Technische Universität Berlin

MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master’s Thesis

„The influence of musical expertise on the neurological processing of the spoken language rhythm“

verfasst von / submitted by

Canberk Turan

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (M.Sc.)

Berlin, 2020

[redacted] [redacted]

Studienrichtung / Degree Program:	Audiokommunikation & -technologie
Betreut von / Supervisor:	Prof. Dr. Stefan Weinzierl
Mitbetreut von / Co-Supervisor:	Dr. Athanasios Lykartsis Dr. Katerina Kandylaki

Abstract

Human beings are naturally inclined to discover, comprehend, and respond to their surroundings in a rhythmic way by focusing on the emergence of patterns and distinct events. The perception of language and music is no exception. Linguists claim that languages get the rhythmic feel from the systematic patterns in the timing with the accent, whereas musical rhythm studies concern a regularity in the beat that a listener can tap their foot to. Yet, it has been an open question in speech rhythm research whether one can identify certain features of stimuli and the listener as correlates of perception. This study incorporates techniques from the field of Music Information Retrieval to quantify rhythmicity in spoken language and validates them with a listening experiment that is conducted on two groups of people: musicians and non-musicians. We hypothesized that rhythmicity perception of subjects with music education background show higher correlation with results of beat detection algorithms. Results have shown that participants with varying music expertise do not tend to rate speech rhythmicity differently. We have also tested the influence of stimulus type and length by presenting stories and poems in whole length as well as chunked form. For this analysis, we predicted that poems and whole length stimuli should result with a higher rating of rhythmicity due to the effect of metrical organization and completeness. As predicted, perceived rhythmicity scores of poems and whole-length stimuli were respectively higher than stories and 10-second chunks although the algorithms have suggested a significant difference of rhythmicity in neither case. Furthermore, comparative evaluation of beat histograms that are extracted from various novelty functions demonstrated different levels of association with the perceived rhythmicity ratings. The broad implication of the present research is that musical expertise solely based on training hours does not play a role in the perception of spoken language rhythmicity and this perception is strongly affected by the contextual factors that beat histograms do not capture and represent. Future research should further develop and confirm these initial findings by applying different set of measure not only to classify test subjects but also for describing the objective rhythmicity in an accurate way.

Zusammenfassung

Menschen sind natürlicherweise geneigt, ihre Umgebung auf rhythmische Weise zu entdecken, zu begreifen und auf sie zu reagieren, indem sie sich auf die Herausbildung von Mustern und eindeutigen Ereignissen konzentrieren. Die Wahrnehmung von Sprache und Musik ist keine Ausnahme. Sprachwissenschaftler argumentieren, dass Sprachen das rhythmische Gefühl aus den systematischen Mustern im Timing mit dem Akzent erhalten, während es bei musikalischen Rhythmusstudien um eine Regelmäßigkeit im Takt geht, auf die ein Zuhörer mit dem Fuß wippen kann. Dennoch war es in der Sprachrhythmusforschung eine offene Frage, ob man bestimmte Merkmale von Stimuli und dem Zuhörer als Korrelate der Wahrnehmung identifizieren kann. In dieser Studie werden Methoden aus dem Bereich des Music Information Retrieval zur Quantifizierung der Rhythmizität in gesprochener Sprache eingesetzt und mit einem Hörversuch validiert, der an zwei Gruppen durchgeführt wird: Personen mit und ohne musikalische Ausbildung. Wir stellten die Hypothese auf, dass die Rhythmizitätswahrnehmung von Probanden mit musikalischem Bildungshintergrund eine höhere Korrelation mit den Ergebnissen von Beat-Erkennungsalgorithmen aufweist. Die Ergebnisse haben gezeigt, dass Teilnehmer mit unterschiedlicher Musikerfahrung nicht dazu neigen, die Sprachrhythmizität unterschiedlich zu bewerten. Wir haben auch den Einfluss von Stimulustyp und -länge getestet, indem wir Geschichten und Gedichte sowohl in ganzer Länge als auch in zerstückelter Form präsentiert haben. Für diese Analyse haben wir vorausgesagt, dass Gedichte und Stimuli in ganzer Länge aufgrund des Effekts der metrischen Organisation und Vollständigkeit zu einer höheren Bewertung der Rhythmizität führen. Wie angenommen, waren die Bewertungen der wahrgenommenen Rhythmizität bei Gedichten und vollständigen Stimuli höher als bei Geschichten und 10-Sekunden-Stücken, obwohl die Algorithmen in beiden Fällen keinen signifikanten Unterschied in der Rhythmizität erkennen konnten. Darüber hinaus zeigte die vergleichende Auswertung von Beat-Histogrammen, die aus verschiedenen Novelty-Funktionen extrahiert werden, unterschiedliche Assoziationsgrade mit den Bewertungen der wahrgenommenen Rhythmizität.Die weitreichende Implikation der vorliegenden Forschung ist, dass Musikerfahrung, die ausschließlich auf Trainingsstunden basiert, keine Rolle bei der Wahrnehmung der Rhythmizität der gesprochenen Sprache spielt und dass diese Wahrnehmung stark von den kontextuellen Faktoren beeinflusst wird, die Beat-Histogramme nicht erfassen und darstellen. Künftige Forschungsarbeiten sollten diese ersten Ergebnisse weiterentwickeln und bestätigen, indem sie verschiedene Maßeinheiten nicht nur zur Klassifizierung der Testpersonen, sondern auch zur genauen Beschreibung der objektiven Rhythmizität anwenden.

Contents

List of Figures

List of Tables

1 Introduction

Psychology studies the behaviour of human beings - their external, directly observable behaviour as well as the inner, psychic processes, including perception, thinking, recognition, feeling, remembering, imagining, learning and motivation. When psychologists deal with rhythm phenomena, the focus is on the inner and outer behaviour patterns with which people create temporally structured sequences of events or react to temporal structures in their environment. People encounter rhythms in many areas of life, for example in processes and movements in the natural environment and in social life, in body movements, in music and while speaking. Nevertheless, there is no generally accepted definition of rhythm in psychology. The respective understanding of rhythm depends on the area of phenomena under investigation as well as on the theoretical and methodological approaches used in the research of these phenomena.

Rhythm is one of the foremost mechanisms for the human brain to translate a piece of auditory information. As is known, phone numbers are preferably separated into shorter groups to make it easier to be remembered with its evenly paced sequence. This is because human brain is more prone to process rhythmic grouping patterns [5]. In this regard, speech having inherently a serial structure of vowels and consonants which could be also considered as a material to employ the tools of musical rhythm analysis. Similar rhythm analysis metrics have been developed in the latter years to decipher the underlying temporal patterns of spoken languages [6, 7, 8, 9]. This leads us to the idea of incorporating the recently developed computer-aided rhythm analysis metrics with a comprehensive study that encompasses both neurobiological and behavioral aspects of language perception. As a part of the EU-funded project "The NEurobiology of RHYthm: effects of MUSical expertise on natural speech comprehension" (NERHYMUS) [1], it is investigated how rhythm expertise of musicians may affect the way they process the rhythm of spoken language. In order to achieve this, both musicians and non-musicians have gone through a variety of tests: an online questionnaire for quantification of musical experience, behavioral measures of rhythmicity perception in language and EEG experiments. Whilst these tasks are conducted at the Maastricht University, the infrastructure of the behavioural experiment and data analysis has been carried out within the framework of this thesis.

Following parts of Chapter 1 provides an overview of the related past research in different

[1] https://cordis.europa.eu/project/id/794455

domains, with the intention to clarify the motivation and the reasoning behind the methodology that is implemented in the thesis. The section on related work consists of three main subsections: the first presents anthropological, evolutionary, and developmental aspects of the comparative manner in language vs. music research; the second presents rhythm definitions attempts from various fields, particularly in the sense of linguistics and music theory; and finally, the third reviews the speech metrics that have been previously developed to obtain rhythmicity level and style in different languages; the fourth summarizes the state-of-the-art regarding automatic beat detection and its applications in speech. In Chapter 2, the primary principles of musical rhythm and signal processing are briefly delivered to explain the implemented methods and tools within the thesis. Chapter 3 describes how the behavioral experiment is designed and executed to gather the desired data set to be used in the statistical analysis. Then in Chapter 4, the results are tabulated, visualized and discussed with respect to the expected outcomes of the thesis. Finally, Chapter 5 states the achievements and future outlook in this theme supported by thoughts and suggestions.

1.1 Related work

1.1.1 A Comparative Approach to Language & Music

Before proceeding to rhythm definitions in speech and music, we initially discuss the relationship between two phenomena in a comparative manner to highlight the psychological motivation behind the research approach. Following compilation of studies has encouraged our paradigm by displaying similarities and differences solidifies.

Evolutionary theories imply that music and speech may have had a common foundation in the structure of an early communication system based on vocalizations and body gestures. Drawing evidence from a wide range of disciplines, Mithen [1] claims that while both language and art are most likely limited to Homo sapiens, musicality has a significantly earlier presence in human evolution and was used by various ancestors and relatives. As shown in Figure 1.1, he names this evolutionary theory as “Hmmmmm” [2] for being the type of communication used by ancestors of Homo sapiens in Africa.

Rousseau was also supporting the idea claiming that human beings were passionate before being rational language grew out of music for the sake of a social organization [10]. Although they may show divergence in the different social contexts, each relies on codes that connect the mechanisms of sound that they use for physical-biological purposes. Both music and language are evident on time and comprise of hierarchically organized components, unlike many other communication modalities and domains of human expertise [11]. Language and music are cross-cultural skills whose use distinguishes humans from other species. In all cultures there are stories, texts, meanings, words. Likewise, each culture has its own music,

[2] holistic, manipulative, multi-modal, musical, and mimetic

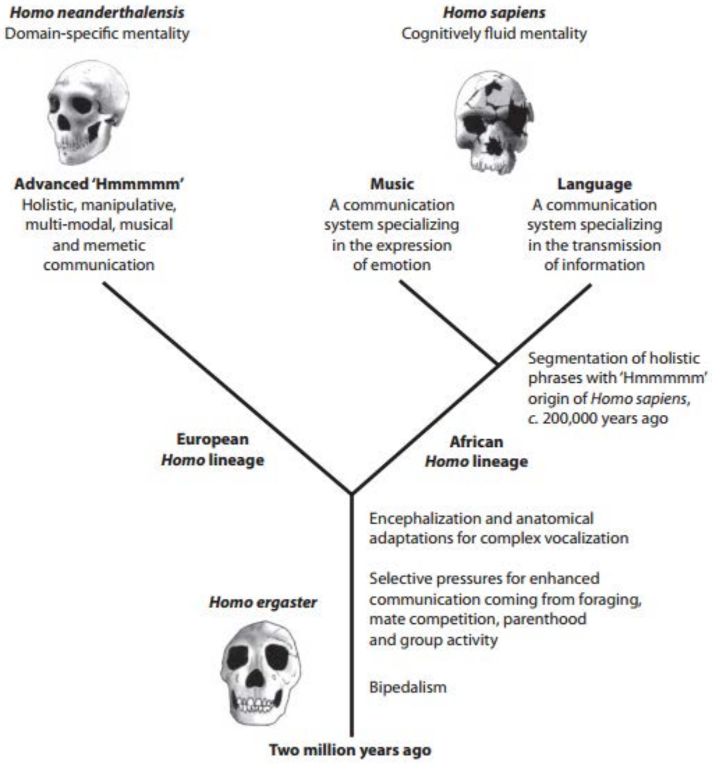

Homo neanderthalensis
Domain-specific mentality
Homo sapiens
Cognitively fluid mentality
Advanced 'Hmmmmm'
Holistic, manipulative, multi-modal, musical and memetic communication
Music
A communication system specializing in the expression of emotion
Language
A communication system specializing in the transmission of information
Segmentation of holistic phrases with 'Hmmmmm' origin of Homo sapiens, c. 200,000 years ago
European Homo lineage
African Homo lineage
Encephalization and anatomical adaptations for complex vocalization
Selective pressures for enhanced communication coming from foraging, mate competition, parenthood and group activity
Bipedalism
Homo ergaster
Two million years ago

Figure 1.1: The evolution of music and language suggested by Mithen [1]

whether instrumental or vocal in the combination of language and music, as songs. It is therefore not surprising that language and music play a decisive role in the communication and coordination of groups and for social cohesion [12]. Social groups are characterized not only by a group-specific language, but often also by group-specific music, which is transported and received in special media formats on radio, internet and television. In evolutionary research there are discussions about common roots of music and language, whereby both phenomena presumably developed as different specializations of a common referential predecessor, the so-called *musilanguage* [13, 14, 15]. In the further development of mankind, music then crystallized and developed as a means of conveying emotions, while language is used for the referential-content-based mediation.

Morever, both phenomena also show similarities in child development. For example, the developmental and learning mechanisms of statistical learning and implicit learning can be found in both [16, 17]. Children do not know anything about the grammatical structure of sentences, about word forms, verbalization and inflection possibilities, yet they usually apply the rules correctly. They also sing the melody of a song without knowing anything about scales, tonality and intervals. Particularly at the beginning of life in the pre-linguistic phase, when the infant begins to focus its attention on language and discover its first words, the brain does not seem to process music and language in different domains. At this stage of development, language is more like music to the brain [12]. At the beginning of speech development, the infant has to pay attention to and analyse the prosodic or musical elements with speech melody, speech contour, speech rhythm, dynamics, intonation, in order to recognise the first sounds and words in the mother tongue. This enables babies to distinguish between languages directly after birth due to the language-typical alternation of stressed and unstressed syllables, and to observe the contours [18, 19]. In this way the infants analyse the linguistic input primarily on the basis of its musical-acoustic characteristics. In the further course of typical language acquisition, special attention to the prosodic elements of language diminishes and the children are increasingly able to analyse the input in terms of language structure. They are no longer dependent on prosodic additional information when recognising syllables, words and sentences, as they can increasingly draw on language-specific knowledge [20, 21].

Nevertheless, the melodies and contours in the spoken utterance differ clearly from sung melodies. For example, the range of a singer's voice with more than two octaves is considerably larger than that of a speaking voice [22]. In addition, when singing, the pitches and tone durations are precisely defined and the ratio of vowel durations to consonants is shifted many times over in favour of these pitches. No matter what the cultural background is, musical pieces follows a scale that incorporates limited group of tones in an octave, where as the speech intonation does not follow harmonic sequences in a tonal way [23]. In contrast, the differences between speaking and singing voices are smaller in the child directed speech, because the persons relating to infants speak in a higher pitch and with a higher speech melody, which brings them very close to singing. The infant uses this clear emphasis of pitches and melody progressions for better speech understanding. Wermke [24] was able to show that infants from different languages already differ in their crying expressions through different contour progressions. In adult speech processing, melodic additional information can be used to learn new words better. For example, in an experiment on language learning, it has been found that adults learn three-syllable artificial words when they are presented to them by singing [25]. Language learning was demonstrated in two different conditions. Firstly, each syllable was assigned a very specific tone and was always sung on exactly this tone. But the words were also learned when the allocation of syllables and pitches was random and different. In contrast, the words were not learned when performed monotonously at one pitch. As possible reasons for these findings, Schön and colleagues [25] mention that singing makes the phonological boundaries between syllables clearer, and thus, increases feeling and attention.

Furthermore, another interesting study was able to address an significant connection between melodic deviations of languages and music. They calculated the acoustic characteristics of English and French language excerpts by analyzing vocal durations and fundamental frequencies for each syllable [26]. They then compared these language-typical characteristics with English and French instrumental music of the late 19th and early 20th centuries. They were able to show that English instrumental music of this period shows a great variability of successive intervals and this is comparable to the English language, where successive vowels also differ significantly in terms of duration and fundamental frequency. In contrast, French music shows a low variability with often only semitone steps. This corresponds to the acoustic value for the French language, where the characteristic values for vowel duration and fundamental frequency are much lower. Essentially, this inherent similarity between two phenomena has become visible through the surface of the music with patterns that reflected rhythm of the native language.

