Speech Rhythm Detection and its Application in Speech Perception

Eleftheria Lydaki

Thesis submitted in partial fulfillment of the requirements for the

Masters' of Science degree in Computer Science and Engineering

University of Crete
School of Sciences and Engineering
Computer Science Department
Voutes University Campus, 700 13 Heraklion, Crete, Greece

Thesis Advisor: Professor *Panagiotis Tsakalides*

This work has been performed at the University of Crete, School of Sciences and Engineering, Computer Science Department. The experiments have been conducted at Apple UK as a part of the internship program.

The work has been supported by the Foundation for Research and Technology - Hellas (FORTH), Institute of Computer Science (ICS).

University of Crete
Computer Science Department

Speech Rhythm Detection and its Application in Speech Perception

Thesis submitted by
Eleftheria Lydaki
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

THESIS APPROVAL

Author:

Eleftheria Lydaki

Committee approvals:

Panagiotis Tsakalides
Professor, Thesis Supervisor

Ioannis Stylianou
Professor, Committee Member

Grigorios Tsagkatakis
Assistant Professor, Committee Member

Departmental approval:

Polyvios Pratikakis
Assistant Professor, Director of Graduate Studies

Heraklion, June 2023

Speech Rhythm Detection and its Application in Speech Perception

Abstract

Speech rhythm refers to the rhythmic patterns and timing variations that occur in spoken language. It encompasses the natural flow, stress patterns, and timing of speech sounds, syllables, and words. Rhythm consists an important dynamic prosodic feature of speech that is linked with speech perception. The detection of speech rhythm is a significant task with diverse applications. In this study, the focus is on using rhythmic measures to estimate voice preference. The motivation behind this research arises from the belief that voices demonstrating specific rhythm patterns are generally preferred by individuals.

In this thesis, speech rhythm was studied as a possible predictor of listener preference. Even though rhythm can be perceived by humans, there is no ubiquitously accepted definition or measure for speech rhythm in the scientific community. In the literature, there is strong evidence that rhythm is encoded in the amplitude envelope of a signal. Mainly, the envelope is decomposed into partials and then the corresponding instantaneous frequency is extracted which is assumed to carry the information regarding the signal's rhythmicity. Two techniques were utilized to achieve the decomposition of the envelope into meaningful components. The first technique, which was proposed in a previous study, includes extracting rhythmic measures via an Empirical Mode Decomposition (EMD) of the envelope. Here, it is suggested to extract the same measures by using an AM-FM decomposition on the envelope instead of EMD. This modification has the potential to improve the accuracy of the resulting values since EMD isn't mathematically robust. The envelope, although informative to some extent, is a simplified representation of the speech signal. It lacks important elements like pitch, which could potentially contribute to the understanding of rhythm. Relying solely on the envelope may overlook relevant rhythmic features present in the speech signal. We hypothesize that the rhythmicity of speech is closely related to the manner in which individuals transition between syllables. Therefore, an approach that directly captures the rhythmicity of speech was introduced by considering the segment of the speech signal associated with syllable transitions. This, effectively addresses the concern of information loss that occurs during envelope extraction.

During this research, data consisting of speech signals from multiple speakers were utilized. The information regarding the preferred speakers, as determined by listeners, was also available. This knowledge allowed the investigation of the underlying factors contributing to voice preference and the analysis of the specific characteristics that make certain speakers more preferred than others. The experiments were extended beyond natural speaking rate, namely for fast speaking style, and the preference and rhythm in fast speech was explored as well.

Statistical analyses were conducted to evaluate the suitability of rhythmic metrics derived from envelope and signal-based techniques for the task at hand. Findings revealed that the envelope-derived metrics are heavily influenced by speech rate and they are not well-suited for accurately capturing rhythm. In contrast, syllables transition derived directly from the speech signal showcased promising results. A satisfactory separation between preferred and non-preferred speakers was achieved, effectively capturing certain characteristics that influence listeners' preference. One-way ANOVA and pairwise comparison tests were preformed to validate the statistical significance of the differences between speakers.

The results based on syllables transition indicate promising avenues for future research. Considering the multi-component nature of preference, the exploration of additional metrics becomes crucial in improving the overall performance which will lead to a comprehensive and reliable evaluation of listener preference.

Ανίχνευση του Ρυθμού Ομιλίας και η Εφαρμογή του στην Αντίληψη της Ομιλίας

Περίληψη

Ο όρος ρυθμός της ομιλίας αναφέρεται στα ρυθμικά μοτίβα και που συμβαίνουν στον προφορικό λόγο. Περιλαμβάνει την φυσική ροή, τα μοτίβα τονισμού και έμφασης και τις χρονικές μεταβολές των ήχων, των συλλαβών και των λέξεων. Ο ρυθμός αποτελεί ένα σημαντικό δυναμικό προσοδιακό χαρακτηριστικό της ομιλίας που συνδέεται με την αντίληψη της. Η ανίχνευση του ρυθμού της ομιλίας έχει πολλαπλές εφαρμογές. Σε αυτήν την μελέτη σκοπός είναι η χρήση ρυθμικών μέτρων για την εκτίμηση της προτίμησης των ακροατών. Η έρευνα αυτή είχε ως βάση την ιδέα ότι φωνές με συγκεκριμένα ρυθμικά μοτίβα είναι εν γένει προτιμότερες.

