosa using parabolic interpolation [27].

For more information on all those features, the interested reader can refer to [28]. Using those features has the goal of capturing energy, tonal and spectral information providing, e.g., clues to the speaker's voice characteristics or their tone of voice (e.g. loud vs. quiet). Those might change during a call or have different values depending on the speaker's cognitive state, which we assume to be related to dialogue success.

All of the above mentioned features are based on either a Short-Time Fourier transform (STFT) or on a frame-based processing with the same parameters. For this we use a *hann* window with length of 32ms, and an overlap of 50%. These values resulted after a parameter grid search, but are also characteristic for similar speech processing tasks, as they capture instantaneous values of the signal in a time frame, where the signal can be deemed to be stationary. All other parameters of the features (wherever that applies) are the default ones provided by the library. The feature values over the whole audio track are then aggregated using their mean and standard deviation values, as well as their skewness and kurtosis (4 subfeatures in total).

Furthermore, we use beat histogram features as in [22], based on different novelty functions. This procedure has the following steps:

Calculate the temporal trajectory of 7 of the instantaneous, frame-based features mentioned (flux, flatness, centroid, rolloff, bandwidth, RMS and pitch);

Compute the first difference function over time for them, stressing changes in the respective features;

Create a beat histogram by using an autocorrelation function on each of the novelty functions, and retaining periodicities in the area between 0.5 and 10 Hz, since that is the approximate expected frequency area for speech rates and rhythms [29].

Equation (4.2) shows the creation of a beat histogram BH_{IF} for an instantaneous feature IF with a novelty function $NF = diff(IF)$, where σ^2 is the sample variance, μ the sample mean, n the index for the total amount of instantaneous frame values for the feature N , and k a shifting index:

$$BH_{IF} = \frac{1}{\sigma^2 N} \sum_{n=1}^{N-k} (NF_n - \mu)(NF_{n+k} - \mu) \quad (9)$$

At the end, from the resulting beat histograms (for each novelty function), we extract the aggregate values and the rate, i.e. the most prominent periodicity value, resulting in 5 subfeatures per novelty function. For brevity reasons, we do not include any figures here concerning those novelty functions and their beat histograms, but the interested reader may refer to [22] for more information. With those features we intend to capture information relating to the speaker's rhythmic voice characteristics, for example the regularity of speech or the speech rate. These might also be influenced through the satisfaction of the user with the SDS, and they are expected to be useful.

The total feature count is 84, comprising $3 * 4 = 12$ MFCCs, $4 * 6 = 24$ spectral, $3 * 4 = 12$ temporal/energy features and $5 * 7 + 1 = 36$ rhythm features (one of them being the tempo value extracted through the tempogram toolbox with the same STFT parameters as before, s. [30]). It is reminded that we consider 4 subfeatures per instantaneous feature and 5 subfeatures for each novelty function based beat histogram, that is why the number of features is multiplied by 4, respectively 5 in the above calculation.

4.3. Classification and Feature Selection

Classification is performed using two classifiers, which are the ones who led to the optimal results: a support vector machine with radial basis function kernel (RBF-SVM) classifier through the sklearn module [31], as well as a deep neural network (DNN) of 2 hidden layers in tensorflow [32].

The mode of classification is binary, i.e. successful/non-successful, for both the objective and the subjective success prediction case. In order to test the stability of the proposed approach, 5-fold cross validation is used and the average accuracy and standard deviation of it across the 5 folds is reported in the next section. We used a development set design: The samples were separated in a training, development and

test set in a ratio of 0.8 – 0.1 – 0.1. The test set samples were kept absolutely separate from the rest, the development set samples were used for the classifier parameter tuning. Prior to classification, the features are z-score standardised, separately for each set.

After performing initial experiments with the 84 features mentioned in 4.2, we decided to employ methods of dimensionality reduction to check if we can achieve the same accuracy, but with a reduced feature set, thus rendering our approach computationally less intensive. To that end, a filter feature selection method [33] is implemented through the sklearn package, namely univariate feature selection, so as to capture both linear and non-linear dependencies between the features and the labels. We experiment with retaining the best n features by gradually reducing their number, resulting to a minimal set of just 5 features.

Concerning the classifier setting, we experimented with several alternatives to reach the best accuracy. For the RBF-SVM, the hyperparameters are determined through grid search in the development set, resulting in $C = 2$ and a γ of 0.2 for the 5 features and 0.012 for the 84 features. To find the optimal architecture for the DNN, several alternatives have been tested. For 2 hidden layers we tried having 68 neurons for the first hidden layer and 24 for the second; or 50 neurons for the first layer and 100 for the second; or finally 24 neurons for the first layer and then 68 for the second. Out all those, the best performing architecture for both cases was found to be the neural network where the first hidden layer had 64 neurons and the second one 28 neurons. Additional architectures having more hidden layers, such as 5, 5, 5, 5 or 89, 28, 11, 5 or finally 68, 24, 12 were suffering from data starvation, thus leading to poor results. Hence, we resorted to the 64, 28 neurons architecture. The Adam algorithm [34] was used as optimiser. A dropout value of 0.9 and a learning rate of 0.01 was used, whereas softsign is the activation function.