1.1.2 Rhythm Definitions and Perception

In order to investigate the rhythm concept and its perception independently of the form and style, it is necessary to take a relatively broad definition of rhythm and to bring it to a manageable level under cognitive and psychological aspects. Approaching it from a post-cognitivist and philosophical way, Christopher Hasty [27] describes it as follows:

> *"Rhythm is significant in that it provokes us to contemplate the problems of temporality and elapsing, not as a mere abstraction, but as felt reality. It could be stated that rhythm is the actual course of things in its execution, a felt execution - a feeling or sense of change and difference. By pointing us to what is actually happening, rhythm always involves movement, change, continuous activity. Rhythm can not be thought of apart from the time passing by. All life can be understood in one way or another as rhythmic."*

Originally developed from the earliest surviving works of dramatic theory, metre and rhythm in particular have a long history of terms and meanings. Already Aristoxenos worked up the claims of Plato ,who defined rhythm as the order of movement, and developed ideas on musical meter, whereby in Greek antiquity rhythm and meter were related to each other as a pair of terms mostly in the context of spoken poetry [28].

Rhythm is an essential element in both music and languages. However, there are few studies that attempt to compare linguistic and musical rhythms. This is mainly due to the fact that it is difficult to define rhythm in a universal way. The term is often used in connection with terms like biorhythm, pulse and brain frequency and equated with periodicity [29]. However, not all rhythmic patterns are periodic, while all periodic patterns are perceived rhythmic. According to Snyder [30], if two or more events take place within the duration of the short-term memory, a rhythm is already established. Snyder combines his definition of rhythm with a perceptually relevant observation by stating that two consecutive acoustic events occur as one after the other, they are psychologically transformed into a metre of impulse and dissolution. An impulse element is interpreted as a tension that dissolves into the following pulse elements. The sum of these elements results in the rhythm as a differentiation in meter. In this process, physical and mental rhythm are inseparably connected with each other in our perception and always include processes of shape perception such as groupings, accentuation, and regularities within the psychic presence time [31].

With regard to musical rhythm perception, most significant works were the ones of Fraisse [11] and Lerdahl and Jackendoff [32]. Nonetheless, several others, such as Essens [33] or Cooper and Mayer [34], also attempted to formulate their own theories on the representations of rhythmic experience and the structure of temporal hierarchies in music and dance. A major difference between models is that rhythm is seen on one hand as a mental construct and on the other hand as an actually existing acoustic event. While Fraisse [11], for example, starts out from the general perception of time and describes rhythmic representations as relations of durations perceived in this way, other studies focus on the measurable characteristics of the rhythmic stimuli that the listener is exposed to. On the other hand, according to a definition provided in [32] rhythm develops over three different structure type:

1. the grouping of the musical sounds such as phrases and sections,
2. the organization of elements with accentual difference on the musical surface (e.g high or low register, harmonic change),
3. structure of alternating strong and weak beat predictions on a well-formed grid.

Concerning language and music, there are also complex models that attempt to distinguish from one another and to relate them to each other. For Patel [35], language and music are similarly characterized as systematic temporal, accentuated and phrasal acoustic patterns and thus rhythmically analyzable. Grouping of linguistic or musical units within phrases is an essential rhythmic characteristic. In music, each piece of music has a certain time signature and a given tempo. The time signature results in tones that are more strongly emphasized (main counts) and unaccented tones (secondary counts). Musical phrases are characterized not only by melodic and harmonic relationships but also by these rhythmic peculiarities and pauses. In language, one speaks of prosodic groupings [36]. These groupings can occur as a result of the interaction of serially ordered segments (e.g. consonants and vowels) or suprasegmental levels that a often extend over syllables, words, or phrases. In this regard, phonological and syntactic structure of a sentence also plays a role, yet it is not solely affected by this [37].

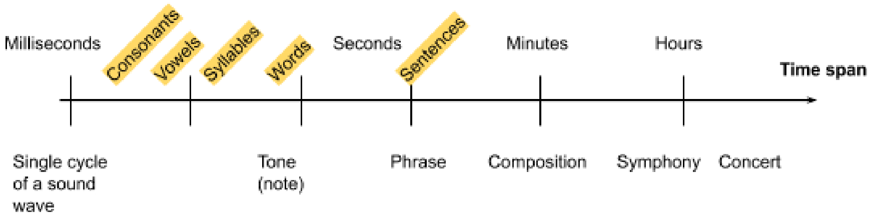


Figure 1.2: Representation of durations for both speech and music units. (Extended from the definitions provided in [2])

Figure 1.2 compares the duration scales of language segments and musical units on the same time span. Greenberg [38] shows in his study that even heavily stressed vocalic segments do not exceed the 500 ms range. Furthermore, on the question of musical rhythm, sounds and sound patterns in the range of 100 ms to 5–6 s draws most of our interest considering the human perception and interaction skills, making this visualization a consistent way of showing the commonalities in the temporal range [39]. As both speech and music are composed by varying temporal intervals, it would be fair to assume that they would both be acoustically marked with tempo-spectral properties leading to anticipations as beats[3]. Although they merely originate from a rather musical point of view, studying beat perception in the context of neural entrainment and resonance theory has received an enthusiastic response from the scientific community [41]. Entrainment describes the phase alignment process when the neural oscillations are driven by a rhythmicity in the environment, which, in a very general sense, enables the sensory interaction with the stimulus [42]. According to neural resonance theory, musical rhythm elements such as meter and tempo are main reference points that enable neural rhythms to synchronize with the auditory stimulus, which then establishes the dynamic attention and gives rise to beat prediction as well as motor coordination [43]. Albeit there is a lack of clear separation between the prelexical units in a continuous speech signal. Therefore, auditory neuroscience and speech processing fields have focused on quasi-rhythmic energy fluctuations (amplitude envelopes) in the speech signal and have taken it into account for the comparison with neural responses [44, 43]. This way, phases of high energy in the speech signal are expected to be associated with a high neuronal response rate. Studies have show that there is a noteworthy correspondence between average durations of speech units and the frequency ranges of cortical oscillations as shown in Table 1.1 [45]:

Table 1.1: Shared time scales of speech units and brain oscillation types

Speech unit		Cortical oscillation	
type	*duration (ms)*	*type*	*Frequency range (Hz)*
Phonetic features	20 - 50	gamma beta	> 50 15 - 30
Syllables & words	~250	theta	4 - 8
Prosodic phrases	500 - 2000	delta	< 3

Neuroimaging studies has remarked a shared mechanism for both music and speech processing that takes place in anatomically overlapping brain structures and highlighted that supplementary motor area as a crucial brain region involved in the temporal processing of speech [46]. Further research has addressed the multitude of regions in both cortical and subcortical parts contribute in the mechanism of temporal grouping [47] revealing the complex interplay

[3] Beat refers to the perceived pulses which are approximately equally spaced and define the rate at which the notes in a piece of music are played[40]. Further information is provided in Section 2.1

between the contextual and temporal brain activities.

Behavioral studies on rhythm perception mostly favored measuring the abilities with motor response tasks such as finger tapping in which subjects were observed while they perform certain physical actions synchronized to given stimuli [48]. Furthermore, various studies involved subjects' judgments on the deviation from isochrony[4] based on single events as well as the tempo changes occurring on an entire auditory sequence [50].

On one hand, most of these approaches had certain drawbacks considering the difficulty level of the tasks demanded from the test persons. In case of lacking cognitive and memory skills required for following instructions and beat production, beat perception induced by the stimulus only is altered by the quality of the task performance [51]. On the other hand, musical forms of stimuli have primarily been more under the spotlight of most studies as opposed to sequences of speech utterances, since it consists of noticeably periodic pulses, which is preferred for motor skill activities and speech has a relatively irregular temporal structure. This also brings up the question about how speech can possess abstract periodic patterns and still induce perceptual regularities in the human brain [52]. Moreover, it is also reported that listeners who grow up in a musical environment with temporally irregular beats are capable of synchronizing with this asymmetric sound structure. This indicates that synchronizing with or detecting periodicities to a large extent depends on familiarity and enculturation rather than on physically measurable periodicity [53]. Poeppel [54] combine this hierarchy in the time scales of brain activity with the hemispheric lateralization. He proposed the idea of asymmetric sampling which points out that temporal functionalites in different frequency ranges spatially distribution in left and right lobes of the brain.

Although psycholinguistic studies (e.g. [55]), have revealed that the human brain responds to the interchange of strong and weak beats under certain conditions, the validity of the idea that assumes brain signals are representing the stimulus envelopes faithfully, is also a topic of discussion. In [43], it has been made evident that time-frequency rhythms happen to be dissociable from the beat perception and therefore it is recommended to combine neuroimaging with creative behavioral paradigms in order to produce more consistent results in understanding the speech rhythm perception.

1.1.3 Speech Rhythm Description

In the field of a linguistics, the rhythm of a particular language is always associated with the pattern of its speech phenomena, which are characteristic of that specific language. The distinction between the languages of the world into two rhythmic classes goes back to Lloyd James [56], recognizes that some languages are similar or dissimilar to English in rhythmic terms. He uses two terms, ”Morse-code rhythm” and ”machine-gun rhythm”. The former referst to the impression of the the Morse-code, in which long and short impulses follow each

[4] Principle by which phonological units tend to be equally spaced in time [49]

other. In languages with ”machine-gun rhythm”, on the other hand, syllable durations seem to be of equal length. This division of languages into one of two rhythmic classes is taken up by Pike [57], who also replaces the somewhat aggressive terms ”Morse code” and "machine-gun" with "stress-timed” and "syllable-timed". Thus, the classification of Abercrombie [58] is adopted and further developed, combined with the assertion that every language falls into one of these two categories. This division is called rhythmic dichotomy. According to this

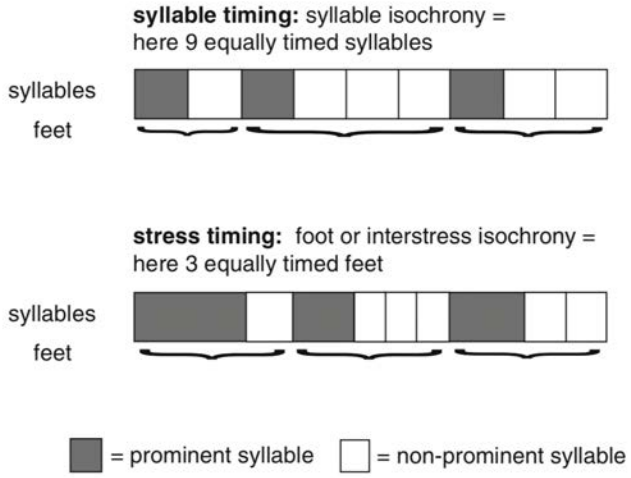

syllable timing: syllable isochrony =
here 9 equally timed syllables
syllables
feet
stress timing: foot or interstress isochrony =
here 3 equally timed feet
syllables
feet
= prominent syllable
= non-prominent syllable

Figure 1.3: Visual representation of rhythmic dichotomy [3]

paradigm, in stress-timed languages such as English and German, stressed syllables were claimed to occur at regular temporal intervals, whereas in syllable-timed languages such as Italian and French, syllable onsets were claimed to be evenly timed. Since the core idea of these assumptions arises from the observation that in spoken language there are basically isochronous intervals of one or the other type of equal length,they are collectively called the isochrony hypothesis.

However, the assumption that each language belongs to one or the other rhythmic class says very little about certain individual languages or language groups. In fact, there has been various phonetic attempts to save the isochrony hypothesis. The most promising of these shift the isochrony concept from the sound event to its perception by and try to explain the divergences between perception and speech signal by means of other intervening factors [59, 60].

Moreover, a substantial amount of evidence is also demonstrated through behavioural experiments that the perceived timing is not necessarily reflecting the measured inter-stress intervals [61, 62]. Morton [63] proposed that the psychological moment of occurrence, in other words P-center (Perceptual Center) does not overlap with the signal onsets as expected. In parallel to all these discussions, Liberman [64] interestingly refuses to concern isochrony. He separates the hierarchically organized strong-weak components from the grid of temporal elements. This induced more motivation to address the rhythm’s role in the field of phonology. On the phonetic-phonological side, the rhythmic classification has also not simply been dropped but has been reformulated. In later studies, it has been suggested that stress-timed and syllable-timed languages differ fundamentally from one another on the basis of several rhythmic factors such as syllable structures, lexical compositions, processes of reduction (both for vowel and consonant) and intonations [65]. Hence, the impression of the regularity in segmental isochrony could be altered by the aforementioned factors, which implies, the idea of strict classification becomes even more obsolete. Eventually, it encouraged others for the development of certain measurement methods based the consonant/vowel segmentation and variation [6, 7]. In the two studies mentioned above, a consonant and a vocal variability measure were proposed as acoustic rhythmic correlates. In [6], these are the percentage duration during which a speech signal is vocal (%V) and the standard deviation of consonant intervals (ΔC). In [7], consonant and vocal variability are measured using a measure that calculates average differences between two consonant or vocal intervals following each other (Pairwise Variability Index; PVI). Since all acoustic rhythm correlates are based on measurements of the duration of vocal and consonant intervals, it can be assumed that these interval durations are influenced by speech rate (fast speech = shorter interval durations, slow speech = longer interval durations). Besides, it will never be possible to take this influence into equation since it is non-linear [66]. Galves [8] introduced an index of local regularity of the speech signal under the name of sonority. This was a mapping of the signal spectrogram into a function of time scaled between 0 to 1. At each time step it computes the relative entropy between adjacent normalized bins of the spectrogram. A local average of these relative entropies is then mapped through a fixed decreasing function to define the current value of the sonority. In comparison to the other linguistic attempts, this procedure has the principal benefit that it can be performed in a completely automatic way, with no need for the prior hand-labeling of the acoustic signal. With the rapid expansion of its commercial applications, rhythm description tools of Music Information Retrieval (MIR) gained traction and has already penetrated the field of speech signal processing. Motivated by the need in automated genre classification beat detection algorithms has been proposed in the literature in varying representation formats such as Beat Histogram [67], Beat Spectrum [68], Tempogram [69]. Their use in language identification systems has already given promising results employing different novelty functions based on multiple audio features [9]. The use of the novelty functions has served as an important source of information for temporal behaviour of onsets, especially in music signals. Bello [70] introduced multiple calculation methods based on the frequency shifts, phase irregularity and amplitude variations. Lykartsis [71] extended these methods by specifying the relevant signal quantities and illustrating their temporal trajectories as most explanatory novelty functions. Relying on satisfactory results achieved in the framework of language identification tasks, we believe it could provide as a foundation for further behavioral experiments and shed light on the ambiguities of previously proposed rhythm typology.

1.2 Thesis Aim and Application

Aforementioned quantification attempts have contributed in the research in conjunction with many neuroimaging and behavioural experiments, yet it has failed to give globally consistent and robust results as discussed in [72]. This brought us to the following questions: Can the evolutionary commonalities of speech music be a motivator for a shared methodology in rhythm analysis? Does spoken language embody a pattern of time intervals within a sequence that constitutes the quasi-rhythmic acoustic cues? To what extent are the humans conscious of the rhythmicity concept in natural speech? If yes, does the musical training play a role in this?