Στην παρούσα εργασία, μελετήθηκε ο ρυθμός της ομιλίας ως μέτρο κατανοησιμότητας και πιθανό μέσο πρόβλεψης της προτίμησης των ακροατών. Αν και η ρυθμικότητα γίνεται αντιληπτή από τους ανθρώπους, δεν υπάρχει καθολικά αποδεκτός ορισμός ή μέτρο ποσοτικοποίησης του ρυθμού της ομιλίας στην επιστημονική κοινότητα. Στη βιβλιογραφία υπάρχουν ισχυρά επιχειρήματα ότι η ρυθμικότητα κωδικοποιείται στη χρονική περιβάλλουσα του σήματος. Συνήθως η περιβάλλουσα αποσυντίθεται σε συνιστώσες και εξάγωνται οι αντίστοιχες στιγμιαίες συχνότητες, οι οποίες υποθέτουμε ότι φέρουν πληροφορίες για την ρυθμικότητα του σήματος. Χρησιμοποιήθηκαν δύο τεχνικές για την ανάλυση της περιβάλλουσας σε ουσιώδεις συνιστώσες. Η πρώτη τεχνική, που προτάθηκε σε προηγούμενη μελέτη, περιλαμβάνει τον 'Εμπειρικό Τρόπο Αποσύνθεσης' (EMD) της χρονικής περιβάλλουσας για την εξαγωγή μετρικών για τον ρυθμό της ομιλίας. Εδώ προτάθηκε η εξαγωγή των ίδιων μετρικών χρησιμοποιώντας μία αποσύνθεση AM-FM στην χρονική περιβάλλουσα αντί για τον EMD. Αυτή η τροποποίηση έχει προοπτικές να βελτιώσει την ακρίβεια των αποτελεσμάτων, αφού ο EMD δεν είναι μαθηματικά αξιόπιστος. Η περιβάλλουσα, αν και περιέχει ορισμένες πληροφορίες, απότελει μια απλοποιημένη μορφή του σήματος της φωνής. Του λείπονται σημαντικά στοιχεία όπως η τονικότητα (pitch), τα οποία πιθανώς συνεισφέρουν στην κατανόηση του ρυθμού. Η αποκλειστική χρήση της περιβάλλουσας μπορεί να οδηγήσει στην παράλειψη ρυθμικών χαρακτηριστικών που περιλαμβάνονται στο σήμα της φωνής. Υποθέτουμε ότι η ρυθμικότητα της φωνής είναι στενά συνδεδεμένη με τον τρόπο με τον οποίο οι ομιλητές μεταβαίνουν από την μία συλλαβή στην επόμενη. Συνεπώς, προτείνουμε μία προσέγγιση που εξάγει απευθείας τη ρυθμικότητα από το σήμα της φωνής αναλύοντας το τμήμα του σήματος της φωνής που σχετίζεται με τις μεταβάσεις μεταξύ συλλαβών. Αυτή η μέθοδος αντιμετωπίζει αποτελεσματικά το πρόβλημα της απώλειας πληροφορίας, που συμβαίνει αναπόφευκτα κατά την εξαγωγή και ανάλυση της περιβάλλουσας.

Σε αυτήν την έρευνα χρησιμοποιήθηκαν δεδομένα που περιείχαν σήματα φωνής από διάφορους ομιλητές. Συχρόνως ήταν ακόμη διαθέσιμες πληροφορίες σχετικά με τους ομιλητές που προτιμήθηκαν από τους ακροατές. Αυτή η γνώση επέτρεψε την διερεύνηση των παραγόντων που συνεισφέρουν στην προτίμηση ορισμένων φωνών και την ανάλυση των χαρακτηριστικών που τις καθιστούν προτιμότερες. Τα πειράματα

επεκτάθηκαν πέρα από την φυσική ταχύτητα ομιλίας, δηλαδή στον γρήγορο τρόπο ομιλίας, και μελετήθηκαν η προτίμηση και ο ρυθμός στον γρήγορο λόγο.

Πραγματοποιήθηκε στατιστική ανάλυση για να αξιολογήσουμε την καταλληλότητα των μετρικών που προήλθαν από τις τεχνικές αποσύνθεσης της περιβάλλουσας και του σήματος στα πλαίσια του σκοπού αυτής της εργασίας. Τα αποτελέσματά μας έδειξαν ότι οι μετρικές που σχετίζονται με την περιβάλλουσα δεν είναι κατάλληλες για την ακριβή αποτύπωση του ρυθμού, καθώς επηρεάζονται σημαντικά από την ταχύτητα του λόγου, με αποτέλεσμα να υπάρχει έλλειψη ακρίβειας στην αναπαράσταση ρυθμικών μοτίβων. Αντιθέτως, η μελέτη των μεταβάσεων μεταξύ συλλαβών απευθείας από το σήμα της φωνής έδειξε υποσχόμενα αποτελέσματα. Επιτεύχθηκε ένας ικανοποιητικός διαχωρισμός ανάμεσα στους ομιλητές που προτιμήθηκαν και τους υπόλοιπους, και επομένως η σύλληψη κάποιων χαρακτηριστικών που διαμορφώνουν την προτίμηση των ακροατών. Η στατιστική σημαντικότητα των διαφορών ανάμεσα στους ομιλητές επιβεβαιώθηκε με στατιστικά τεστ ANOVA (Ανάλυση της Διασποράς).

Τα αποτελέσματα από τις μεταβάσεις των συλλαβών υποδεικνύουν υποσχόμενες ευκαιρίες για μελλοντική έρευνα. Λαμβάνοντας υπόψη την πολυδιάστατη φύση της προτίμησης, η διερεύνηση επιπλέον μετρικών γίνεται αναγκαία στην βελτίωση της απόδοσης που θα οδηγήσει σε μία πιο εμπεριστατωμένη και αξιόπιστη εκτίμηση της προτίμησης των ακροατών.