5. EXPERIMENTAL RESULTS

Figures 2 and 3 show the mean accuracy and standard deviation over the 5 folds of the cross validation achieved for the SVMs and the DNNs using the minimal set of 5 features, for an increasing amount of turns (one to five, where the number of turns is cumulative, i.e., the 2 turn case is turn 1 and turn 2 together; and then altogether), for the subjective- and the objective-label data. These results refer to the performance on the *test* set. Finally, Table 3 provides an overview of the results mentioned before, but also includes the accuracy results on the development set. Concerning the performed feature selection, the most informative features are the standard deviation of the three MFCCs and the tempo and mean of the RMS-based beat histograms.

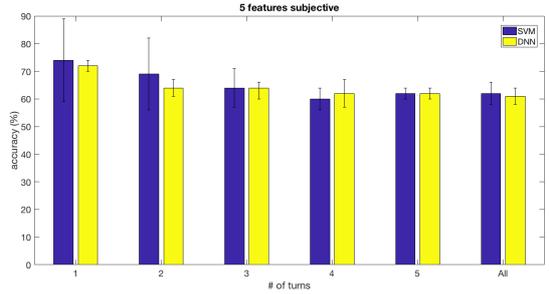


Fig. 2. SVM and DNN mean and standard deviation of accuracy for different turns, subjective labels, test set.

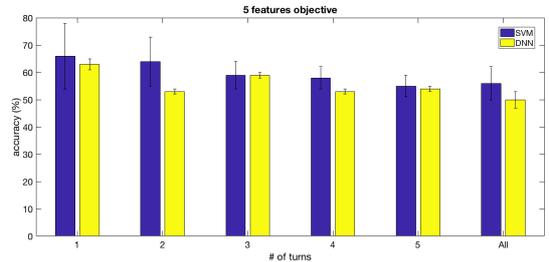


Fig. 3. SVM and DNN mean and standard deviation of accuracy for different turns, objective labels, test set.

6. DISCUSSION

The best performance achieved for the both the SVM and the DNN case is that of considering only one turn for the subjective labels. In that case, the reported SVM accuracy is $74\% \pm 15\%$, whereas the corresponding accuracy for the DNN is $72\% \pm 2\%$. For the objective case, the best SVM accuracy is $66\% \pm 12\%$ when only one turn is considered, for the DNN the best performance is achieved when also 1 turn is considered and is equal to $63\% \pm 2\%$. Hence, it can be deduced that the objective success of the dialogue is more difficult to predict. This could be partially due to the acoustic features capturing more information about the speaker’s cognitive state or voice characteristics, which could give indications as to whether they are satisfied or not. Additionally, the DNN architecture applied for the subjective results seems not to provide optimal results for the objective case. Also, the SVM exhibits much larger variance for the accuracy between the folds. Also, the results on the development set (seen in table 3) are in almost any case worst than on the test set.

Regarding the turns used, it can be observed that there is not a very pronounced effect of the audio information size used, but it can be said that accuracy decreases with the rising amount of turns generally. This may be partially attributed to the telephone speech quality of the data introducing more redundant information the longer the processed become, which

Setting/Number of Turns	1	2	3	4	5	All
SVM development set, subjective labels	68 (12)	64 (11)	60 (8)	59 (5)	58 (5)	59 (5)
SVM development set, objective labels	66 (10)	62 (8)	58 (7)	56 (6)	55 (4)	55 (4)
DNN development set, subjective labels	67 (1)	61 (1)	60 (3)	59 (4)	60 (2)	64 (1)
DNN development set, objective labels	60 (1)	61 (1)	50 (1)	50 (1)	50 (1)	50 (1)

Table 2. Classification results, development set, average accuracy and standard deviation in percent (rounded). The number of turns reported is cumulative.

Setting/Number of Turns	1	2	3	4	5	All
SVM test set, subjective labels	74 (15)	69 (13)	64 (7)	60 (4)	62 (2)	62 (4)
SVM test set, objective labels	66 (12)	64 (9)	59 (5)	58 (4)	55 (4)	56 (6)
DNN test set, subjective labels	72 (2)	64 (3)	64 (3)	62 (5)	62 (2)	61 (3)
DNN test set, objective labels	63 (2)	53 (1)	59 (1)	53 (1)	54 (1)	50 (4)

Table 3. Classification results, test set, average accuracy and standard deviation in percent (rounded). The number of turns reported is cumulative.

might be reflected into their aggregate values; it is, however, a very positive result, since one quick interaction part can be almost as meaningful as a longer file, essentially reducing substantially the amount of data necessary to make good predictions. This is most important for on-line, real-time systems since meaningful predictions can be reached using just some seconds of the captured speech. This practically means that shorter interaction can capture the cognitive state of the speaker more effectively. A second possible reason could be that over the whole phone call, the cognitive state of the user may vary significantly. For example, the call may start with a neutral user that gets satisfied for the first few turns and then dissatisfied towards the end of the call. User cognitive state evolves over time during a phone call and this is what is expected. For example, in [35] the authors found that emotions evolve as time progresses, whereas the authors of [36] that analyse call centre dialogues, try to predict a category per utterance, rather than over the whole dialogue.

Regarding feature selection, experiments also took place when all the 84 features are used. In that case, the results are deteriorated for most of the cases. For example, for the already reported case of subjective results when 1 turn is considered, the SVM accuracy is $65\% \pm 10\%$, whereas the DNN accuracy is $72\% \pm 2\%$. Moreover, the feature selection results corroborate what has been found in previous studies [22], namely that basic statistics of spectral features and the tempo are informative descriptors of the speaker characteristics. For example, speech rate as predicted through the RMS tempo, seems to provide important information, as it can probably capture partially if the speaker becomes agitated or not [37].