The scope of the thesis is designated to complement the neurobiological findings that are achieved as a part of the NERHYMUS project. In specific, the remaining part of the project will focus on brain activation with electroencephalography (EEG), in order to find out whether musical expertise affects beat perception in language processing. The results of the proposed project can develop theoretical accounts of rhythm perception and inspire experts to use targeted rhythm-based therapies to treat stroke-induced or developmental language deficiencies. In this regard, we have focused solely on the behavioral experiment side of the entire project. Psychophysical responses of the subjects are collected to specify a reference variable for rhythmicity in speech perception. The control variables are identified in a way that results can serve for the benefit of all stakeholders of the multifaceted project. On one hand, speech rhythm analysis suggested in [73] are put to test. On the other hand, the influence of the musical expertise, stimulus type and stimulus length are taken as control factors on the determination of speech rhythmicity. Accordingly, the the scope of the thesis is defined as:

- Programming and technical support of the listening experiment
- Data analysis and reporting following the completion of the experiment

In order to deliver the expected outcomes successfully, we applied the following steps:

1. Writing a program script that can be operated in the circumstances of the laboratory of the fellow research team
2. Specifying a practical and scalable data storage format to be used later stages of the analysis
3. On-site technical support to secure the right implementation of the experiment practices
4. Establishing a trackable sheet where experiment participation can be remotely monitored and/or logged
5. Compiling data segments into a single finalized dataframe
    - Post-processing of online musical expertise questionnaire

- Definition and computation of the set of subfeatures and novelty functions that could reflect the concepts of interest (such as rhythmicity, speech rate) accurately from the beat histogram

6. Selection and implementation of the statistical measures for feature association and group comparison

Although this topic extends from neural fundamentals of timing to novel derivations of linguistic and musical metrics, it is intended to give an overview of the current state of psycho-linguistic research on natural speech comprehension as well as it is correlation with the musical rhythm analysis.

Using these as our starting point, we hypothesized that people who had musical training are more capable to sense periodicity in speech than the ones without music training. Assuming that there are periodicities of anticipated cues in speech, we have specified beat histograms as our central tool to generate subfeatures for defining rhythmicity and speech rate in a statistical sense. In this regard, we also hypothesized that methods based on features with perceptual relevance outperform others in correlating with people's conscious evaluation of speech rhythmicity. Finally, on the stimulus side, we predicted to have results showing how people are being affected by the context and length of the stimulus types and rate them as more rhythmic. In this way, we aimed to identify the cognitive network of information when a speech rhythm processing task is translated to a response. Results are expected to motivate further research that combines psycho-linguistics and MIR to improve the accuracy in speech rhythm definition.

2 Background Theory

2.1 Elements of Musical Rhythm

Before explaining the rhythm analysis procedure, it is necessary to define the basic musical terms. Beat, meter, and partly also tempo are often used in everyday life in an undifferentiated way, sometimes even synonymously. However, as the terms differ in their actual meaning, an exact definition is essential as a basis for further work with the terms. These are not only phenomena of music, but also occur in contexts such as poetry, language itself, architecture, drama, in films, athletics, the dance, in body movements or biological processes like the heartbeat. In the following, the relevant terms are described with regard to their use in Music Information Retrieval, and thus, the focus is on the context of Western music theory as well as general signal processing practices.

Beat The basic element of rhythmic structures in music is the beat. On the level of musical description, it is understood as a singular point without a temporal extension. Physically seen, however, it is a sound having a certain duration that indicates a musical event. The point in time when the musical "event" occurs, in other words when there is a change in signal characteristics, is also referred to as the onset. It is the earliest cue point the event can possibly be detected at. The distance between two beats is defined as Inter-Beat interval (IBI), Inter-Onset Interval (IOI, or Inter-Stimulus-Onset Interval (ISI) and usually refers to a sequence of beats [74]. The beat is usually not continuously audible in music but is indirectly generated by the rhythm of the music. Rhythm and beat are closely related. The rhythm is based on the beat, but the latter is only established by the respective rhythm [31].

Pulse A series of auditory events, in this case a sequence of beats that occur at a recurring time interval from one another, is called a pulse [34]. The pulse is derived from a multitude of events and as a first rhythmic layer forms the most important reference and orientation point for the perception of other rhythmically relevant elements [74]. With a pulse, all beats have the same emphasis. Since this occurs in the rarest cases of musical events, it is therefore an abstract description of the temporal course. As a basic process of rhythm perception, the recognition of a pulse indicates the periodicity in this course.

Meter Meter is used to define the equal subdivisions of the pulse with the marking of the strongly accented pulse that signifies the beat [37]. Thus, it represents a structure of the pulses, where this structure is realized by the beats. The hierarchical structure is represented in musical pieces with time signature notation which also informs about the locations of the beat without demonstrating durational value.

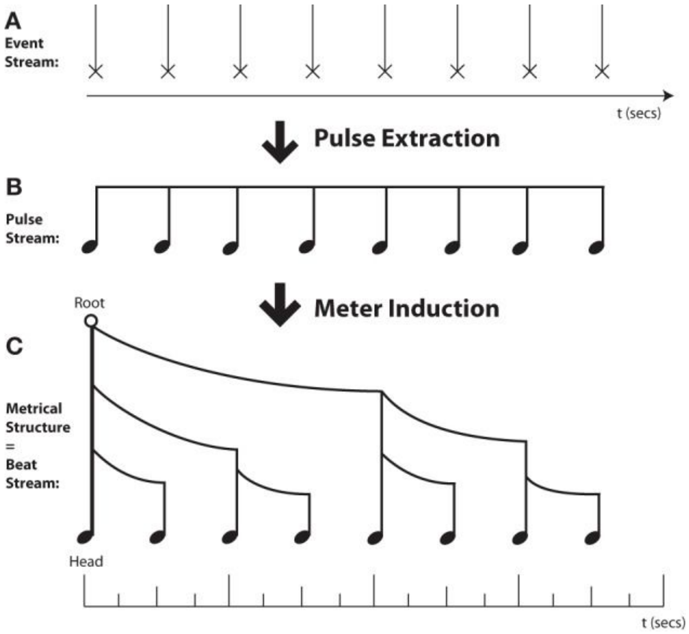


Figure 2.1: Musical rhythm units according to the cognitive processing order [4]

Tempo It is the rate at which perceived pulses with equal duration units occur at a moderate and natural rate [75]. This perceived tempo is called the "tactus" and is sometimes simply referred to as the foot tapping rate [32]. For segments of music with constant tempo, the tempo $\mathfrak{T}$ in BPM can be computed using the length of the segment Δt_s in seconds and the number of beats $\mathcal{B}$ in the segment:

$$\mathfrak{T} = \frac{\mathcal{B} \cdot 60\text{s}}{\Delta t_\text{s}} [\text{BPM}]$$

If you represent rhythms on a time axis, a faster tempo only compresses this axis, while a slower tempo stretches it. The tempo usually has no influence on the underlying metre or rhythm. However, despite the same rhythmic notation, tempo variations can occur within a piece and greatly change the character of the music. The local tempo $\mathfrak{T}_l$ can be determined by identifying the time of the occurrence of each beat t_b and the calculation of the tempo between successive beats i and $i+1$:

$$\mathfrak{T}_l(i) = \frac{60s}{t_b(i+1) - t_b(i)}$$

For the calculation of the total pace of the piece, two terms must be distinguished: Firstly, there is the average tempo, which is represented by the average of all local tempi. This variant often does not correspond to the perceived tempo. Second one is the main tempo which is specified with the highest frequency of occurrence. This frequency distribution is visualized usually in a histogram and the the tempo is determined from its peak point.

2.2 Audio Signal Processing

The practice of beat histogram extraction does not work directly with the audio signal, but with feature trajectories extracted from the audio signal $x(t)$. The audio signal is a continuous sequence of samples $x(i)$ in the digital domain. The sampling rate is $f_s = 44100$ Hz for all time signals treated in this thesis. However, as a part of the pre-processing, all the signals were converted to mono and downsampled to 22500 Hz to increase the computation speed. Short-time signal sections (frames) are cut out of $x(i)$ and analyzed for certain characteristics to generate the novelty functions. For the following parts of this chapter, Alexander Lerch's *Audio Content Analysis* is taken as the main reference to compile relevant definitions and notations [75].

Short-time Fourier Transform Short-time signal analysis corresponds to a data transformation which reduces the amount of data to be processed and considers only a small section of the signal. A short-time Fourier transform is a Discrete Fourier Transform (DFT) windowed in the time domain. Based on the Fourier transform, it is the most convenient method to determine the frequency response of a signal over time. The signal under consideration is divided into individual time frames by means of an analysis window. A frame length with an acceptable computational effort is chosen and it is assumed that the spectral characteristic remains reasonably stationary with in that frame. This section of the signal is also called a window and means a multiplication of the signal section with a window function. The window of length $\mathcal{K}$ is denoted as:

$$\mathcal{K} = i_\mathrm{e}(n) - i_\mathrm{s}(n) + 1$$

where i_e and i_s denote the sample index of the last and the first samples of the signal frame having the index n. This window is shifted along the signal with a certain hop size and the DFT is calculated in each step. This gives the STFT with the formula:

$$X(k,n) = \sum_{i=i_{\mathrm{s}}(n)}^{i_e(n)} x(i) \exp\left(-\mathrm{j}k \cdot (i - i_{\mathrm{s}}(n)) \frac{2\pi}{\mathcal{K}}\right)$$

with the frequency expressed with bin k:

$$f[k] = \frac{kf_s}{\mathcal{K}}$$

2.2.1 Frame-based Features

In the context of this thesis, the speech signals are processed based on trajectories of some of the conventional instantaneous features. As mentioned in Section 1.1.2, rhythmic groupings and patterns manifest over several types of occurrences which can be identified by using both statistical and spectral quantities. A selection of these quantities are presented in this section. Different congregations of these the statistical and spectral features will then form the rhythm analysis techniques such as onset detection (novelty function) and beat histogram. Therefore, the definitions and calculation methods are simply given to explain the building blocks of the rhythm analysis procedure of the respective signals.

2.2.1.1 Statistical Features

These measures can be applied to both, the time- domain signal block as well as the spectrum. While the definitions below use $x(i)$ as input signal, it could be substituted by $X(k,n)$, by a series of feature values $v(n)$ or by any other signal of interest. Theoretically, the statistical properties presented below require a signal of infinite length, however, in practical applications they can be assumed to be sufficiently accurate as long as the block length is adequate.

Arithmetic Mean It is calculated by simply taking the sum of a all the samples in the block, then dividing that sum by the count of the samples in that block:

$$\mu_X(n) = \frac{1}{\mathcal{K}} \sum_{i=i_s(n)}^{i_e(n)} x(i)$$

Geometric Mean The geometric mean is an average measure for array of positive numbers that are ordered on a varying scale,such as logarithmic. It represents the central tendency in skewed datasets where arithmetic fail to help a valid interpretation. Unlike arithmetic mean, here, the value of samples are multiplied and divided by the count of samples, which can also be expressed in logarithmic sum.

$$
\begin{aligned}
M_x(0,n) &= \sqrt[x]{\prod_{i=i_s(n)} x(i)} \\
&= \exp\left(\frac{1}{\mathcal{K}} \sum_{i=i_s(n)}^{i_e(n)} \log[x(i)]\right)
\end{aligned}
$$

Variance Variance is a measure of the dispersion of the probability density around its expected value. Mathematically it is defined as the mean square deviation of a real random variable from the arithmetic mean:

$$\sigma_X^2(n) = \frac{1}{\mathcal{K}} \sum_{i=i_{\mathrm{s}}(n)}^{i_{\mathrm{e}}(n)} (x(i) - \mu_x(n))^2$$

2.2.1.2 Instantaneous Features

Root Mean Square (RMS) Intensity and loudness of the piece play a major role, however they they differ on the physical and perceived level. Intensity is a measurable unit which describes the strength of a sound, whereas the loudness is characterized only by a human listener. Although they hold very close meanings, then there is there is a non-linear connection between the two expression [76]. Therefore, in audio feature analysis computations one should always refer to the intensity. As a measure of intensity, the effective value, in other words Root Mean Square (RMS), should be introduced as one of the major audio features. The calculation is based on an audio block $\mathcal{K}$ with the length of several hundred milliseconds:

$$v_{\mathrm{RMS}}(n) = \sqrt{\frac{1}{\mathcal{K}} \sum_{i=i_{\mathrm{s}}(n)}^{i_{\mathrm{e}}(n)} x(i)^2}$$

Spectral Flux (SF) The spectral flux measures the amount of change of the spectral shape. It is defined as the average difference between consecutive STFT frames:

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

Spectral Centroid (SC) The Spectral Centroid is one of the statistically determinable values of the spectrum and defines the center of gravity of the spectral energy of an audio signal. In the form of a formula, it describes the frequency-weighted sum of the spectrum, normalized with its unweighted sum:

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

The resulting result is usually converted to the unit Hertz. High values indicate a large proportion of high frequencies in the signal. Within the analysis section of this thesis, Spectral Centroid was not used for the determination of novelty functions. However, its formula was utilized in the extraction of beat histogram subfeatures to be used as a descriptive of speech rate.

Spectral Flatness (SFL) The ratio between the geometric and arithmetic mean of the magnitude spectrum indicates the spectral flatness:

$$v_{\mathrm{SFL}}(n) = \frac{\sqrt[\kappa/2]{\prod_{k=0}^{\kappa/2-1} |X(k,n)|}}{2/\kappa \cdot \sum_{k=0}^{\kappa/2-1} |X(k,n)|}$$

This feature belongs to the group of measures that represent the signal tonalness. Tonalness refers to the existence of predominant periodicities regardless of the musical harmony. For that reason, high values of spectral flatness signify a high amount of noise components and thus could be also interpreted as an audio quality measure. It ranges between zero and one, where "0" theoretically is descriptive of a perfect sinusoidal signal and the latter is of white noise [75].

Mel Frequency Cepstral Coefficients (MFCC) MFCC is also a measure to describe the spectral progression with additional recognition of human perception range. They have been used frequently in connection with the classification of genres [77, 67]. When determining the coefficients, the spectrum of the audio signal is calculated first. This is followed by the logarithmization of the magnitude values and a mapping of these to the Mel-Scale [78], which approximates the frequency distribution to the human auditory system. Finally, a Discrete Cosine Transform (DCT) is performed, which is comparable to the DFT, but only the real components are included in the calculation. The result corresponds to a strong energy compression, which is why most signal information is concentrated in a few low frequency components of the DCT. In the current study, we have used thirteen (13) coefficients that are denoted by the coefficient index j. The calculation steps have been implemented according to the formula below:

$$v_{\mathrm{MFCC}}^{j}(n) = \sum_{k'=1}^{K'} \log\left(\left|X'\left(k',n\right)\right|\right) \cdot \cos\left(j \cdot \left(k' - \frac{1}{2}\right) \frac{\pi}{\mathcal{K}'}\right)$$

where $\left|X'\left(k',n\right)\right|$ denotes the Mel-warped magnitude spectrum of the signal frame.

Spectral Pitch Chroma (SPC) Human brain is capable of detect the similarity between two musical pitches if they are exactly octave apart from each other, such as C1 and C2. Pitch representations consist of two components: chroma (letter) and the tone height (number). One octave is divided into 12 equal semitones which are represented by the western musical notations: [C, C$^\sharp$, D, D$^\sharp$, E ,F, F$^\sharp$, G, G$^\sharp$, A, A$^\sharp$, B]. Tonality and harmony, however, can not be read directly from the spectrum or spectrogram of a signal, since spectral components of a sound can be distributed over several octaves. Therefore, the spectrogram is reduced to twelve-dimensional vector. When applied to the spectrogram, all frequency bins that correlate to a given pitch class are aggregated into a single coefficient for a given local time frame. Being perceptually driven, it allows us to represent spectral change on a different frequency scale.