Acknowledgements

I am deeply grateful to my supervisor, Prof. Panagiotis Tsakalides, and co-supervisor, Prof. Yannis Stylianou, for their invaluable guidance and support throughout my master's studies. Additionally, I extend a heartfelt thanks to my manager, Dr. Petko Petkov, for the insightful discussions and ideas shared during my internship. Their significant contributions were indispensable in the completion of this work. I shouldn't forget to thank Dr. Ioannis Pantazis and Dr. Maria Koutsogiannaki for kindly providing their code for the AM-FM algorithm.

I would also like to express my sincere appreciation to my family for their unwavering support and encouragement that have brought me to this point in life. A special thanks goes to my brother, Giorgos, for all the help he has provided over the years of my studies. I would also like to acknowledge the assistance and kindness of my colleagues and friends, who were always willing to lend a helping hand. A separate note of gratitude goes to Dipjyoti Paul for his invaluable assistance in the writing process of this thesis, which made it easier and more enjoyable. Finally, I want to thank Yannis for motivating me and never leaving my side throughout my studies.

στους γονείς μου

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Speech is the most direct and informative method of communication among humans. Its power and uniqueness can be attributed to how it delivers information not only through its content, but also through its prosodic features that contain speaker specific information. The term prosodic features refers to the variations in pitch, rhythm, loudness and timing that give spoken language additional information. Every speaker possesses unique characteristics and speaking style, resulting in distinct prosodic variations that can be used to distinguish them from other speakers. Furthermore, individuals can adopt various speaking styles in order to express their current emotional state, emphasize certain words or phrases, or indicate their intentions. Hence, a lot of research has been conducted to study these prosodic features of speech, for applications such as voice conversion, speaking assistance, speaker recognition and many others. Despite that, speech rhythm, while being an important prosodic feature, has not yet been studied to a satisfactory level.

When it comes to the definition of rhythm, there is no consensus among the scientific community. In [95] the authors make a detailed review on different definitions that exist in bibliography. Intuitively, the word rhythm implies the existence of periodicity which can be tracked down either on the production or the perception of speech. According to [95], the presence of periodicity in the surface of the speech signal is unsupported. At the same time, even though evidence of periodic movements in motor control, e.g. chewing exist, there are still doubts regarding the importance of periodicity in speech motor control. Further complications in the study of rhythm arise also due to the difficulty of its quantification and the absence of unbiased and universally accepted metrics to model it.

In this thesis, the primary focus revolves around exploring the correlation between speech rhythm and listeners' preferences. It is important to note that preference is inherently subjective, and listeners often struggle to precisely articulate the reasons behind their preference for one speaker's voice over another's. Nevertheless, numerous ABX tests have revealed that listeners frequently reach a consensus on the voice they prefer. This observation provides motivation to identify measures

that could potentially indicate or even predict this preference. We propose that rhythm is one of the contributing factors influencing listeners' choices. While a significant body of literature examines rhythm across different languages, our study delves specifically into the comparison of speech rhythm among English speakers and its impact on intelligibility.

As technology progresses, the prevalence of voice systems in daily life continues to expand. Hence, the selection of speakers that are appealing to the public is an important undertaking, which is both time-consuming and resource-demanding. Consequently, it is evident that discovering an automated way to estimate the evaluation of different voices among the broader population would yield significant benefits. At the same time virtual assistants for visual impaired people are not only a convenience or a preference but a necessity. In modern societies, accessibility should be a priority and thus finding voices that fulfill the needs of this part of the population is a significant task. Similarly, research and experiments have demonstrated that individuals with visual impairments exhibit an enhanced auditory perception and are capable of comprehending speech at accelerated rates [11] [37] [14]. Perfecting speed up methods is a crucial step towards accessibility. Therefore in this thesis special attention was given to fast speech. The results of this study would be enlightening since the identification of certain rhythm characteristics that render one speaker's fast speech more intelligible and preferred over another's, could help perfect speed up methods.

Selecting the appropriate approach for addressing the preference problem is a nontrivial decision. Despite the challenges involved in studying and modeling rhythm, it holds the potential to provide a meaningful representation for preference. In the field of speech analysis, it is a common practice to divide the signal into smaller frames to leverage the stationarity it offers. Techniques like Fourier Transform rely on this segmentation to yield desired outcomes. However, when exploring perception, it may be necessary to examine the speech signal on a larger time scale. Rhythmicity, in contrast, pertains to larger speech units such as syllables or stress intervals, as well as their transitions. Hence, it is plausible that preference is more closely related to rhythmic features extracted from these larger segments of speech rather than from smaller frames. Furthermore, speech rhythm is often connected to intelligibility [73] [71], which, especially for fast speech, plays a key role in the selection of a preferred speaker.

In this thesis we consider that speech rhythm can be extracted from the amplitude envelope of the speech signal. We adopt the method proposed in [89], which involves decomposing the envelope using Empirical Mode Decomposition (EMD). For each component, we calculate the instantaneous frequency, assuming that the variance of this frequency reflects the rhythmic stability across different time scales. Building upon the work of Tilsen & Arvaniti, we also propose utilizing an AM-FM decomposition of the envelope, using the algorithm in [68]. This method is a mathematically robust framework that potentially yields more reliable results than EMD. By employing both methods, we extract rhythmic features and compare their performance in terms of predicting preference. Additionally, we propose

a method that directly extracts rhythmicity from the speech signal itself. We recognize that while the envelope provides some information about the speech signal, it falls short in capturing essential elements like pitch, which could be crucial for a comprehensive understanding of rhythm. By addressing the limitations associated with envelope-based approaches, our method effectively overcomes the issue of information loss during envelope extraction. Specifically, our research delves into the segment of the speech signal that includes syllable transitions, hypothesizing that the rhythmicity of speech is intimately intertwined with the manner in which individuals transition between syllables.