In total, it could be argued that the findings can be very helpful in providing assistance to an SDS in a fully automatic way: Solely by using a pre-trained model on a medium-sized amount of data, a real-time analysis of the audio signal (using fast algorithms and few features) could allow an estimate of the dialogue success with high probability (74%) only after

one short interaction unit. Those results, although not directly comparable due to the different datasets and/or features used in other studies [8], [10], [7], provide a higher improvement compared to the baseline, given the dataset priors. Specifically, [8] reported results 6% better than the prior for the subjective labels as well as 6.5% for the objective labels. For [10] only the objective labels are provided, with a performance 2% higher than the prior, while the results from [7] are not comparable, as no prior is given and the evaluation measure is F-Score. In our case, we achieve a 24% increase (compared to the 50% prior) in accuracy for the subjective labels, and 16% for the objective ones.

7. CONCLUSIONS AND FUTURE WORK

In this paper various acoustic features have been extracted, some of them used for the first time in this context, and then selected just a small subset of 5 of them for the prediction of dialogue success. It was possible to show that a satisfactory classification accuracy can be achieved when using even one dialogue turn and that the subjective success is more accurately predicted. The features related to RMS and MFCCs are the most informative. Our results are slightly better than the state-of-the-art for dialogue success prediction based on audio features only, but there is room for further improvement.

In the future, we aim to consider more data and extract further acoustic features in real-time, using the essentialia library [38], as well as state-of-the-art methods such as i-vectors which might be useful in identifying speaker characteristics even for short utterances and telephone quality speech [39, 40]. Another possibility is to use PCA-reduced (mel-)spectrograms as input to a convolutional neural network (CNN), as was similarly done in [41] for onset detection. Furthermore, information pertinent to the emotional state of the user [42] could provide cues to user satisfaction.

8. REFERENCES

- [1] François Portet, Michel Vacher, Caroline Golanski, Camille Roux, and Brigitte Meillon, “Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects,” *Personal and Ubiquitous Computing*, vol. 17, no. 1, pp. 127–144, 2013.
- [2] Bayan AbuShawar and Eric Atwell, “Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems,” *International Journal of Speech Technology*, vol. 19, no. 2, pp. 373–383, 2016.
- [3] Yoichi Matsuyama, Arjun Bhardwaj, Ran Zhao, Oscar Romeo, Sushma Akoju, and Justine Cassell, “Socially-aware animated intelligent personal assistant agent,” in *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2016, pp. 224–227.
- [4] Mandy Korpusik and James Glass, “Spoken language understanding for a nutrition dialogue system,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 7, pp. 1450–1461, 2017.
- [5] David Reitter and Johanna D Moore, “Predicting success in dialogue,” 2007.
- [6] Stefan Steidl, Christian Hacker, Christine Ruff, Anton Batliner, Elmar Nöth, and Jürgen Haas, “Looking at the last two turns, i’d say this dialogue is doomed—measuring dialogue success,” in *International Conference on Text, Speech and Dialogue*, 2004, pp. 629–636.
- [7] David Vandyke, Pei-Hao Su, Milica Gasic, Nikola Mrkšić, Tsung-Hsien Wen, and Steve Young, “Multi-domain dialogue success classifiers for policy training,” in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 763–770.
- [8] Alexandros Papangelis, Margarita Kotti, and Yannis Stylianou, “Predicting dialogue success, naturalness, and length with acoustic features,” in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, 2017, pp. 5010–5014.
- [9] Sebastian Möller, Klaus-Peter Engelbrecht, and Robert Schleicher, “Predicting the quality and usability of spoken dialogue services,” *Speech Communication*, vol. 50, no. 8-9, pp. 730–744, 2008.
- [10] Margarita Kotti, Alexandros Papangelis, and Yannis Stylianou, “Will this dialogue be unsuccessful? prediction using audio features,” in *ICTIR 2017 Workshop on Search-Oriented Conversational AI*, 2017.
- [11] Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan, “Automatic online evaluation of intelligent assistants,” in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 506–516.
- [12] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Aidan C. Crook, Imed Zitouni, and Tasos Anastasakos, “Understanding user satisfaction with intelligent assistants,” in *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval*, New York, NY, USA, 2016, pp. 121–130.
- [13] Francesc Alías, Joan Claudi Socoró, and Xavier Sevilano, “A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds,” *Applied Sciences*, vol. 6, no. 5, pp. 143, 2016.
- [14] Stefan Ultes, Paweł Budzianowski, Inigo Casanueva, Nikola Mrkšić, Lina Rojas-Barahona, Pei-Hao Su, Tsung-Hsien Wen, Milica Gašić, and Steve Young, “Domain-independent user satisfaction reward estimation for dialogue policy learning,” in *Eighteenth Annual Conference of the International Speech Communication Association*, 2017, pp. 1721–1725.
- [15] Marilyn Walker, Irene Langkilde, Jerry Wright, Allen Gorin, and Diane Litman, “Learning to predict problematic situations in a spoken dialogue system: experiments with how may i help you?,” in *1st North American chapter of the Association for Computational Linguistics conference*, 2000, pp. 210–217.
- [16] Joseph Tepperman, David Traum, and Shrikanth Narayanan, “Yeah right: Sarcasm recognition for spoken dialogue systems,” in *Ninth International Conference on Spoken Language Processing*, 2006.
- [17] Narichika Nomoto, Hirokazu Masataki, Osamu Yoshioka, and Satoshi Takahashi, “Detection of anger emotion in dialog speech using prosody feature and temporal relation of utterances,” in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.
- [18] José Lopes, Giampiero Salvi, Gabriel Skantze, Alberto Abad, Joakim Gustafson, Fernando Batista, Raveesh Meena, and Isabel Trancoso, “Detecting repetitions in spoken dialogue systems using phonetic distances,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [19] Alessandra Cervone, Catherine Lai, Silvia Pareti, and Peter Bell, “Towards automatic detection of reported speech in dialogue using prosodic cues,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