2.2.2 Rhythm Analysis Features

Novelty Function In event based signals like music or speech, transients or onsets are great source of information to identify energy changes as a temporal array. Even in non-note based music, trajectories are manifested by spectral changes which can function as marker. To follow these markers, novelty functions are generated by calculating the specified audio feature through a frame-based processing. Novelty functions are the basis of periodicity detection and thus the beat identification. As an example, novelty functions based on spectral flux of a speech signal and a music signal is compared in Figure 2.2. For the speech signal, we picked one of the spoken poems that was used in the current study and a well-known disco song “Stayin Alive” by BeeGees. Both signals are 10 seconds long. This figure shows the how accurate novelty functions can depict the periodicity in music signal, while for speech signals hardly demonstrate periodic changes.

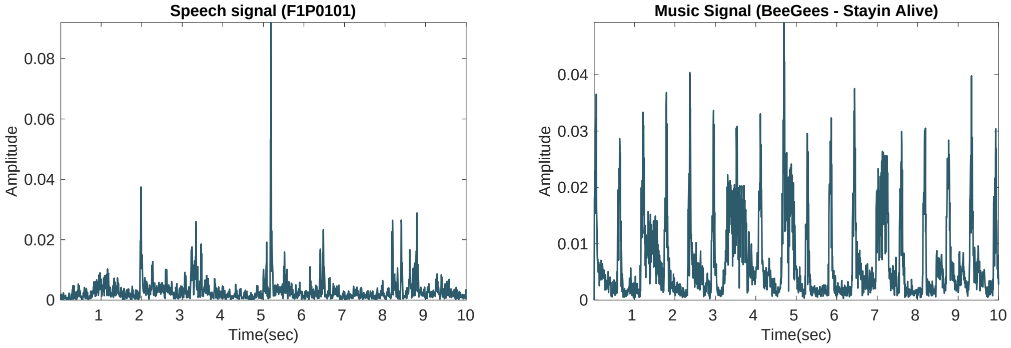


Figure 2.2: Novelty functions of a speech signal and a music signal

Autocorrelation Autocorrelation function (ACF) is a method that can help reveal overarching structure within a pattern, specifically repeating temporal sub-patterns. In a general sense, it is the correlation of a time series data, such as novelty function, with a delayed copy of itself at different time lags and shown as a function of time. The correlation describes the similarity of two signals at different displacements to each other. For finite, discrete signals $x(i)$ and $y(i)$, the products of all samples lying on top of each other are summed up for every possible shift η of the signals. ACF is a special case of the correlation function when the input signals are identical. It becomes a measure of self-similarity and used for expressing the periodicity in the signal. Using the notation above, it is referred as:

$$r_{xx}(\eta, n) = \sum_{i=i_{\mathrm{s}}(n)}^{i_{\mathrm{e}}(n)-\eta} x(i) \cdot x(i+\eta)$$

The result of the process is a vector of these sums at the different time lags between the two signals:

$$v^{\eta}_{\mathrm{ACF}}(n) = r_{xx}(\eta, n) \quad \text{with } \eta = 1, 2, 3, \ldots$$

When one or more events in the two signals overlap a peak appears in the vector. The more the events overlap at a certain lag, the higher the peak in the function, suggesting that a sub-pattern might repeat after a duration that equals this lag. Real-world patterns, however, are often not isochronous and any rhythmic structure usually contains some amount of swing or error. Therefore, these pulse streams are represented on histograms based on occurrence rates.

Beat Histogram The Beat Histogram (BH) is graphical display of the distribution of multiple beat periodicities in a signal.In speech signals and music signals in which no clear periodicity is evident, use of a global tempo is redundant. At that point, we refer to beat histograms to elaborate the rhythmic characteristics of the signal. One should note that BH is essentially a empirical representation of various periodicities in the signal and can be potentially calculated using various methods. Analogous to Discrete Fourier Transform, the horizontal axis is denoted by a unit of frequency, BHs can be further analyzed with the help of spectral audio features. Here, we extract the beat histograms from the auto-correlated novelty functions. In Figure 2.3, the beat histograms derived from the novelty functions in Figure 2.2 are presented. As it is seen from the plots, histogram curve for the music signal is highly smooth in comparison with the speech signal. Such signals are classified as “rhythmic”, which is almost impossible to see in natural speech. In the present study, we quantify their closeness to this ultimate curve trend by defining statistical or spectral subfeatures. These subfeatures are explained in detail under Section 3.2.

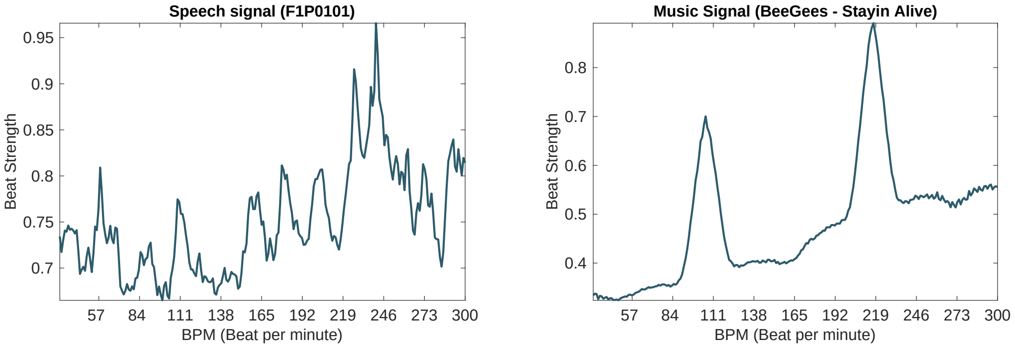


Figure 2.3: Beat histograms of a speech signal and a music signal

3 Methodology

In this chapter, we present an overview of the empirical studies that we have accomplished in the framework of the behavioural experiment we have conducted in the facilities of the fellow research team in Maastricht. These studies comprise of the design, conduct and data analysis of the experiment following sections summarize the practices employed in each aspect with providing details about the materials and tools that have been used for each step.

3.1 Behavioural Experiment

Design The experiment design was planned by the fellow supervisor in Maastricht and provided as a set of requirements. These requirements included the user interface, scoring scale for the participant, lists of audio stimuli showing the the playback order. The listening experiment is programmed accordingly using the `MATLAB version R2012B` (The MathWorks Inc., US) and `Psychophysics Toolbox Version 3` [79]. This free toolbox provides an exhaustive set of functions that are employable for neuroscience research offering a high level of control on auditory stimuli and the design of experiment interface. Later on, the technical maintenance of the experiment setup will be handled remotely.

Participants A total of twenty six (26) undergraduate participants were recruited for the study, all of which were undergraduate students in the Faculty of Psychology and Neuroscience of Maastricht University. Mean age was 21,5 years and there were more females (n = 19; 76%) than males (n = 6; 24%). All participants completed a pre-screening form containing the inclusion and exclusion criteria and they were all found eligible for the study. They were German natives and right-handed. None of them had a neurological or psychiatric record or had been diagnosed with any language or other developmental difficulty. Participants were informed about the study, registered for their first session and where granted with credits through the Sona System, a cloud-based participant pool management software. At each one of the three sessions, the participant's informed consent was obtained.

Music Background Questionnaire Music and dancing background was measured with the MMHQ (Montreal Music History Questionnaire) [80]. In order to group participants into musicians and non-musicians, we collected information on the participants' past experience in vocal, instrumental or dance training. Additionally ,we used a modified version with additional questions assessing language skills. The questionnaire was administered on the online survey tool `LimeSurvey` running on the server of the International Laboratory for Brain, Music and Sound Research (Brams). It consisted of a total of 299 questions; subjects had to answer only these questions which were relevant to their experience. Participants were registered in the `LimeSurvey` environment using their ID number given from the researchers to keep the personal data protected.

Audio Stimuli According the project requirements, stimuli should comprise two male and two female recordings, each in a clear background-free sound environment for specific text types (poetry and short stories) in German. Nine poems and six stories were selected and studied in written format.

Table 3.1: Titles and the authors of the poems used in the experiment

Poem Title	**Author**	**List code**
Der faule Hanns	Felix Dahn	P01
Der Streit um die Krone	Felix Dahn	P02
Das Lied vom blöden Ritter	Heinrich Heine	P03
Wie der Teufel den Schwanz verlor	Heinrich Hoffmann	P05
Jetzt wohin?	Heinrich Heine	P07
In der Fremde	Heinrich Heine	P08
Wir saßen am Fischerhaus	Heinrich Heine	P09
An Luna (Schwester von dem ersten Licht)	J. W. von Goethe	P10
Guter Rat	Heinrich Heine	P12

Table 3.2: Titles and the author of the stories used in the experiment

Story Title	**Author**	**List code**
Monolog eines Kellners	Heinrich Böll	S01
Es wird etwas geschehen	Heinrich Böll	S02
Das Leben ist ein Würfelspiel	Francois Loeb	S04
Nachspielzeit	Francois Loeb	S05
Stehauf-Mädchen	Petra Müller	S08
Weihnachtsstress	Konni Mente	S10

Fellow researchers recruited four students, who were semi-professional actors and actresses. The recordings were completed in the sound proof EEG chamber of the Faculty for Psychology and Neuroscience in Maastricht University. Recorded stimuli were then normalised and segmented with the track codes listed in Table 3.3 to ease the designation of material in simultaneously running experiments. To keep the total duration of each session similar, these segments were planned to be presented in three lists (List A, List B, List C). However, the playback order was subject to change according the prompt input in the very beginning of the each experiment session. Since the experiment design follows a within-subject assessment, six pseudo-randomized orders of stimuli are created from which the the user could select blindly on the opening screen by putting in a number. This way, it was aimed to mitigate the sequence effect by partially counterbalancing [81].

Table 3.3: List of stimuli with coded naming

List 1	Speaker	List 2	Speaker	List 3	Speaker
P0101	F1	P0201	M1	P0501	F1
P0102	F2	P0202	M2	P0502	F2
P0103	M1	P0203	F1	P0503	M1
P0104	M2	P0301	F2	P0504	M2
P0105	F1	P0701	M1	P0801	F1
P0106	F2	P1201	M2	P0901	F2
S0401	M1	S0201	F1	P1001	M1
S0402	M2	S0202	F2	S0801	M2
S0501	F1	S0203	M1	S0802	F1
S0502	M1	S0204	F1	S1001	F2
		S0101	M2	S1002	M1
		S0102	F2	S1003	M2
				S1004	F1

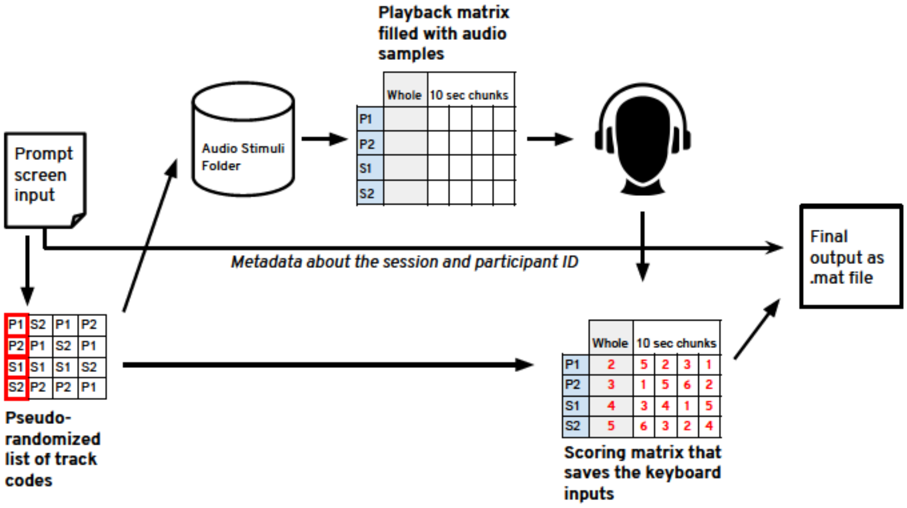
Instruments The experimental sessions were conducted in Audiolab 1 of the Faculty of Psychology and Neuroscience in Maastricht University. The actual setup was placed in an acoustically isolated cabinet room, where the participant screen was duplicated and also streamed on one screen next to the control computer where the session was administered. The research assistant was able to monitor the screen of the participant as the session continued. The desktop computer was connected to the outboard audio interface M-Track Eight (M-Audio, US). The scripts were running in `MATLAB version R2012B` (The MathWorks Inc., US). Audio samples are trasmitted are transmitted to the audio interface which converts from digital domain to analog and sends out the signal. For the experiment playback the headphone outputs of the interface was used. The sound goes directly to the headphones Sennheiser HD 600 (Sennheiser electronic GmbH, Germany).

Procedure For each participant, three sessions were scheduled. Subjects had to listen German poems and stories in two ways; the first time they listened to the whole story or poem and the second time they listened to the same story or poem cut in 10 sec chunks. After each poem or story and after each chunk of them as well, the participant was asked to evaluate the rhythmicity of the part they had just heard. Participants rated the rhythmicity of German poems and stories on a Likert scale ranging from 1 to 7 [82]. Meanings of these responses are

Table 3.4: Descriptions of Likert scale ratings as they appear on the experiment screen

Input	Response	
	German(original)	English(translation)
1	gar nicht rhythmisch	not rhythmic at all
2	kaum rhythmisch	barely rhythmic
3	ein wenig rhythmisch	a little bit rhythmic
4	mittelmässig rhythmisch	moderately rhythmic
5	gut rhythmisch	well rhythmic
6	sehr gut rhythmisch	very good rhythmic
7	perfekt rhythmisch	perfectly rhythmic

as in Table 3.4. At the beginning of every session, participant was instructed both orally and in writing. The sound pressure was adjusted by the subject to a comfort level and remained unchanged throughout the experiment. Then, the participant was asked to remain seated in the acoustically treated room and waited for the prompt screen to appear on their screen.