Our findings suggest that envelope-based rhythm metrics are inadequate for accurately predicting preference. The influence of speech rate and their limited ability to differentiate between speakers undermine their effectiveness for this purpose. We found that some of the envelope-based metrics have up to 0.91 correlation with the rate of the speech. Speakers can have varying natural speech rates and these variations alone do not dictate preference or intelligibility. Hence, it is essential to acknowledge that any bias stemming from differences in speech rate has the potential to result in misleading conclusions.

The approach of directly extracting rhythm from the speech signal yielded intriguing findings. Although the conclusion was not definitive, it demonstrated a partial capacity to distinguish preferred speakers in a statistically significant manner. This suggests that we have identified a metric that reflects certain characteristics associated with preference, albeit not in an absolute manner. This outcome opens up avenues for future investigations aiming to identify a measure or a combination of measures that can reliably estimate preference.

The structure of this thesis is as follows: Chapter 2 provides a comprehensive overview of the existing literature encompassing speech rhythm, speech rhythm metrics, and preference. Chapter 3 offers a meticulous description of the methodology and data employed in this study. In Chapter 4, the experimental findings are presented, while Chapter 5 combines a discussion of the results, along with the conclusions drawn and suggestions for future research.

Chapter 2

Related Work

The topic of speech rhythm and its quantification has been in the scope of a lot of research for many years. Despite the consensus among scientists that speech can be characterized as quasi-rhythmical [73] [71] [32], the endeavor of examining and deciphering this rhythmicity proves to be a complex and demanding task. It necessitates the utilization of multiple techniques and approaches to effectively illustrate and comprehend the intricacies of speech rhythm.

Quasi-rhythmical speech refers to the phenomenon where speech exhibits rhythmic patterns that are not strictly regular but possess an inherent sense of rhythm. This characteristic of human communication has garnered significant attention in research over the years. Unlike rhythmic patterns found in music, which often follow precise and predictable patterns, quasi-rhythmical speech demonstrates a more fluid and variable nature. It is influenced by a variety of factors, including linguistic structures, phonetic features, and individual speaking styles. These factors contribute to the unique rhythmic qualities observed in speech. The presence of quasi-rhythm in speech has been acknowledged and studied by scientists across disciplines such as linguistics, psychology, and neuroscience. Researchers have employed various techniques to explore and quantify these rhythmic patterns. These techniques include acoustic analysis, linguistic modeling, and statistical approaches to capture the dynamic nature of quasi-rhythm in speech.

2.1 Speech Rhythm and Rhythmic Measures

In the existing bibliography, a variety of speech rhythm metrics have been proposed and used in different contexts. In [54] [93], the idea of a curve that represents rhythm, namely the rhythmogram, was introduced. This rhythm visualization technique can provide valuable insights into the temporal structure and organization of musical and linguistic performances. The process for the extraction of a rhythmogram curve involves three main steps. First, the algorithm uses a cochlear filter based on the gammatone filter-bank to filter the sound. Second, a Gaussian low-pass modulation filter-bank is applied to perform modulation filtering. Finally,

the output from the modulation filter-bank is integrated after detecting the peaks. Despite the limited dataset employed in this study, the authors have demonstrated that the visualization of rhythm can be applied at all levels of rhythmic structures from individual phonemes to the structure of a complete poem.

The features that are most commonly used to measure speech rhythm, are summarized in [7] [31]. These metrics can be separated into two categories. The first category includes rhythm interval measures (IM), namely the percentage of vocalic intervals in an utterance (%V), the average standard deviation of consonantal -or intervocalic- intervals (ΔC), and the average standard deviation of vocalic duration (ΔV) [78]. The IM were originally introduced by Ramus et al in order to distinguish the differences in rhythm among languages. However, in follow-up research, IM have been employed for various applications. The shortcomings of these metrics arise from their sensitivity to speech rate [9] [19]. In fast speech, there is a tendency to decrease the duration of consonantal intervals. As a result, the measurement of ΔC can yield biased outcomes that primarily reflect the speech rate rather than the underlying rhythm. In [18], the author introduced an alternative approach to ΔC. This variation incorporates a normalization technique, which considers the speech rate in order to prevent the occurrence of misleading outcomes. The mathematical formula is given by:

$$Var\Delta C = \frac{\Delta C * 100}{meanC} \tag{2.1}$$

where C denotes the duration of consonontal intervals. The same normalization can also be applied for vocalic intervals. The second set of rhythm metrics includes the raw Pairwise Variability Index (rPVI) [38] which quantifies the difference in length of consecutive syllables or other speech units. It calculates the mean of these differences over a speech utterance. In [49] the authors compare the vowel variability index to the interval measures, proposed by Ramus et al., and conclude that the variability index is more robust in capturing rhythmic patterns. Grabe & Low in [38] also propose a normalized version of the PVI, namely the nPVI, that eliminates the bias of speech rate. The use of nPVI is recommended for vocalic intervals while the rPVI is mostly used for consonontal as they are less influenced by speech rate. The indices are defined mathematically by:

$$rPVI = \sum_{k=1}^{N-1} \frac{|d_k - d_{k+1}|}{(N-1)} \tag{2.2}$$

$$nPVI = 100 * \sum_{k=1}^{N-1} \frac{\frac{|d_k - d_{k+1}|}{(d_k + d_{k+1})/2}}{(N-1)}, \tag{2.3}$$

where d_k and d_{k+1} denote the duration of a vocalic or consonontal interval and its successive and N is equal to the total number of intervals in a speech chunk. Some alternative forms of the PVI are also presented in [28]. Bertinetto & Bertini in [10] propose a novel rhythm metric, based on the PVI, which they

call Control/Compensation Index (CCI). CCI is differentiated from PVI by taking into consideration the number of segments (n_k) that compose the vocalic and consonontal intervals as following:

$$CCI = \frac{100}{N-1} \sum_{k=1}^{N-1} |\frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}}| \tag{2.4}$$