- [20] Raul Fernandez and Rosalind W Picard, "Dialog act classification from prosodic features using support vector machines," in *Speech Prosody 2002, International Conference*, 2002.
- [21] Leimin Tian, Johanna Moore, and Catherine Lai, "Recognizing emotions in spoken dialogue with hierarchically fused acoustic and lexical features," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*, 2016, pp. 565–572.
- [22] Athanasios Lykartsis and Stefan Weinzierl, "Using the beat histogram for speech rhythm description and language identification," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [23] Mayank Bhargava and Tim Polzehl, "Improving automatic emotion recognition from speech using rhythm and temporal feature," *arXiv preprint arXiv:1303.1761*, 2013.
- [24] Athanasios Lykartsis, Stefan Weinzierl, and Volker Dellwo, "Speaker identification for swiss german with spectral and rhythm features," in *Audio Engineering Society Conference: 2017 AES International Conference on Semantic Audio*, 2017.
- [25] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, "librosa: Audio and music signal analysis in python," in *14th python in science conference*, 2015, pp. 18–25.
- [26] Dan-Ning Jiang, Lie Lu, Hong-Jiang Zhang, Jian-Hua Tao, and Lian-Hong Cai, "Music type classification by spectral contrast feature," in *Multimedia and Expo, 2002. ICME'02. Proceedings. 2002 IEEE International Conference on*. IEEE, 2002, vol. 1, pp. 113–116.
- [27] Julius O Smith et al., *Spectral audio signal processing*, vol. 1334027739, W3K, 2011.
- [28] Alexander Lerch, *An introduction to audio content analysis: Applications in signal processing and music informatics*, John Wiley & Sons, 2012.
- [29] Aniruddh D Patel, *Music, language, and the brain*, Oxford university press, 2010.
- [30] Peter Grosche and Meinard Müller, "Tempogram toolbox: Matlab implementations for tempo and pulse analysis of music recordings," in *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR), Miami, FL, USA*, 2011, pp. 24–28.
- [31] Fabian Pedregosa et al., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
- [32] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al., "Tensorflow: A system for large-scale machine learning," in *OSDI*, 2016, vol. 16, pp. 265–283.
- [33] Isabelle Guyon and André Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.
- [34] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization. arxiv.org," 2014.
- [35] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey, "Toward machine emotional intelligence: Analysis of affective physiological state," *IEEE transactions on pattern analysis and machine intelligence*, vol. 23, no. 10, pp. 1175–1191, 2001.
- [36] Laurence Devillers and Laurence Vidrascu, "Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs," in *Ninth International Conference on Spoken Language Processing*, 2006.
- [37] Volker Dellwo, Adrian Leemann, and Marie-José Kolly, "Speaker idiosyncratic rhythmic features in the speech signal," in *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.
- [38] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez Gutiérrez, Sankalp Gulati, Perfecto Herrera Boyer, Oscar Mayor, Gerard Roma Trepas, Justin Salamon, José Ricardo Zapata González, and Xavier Serra, "Essentia: An audio analysis library for music information retrieval," in *Britto A, Gouyon F, Dixon S, editors. 14th Conference of the International Society for Music Information Retrieval*, 2013, pp. 493–498.
- [39] Ahilan Kanagasundaram, Robbie Vogt, David B Dean, Sridha Sridharan, and Michael W Mason, "I-vector based speaker recognition on short utterances," in *12th Annual Conference of the International Speech Communication Association*, 2011, pp. 2341–2344.
- [40] Mohammed Senoussaoui, Patrick Kenny, Najim Dehak, and Pierre Dumouchel, "An i-vector extractor suitable for speaker recognition with both microphone and telephone speech," in *Odyssey*, 2010, p. 6.
- [41] Jan Schlüter and Sebastian Böck, "Musical onset detection with convolutional neural networks," in *6th International Workshop on Machine Learning and Music (MML), Prague, Czech Republic*, 2013.
- [42] Dimitrios Ververidis and Constantine Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech communication*, vol. 48, no. 9, pp. 1162–1181, 2006.