Playback matrix filled with audio samples
Prompt screen input
Audio Stimuli Folder
Metadata about the session and participant ID
Final output as .mat file
Pseudo-randomized list of track codes
Scoring matrix that saves the keyboard inputs

P1	S2	P1	P2
P2	P1	S2	P1
S1	S1	S1	S2
S2	P2	P2	P1

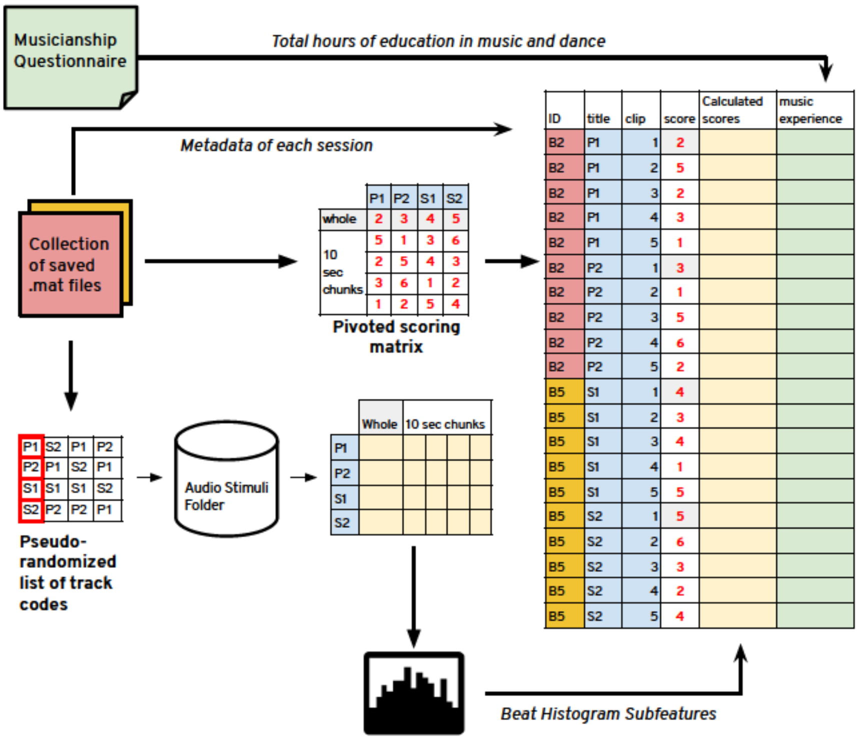
	Whole	10 sec chunks			
P1					
P2					
S1					
S2					

	Whole	10 sec chunks			
P1	2	5	2	3	1
P2	3	1	5	6	2
S1	4	3	4	1	5
S2	5	6	3	2	4

Figure 3.1: Concept model of the data collection process

The prompt screen launches once the experiment script is activated in MATLAB. It asks for the participant ID, the age, the list number and the session number. Participant ID is the pre-allocated code for each participant. The session number ranges from 1 to 3 and indicates the list of stimuli that are for the playback (1: List A, 2: List B, 3: List C). The list number ranges between 1 - 6 and calls the order of file names that are pseudo-randomized in a matrix. Thereafter, files that carry these names, are collected from the storage folder and placed into a cell array, where each row is dedicated to one audio file respectively. The first column includes the whole-length audio samples. Audio samples are then spliced into 10 second long chunks using the buffer function with zero overlap and distributed in the following columns. Based on the approximate durations of the spoken language units (see Figure 1.2 in Section 1.1.2), we assumed that 10 seconds will be adequate to cover multiple sentences, and thus, allow to listener larger prosodic groupings. This cell array is used for the audio playback. Simultaneously, the selection in the pseudo-randomized list also generates another matrix with the same structure and size in order to store the rhythmicity ratings of the test subject. Likewise, the reaction time for each scoring is also saved in a separate matrix. However, this information is not used within the framework of the thesis. After the completion of the session, the scoring matrix is saved in combination with the metadata collected in the prompt screen as well as the session date. This workflow is illustrated in Figure 3.1 with exemplary arrays of numbers.

3.2 Data Handling


Musicianship Questionnaire
Total hours of education in music and dance
Metadata of each session
Collection of saved .mat files
Pivoted scoring matrix
Pseudo-randomized list of track codes
Audio Stimuli Folder
Beat Histogram Subfeatures

	P1	P2	S1	S2
whole	2	3	4	5
10 sec chunks	5	1	3	6
	2	5	4	3
	3	6	1	2
	1	2	5	4

P1	S2	P1	P2
P2	P1	S2	P1
S1	S1	S1	S2
S2	P2	P2	P1

	Whole	10 sec chunks			
P1					
P2					
S1					
S2					

ID	title	clip	score	Calculated scores	music experience
B2	P1	1	2		
B2	P1	2	5		
B2	P1	3	2		
B2	P1	4	3		
B2	P1	5	1		
B2	P2	1	3		
B2	P2	2	1		
B2	P2	3	5		
B2	P2	4	6		
B2	P2	5	2		
B5	S1	1	4		
B5	S1	2	3		
B5	S1	3	4		
B5	S1	4	1		
B5	S1	5	5		
B5	S2	1	5		
B5	S2	2	6		
B5	S2	3	3		
B5	S2	4	2		
B5	S2	5	4		

Figure 3.2: Concept model of the data handling process

Collected .mat files are decomposed to manipulate the scoring matrix and the extraction of metadata. Variables such as item, participant, chunk, rhythmCond, session, Age_in_years and list were extracted from this metadata. Especially rhythmCond (poem vs. story) and chunk (whole vs. 10 sec) are thoroughly used in the later stages of the analysis.

The rhythmicity scorings were given by the participants using the number keys from 1 to 7 on the keyboard and therefore saved in the scoring matrix as integer. The first step of data handling process started with pivoting of this matrix and was followed by concatenation of the stimuli in order to have one scoring in each data row. Same procedure was applied at all sessions and then dataframes of all sessions are joined vertically to construct the final dataframe. Concurrently, a separate cell array was generated for each of three lists to store the audio samples similarly to the previous step in the experiment playback. This one, however, was used to serve as a collection of input files to enable the calculation of beat histogram subfeatures. Subfeatures were calculated in two steps:

1. Extraction of beat histogram for each audio sample using the base algorithm that has been implemented in the procedures of [83].

2. Computation of subfeatures from each beat histogram analogous to the audio feature extraction routines.

The subfeatures are as follows:

- **Peakiness** represents the inverse of the spectral flatness. It is calculated simply by subtracting the spectral flatness from 1, since it denotes the distribution characteristics with a value between 0 and 1. In the context of beat histograms, It allows us to see if certain BPMs behave “peaky”, in other words, stand out.

- **Variance** gives of the squared difference of the occurrence values of each BPM in histogram from the mean. In this regard, it estimates how far y-axis values are spread out from their average value. As a boundary condition, one could say the variance equals to zero if all BPMs had the same number of occurrences. In contrast, the rhythmicity of the signal should raise as the variance in its beat histogram increases.

- **Centroid** marks the center-of-mass for of the BPM distribution and indicates dominant region of BPMs. In the lack of a single dominant BPM, the centroid gives a more robust information about the speech rate than the peak of the beat histogram. This is, therefore, rather a measure of the speech rate and not of rhythmicity. It is included in order to enable the comparison that was stated in the Section 1.2.

The calculations of these subfeatures are repeated on five beat histogram versions having five different novelty functions (*MFCC, RMS, SF, SFL, SPC*). For the selection of the novelty functions, it was aimed to include one from each group of audio features. With this list of features, we consider trajectories of intensity, spectral shape(both linear and Mel-scale), tonalness, and perceptual tonality. Before starting with the comparison of these variables, on each variable Z-score standardization is applied in order to avoid outlier issues. This way, data is centralized around zero with a rescaled distribution.Since mutual information calculation requires discrete data types, subfeatures are, then, also recoded into new subfeatures arrays with a 7-bin discretization. Subsequently, 15 arrays of numeric and 15 arrays of integer data are pivoted and merged with the main dataframe.

Table 3.5: Complete set of variables categorized under data type

factor	**integer**	**numeric**
item	session	total_musical_expertise_hours
participant	list	peakiness_MFCC
chunk	clipindex	peakiness_RMS
rhythmCond	Age_in_years	peakiness_SF
musicianship	Amount_of_Languages	peakiness_SFL
	P_MFCC	peakiness_SPC
	P_RMS	var_MFCC
	P_SF	var_RMS
	P_SFL	var_SF
	P_SPC	var_SFL
	V_MFCC	var_SPC
	V_RMS	centroid_MFCC
	V_SF	centroid_RMS
	V_SFL	centroid_SF
	V_SPC	centroid_SFL
	C_MFCC	centroid_SPC
	C_RMS	
	C_SF	
	C_SFL	
	C_SPC	

As the final step, the `LimeSurvey` output of the MMHQ was processed to calculate the total hour spent in the training in vocal, instrumental and dance performance. Resultant number is saved under the variable name `total_musical_expertise_hours`. Separation of focus groups in musicianship is accordingly defined as another variable. In addition to that, the number of languages subject can speak is also captured.

Eventually, we have generated a [1269 x 49] dataframe storing the relevant variables we aimed to have for further evaluation. This dataframe is then processed and analyzed in different software units depending on the purpose. Calculation of non-linear associations are done with using the dedicated files [84]. Pattern Recognition and Machine Learning Toolbox [85] Statistical analysis was conducted in `SPSS Statistics - Subscription Build Number 1.0.0.1347` (IBM Corp.,US) and `R` [86] using `Rstudio version 1.2.1335` (RStudio PBC, US) as the IDE. Figures were produced using the package `ggplot2` [87].

4 Evaluation

In the first section of the chapter, it is aimed to put perceptual assessment comparison versus automatic machine-retrieved information and to test the plausibility of computational methods in speech rhythm description. In the second part, we assessed the significant differences in perceptual scorings using the control factors based on the poem-story and whole-10 seconds length comparison. The results of the statistical analysis are shown in following Tables (4.1 - 4.5) as well as Figures (4.1 - 4.5) and interpreted accordingly.

There are number of measurement methods to seek that relationship, however they differ from each other in their basic assumptions and implementations. First part comprises linear correlation, mutual information and cosine similarity [88, 89]. In the second section, the non-parametric tests are applied and tabulated to show the influence of the stimulus properties on the rhythmicity scorings. As the non-parametric test, we have used Mann-Whitney test, since it delivers consistent results for ordinal variables and unequal sizes of groups [90].

The mean rhythmicity scoring from the behavioural experiment for musician group ($M = 4,44; SD = 1,64$) was greater than the mean of non-musician group ($M = 4,37; SD = 1,59$). There has been sixteen ($n = 16; \%64$) non-musicians and nine ($n = 9; \%36$) musicians identified between the test subjects. Hence, we collected 8051 rows of the data frame from the non-musicians and 4644 rows from the group that is labeled as musician. If we look from the stimulus side, 7048 rows of the dataframe is coming from the poems and 5647 are from the stories. The mean rhythmicity scoring from the behavioural experiment for for poems ($M = 5,22; SD = 1,184$) was greater than the mean of non-musician group ($M = 1,463; SD = 1,59$). More extensive comparison of the groups is provided in the following sections.

4.1 Validation of Rhythmicity Subfeatures

4.1.1 Results

As described in Section 3.2, we have attempted to introduce various subfeatures of the beat histogram. With a hypothesis driven approach, these subfeatures were defined to describe rhythmicity wit regard to the shape of the histogram. Peakiness and Variance were expected to capture whether signal has emphasized BPMs, in other words periodities. An increasing Centroid, on the other hand, indicates the presence of rather higher BPMs or small IOIs. Since

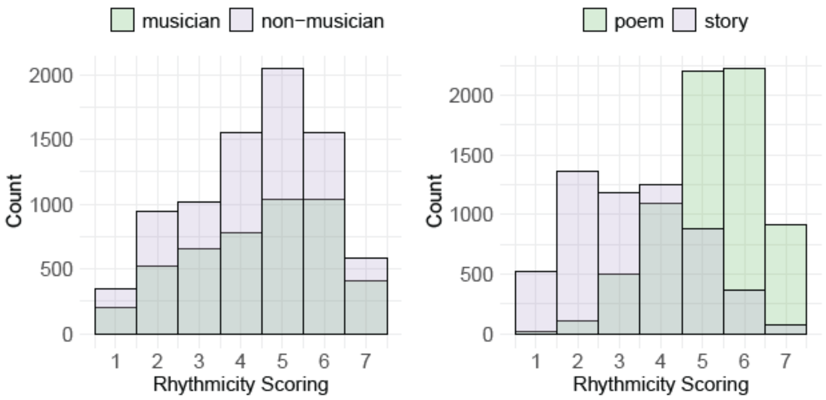


Figure 4.1: Rhythmicity scoring distribution for two main factors of interest: musicianship (left) and stimulus type (right)

we have extracted beat histograms from five different novelty function, we ended in having fifteen different subfeatures to be compared with perceptual scorings in order to measure their relative strength in representing the perceived rhythmicity. The tables in this section are constructed in a way that each cell corresponds to the measure of association between the calculated subfeature and the scores collected from behavioral experiments. Since the behavioral side of the compared values always stays the same, the change in the measured value differs solely with the change of the other variable that is incorporated in the analysis. Thus, we decided to list the values in a 2 dimensional matrix form to enable the reader to visually group similar values with respect to novelty functions and subfeatures. For each cell, the subfeature's name and its novelty function can be read from the very left column and the top row respectively. Additionally, the investigated group of subjects is on the top left corner of each table. For mutual information and cosine similarity, the values listed in Tables 4.2 and 4.3 are visualized with stacked line charts in Figures 4.2 and 4.3 to compare the accuracy of all subfeatures and experiment performance of each group of interest.

Given the variables are ordinal, Spearman's ρ was used in the linear correlation analysis. As it is shown in the Table 4.1, there is a statistically significant ($p < 0,01$) correlation for all novelty functions except for MFCC ($p < 0,05$) for the analysis for all and non-musician groups) and SF (for Centroid). The correlation factors for SF, SFL, and SPC appear to be negative, meaning a inverse linearity. Nevertheless, for MFCC and RMS, the values are positive. Generally, the values tend to increase for the non-musician subjects as opposed to musicians. Moreover, the values for Centroid show a better performance is showing a linear relationship with the scoring.

Figure 4.2 shows the variation in the mutual information that each subfeature has with scorings from the experiment. In all three subfeatures, novelty function based on MFCC appears to demonstrate relatively higher mutual information. SFL. For Variation, SFL reaches to 0,12 for *all subjects*, nearly two times the average trend that is seen in the other

Table 4.1: Linear correlation between rhythmicity scorings and the subfeatures extracted from different Beat Histograms versions

Beat Histogram Subfeatures	**Novelty function versions**				
	MFCC	RMS	SF	SFL	SPC
(For all participants)					
Peakiness	**,021***	**,041****	-,142**	-,194**	-,080**
Variance	**,021***	**,045****	-,138**	-,352**	-,157**
Centroid	**,202****	**,143****	-,013	-,192**	-,088**
(For non-musicians only)					
Peakiness	**,024***	**,038****	-,176**	-,228**	-,108**
Variance	**,024***	**,042****	-,173**	-,403**	-,202**
Centroid	**,231****	**,155****	-,02	-,222**	-,115**
(For musicians only)					
Peakiness	,015	**,043****	-,085**	-,137**	-,034*
Variance	,015	**,047****	-,080**	-,268**	-,082**
Centroid	**,151****	**,121****	-,004	-,142**	-,043**

$p < 0,05$ *, $p < 0,01$ **

Table 4.2: Mutual Information between rhythmicity scorings and the subfeatures extracted from different Beat Histograms versions

Beat Histogram Subfeatures	**Novelty function versions**				
	MFCC	RMS	SF	SFL	SPC
(For all participants)					
Peakiness	**0,0754**	0,0413	0,0424	0,0430	0,0412
Variance	0,0773	0,0471	0,0319	**0,1167**	0,0513
Centroid	**0,0829**	0,0324	0,0350	0,0549	0,0480
(For non-musicians only)					
Peakiness	**0,1030**	0,0595	0,0588	0,0564	0,0542
Variance	0,1037	0,0652	0,0447	**0,1615**	0,0861
Centroid	**0,1134**	0,0442	0,0481	0,0743	0,0795
(For musicians only)					
Peakiness	**0,0552**	0,0418	0,0323	0,0361	0,0400
Variance	0,0581	0,0440	0,0291	**0,0688**	0,0364
Centroid	**0,0660**	0,0358	0,0265	0,0432	0,0303

subfeatures. When the musicianship considered, mutual information curves follow a fairly similar trajectory. Nevertheless, the level of the curve is higher for non-musician on the entire range of subfeatures.