Scott in [81] proposes an index of irregularity defined by a set of taps. These taps separate a speech utterance into units or intervals and the irregularity measure reflects the similarity in terms of duration among the intervals. If we denote the duration of the i-th interval as I_i, the rhythmic irregularity measure (RIM) can be expressed as follows:

$$RIM = \sum_{i \neq j} log(\frac{I_i}{I_j}) \tag{2.5}$$

The literature demonstrates the value of capturing speech rhythm through its diverse applications and purposes. Multiple studies have demonstrated the association between rhythm and different attributes of the speaker. The authors in [87] utilize rhythmic metrics in order to automatically estimate the severity of Parkinson's Disease. These measures include the duration of speech, pauses, phonemes and vowels along with their ratios to the total duration. The classification models that were trained with these features were highly successful especially within the male patients. This result suggests that speech rhythm is affected by the disease and at the same time, it confirms that the employed measures are effective indicators of rhythm. Likewise, in [6], rhythm is modeled with the use of IM and PVI. The authors deploy this group of features to examine the correlation between speech rhythm and the speaker's gender and/or social environment. In order to substantiate these assertions, the obtained metrics from an Arabic Corpus were utilized to train a neural network. This network was then employed to predict the gender and social environment of the speakers. The moderately high predictive power of their model indicates the existence of correlation between rhythm and these characteristics. Meftah et al. in [56] investigate the usefulness of rhythm metrics as features to classify speech emotion, gender and accent.The classification task involved utilizing both an MLP and an SVM classifier. The MLP is a type of artificial neural network that consists of multiple layers of interconnected nodes, while the SVM is a supervised learning model that separates data points into different classes using hyperplanes. Initially, the classifiers were trained using solely rhythmic features. Subsequently, they were trained using other acoustic features. These features are basically low-level acoustic and prosodic features that are extracted by PRATT toolkit. The features are as follows: Pitch, intensity, formants, jitter, shimmer, and speech rate. Finally, the network was trained with a combination of both rhythmic and acoustic features. The authors reach to the conclusion that relying only on rhythmic measures is inadequate for the prediction of speech emotion, gender and accent. The network trained on rhythmic features

had the lowest accuracy. However, the accuracy of the classifier improved when combined with other acoustic features, indicating a relationship between speech rhythm and the predicted characteristics.

A lot of research in the literature connects rhythm to language. The work of Abercombie in [1] [2] originated the idea that languages can be categorized to stress-timed and syllable-timed ([72], [78], [45]), based on the unit of segmentation of speech. The assumption is that the timing between units - stress beats and syllables respectively- is isochronous. Later on the term of mora-timed was added to describe languages such as Japanese. This concept spurred extensive research around the relation of speech rhythm and language. In [51] machine learning techniques are deployed to perform language identification using rhythm metrics as features. The findings contradicted the theory of three language classes, and rather indicated that different metrics lead to different groupings of languages. Thus the various aspects of a language may categorize it contrastingly. It is still unclear in the scientific community which rhythm features separate languages optimally. Liu and Takeda in their recent work [50] suggest that two measures are more reliable for this task, namely $Varco\Delta V$ and vocalic nPVI. Even though this classification is a challenging problem, experiments have shown that infants can recognise their mother tongue among languages that belong to different rhythmic classes, and react to it, despite their inability to speak or understand the actual words [62] [96] [24] [22]. However, they are unsuccessful to discriminate languages within the same rhythmic class. This indicates that languages differ phonologically in such a way that can be perceived even by non-verbal individuals. Other studies investigate the presence of rhythm in non-verbal lip movements of primates [29] [58]. Their discoveries offer valuable perspectives on the evolutionary roots and possible shared elements of rhythmic communication among humans and primates. Further research on simians was conducted by Tincoff et al [91] and concluded that tamarins, just like newborns, possess the ability to discriminate among languages of different rhythmic classes, which is not extended though to distinguishing within the same class. Even though these findings may lead to the conclusion that rhythm is part of human nature, [74] [12] [70] show that native English children's speech gradually becomes more stress - timed with age, and thus rhythm can be considered acquired.

Multiple studies also investigate the progress of speech rhythm after acquiring a second language. In [20] speech rhythm in English utterances is measured before and after native Spanish speakers study English for a year. The results of this study were not definitive, which is not unexpected, since as described above quantifying rhythm is a difficult and yet to be solved task. More detail regarding the challenge of capturing speech rhythm in L2 (second language) is displayed in [39]. Gut examines the effectiveness of different rhythmic metrics to separate non-native speakers with respect to their varying levels of proficiency. The results of this study culminate in further doubts for the validity of the rhythm metrics, since the influence of speech rate renders some of them unreliable. At the same time, the author questions the existence of L2 rhythm altogether. A similar analysis

is presented in [63]. The authors compare the development of speech rhythm in English between native German and French speakers. German and French were chosen, due to the fact that it has been shown that German is rhythmically closer to English than French, with respect to the variability measures. It is concluded that the development of speech rhythm in L2 acquisition appears to follow a universal trend regardless of the learner's native language. However, the specific rhythmic patterns observed at a particular stage of development exhibit distinct characteristics influenced by the learner's L1. In [97] Whitworth makes a detailed investigation of speech rhythm in bilingual children in English and German. The study suggests that their speech, despite its fluency, resembles rhythmically L2 speakers rather than native speakers.