Analysis of Speech Rhythm for Language Identification Based on Beat Histograms

Athanasios Lykartsis¹, Alexander Lerch², Stefan Weinzierl³

¹ *Audio Communication Group, TU Berlin, 10587 Berlin, Germany, Email: athanasios.lykartsis@tu-berlin.de*

² *Georgia Institute of Technology, Center for Music Technology, GA 30332 Atlanta, USA, Email: alexander.lerch@gatech.edu*

³ *Audio Communication Group, TU Berlin, 10587 Berlin, Germany, Email: stefan.weinzierl@tu-berlin.de*

Introduction

Rhythm is a basic property of acoustic signals [1][2], with a presumed common basis for its perception grounded both in speech and music [3], hinting towards a similarity which can be tracked in the acoustic signals as well. For speech signals, rhythm analysis can provide relevant conclusions both with respect to linguistic questions (e.g. language rhythm typology) and for applications in speech technology (e.g. in multilingual dialogue systems). However, speech rhythm is difficult to analyze, since its modeling or measurement are not straightforward.

In phonetics, the measurement of speech rhythm has mainly been performed by the development of statistical measures (known as *rhythm metrics*) that capture the patterns of intervals of and between salient speech elements such as vowels, consonants and syllables. Such metrics include the standard deviation of consonant intervals ΔC , the percentage of vocalic intervals %V and the Pairwise Variability Index (PVI) [4][5][6]. Although they have been used extensively for speech rhythm description and the investigation of rhythmical differences between languages, those measures have also been criticized [7] for lack of robustness and for producing inconsistent results with respect to the rhythm class hypothesis, which states that languages belong either to a stress-timed or to a syllable-timed group [8]. Further problems include the manual or automatic annotation of speech elements which is required in order to perform the analysis, as well as that the focus lies only on high-level language elements (such as syllables or consonants-vowels) and their duration patterns for rhythm description instead of examining directly measurable signal properties. Various technical attempts to model rhythm were also undertaken in the field of rhythm-based language identification (LID). A number of studies ([10][11][12][13]) have extracted rhythmic units by using the concept of automatic segmentation in pseudosyllables (structures of the form C^nV , where C is a consonant and V a vowel) and calculating parameters concerning duration and properties of speech elements such as fundamental frequency or energy. Such studies have achieved satisfactory results (60 – 80%) in rhythm-based LID for a number of speech corpora, which shows the importance of rhythm and prosody based features for the LID task. They still, however, bear the disadvantage of taking into account higher-level language units such as syllables to extract speech rhythm.

In order to overcome these problems, we propose an alternative approach for rhythm extraction and modeling for LID. We draw inspiration from the field of Music Infor-

mation Retrieval (MIR), where there have been numerous approaches for rhythm extraction, for instance for the problem of automatic musical genre classification. One of the widely used representations is the *Beat Histogram*, which has emerged as a method for rhythmic content description for audio classification and has been described in [15][16][17]. Its basic premise is that the rhythm of an audio excerpt can be described through creating a representation of the distribution of its periodicities in a very low frequency area and extracting relevant statistical and other properties from it. A similar approach has been recently presented by Tilsen & Arvaniti [9], who modelled speech rhythm by extracting periodicities and from the signal envelope and analyzing their relationships.

This paper describes the use of the beat histogram for the creation of speech rhythm features for LID by using several relevant signal properties as the basis for its creation. The goals are the evaluation of those novel features for rhythm-based LID and the analysis of speech rhythm through investigation of the rhythm class hypothesis. In the following, the methods for speech rhythm feature extraction are described. The classification setup with two supervised learning algorithms as well as the experimental results for one multilingual speech corpus are presented and discussed.

Method

The beat histogram is created through the extraction of the temporal trajectory of a given signal quantity or its difference (also known as a Novelty Function [18]). After the signal is preprocessed (mean removal, filtering etc.), the novelty function of the signal amplitude or its envelope is calculated, half-wave rectified and periodicities are represented for an area typically between 30 and 300 BPM, by using a method such as the Autocorrelation Function (ACF) [15], the Discrete Fourier Transform (DFT) or the comparison with a filter bank of tuned bandpass filters [14][16]. The end result is a compact representation of the magnitude and value of all important signal periodicities, where for example the tempo (main periodicity) of the analyzed track can be observed. The properties of the rhythmic content of the excerpt can be then extracted with the use of descriptors such as the mean, standard deviation and other distribution statistics, as well as more specific descriptors such as the amplitude and frequency of the most salient peak.

In the context of rhythm description and musical genre classification, most of the studies have used the beat histogram with the signal amplitude envelope as a novelty function. This approach, however, does not take into ac-

count changes in other signal properties such as tonal or general spectral changes which might have other periodicities. Therefore, it is sensible to expand the beat histogram by taking into account novelty functions of other signal properties whose change over time is relevant. Experiments in musical genre classification using amplitude, tonal and spectral shape novelty functions have shown promising results for a wide range of datasets [19]. This approach is therefore adapted here for speech: We use three categories of novelty functions so as to capture the characteristics of the most important temporal trajectories in the signal:

- **amplitude-based**, accounting for changes in signal energy or loudness reflecting changing intonation,
- **fundamental frequency-based (F0)**, tracking changes in speech prosody and
- **spectral shape-based**, accounting for changes in spectral content which reflect changes of speech elements (consonants/vowels) or phoneme position.