In a similar manner, Cosine Similarity values from the Table 4.3 are also plotted against the discretely grouped subfeatures. In this graph, there is no clear variation between the groups of musical expertise. Nevertheless, there is a distinctive increase in the score for MFCC and

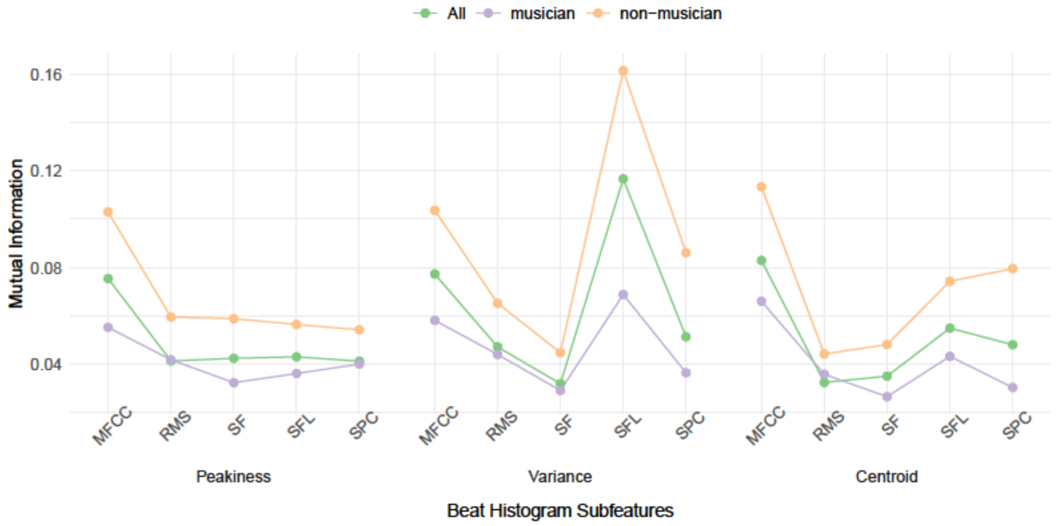


Figure 4.2: Mutual Information comparison of beat histogram subfeatures based on different novelty functions

RMS and a drop for SF in all three subfeatures. Besides for every novelty function, Centroid appears to be relatively higher than the values of Peakiness and Variance.

Table 4.3: Cosine similarity between rhytmicity scorings and the subfeatures extracted from different Beat Histograms versions

Beat Histogram subfeatures	**Novelty function versions**				
	MFCC	RMS	SF	SFL	SPC
(For all participants)					
Peakiness	**0,8409**	0,8093	0,6964	0,7523	0,7712
Variance	**0,8357**	0,8083	0,7498	0,7927	0,8190
Centroid	0,9089	**0,9111**	0,7571	0,8522	0,7961
(For non-musicians only)					
Peakiness	**0,8435**	0,8115	0,6914	0,7494	0,7727
Variance	**0,8388**	0,8106	0,7452	0,7849	0,8142
Centroid	0,9113	**0,9121**	0,7548	0,8490	0,7938
(For musicians only)					
Peakiness	**0,8365**	0,8057	0,7050	0,7574	0,7687
Variance	0,8307	0,8045	0,7579	0,8060	0,8274
Centroid	0,9050	**0,9096**	0,7612	0,8578	0,8001

4.1.2 Discussion

The calculated rhythmicity ratings Peakiness and Variance were hypothetically defined to represent rhythmicity in Beat Histogram. Nonetheless, their similarity in all Tables has verified the correctness of our approach.

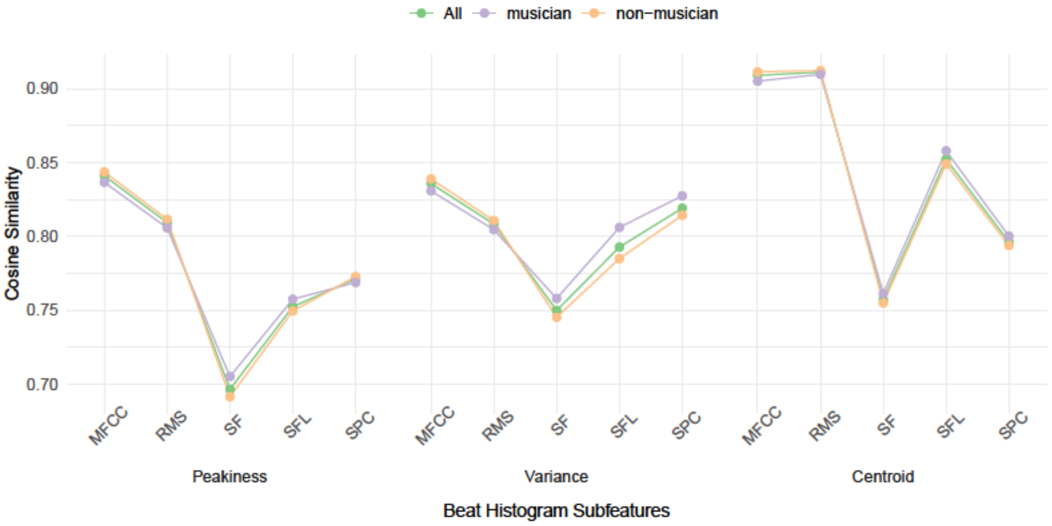


Figure 4.3: Cosine Similarity comparison of beat histogram subfeatures based on different novelty functions

The values for the linear correlation results could be interpreted with their absolute values ranging between -1 and 1 in a similar way with Pearson's correlation factor, whereas the values for mutual information and cosine similarity should be read always in a relative way, not as a their absolute values. Therefore in the following discussion, the features are always compared with each other within the similarity measure they result from.

We have applied three different methods to display the strength of association between the calculated and perceived ratings. We have predicted to see positive linear correlation for the selected subfeatures, which however the most of the results fail to deliver as show in the Table 4.1. Although most of the subfeature versions have a positive correlation factor, only MFCC and RMS resulted in having a positive correlation significance ($p < 0{,}05$). Particularly, the third subfeature, Centroid, has shown a relatively higher correlation for the aforementioned two novelty versions. Again, Centroid was implemented analogous to the audio feature Spectral Centroid. One can consider an increasing Centroid as more presence of higher tempi and therefore shorter IOIs. Thus, it would be fair to conclude that test subjects tended to associate the signals having “on average” denser pulse streams with mental abstraction of rhythmicity. This result ties well with previous studies wherein high-gamma autocorrelations in auditory areas in the superior temporal gyrus and in frontal areas of the brain significantly matched the both attended and imagined musical rhythms [91]. This is an important finding in the understanding of the relevant frequency ranges of speech units that engages neural oscillations, as previous studies have highlighted the dominance of lower frequency patterns that are induced by the signal envelope [45].

It is well known among empirically working social scientists that the reference of 'correlative findings' to 'causal relationships' poses many problems. The hint that a causal relationship can also be concealed by a third variable that is heard less frequently [92]. In this case a causal relationship only manifests itself in a correlation if the third variable is controlled experimentally or statistically, which becomes difficult where rhythm perception is proven to be a complex network of cognitive tasks [43, 47]. Valsiner [93] emphasizes, correlations provide information about samples or, in generalized terms, about populations, but not about the individual observation units. At this point, mutual information (MI) becomes useful, since it is rather a favored measure of the relatedness between two random variables. It is conventionally calculated by their joint probabilities estimated from the frequency of observed samples between some set of response alternatives (e.g Likert scale) [94]. Nevertheless, the estimation of MI, especially for continuous data, relies strongly on the selection of the partitioning parameters involved such as the number of bins [95]. Considering the potential instability of the method, cosine similarity has been employed to crosscheck the earlier results. Hybridization of two methods has been suggested in the machine learning literature in the context of classification accuracy improvement [89]. If we look at their results in Figures 4.2 and 4.3, it can be argued that MI gives more insight about the difference between groups, whereas Cosine Similarity displays the performance difference between subfeatures more clearly. In Figure 4.2, the yellow curve, which represents non-musicians, lies above the purple curve that indicates the values for musicians, with a clear separation. This means the correlation of subfeatures with the perceptual ratings are even higher for non-musicians which contradicts the initial hypothesis. In Figure 4.3, it gets even more difficult to speak of a group's superiority, as curves of two groups follow very similar paths. The ranges do not vary for different groups, and therefore the differentiation between the cosine similarities of subfeatures is more pronounced and generalizable.

Regarding the influence of musical experience, there was no increase in relatedness for subjects having music training in their past. In contrast, previous studies has shown significant influence of music experience on speech rhythm perception and suggest shared neurocognitive resources for rhythm perception in speech and language [96, 97, 98]. Most of these studies were experimenting with highly controlled and reduced experimental designs in terms of listening conditions and stimulus types. Yet, in line with the ideas of [99], we incorporated more naturalistic experimental designs. Besides, the present experiment design has not demanded any motor tasks. For that reason, these results still do not disprove the theories stating that musical experience modulates perceptual effects of speech rhythm. It is important to note that subjective judgement of participant does not necessarily represent their skills of identifying the rhythmic pattern. Conversely, it might just be that these specific features (novelty functions, beat histograms and subfeatures) do not capture rhythm in the same way as participants do. People might indeed have a common processing of rhythm for speech and music on a higher cognitive level, but it is not traceable by these features.

Concerning the parametrization efforts of rhythmicity, MFCC-based subfeatures outperformed others in all relatedness measures. Since MFCCs are known with its strength in representing perceptual properties of a signal, it was hypothesized that they will end in a better matching with the human evaluation. Initially, they were introduces by Davis [100] as a tool for speech recognition systems. MFCCs have been classified as a perceptual feature due to the fact that it approximates some parts of both the speech production (source-filter model) and the auditory processing in the cochlea [101]. Without applying a rhythm analysis, MFCCs were already a tool in use for language classification [102]. On the rhythm-based language classification studies the focus has lied rather on amplitude and spectral shape features (e.g RMS, SF) [9]. However on the rhythm-based genre classification side, results have shown high accuracy rates for MFCCs [103], which was already a great incentive for the current study. As shown in Figure 4.3, RMS novelty function remains as a good indicator of the rhythmic progress of a signal, however SF novelty functions exhibited lower similarity with the perceived rating. Apparently, human hearing mechanism for speech is not as sensitive as it is for others, to the dynamic range of Spectral Flux, although it stood out as one of the most accurate features in [9]. Another interesting point in Figure 4.2 was that the drastic increase of MI for Variance based on Spectral Flatness. A similar peak for SFL was observed in the context of language accuracy in [9].

We have also had difficulties with recruiting professional musicians in Maastricht, and thus the musicianship of the participants were solely judged by the questionnaire results. The separation between non-musicians and musician was adequate in terms of musical training hours. Although the music questionnaire was collecting the musical experience as a composite score that combined information about hours spent in vocalic, instrumental and dance training, it suffers from a narrow perspective on quantifying the musical experience. We believe, in times where people have access to information and artistic content more than before, the instrumental and vocal training surveys are slightly outdated to reflect the modern ways of self-education. A recent study on the learning practices of electronic music producers highlights dedicated practice, extended listening, and motivation without having any connection formal practices of musical training [104]. In music genres like hip-hop or electronic dance music, where rhythmic instruments and vocal use plays the biggest role, most people become professionals without having a formal music training. Especially when the the concept of rhythm is investigated, these individuals should become more and more a point of interest.

4.2 Influence of Stimuli Properties

4.2.1 Results

In this section, we examined the effect of stimuli properties on the rhythmicity scorings and considered stimulus type and length as the control factors. Stimulus type corresponds the separation of data into poems and stories, whilst the length factor is evaluated with the differentiation between whole length and 10 second chunks. In order to compare mean rank we applied non parametric tests in both aspects of the analysis. Additionally, the distribution of scorings are visualized in Figures 4.4 and 4.5.

Firstly, for the variable `scoring`, we have observed whether the types poem and story play a role; the distributions in the two groups differed significantly (Mann–Whitney U = 6862746, n_{poem} = 7049, n_{story} = 5647, p < 0,01). A substantial difference has been found between the mean ranks showing that poems (8197,78) were rated higher that the stories (4039,29). Same procedure has been applied for the calculated subfeatures to see the same sort of differentiation was reflected in them as well. The mean ranks for all variables mentioned are listed in Table 4.4. Secondly, similar Mann-Whitney test was conducted for the the stimulus length. The distributions in the two groups differed significantly (Mann–Whitney U = 4091865, n_{whole} = 763, n_{10sec} = 11932, p < 0,01). However, the differentiation was not significant for the calculated subfeatures. The result can be seen in detail in Table 4.5.

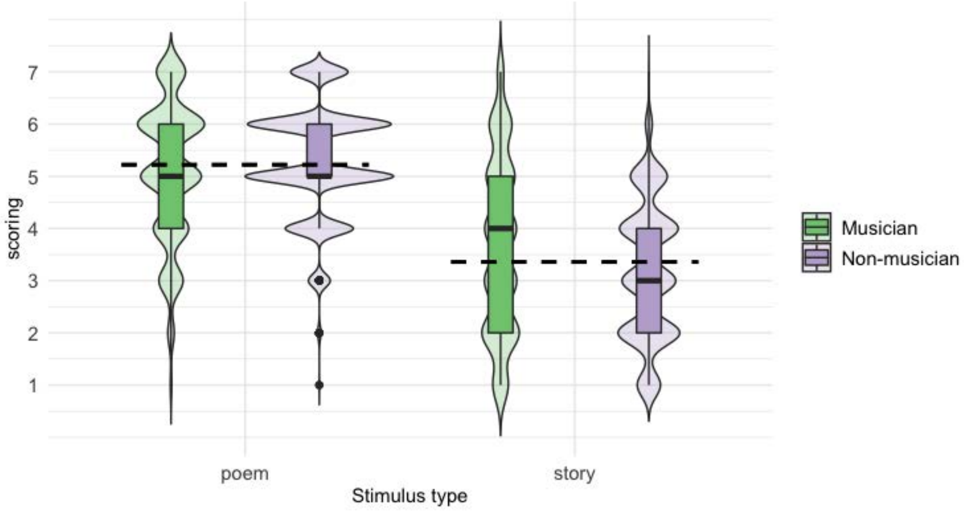


Figure 4.4: Violin and box plot graphs showing the influence of the stimulus type with the consideration of musicianship

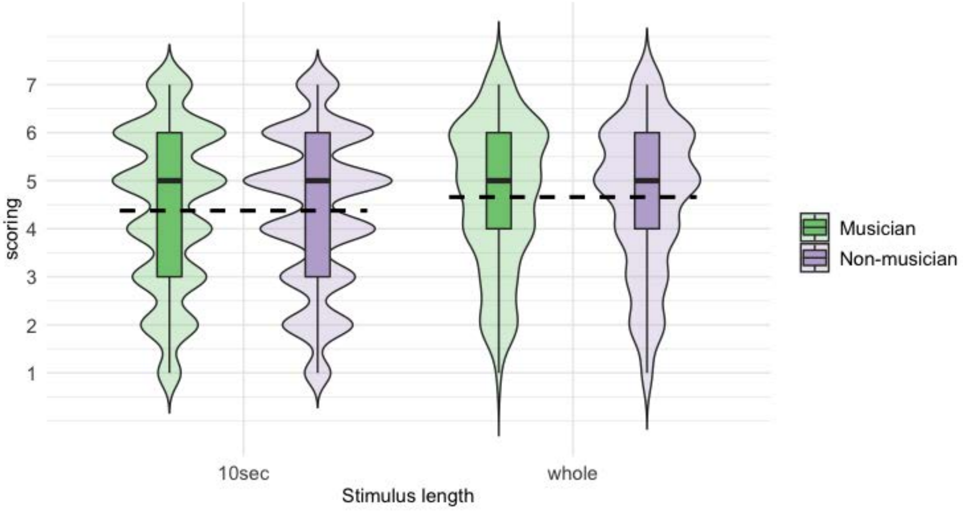


Figure 4.5: Violin and box plot graphs showing the influence of the stimulus length with the consideration of musicianship