All previous work mentioned above, even though it produced interesting and enlightening results, models rhythm using duration or variability metrics. Despite their popularity, it is not indisputable that these metrics can indeed give a desired representation of rhythm and have received criticism. The findings of [7] indicate that the described metrics may not reliably capture cross-linguistic distinctions, as the scores vary significantly within a language and are susceptible to diverse methodological choices. This raises concerns about the reliability of cross-linguistic comparisons and the use of metrics for rhythmic classifications. Therefore another perspective on rhythm and rhythmicity of speech would be insightful.

Poeppel and Assaneo in [73] delve into the importance of the temporal structure (for more information [99] [53] [64]) of the speech signal in the study of rhythm. The term temporal structure of speech refers to the organization and patterning of sounds and pauses in spoken language over time. It involves the timing, duration, and rhythm of individual sounds, syllables, words, and phrases. The authors in [73] emphasize that while speech is not a strictly periodic signal, it exhibits regularities in its temporal structure that indicate the presence of rhythmicity. According to Poeppel and Assaneo, these regularities in speech rhythm are apparent in both production and perception, and they have a notable impact on speech intelligibility. In their review, they analyze rhythmicity from multiple perspectives, including the acoustic, articulatory, and linguistic domains. However, for the purpose of this thesis, the focus will be on the acoustic domain, which is primarily concerned with the properties of sound waves and their perception. For a more comprehensive understanding of the articulatory and linguistic domains and its role in speech rhythm, further details can be found in the works of Stevens [86] and Titze [92].

Rhythmicity in the acoustic domain can be captured by the regularities in the amplitude envelope of the speech signal. The amplitude modulation of the envelope refers to the changes in the intensity or energy level of the speech signal over time. The speech envelope represents the overall shape of the waveform, capturing the rapid variations in amplitude that occur during phonetic segments, syllables, and phrases. Amplitude modulations of the envelope play a crucial role in conveying rhythmic information in speech [73] [89] [47]. These modulations reflect the prominence and rhythmic patterns of syllables and stress patterns within words and sentences. By analyzing the amplitude modulations, researchers can examine

the temporal dynamics and rhythmic structure of speech. Studies have shown that amplitude modulations of the envelope provide cues for perceiving and understanding speech.[71]. The rise and fall in energy levels within the envelope contribute to the perception of stressed and unstressed syllables, as well as the grouping and segmentation of linguistic units. These modulations also influence the perception of prosody, including aspects such as intonation, timing, and emphasis, which are integral to the rhythmic characteristics of speech.

Goswani et al. in [35] [36] examine the correlation between dyslexia and the perception of the amplitude modulations of speech. The experiment that was conducted to test how well children with dyslexia can perceive these modulations evaluated their performance in beat detection. The conclusion the authors draw is that children with dyslexia lack the ability of identifying amplitude modulations and hence they have impaired rhythm detection skills. This finding supports the connection between speech rhythm and intelligibility and demonstrates the significance conducting additional research on this prosodic characteristic. More research regarding the connection of rhythm perception and the ability to read can also be found in [98] [26] [88]. There is also evidence that rhythm can play a role to language acquisition, whether that refers to first language or second. Recent works of Goswami et al [34] [33] examine how rhythm perception can be an indicator of a normal development in terms of language acquisition. The author discusses the potential modeling of speech rhythm from an amplitude modulation (AM) perspective. The analysis of the perception of AM information by children with language acquisition disorders suggests a correlation between rhythm and language development. The importance of these findings lie to the fact that understanding the role of rhythm in typical and atypical language development can contribute to improving interventions and support for individuals with developmental language disorders.

One important aspect in the detection of rhythm constitutes the location of the syllabic beat [3] [5]. Simply put the syllabic beat involves the organization and grouping of syllables into rhythmic units, with certain syllables receiving more prominence or stress than others. The stressed syllables typically receive more emphasis or prominence in terms of pitch, duration, and loudness. The alternation between stressed and unstressed syllables gives rise to a rhythmic pattern, commonly referred to as the syllabic beat. According to [3], the accuracy of a listener in locating the rhythmic beat of a syllable was found to be positively correlated with the level of stress on that syllable. In other words, the greater the stress on a syllable, the more reliable the subject's responses in determining the beat of that syllable. Another very similar concept is that of the perceptual center or p-center [[41] [27] [76] [40] [75]]. The p-centers refer to the points of prominence within a rhythmic unit. These points correspond to moments in speech where there is a perceived increase in prominence or stress. Hoequist and Charles in [41] establish that the concept of p-centers generalizes well across all language rhythmic classes rather than being applicable only for English. This suggests that the

presence and perception of p-centers are not limited to a specific linguistic context. According to [59] and [4] the p-center can be detected close to the onset of voicing, and techniques for determining its location can be found in [52] [82] [16]. The exploration of p-centers contributes to our understanding of speech rhythm by identifying specific points of prominence within rhythmic units. These perceptual markers play a crucial role in shaping the rhythm of speech across different languages and aid in the analysis and characterization of linguistic patterns.

Tilsen & Johnson in [90] propose a frequency-domain method for rhythm quantification, by studying the spectral content of the amplitude envelope. They claim that such an analysis gives a better representation of rhythm compared to the interval measures mentioned above. In [65] the authors present a mathematical modelling of speech rhythm with the utilization of a coupled oscillator model. (a bit more)

2.2 Listening Preference

Previous work that links speech rhythm to listeners' preference is presented in [77]. Proctor & Katsamanis examine what specific characteristics of speech affect the preferences and perceptions of listeners. For this cause they utilize a set features that also include rhythm metrics (interval measures and variability indices) and experiment with an audiobook corpus. However, their findings led them to the conclusion that the features and data they examined were insufficient to obtain a clear result regarding the prosodic characteristics that make a speaker more appealing to listeners.