The F0 is extracted through the use of a harmonic product spectrum algorithm [20] on a filtered version of the speech signal (4th-order Butterworth lowpass with a cut-off frequency at 800 Hz) so as to ensure tracking of the fundamental frequency alone. Three established features are extracted to track spectral changes: the **spectral flux** (indicating general spectral change), the **spectral flatness** (indicating tonalness/noisiness) and the **spectral centroid** (a measure of the spectral centre-of-weight), the latter also on a filtered version of the signal (4th-order Butterworth bandpass between 300 Hz and 3300 Hz) to insure that only formant area frequencies are considered. From the corresponding beat histograms, we then extract a list of standard features, relating to periodicity distribution statistics and to the position and salience of the beat histogram peaks, which can be seen in table 1. More information on the extracted novelty functions as well as on the subfeatures listed in Table 1 can be found in [21]. For the beat histograms here, a periodicity range from 0.5 Hz to 10 Hz was selected as characteristic of the most important speech rhythms. All features are extracted over a frame-by-frame basis: The speech files are separated in short frames of 3 s with an overlap of 50%. In the case of the rhythmic features, the beat histograms are averaged over all frames and the subfeatures extracted from them. In total, the rhythm feature set comprises 5 features times 19 subfeatures = 95 features.

Table 1: Subfeatures extracted from Beat Histograms.

Distribution	Peak
Mean (ME)	Saliency of Strongest Peak (A1)
Standard Deviation (SD)	Saliency 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)	

Experiments and Results

Experiments

In order to be able to conduct an evaluation of the proposed beat histogram features, a baseline feature set needed to be established. To that purpose, extraction of a series of non-rhythmic features was undertaken, by calculating the feature values over all texture frames (by keeping the average value inside an analysis window) of a speech file. A number of acoustic features such as MFCCs, LPCs and SDCs have been used widely for non-rhythmic language identification [22][23]. In order to maximize comparability with our rhythmic features and to be able to estimate the merit introduced by the use of the beat histograms, we used as a baseline feature set all five novelty functions listed in the previous section. The features on each novelty function can be seen in the *Distribution* column of Table 1. In total, the baseline feature set comprises 5 features times 11 subfeatures = 55 features. For supervised classification, we use the established Support Vector Machines (SVM) [24], which have been used extensively and has shown good results in many classification problems up to date. For comparison, we also use the basic and simple k-Nearest-Neighbors algorithm. For the SVM algorithm the Radial Basis Function (RBF) Kernel is used with the parameters C and γ determined through grid search, while for the kNN algorithm the euclidean distance was used with $k = 1, 3, 5$. All experiments take place as multiclass one-vs-one classification problems with 10-fold cross validation and 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, defined as the proportion of correctly classified samples to all samples classified, which can be easily derived from the confusion matrices as the sum of the diagonal to the total samples count.

We tested our features on one established multilingual speech corpus, MULTEXT [25]. This is a read speech dataset, comprising five indoeuropean languages (English, French, German, Italian and Spanish) with high signal quality (20 kHz sample rate, 16 bit quantization depth). The dataset contains between 10 and 20 passages with an average length of 20 s from 10 speakers per language (5 male and 5 female). The choice of this dataset is of importance, since it has been used extensively for rhythm-based LID and can allow conclusions both to rhythm-based automatic LID performance and language typology, since the languages contained are those prototypically belonging to the two basic groups after rhythm class hypothesis [8]: English and German to the stress-timed, French, Spanish and Italian to the syllable-timed.

Results

Results of classification can be seen in Fig. 1 and Tables 2 and 3. Concerning classification accuracy, two tendencies can be observed: First, concerning overall accuracy, the SVM algorithm outperforms the kNN in all cases, with the kNN showing very low scores (even for $k = 5$ which was the best case). Second, for the SVM, the rhythmic

feature set has slightly better accuracy than the baseline, whereas for the kNN, the baseline set shows moderately better performance than the rhythmic set. Furthermore, for the SVM results are clearly above the average prior (Pr.) of 20% (the percentage of the samples of each class in the dataset) and satisfactory, whereas for the kNN, accuracy is low and below the prior for both feature sets. With regards to language rhythm typology, the pure form of the rhythm class hypothesis does not seem to be confirmed in either case: for the kNN, all languages are confused with English, except for English itself which is classified as German, hinting towards a rhythmic similarity only between stress-timed languages. For the SVM, all languages are confused with French, but neither Italian and Spanish nor German and English are confused with each other more than with other languages.

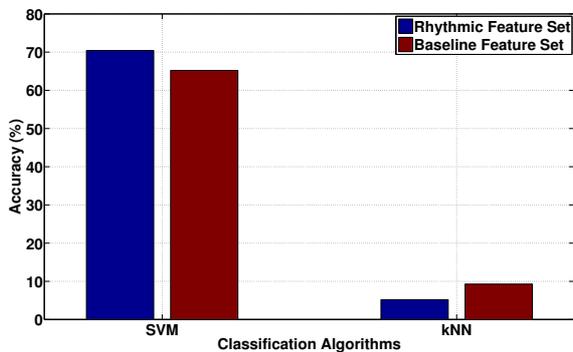


Figure 1: Classification results

Table 2: Confusion Matrix for the kNN algorithm, rhythmic feature set. All numbers indicate sample count, rows Acc. and Pr. are given as percentages, average accuracy 7.2%

True	Predicted				
	Eng.	Fre.	Ger.	Ita.	Spa.
Eng.	12	27	66	24	21
Fre.	62	2	33	2	1
Ger.	121	23	40	12	4
Ita.	97	14	38	0	1
Spa.	90	15	42	0	1
Acc.	8	2	20	0	1
Pr.	20	13.7	27.3	20	20