Table 4.4: Nonparametric Test on the stimulus type

Variable name	**Mann-Whitney U**	**Z**	**Sign. (2-tailed)**	**Mean Rank**	
				poem	**story**
scoring	6862746	-64,625	0	**8197,78**	**4039,29**
P_MFCC	19462962,5	-2,203	0,028	6285,99	6425,4
P_RMS	19791570	-0,544	0,587	6332,61	6367,21
P_SF	15883903	-25,238	0	5778,18	7059,2
P_SFL	17045569	-15,168	0	5943	6853,48
P_SPC	19029682	-4,525	0	6471,49	6193,87
V_MFCC	19628967,5	-1,357	0,175	6309,54	6396
V_RMS	19821053	-0,395	0,693	6359,21	6334,01
V_SF	16746142	-16,476	0	5900,51	6906,51
V_SFL	8176987	-64,347	0	4684,69	8423,98
V_SPC	14433582	-27,561	0	5572,4	7316,03
C_MFCC	12944110,5	-34,755	0	7334,93	5116,21
C_RMS	15847710	-20,721	0	6922,96	5630,39
C_SF	19067757	-4,341	0	6229,91	6495,38
C_SFL	14595609,5	-27,292	0	5595,39	7287,33
C_SPC	16745212,5	-15,755	0	5900,38	6906,67

Table 4.5: Nonparametric Test on the stimulus length

Variable name	**Mann-Whitney U**	**Z**	**Sign. (2-tailed)**	**Mean Rank**	
				whole	**10-sec**
scoring	4091865	-4,77	0	**6951,14**	**6309,43**
P_MFCC	4508735	-0,457	0,648	6291,22	6351,63
P_RMS	4548603	-0,036	0,971	6352,53	6347,71
P_SF	4390495,5	-2,123	0,034	6559,75	6334,46
P_SFL	4475231	-0,854	0,393	6448,69	6341,56
P_SPC	4317392	-2,551	0,011	6655,56	6328,33
V_MFCC	4510441	-0,436	0,663	6293,46	6351,49
V_RMS	4517370	-0,363	0,717	6393,46	6345,09
V_SF	4445671,5	-1,162	0,245	6487,43	6339,08
V_SFL	4493186	-0,676	0,499	6270,84	6352,93
V_SPC	4343875	-2,195	0,028	6620,85	6330,55
C_MFCC	4496318,5	-0,582	0,56	6421,05	6343,33
C_RMS	4473892,5	-0,836	0,403	6450,44	6341,45
C_SF	4424938,5	-1,386	0,166	6514,6	6337,35
C_SFL	4510141,5	-0,451	0,652	6402,94	6344,49
C_SPC	4312352	-2,503	0,012	6662,16	6327,91

4.2.2 Discussion

Results show that type of the stimulus creates a contextual illusion in rhythm perception. Although measured variables do not display a similar differentiation, perceptual scoring appears to differ strongly. This is a strong indicative for the fact that the listeners classify the stimulus in their minds as poems once they hear the rhymes and metrical structure. We relate this to the incentive triggered by the potential variations of speed in which a poem is likely to be read. This speed is influenced particularly by certain pauses, expansions, vowel lengths and consonant clusters. In the recordings, we aimed to minimize the theatrical effects to avoid periodic patterns of build-ups and releases. Loots [105] suggests that, in poetry reading, once the metrical pattern is recognized, the reader retraces the grid that author sets. He also emphasis that the foot boundaries could be minimally realized in the sense that they were triggered by lengthened syllables even if the speech pauses or pitch changes do not take place. When we compare the mean ranks for the other variables that are listed in Table 4.4, it is clear that the scorings are biased in a way that the poems are score with higher ratings, although no difference was in computed values. Removal of linguistic components by low pass filtering has been introduced in infant studies and provided consistent results [106]. Although it contradicts with the naturalistic experiment paradigm, overlaying the signals' envelope curves with sine and noise carrier signals has helped other behavioral studies to get unbiased results [107, 108]. However, other studies like [109] points out the pragmatic sides of combining rich contextual setting in experiments when neurobiology and psychology of language is concerned. Hence, there is still no consensus on the right methodology of behavioral testing of language rhythm. Similarly, a significant mean rank difference was found in the second part of the non-parametric test. The idea that longer segments establishes prosodic groupings has been a topic of discussion [36, 37]. Although, some studies propose that human perception for grouping is limited in duration [39], there is also substantial amount of evidence showing that longer sequences of words enables the listener to predict the coming speech segments [110]. Therefore, it could be deemed likely that increase in the feeling of predictability has lead to these results in the experiment. All in all, our hypothesis on how people's decision may influenced by stimulus' properties has been validated.

5 Conclusion

5.1 Contribution

Rhythm points to a sensual or felt course of events. Events must be distinguishable and at the same time create a flow or sequence in which difference in the course of events is created. There is no rhythm without articulation or difference and no rhythm without continuity or flow, in which difference arises precisely from the process of connecting. The articulation and flow of the speech rhythm not only takes place in the bodily movements of the speaker, but also in the experience or feeling of the listener. With this thesis, we have drawn attention to the usability of novel approaches introduced in music analysis domain in order to set mathematical descriptions for this so-called experience. Aspects of speech rhythm can be quantifiable in different ways. We propose to examine the frequency distribution of periodicities in a speech signal using beat histograms and to evaluate if the way through the suggested features is sensible. The change of instantaneous values and intensities in the signal guides our perception to a certain extent. Yet, how exactly human brain processes and classifies remain them as a trivial topic - for instance, speech can be seen as more or less rhythmic; where there is less change, less variation, one may think it still constitutes less rhythm. This consideration needs to be continuously validated with innovative psycho-physical paradigms. Only if we grasp the concept of the rhythmicity as perception in a broader sense, both human-human and computer-human communication can go far beyond the question whether rhythm is artistic or internal in the strict sense.

Throughout this study, we have analyzed speech rhythm perception of musicians and non-musicians with a behavioural experiment and results have shown that that the conscious judgement of speech rhythmicity does not vary with the factor of being a musician. Nevertheless, in parallel to the main goal of the thesis we have investigated different variations of rhythm analysis methods by including novelty functions based on different audio content features and perceptually measured the relevance of these methods through hypothetically defined rhythmicity subfeatures. We have implemented tools of music data analysis for a problem that was analysed in the framework of psychology and linguistics and brought a new perspective to the speech rhythm description. As we provide our motivation for our research on rhythm, we have also provided a comprehensive review of literature that pertains to the commonalities of language and music across different social and technical fields and collected evidence to show why further comparative research on these two phenomena should continue.

With regard to the effect of music training, we have not identified any difference of correlation between the musicians and non-musicians. For the most part of the results, MFCC and RMS novelty functions were showing better performance in reflecting the rhythmicity. Unlike other spectral novelty functions, variance of the of SFL based beat histograms was interestingly showing high mutual information with people's responses. Finally, we have seen how test subjects have been influenced by the context and the meaning of the stimulus and consider poems as more rhythmic when they were asked to rate rhythmicity for poetry and story. We have also seen that long an complete stimuli were scored with higher rating as opposed to 10 second chunks.

Despite practical limitations, this thesis has contributed to speech rhythm research by introducing the strength of automatic beat detection algorithms to the field of spoken language rhythm which has remained simplistic for a long time by usually focusing on the duration changes. We have also suggested multiples novelty functions, which were expected to convey information on the energy changes and intonations of speech with a more sophisticated formulation that the features that are usually used in linguistics studies. This approach not only adds more depth to the acoustic dimensions of the language analysis, but also eliminates the manual annotation of segments that has been a necessity as well as a burden in linguistic research.

5.2 Future Outlook

We believe findings in the relation to musical rhythm can foster many new ideas for further applications. Though, the grouping of the test subjects with respect to their musicianship requires primarily a better definition of musicianship notion per se. Especially, when referring to the rhythm cognition and its relation to the human speech production skills, effects of close engagement with music becomes a great point of interest. However this engagement includes multiple layers such as dancing, listening, training, and performing. Furthermore these layers also differ in cultures, genres and platform. Therefore, limiting the musicianship only to the measures of formal training hours fails to describe the actual engagement level. Experiments targeting different levels of music engagement could could get more specific by applying more specific questionnaires or musical aptitude tests.

Except for infant experiments, the native language of the test person becomes a crucial factor in their perception. In the naturalistic experiments where realistic stimuli are used, native speakers would always apply their knowledge to the processing of which cues are more important than others and accompany the noticeable regularity. Even the adults who are not familiar with the language in use, would perceive the the regular pattern by comparing with their own language phonology. Therefore, a complex effect of native language is undeniable and further work is certainly required to disentangle these complexities of language-specific effects. Test subjects who were recruited for the present study could speak at least two languages, however these languages were limited to German, English, and Dutch. Clearly, all three of these languages are very similar in term of rhythm structure (stress-timed). Experimental studies could be widened with people who speak only syllable-timed languages (e.g. French, Italian, Spanish, Turkish) and people who speak at least one from each group of languages. Since syllable-timed languages demonstrate less vocal reduction, it may conform feature-based beat detection algorithms. This provides a good starting point for further research.

The selection of instantaneous features and rhythm analysis features were adopted from the recent language identification studies [71]. However, in the music rhythm analysis there have been vast amount of methods developed to capture the rhythmicity [111]. Future studies should aim to replicate results in a larger set of features and look into rhythm descriptors that represents rhythm changes over time. Analogous to the the transformation between DFT Magnitude Spectrum and STFT Spectrogram, time dependant representations of Beat Histogram could increase the set a new ground of rhythm information, especially for highly irregular signal types like speech.

Bibliography

[1] Mithen, Steven J (2007): *The Singing Neanderthals: the Origins of Music, Language, Mind and Body.* Harvard University Press.

[2] Levitin, Daniel J. and Susan E. Rogers (2005): “Absolute pitch: perception, coding, and controversies.” In: *Trends in Cognitive Sciences*, **9**(1), pp. 26–33.

[3] Fuchs, Robert (2016): *Speech Rhythm in Varieties of English.* Prosody, Phonology and Phonetics. Berlin, Heidelberg: Springer Berlin Heidelberg.

[4] Fitch, W (2013): “Rhythmic cognition in humans and animals: distinguishing meter and pulse perception.” In: *Frontiers in systems neuroscience*, **7**, p. 68. Publisher: Frontiers.

[5] Fonollosa, Jordi; Emre Neftci; and Mikhail Rabinovich (2015): “Learning of Chunking Sequences in Cognition and Behavior.” In: *PLOS Computational Biology*, **11**(11), p. e1004592.

[6] Ramus, Franck; Marina Nespor; and Jacques Mehler (1999): “Correlates of linguistic rhythm in the speech signal.” In: *Cognition*, **73**(3), pp. 265–292. Publisher: Elsevier.

[7] Grabe, Esther and Ee Ling Low (2002): “Durational variability in speech and the rhythm class hypothesis.” In: *Papers in laboratory phonology*, **7**(515-546).

[8] Galves, Antonio; Jesus Garcia; Denise Duarte; and Charlotte Galves (2002): “Sonority as a Basis for Rhythmic Class Discrimination.” In: Isabel Marlien and Bernard Bel (Eds.) *Proceedings of the speech prosody 2002 conference.* Aix-en-Provence: Laboratoire Parole et Langage.

[9] Lykartsis, Athanasios and Weinzierl Weinzierl (2015): “Using the beat histogram for speech rhythm description and language identification.” In: *INTERSPEECH 2015.* Dresden: International Speech Communication Association, pp. 1007–1011.

[10] Rousseau, J.J.; J.T. Scott; R.D. Masters; and C. Kelly (1990): *The Collected Writings of Rousseau: Essays on the origin of languages and writings related to music.* Essay on the Origin of Languages and Writings Related to Music. University Press of New England.

[11] Fraisse, Paul (1984): “Perception and Estimation of Time.” In: *Annual Review of Psychology*, **35**(1), pp. 1–37.

[12] Koelsch, Stefan and Walter A. Siebel (2005): "Towards a neural basis of music perception." In: *Trends in Cognitive Sciences*, **9**(12), pp. 578–584.

[13] Brown, Steven; Peter Q. Pfordresher; and Ivan Chow (2017): "A musical model of speech rhythm." In: *Psychomusicology: Music, Mind, and Brain*, **27**(2), pp. 95–112.

[14] Hauser, Marc D and Josh McDermott (2003): "The evolution of the music faculty: a comparative perspective." In: *Nature Neuroscience*, **6**(7), pp. 663–668.

[15] Fitch, W. Tecumseh (2005): "The Evolution of Language: A Comparative Review." In: *Biology & Philosophy*, **20**(2-3), pp. 193–203.

[16] Mcmullen, Erin and Jenny R. Saffran (2004): "Music and Language: A Developmental Comparison." In: *Music Perception*, **21**(3), pp. 289–311.

[17] Tillmann, Barbara; Jamshed J. Bharucha; and Emmanuel Bigand (2000): "Implicit learning of tonality: A self-organizing approach." In: *Psychological Review*, **107**(4), pp. 885–913.

[18] Werker, Janet F. and H. Henny Yeung (2005): "Infant speech perception bootstraps word learning." In: *Trends in Cognitive Sciences*, **9**(11), pp. 519–527.

[19] Kuhl, Patricia K. (2004): "Early language acquisition: cracking the speech code." In: *Nature Reviews Neuroscience*, **5**(11), pp. 831–843.

[20] Sallat, S. (2011): "Prosodische und musikalische Verarbeitung im gestörten Spracherwerb." In: *Sprache · Stimme · Gehör*, **35**(03), pp. e105–e111.

[21] Höhle, Barbara (2005): "Der Einstieg in die Grammatik : Spracherwerb während des ersten Lebensjahres." In: *Forum Logopädie*, **6**, pp. 16–21.

[22] Fry, Dennis Butler (1979): *The physics of speech.* Cambridge University Press.

[23] Ross, D.; J. Choi; and D. Purves (2007): "Musical intervals in speech." In: *Proceedings of the National Academy of Sciences*, **104**(23), pp. 9852–9857.

[24] Wermke, Kathleen (2008): "Melodie und Rhythmus in Bablauten und ihr potenzieller Wert zur Frühindikation von Sprachentwicklungsstörungen." In: *Logos Interdisziplinär*, **16**, pp. 190–195.

[25] Schön, Daniele; et al. (2008): "Songs as an aid for language acquisition." In: *Cognition*, **106**(2), pp. 975–983.

[26] Patel, Aniruddh D.; John R. Iversen; and Jason C. Rosenberg (2004): "Comparing rhythm and melody in speech and music: The case of English and French." In: *The Journal of the Acoustical Society of America*, **116**(4), pp. 2645–2645.

[27] Hasty, Christopher (2014): "Rhythmusexperimente – Halt und Bewegung." In: Christian

Grüny and Matteo Nanni (Eds.) *Rhythmus - Balance - Metrum.* Bielefeld: transcript Verlag.

[28] Marchetti, Christopher C (2009): *Aristoxenus Elements of rhythm: text, translation, and commentary with a translation and commentary on POxy 2687.* Ph.D. thesis. Publisher: Rutgers University-Graduate School-New Brunswick.

[29] Cosic, Irena; Qiang Fang; Elena Pirogova; and Heiko Rudolph (2005): "Biorhythms - Periodicity in living systems." In: *Proceedings of the 3rd IASTED International Conference on Biomedical Engineering 2005*, pp. 468–473.

[30] Snyder, Bob (2000): *Music and memory: an introduction.* Cambridge, Mass: MIT Press.

[31] Gabrielsson, Alf (1993): "The complexities of rhythm." In: *Psychology and music: The understanding of melody and rhythm*, pp. 93–120.

[32] Lerdahl, Fred and Ray Jackendoff (2010): *A generative theory of tonal music.* Repr. Cambridge, Mass.: MIT Press.

[33] Essens, Peter J. and Jan Povel (1985): "Metrical and nonmetrical representations of temporal patterns." In: *Perception & Psychophysics*, **37**(1), pp. 1–7.

[34] Cooper, Grosvenor and Leonard Meyer (1960): *The rhythmic structure of music.* University of Chicago Press.