In the study conducted by Kinoshita and Sheppard [44], the focus is on investigating how rhythm influences the ratings provided by native speakers when evaluating non-native (L2) speakers. The authors employed the pairwise variability index as a metric and successfully demonstrated its effectiveness as an indicator for this particular task. In [21] Ding et al. aim to validate whether glottalization has a negative effect on preference across multiple languages. The authors define glottalization "as a region in the speech signal with irregular pitch periods, and often accompanied by extremely low f0, voicelessness or pause". They conclude that preference is not negatively correlated with the frequency of glottalization and the listeners were receptive towards this characteristic.

In [61] the authors investigate the relationship between perceptual judgments and acoustic measures of normal and pathologic voices. Four acoustic parameters were analyzed: harmonics-to-noise ratio, autocorrelation function, average jitter, and standard deviation of the fundamental frequency. Correlations were found between perceptual preferences and measured results, indicating that certain acoustic characteristics are associated with voice preference. However, the acoustic-perceptual relationships differed between normal and pathologic voices, suggesting that additional factors contribute to overall voice preference in pathologic voices. In [8] the correlation between listeners' preference and fundamental

and formant frequencies is explored. The authors reach the conclusion that listeners' assessments of voice quality are impacted by the statistical distribution of mean fundamental frequency and formant frequencies in human voices. Coelho et al. in [15] utilize machine learning to build a model that automatically predicts voice preference. The model was trained to classify speakers across six languages to two classes, namely preferred and not preferred. The final classifier exhibited a sufficiently negligible error, and the results of feature selection provide valuable insights into the characteristics that influence preference. Nevertheless the model was trained and tested with a relatively small dataset (thirty voices), and therefore it is crucial to verify its generalization. Listeners' preference has also been studied in different contexts such as age and gender of listener [57], speech rate [85] and accent [60]. However, the accurate prediction of preference remains a hard and unsolved task in the contemporary scientific community.

2.3 Empirical Mode Decomposition

The concept of the Empirical Mode Decomposition (EMD) was originated by Huang et al. in [42]. The authors introduce EMD as a non-linear method to decompose a signal into meaningful, in terms of frequency, zero-mean AM-FM components. These components are called intrinsic mode functions (IMF) and they are extracted through a sifting process so that they satisfy certain desired properties. Each IMF contains valuable information regarding the oscillations of the signal within distinct and non-overlapping frequency ranges. More information about the EMD algorithm is provided in Chapter 3.

There is a lot of research in the bibliography that aims to the optimization and the theoretical foundations of EMD. Sharma et al. [84] conducted an extensive review of signal decomposition methods, with a particular focus on the advantages of using Empirical Mode Decomposition (EMD) for speech analysis compared to other methods such as Linear Prediction, Short Time Fourier Transform, and Mel Filterbank Cepstral Coefficients. The authors highlighted the strengths of EMD and also proposed techniques to enhance the EMD algorithm. However, they acknowledged the limitations and shortcomings of the method, primarily its lack of mathematical robustness. The paper provides valuable insights into the benefits and challenges associated with employing EMD for speech analysis. Some algorithmic variations of EMD are also proposed and evaluated in [80]. The authors emphasize that their work is primarily experimental in nature and acknowledged the need for further theoretical development of the empirical mode decomposition (EMD) method. In [79] the authors discuss the need for multivariate extensions of empirical mode decomposition (EMD) to enable direct analysis of multichannel data. They propose a method that utilizes real-valued projections on hyperspheres to compute the envelopes and local mean of multivariate signals, demonstrating its effectiveness through simulations on synthetic and real-world human motion data. An optimization for the decomposition process of EMD is also presented in [17].

They authors introduce an analytical solution using a parabolic partial differential equation (PDE)-based approach, to avoid computing the mean envelope during the sifting process. Flandrin et al. in [25] present initial numerical experiments suggesting that EMD behaves as a "wavelet-like" filter bank for structured broadband stochastic processes like fractional Gaussian noise. The findings highlight the need for theoretical explanations of EMD's observed behaviors and its potential for analyzing self-similar processes, while acknowledging the ongoing comparisons with other methods. The study aims to contribute to a better understanding of EMD's decomposition of broadband noise, bridging the gap between the lack of a formal theory and its practical application in real-world scenarios.

2.4 AM-FM Decomposition

AM-FM signals refer to signals that exhibit both amplitude modulation (AM) and frequency modulation (FM). AM-FM signals are used to model and analyze speech signals, considering the variations in both amplitude and frequency over time. Speech signals can be decomposed into their AM-FM components, where the AM component represents the envelope or the changes in the signal's amplitude, and the FM component represents the variations in the signal's frequency. This decomposition allows for a more detailed analysis of speech characteristics, such as the dynamics of speech production and perception.