Table 3: Confusion Matrix for the SVM algorithm, rhythmic feature set. All numbers indicate sample count, rows Acc. and Pr. are given as percentages, average accuracy 70.4%

True	Predicted				
	Eng.	Fre.	Ger.	Ita.	Spa.
Eng.	110	22	5	5	8
Fre.	4	76	5	9	6
Ger.	15	22	121	23	19
Ita.	3	21	5	114	7
Spa.	7	25	5	6	107
Acc.	73.3	76	60.5	76	71.3
Pr.	20	13.7	27.3	20	20

Discussion

The results presented in the previous Section are promising for further research: it is clear that the use of beat histogram features can be useful for rhythm-based LID: The identification accuracy using the SVM algorithm (70.4%) lies in the same range or is better than the ones achieved in other studies [13] ($67 \pm 8\%$). An interesting point is the performance of the baseline set, which is comparable to that of the rhythmic feature set, showing that the rhythmic features can explain as much language-specific variability in the speech signal as simple, more general features. However, it should be noted that the use of other general features can achieve even higher performance scores [22][23], indicating that there is room for improvement, e.g. through the use of other novelty functions or features on the beat histogram.

With respect to classification algorithms, the SVMs are definitely advantageous in performance for the rhythmic feature set. Indeed, the very low performance of the kNN (in contrast to the comparable performance for the baseline set) is an indication that rhythmic features require more robust machine learning algorithms for identification, a result which has been confirmed through other studies [11]. Another possible reason for the lower performance of the kNN is the lack of a sufficient number of training samples or the relatively high number of features, resulting in the curse of dimensionality [21]. A possible amendment would be to perform a principal component analysis (PCA) and use the most important components as features.

Concerning the speech corpus itself, the attained performance shows that for read speech of good quality, features can be extracted which are informative of the rhythmic content and can be used to identify languages on that basis. With respect to the different languages of this dataset, it is interesting to observe from the confusion matrix (Table 2) how French seems to act as a "universal attractor" for all other classes. This effect could be due to actual difference of the french language rhythm in comparison to other languages, or to particular characteristics of this specific speech corpus (such as it containing spontaneous speech). In general, the rhythm class hypothesis, which would classify English and German together as stress-timed and the other as syllable-timed languages, does not seem to be corroborated on basis of those data.

Conclusions

In this paper we presented first results on the use of novel features for rhythm analysis and rhythm-based LID. The use of the beat histogram for speech rhythm analysis is innovative and results are promising, harboring their further use. For the rhythm descriptors, not only the signal amplitude but also other rhythm-relevant signal quantities were used as basis for the creation of the beat histogram. Furthermore, a comprehensive array of sub-features was extracted from the beat histograms, which provides ample information about the periodicities in the signal and their patterns. We could show that classifica-

tion performance for one multilingual speech corpus using the SVM algorithm is comparable to that of similar studies and close to that when using other basic, non-rhythmic features. The proposed method has the advantage that it takes into account the rhythmic on the signal and not on the speech element level, which throws a new light on speech rhythm and allows its analysis from different aspects. Another important advantage of the proposed method for speech rhythm analysis is that it is fully automatic and can be extended for larger datasets, while providing significant information on speech periodicities. This provides another aim for further research: The application of the method to speech corpora with different content (such as the OGI-MLTS [26], which contains more languages and spontaneous speech) and which are much more comprehensive (such as the GLOBALPHONE [27]) is scheduled. 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. Further future goals include the investigation of optimal parameter settings for feature extraction and the conduct of feature selection to identify the most informative features, as well as the utilization of unsupervised classification methods with the focus on evaluating the method and clarifying its merits for rhythm-based LID.