[35] Patel, Aniruddh D. (2008): *Music, language, and the brain.* New York: Oxford University Press.

[36] Nespor, Marina and Irene Vogel (1983): "Prosodic Structure Above the Word." In: Anne Cutler and D. Robert Ladd (Eds.) *Prosody: Models and Measurements.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 123–140.

[37] London, Justin (2004): *Hearing in time: psychological aspects of musical meter.* Oxford ; New York: Oxford University Press.

[38] Greenberg, Steven; Hannah Carvey; Leah Hitchcock; and Shuangyu Chang (2003): "Temporal properties of spontaneous speech—a syllable-centric perspective." In: *Journal of Phonetics*, **31**(3-4), pp. 465–485.

[39] Levitin, Daniel J.; Jessica A. Grahn; and Justin London (2018): "The Psychology of Music: Rhythm and Movement." In: *Annual Review of Psychology*, **69**(1), pp. 51–75. Publisher: Annual Reviews.

[40] Dixon, Simon (2001): "Automatic Extraction of Tempo and Beat From Expressive Performances." In: *Journal of New Music Research*, **30**(1), pp. 39–58.

[41] Large, Edward W (2008): "Resonating to musical rhythm: theory and experiment." In: *The psychology of time*, pp. 189–232.

[42] Lakatos, Peter; Joachim Gross; and Gregor Thut (2019): "A New Unifying Account of the Roles of Neuronal Entrainment." In: *Current Biology*, **29**(18), pp. R890–R905.

[43] Henry, Molly J.; Björn Herrmann; and Jessica A. Grahn (2017): "What can we learn about beat perception by comparing brain signals and stimulus envelopes?" In: *PLOS ONE*, **12**(2), p. e0172454.

[44] Blandini, Fabio; Giuseppe Nappi; Cristina Tassorelli; and Emilia Martignoni (2000): "Functional changes of the basal ganglia circuitry in Parkinson's disease." In: *Progress in Neurobiology*, **62**(1), pp. 63–88.

[45] Ghitza, Oded and Steven Greenberg (2009): "On the Possible Role of Brain Rhythms in Speech Perception: Intelligibility of Time-Compressed Speech with Periodic and Aperiodic Insertions of Silence." In: *Phonetica*, **66**(1-2), pp. 113–126.

[46] Grahn, Jessica A. and Matthew Brett (2007): "Rhythm and Beat Perception in Motor Areas of the Brain." In: *Journal of Cognitive Neuroscience*, **19**(5), pp. 893–906.

[47] Kotz, Sonja A. and Michael Schwartze (2010): "Cortical speech processing unplugged: a timely subcortico-cortical framework." In: *Trends in Cognitive Sciences*, **14**(9), pp. 392–399.

[48] Iversen, John and Aniruddh Patel (2008): "The Beat Alignment Test (BAT): Surveying beat processing abilities in the general population." In: *Proceedings of the 10{th International Conference on Music Perception and Cognition (ICMPC10).*

[49] Matthews, Peter Hugoe and Peter Hugoe Matthews (2014): *The concise Oxford dictionary of linguistics.* Oxford University Press.

[50] Fujii, Shinya and Gottfried Schlaug (2013): "The Harvard Beat Assessment Test (H-BAT): a battery for assessing beat perception and production and their dissociation." In: *Frontiers in Human Neuroscience*, **7**.

[51] Grahn, Jessica A. and Matthew Brett (2009): "Impairment of beat-based rhythm discrimination in Parkinson's disease." In: *Cortex*, **45**(1), pp. 54–61.

[52] Schmidt-Kassow, Maren and Sonja A. Kotz (2009): "Attention and perceptual regularity in speech:." In: *NeuroReport*, **20**(18), pp. 1643–1647.

[53] Hannon, Erin E. and Sandra E. Trehub (2005): "Metrical Categories in Infancy and Adulthood." In: *Psychological Science*, **16**(1), pp. 48–55.

[54] Poeppel, David (2003): "The analysis of speech in different temporal integration windows:

cerebral lateralization as ‘asymmetric sampling in time’.” In: *Speech Communication*, **41**(1), pp. 245–255.

[55] Kandylaki, Katerina D.; et al. (2017): “Where Is the Beat? The Neural Correlates of Lexical Stress and Rhythmical Well-formedness in Auditory Story Comprehension.” In: *Journal of Cognitive Neuroscience*, **29**(7), pp. 1119–1131.

[56] James, Arthur Lloyd (1940): *Speech signals in telephony.* Sir I. Pitman & sons, Limited.

[57] Pike, Kenneth L. (1945): *The intonation of American English.* No. Vol. 1 in University of Michigan publications Linguistics. Ann Arbor: Univ. of Michigan Press.

[58] Abercrombie, David (1967): *Elements of general phonetics.* Edinburgh University Press.

[59] Darwin, C. J. and Andrew Donovan (1980): “Perceptual Studies of Speech Rhythm: Isochrony and Intonation.” In: J. C. Simon (Ed.) *Spoken Language Generation and Understanding.* Dordrecht: Springer Netherlands, pp. 77–85.

[60] Pompino-Marschall, Bernd (1990): *Die Silbenprosodie: ein elementarer Aspekt der Wahrnehmung von Sprachrhythmus und Sprechtempo.* No. 247 in Linguistische Arbeiten. Tübingen: M. Niemeyer.

[61] Allen, George D. (1972): “The Location of Rhythmic Stress Beats in English: an Experimental Study I.” In: *Language and Speech*, **15**(1), pp. 72–100.

[62] Fowler, Carol A. (1979): ““Perceptual centers” in speech production and perception.” In: *Perception & Psychophysics*, **25**(5), pp. 375–388.

[63] Morton, John; Steve Marcus; and Clive Frankish (1976): “Perceptual centers (P-centers).” In: *Psychological review*, **83**(5), p. 405.

[64] Liberman, Mark and Alan Prince (1977): “On Stress and Linguistic Rhythm.” In: *Linguistic Inquiry*, **8**(2), pp. 249–336.

[65] Dauer, R.M. (1983): “Stress-timing and syllable-timing reanalyzed.” In: *Journal of Phonetics*, **11**(1), pp. 51–62.

[66] Dellwo, Volker (2010): *Influences of speech rate on the acoustic correlates of speech rhythm: An experimental phonetic study based on acoustic and perceptual evidence.* Ph. D. Dissertation, Universität Bonn, Bonn, Germany.

[67] Burred, Juan Jose (2004): “Hierarchical Automatic Audio Signal Classification.” In: *J. Audio Eng. Soc.*, **52**(7), p. 17.

[68] Foote, J. and S. Uchihashi (2001): “The beat spectrum: a new approach to rhythm analysis.” In: *IEEE International Conference on Multimedia and Expo, 2001. ICME 2001.* Tokyo, Japan: IEEE, pp. 881–884.

[69] Grosche, Peter; Meinard Muller; and Frank Kurth (2010): “Cyclic tempogram—A mid-level tempo representation for musicsignals.” In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing.* Dallas, TX: IEEE, pp. 5522–5525.

[70] Bello, Juan Pablo; et al. (2005): “A tutorial on onset detection in music signals.” In: *IEEE Transactions on speech and audio processing*, **13**(5), pp. 1035–1047.

[71] Lykartsis, Athanasios and Stefan Weinzierl (2016): “Rhythm Description for Music and Speech Using the Beat Histogram with Multiple Novelty Functions: First Results.” In: *Proceedings of the Inter-Noise 2016 : 45th International Congress and Exposition on Noise Control Engineering.* Hamburg, Germany: Deutsche Gesellschaft für Akustik e.V, pp. 964–967.

[72] Turk, Alice and Stefanie Shattuck-Hufnagel (2013): “What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong.” In: *Laboratory Phonology*, **4**(1).

[73] Lykartsis, Athanasios; Alexander Lerch; and Stefan Weinzierl (2015): “Analysis of Speech Rhythm for Language Identification Based on Beat Histograms.” Deutsche Gesellschaft für Akustik e.V., pp. 1019–1022.

[74] Fischinger, Timo (2009): *Zur Psychologie des Rhythmus: Präzision und Synchronisation bei Schlagzeugern.* Kassel University Press GmbH.

[75] Lerch, Alexander (2012): *Audio content analysis: an introduction.* Hoboken, N.J: Wiley.

[76] Weinzierl, Stefan (Ed.) (2008): *Handbuch der Audiotechnik.* Berlin, Heidelberg: Springer Berlin Heidelberg.

[77] Li, Tom L. H. and Antoni B. Chan (2011): “Genre Classification and the Invariance of MFCC Features to Key and Tempo.” In: Kuo-Tien Lee; Wen-Hsiang Tsai; Hong-Yuan Mark Liao; Tsuhan Chen; Jun-Wei Hsieh; and Chien-Cheng Tseng (Eds.) *Advances in Multimedia Modeling.* Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 317–327.

[78] Stevens, Stanley Smith; John Volkmann; and Edwin B Newman (1937): “A scale for the measurement of the psychological magnitude pitch.” In: *The Journal of the Acoustical Society of America*, **8**(3), pp. 185–190. Publisher: Acoustical Society of America.

[79] Brainard, David H. (1997): “The Psychophysics Toolbox.” In: *Spatial Vision*, **10**(4), pp. 433–436.

[80] Coffey, EBJ; SC Herholz; S Scala; and RJ Zatorre (2011): “Montreal Music History Questionnaire: a tool for the assessment of music-related experience in music cognition research.”

[81] Goodwin, C James and Kerri A Goodwin (2016): *Research in psychology methods and design.* John Wiley & Sons.

[82] Likert, Rensis (1932): “A technique for the measurement of attitudes.” In: *Archives of psychology.*

[83] Lykartsis, Athanasios and Alexander Lerch (2015): “Beat histogram features for rhythm-based musical genre classification using multiple novelty functions.” In: Peter Svensson (Ed.) *Proceedings of the 18th International Conference on Digital Audio Effects.* Trondheim: Norwegian University of Science and Technology, Department of Music and Department of Electronics and Telecommunications.

[84] Chen, Mo (2020): “Pattern Recognition and Machine Learning Toolbox.” `https://github.com/PRML/PRMLT`. Access: 15.06.2020.

[85] Bettinardi, Ruggero G. (2020): “getCosineSimilarity(x,y).” `https://www.mathworks.com/matlabcentral/fileexchange/62978-getcosinesimilarity-x-yT`. Access: 15.06.2020.

[86] R Core Team (2020): “R: A Language and Environment for Statistical Computing.” `https://www.R-project.org`. Access: 15.06.2020.

[87] Wickham, Hadley (2016): *ggplot2: elegant graphics for data analysis.* Springer-Verlag New York.

[88] Smith, Reginald D. (2015): “A Mutual Information Approach to Calculating Nonlinearity.” In: *Stat*, **4**(1), pp. 291–303.

[89] Dubey, Vimal Kumar and Amit Kumar Saxena (2017): “A Cosine-Similarity Mutual-Information Approach for Feature Selection on High Dimensional Datasets:.” In: *Journal of Information Technology Research*, **10**(1), pp. 15–28.

[90] Mann, Henry B and Donald R Whitney (1947): “On a test of whether one of two random variables is stochastically larger than the other.” In: *The annals of mathematical statistics*, pp. 50–60.

[91] Herff, Steffen A; et al. (2020): “Prefrontal High Gamma in ECoG tags periodicity of musical rhythms in perception and imagination.” In: *bioRxiv*, p. 784991. Cold Spring Harbor Laboratory.

[92] Renkl, Alexander (1993): “Korrelation und Kausalität: Ein ausreichend durchdachtes Problem in der pädagogisch-psychologischen Forschung?” In: Christian Tarnai (Ed.) *Beiträge zur empirischen pädagogischen Forschung.* Münster: Waxmann, pp. 115 – 123.

[93] Valsiner, Jaan (1986): “Between groups and individuals: Psychologists’ and laypersons’ interpretation of correlational findings.” In: Jaan Valsiner (Ed.) *The Individual Subject and Scientific Psychology.* Boston, MA: Springer US, pp. 113–151.

[94] Fan, Jin (2014): “An information theory account of cognitive control.” In: *Frontiers in human neuroscience*, **8**, p. 680.

[95] Zeng, Xianli; Yingcun Xia; and Howell Tong (2018): "Jackknife approach to the estimation of mutual information." In: *Proceedings of the National Academy of Sciences*, **115**(40), pp. 9956–9961.

[96] Boll-Avetisyan, Natalie; Anjali Bhatara; Annika Unger; Thierry Nazzi; and Barbara Höhle (2016): "Effects of experience with L2 and music on rhythmic grouping by French listeners." In: *Bilingualism: Language and Cognition*, **19**(5), pp. 971–986.

[97] Kolinsky, Réégine; Héélééne Cuvelier; Vincent Goetry; Isabelle Peretz; and Joséé Morais (2009): "Music Training Facilitates Lexical Stress Processing." In: *Music Perception*, **26**(3), pp. 235–246.

[98] Patel, Aniruddh D. (2014): "Can nonlinguistic musical training change the way the brain processes speech? The expanded OPERA hypothesis." In: *Hearing Research*, **308**, pp. 98–108.

[99] Kandylaki, Katerina D. and Ina Bornkessel-Schlesewsky (2019): "From story comprehension to the neurobiology of language." In: *Language, Cognition and Neuroscience*, **34**(4), pp. 405–410.

[100] Davis, Steven and Paul Mermelstein (1980): "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences." In: *IEEE transactions on acoustics, speech, and signal processing*, **28**(4), pp. 357–366. Publisher: IEEE.

[101] Richard, Gael; Shiva Sundaram; and Shrikanth Narayanan (2013): "An Overview on Perceptually Motivated Audio Indexing and Classification." In: *Proceedings of the IEEE*, **101**(9), pp. 1939–1954.

[102] Li, Ming; Hongbin Suo; Xiao Wu; Ping Lu; and Yonghong Yan (2007): "Spoken language identification using score vector modeling and support vector machine."

[103] Gouyon, Fabien; Simon Dixon; Elias Pampalk; and Gerhard Widmer (2004): "Evaluating rhythmic descriptors for musical genre classification." pp. 196–204.

[104] Thompson, Paul (2012): "An empirical study into the learning practices and enculturation of DJs, turntablists, hip hop and dance music producers." In: *Journal of Music, Technology and Education*, **5**(1), pp. 43–58.

[105] Loots, M.E. (1980): *Metrical Myths.* Dordrecht: Springer Netherlands.

[106] Ramus, Franck and Jacques Mehler (1999): "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.

[107] Stanev, Madeleine; Johannes Redlich; Christian Knörzer; Ninett Rosenfeld; and Athanasios Lykartsis (2016): "Speech and music discrimination: Human detection of differences

between music and speech based on rhythm." In: *Speech Prosody 2016*. International Speech Communication Association, pp. 222–226.

[108] Goswami, Usha; et al. (2016): "Perception of Filtered Speech by Children with Developmental Dyslexia and Children with Specific Language Impairments." In: *Frontiers in Psychology*, **7**.

[109] Hasson, Uri; Giovanna Egidi; Marco Marelli; and Roel M. Willems (2018): "Grounding the neurobiology of language in first principles: The necessity of non-language-centric explanations for language comprehension." In: *Cognition*, **180**, pp. 135–157.

[110] Weissbart, Hugo; Katerina D. Kandylaki; and Tobias Reichenbach (2020): "Cortical Tracking of Surprisal during Continuous Speech Comprehension." In: *Journal of Cognitive Neuroscience*, **32**(1), pp. 155–166.

[111] Gouyon, Fabien and Simon Dixon (2005): "A review of automatic rhythm description systems." In: *Computer music journal*, **29**(1), pp. 34–54.