References

- [1] Patel, A. D.: Music, language, and the brain. Oxford university press, Oxford, 2008.
- [2] London, J.: Hearing in time. Oxford University Press, New York, 2012.
- [3] Hübler, S., Hoffmann, R.: Comparing the rhythmic characteristics of speech and music – Theoretical and practical issues. *Toward Autonomous, Adaptive, and Context-Aware Multimodal Interfaces. Theoretical and Practical Issues* (2011), 376-386.
- [4] Ramus, F., Nespore, M., Mehler, J.: Correlates of linguistic rhythm in speech signal. *Cognition* 73.3 (1999), 265-292.
- [5] Grabe, E., Low, E. L.: Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology 7*. Cambridge University Press, Cambridge, 2002.
- [6] Dellwo, V.: Rhythm and speech rate: A variation coefficient for ΔC . *Language and language-processing* (2006), 231-241.
- [7] Arvaniti, A.: The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics* 40.3 (2012), 351-373.
- [8] Abercrombie, D. (ed.): *Elements of General Phonetics*. Edinburgh University Press, Edinburgh, 1967.
- [9] Tilsen, S., Amalia A.: Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America* 134.1 (2013), 628-639.
- [10] Farinas, J., Pellegrino, F.: Automatic rhythm modeling for language identification. *INTERSPEECH* (2001), 2539-2542.
- [11] Pellegrino, F., Chauchat, J.-H., Rakotomalala, R., Farinas, J.: Can automatically extracted rhythmic units discriminate among languages?, *Speech Prosody* (2002).
- [12] Rouas, J.-L., Farinas, J., Pellegrino, F.: Automatic modelling of rhythm and intonation for language identification. *15th ICPHS* (2003), 567-570.
- [13] Rouas, Jean-Luc, Farinas, J., Pellegrino, F., André-Obrecht, R.: Rhythmic unit extraction and modelling for automatic language identification. *Speech Communication* 47.4 (2005), 436-456.
- [14] Scheirer, E. D.: Tempo and beat analysis of acoustic musical signals. *The Journal of the Acoustical Society of America* 103.1 (1998), 588-601.
- [15] Tzanetakis, G., Cook, P.: Music genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing* 10.5 (2002), 293-302.
- [16] Burred, J. J., Lerch, A.: A hierarchical approach to automatic musical genre classification. *Proceedings of the 6th Int. Conference on Digital Audio Effects* (2003), 8-11.
- [17] Gouyon, F., Dixon, S., Pampalk, E., Widmer, G.: Evaluating rhythmic descriptors for musical genre classification. *Proceedings of the 25th International AES Conference* (2004), 196-204.
- [18] Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M., Sandler, M. B.: A tutorial on onset detection in music signals. *IEEE Trans. on Speech and Audio Processing* 13.5 (2005), 1035-1047.
- [19] Lykartsis, A.: Evaluation of Accent-Based Rhythmic Descriptors for Genre Classification of Musical Signals. Master's Thesis, Technische Universität Berlin, Berlin, 2014.
- [20] Noll, M.: Pitch Determination of Human Speech by the Harmonic Product Spectrum, the Harmonic Sum Spectrum and a Maximum Likelihood Estimate. *Proceedings of the Symposium on Computer Processing in Communications* 19 (1970), 779-797.
- [21] Lerch, A.: *An introduction to Audio Content Analysis*, Wiley, New York, 2012.
- [22] Singer, E., Torres-Carrasquillo, P.A., Gleason, T.P., Campbell, W.M., Reynolds, D. A.: Acoustic, phonetic, and discriminative approaches to automatic language identification, *INTERSPEECH* (2003).
- [23] Campbell, W. M., Singer, E., Torres-Carrasquillo, P.A., and Reynolds, D. A.: Language recognition with support vector machines, *ODYSSEY04* (2004).
- [24] Vapnik, V. N.: *Statistical Learning Theory*. Wiley, New York, 1998.
- [25] Campione, E., Véronis, J.: A multilingual prosodic database. *ICSLP* (1998), 3163-3166.
- [26] Muthusamy, Y. K., Cole, R. A., Oshika, B. T.: The OGI multi-language telephone speech corpus. *ICSLP* (1992), 895-898.
- [27] Schultz, T.: Globalphone: a multilingual speech and text database developed at karlsruhe university. *INTERSPEECH* (2002).

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

Athanasios Lykartsis¹, Stefan Weinzierl¹

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 sig-

nal'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)	Saliency of Strongest Peak (A1)
Standard Deviation (SD)	Saliency 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.

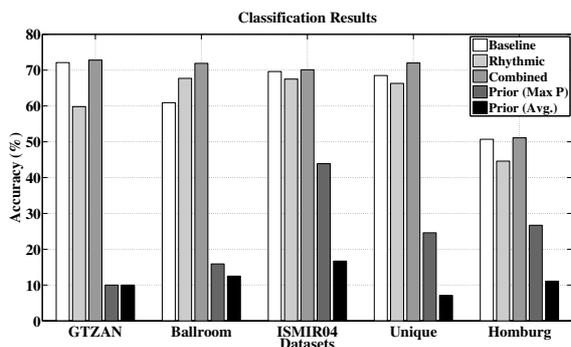


Figure 1: Classification results, comparison between datasets (music). Figure from [10].

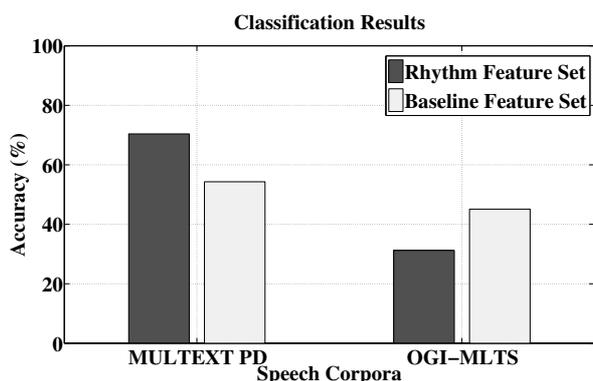


Figure 2: Classification results, comparison between datasets (speech). Figure from [13].

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	PI.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 re-

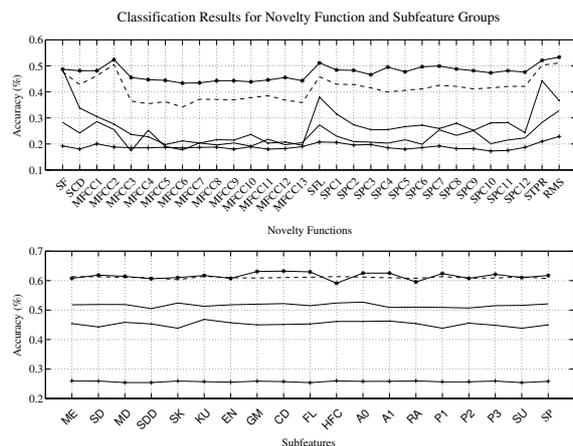


Figure 3: Feature group comparison (music). Table from [13].

sults. 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.

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 patterns 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*.