Numerous algorithms for AM-FM decomposition can be found in the literature. In [83] the authors introduce a novel algorithm for decomposing a bandpass signal into its amplitude modulation and frequency modulation components. The algorithm utilizes zero-crossing instant information in a k-nearest neighbor (k-NN) framework to estimate the FM component, while the AM component is estimated using coherent demodulation and a time-varying lowpass filter based on the estimated instantaneous frequency. Gianfelici et al. in [30] propose a rigorous mathematical formulation for a multicomponent sinusoidal model of speech signals, enabling accurate reconstruction of nonstationary signals with transients, voiced segments, or unvoiced segments. The proposed method utilizes the Hilbert transform iteratively and includes an adaptive segmentation algorithm to compute instantaneous frequencies from unwrapped phases, ensuring a complete AM-FM model with analytically investigated and empirically tested convergence properties. In [23] an adaptive maximum windowed likelihood algorithm is presented and customized for decomposing speech signals into AM-FM components. By adjusting the window length and type, the algorithm can effectively decompose different segments of speech. The algorithm's ability to track formant frequencies is demonstrated through simulations of phonemes and voiced speech, except in cases of highly fluctuating formants where alternative approaches are recommended. The authors in [48] address the limitation of previous AM and FM decomposition methods that assume non-negativity of the AM component, which is not always valid

in over-modulated signals. The proposed two-step algorithm utilizes coherent demodulation to accurately estimate the AM and FM components, correcting phase discontinuities and improving instantaneous frequency estimation. The authors utilize evaluation results to demonstrate the algorithm's effectiveness in synthetic signals and band-passed speech, while also discussing its limitations and potential areas for future improvement.

In this thesis we utilize the AM-FM decomposition algorithm as proposed in [68]. The algorithm is iterative and adaptive and it is based on the Quasi-Harmonic Model (QHM), that was first introduced by Laroche in [46]. In each iteration, the authors propose updating the basis functions used for projecting the input signal by incorporating the information from the estimated instantaneous phase obtained in the previous iteration. The authors also present the validation of the method with synthetic data, which shows a significant improvement compared to the standard sinusoidal representation. More details can be found in [67] [66]. In [43] Kafetzis et al. present an extension of the model, where both the amplitude and frequency of the signal are incorporated into the adaptation process in a direct and straightforward manner. This alternation increases the performance in terms of Signal-to-Reconstruction-Error Ratio (SRER) by more than 2 dB, on average. In [69] the AM-FM decomposition algorithm is used to extract tremor. Tremor is defined as slow modulation of fundamental frequency or its amplitude. The advantage of the AM-FM decomposition is that it can adapt to non-stationarity, and thus it is not necessary to segment the signal into small intervals. The authors claim that this segmentation leads to the loss of important tremor information, and therefore the ability to extract the amplitude and frequency modulations in large speech signals is crucial. Their results indicate that the proposed method can robustly estimate the time-varying modulation frequency and the time-varying modulation level of vocal tremor.

Chapter 3

Methodology

3.1 Data

For our experimental evaluation, we used two Apple internal datasets. The first dataset consisted of recordings from 42 speakers, comprising 25 females and 17 males. All individuals were native English speakers with American English accent. Each speaker was asked to record speech files with four different speech rates: slow, normal, fast, and fastest. Within these speech files, all speakers delivered the same set of 25 sentences with small pauses in between. However, it is worth noting that not all speakers were able to provide all the required files. As a result, the dataset comprised 41 speech files for the slow rate, 41 for the normal rate, 39 for the fast rate, and 40 for the fastest rate. It is important to note that the audio quality of the recordings was not consistent across all files, as they were captured by the speakers themselves using personal devices. Additionally, there was variability in speech rates among the different speakers. Despite the slight variations in the number of files, these recordings formed the basis for our subsequent analyses and investigations. To facilitate analysis, the files were segmented into 25 smaller ones and any initial and final pauses were trimmed. The transcripts of all the speech files were also provided.

The talent selection process involved reaching out to an agency with specific requirements for American English-speaking actors, aged approximately 25-55 years old, who could perform at different speeds. The actors were provided with a character guide and script to perform in four different speed categories. The target speeds were 80 wpm, 150 wpm, and 300 wpm, although hitting these exact marks was not necessary. Voice sound, ability to speak at different rates, voice quality, and cold reading skills were among the criteria considered during the initial evaluation. The top 11 talents were selected for a 45-minute Skype callback session, where they were assessed on various aspects including cold reading, articulation at different speeds, ability to follow directions, and vocal stamina. After this phase, the top 4 talents were chosen for phase 1 recording, followed by further testing

and evaluation, including preference and intelligibility tests. Finally, the top 2 best-performing talents were selected for phase 2.

A second dataset was also provided, consisting solely of recordings provided by the four initially selected speakers. Each speaker contributed 120 speech files, comprising three different speech rates (slow, normal, fast), with 40 utterances per rate.

3.2 Rhythm Extraction with Envelope Decomposition

As a first step, we adopted the methodology proposed in [89] to extract the rhythm. To determine the rhythmic measures for each speaker, the following procedure was followed. The utterances were first segmented into chunks, with a duration of 1500 ms and an overlap of 750 ms. This choice was based on [89], acknowledging that it carries some degree of arbitrariness. For each speech chunk, the amplitude envelope was extracted and subjected to further processing. Three techniques were employed to extract the rhythmic features from the envelope: spectral analysis, Empirical Mode Decomposition (EMD), and AM-FM decomposition.

3.2.1 Envelope Extraction

The initial step of this analysis is to extract the amplitude envelope from the given speech signal, regardless of its size. To capture the variation in signal amplitude caused by the alternation between vocalic nuclei and consonantal margins, the speech signal is subjected to bandpass filtering. This filtering is achieved using a fourth-order Butterworth filter with a range of [400, 4000] Hz, serving two purposes. The low-frequency cutoff deemphasizes the contribution of fundamental frequency (F0), reducing the direct representation of voicing in the signal. This is done to make voiced consonants more similar to voiceless consonants and distinguish them from vocalic nuclei, which preserve resonances in the passband. The high-frequency cutoff decreases the representation of sibilant consonants and bursts, preventing them from being associated with peaks in the envelope. The authors in [89] explain that the specific values chosen for the passband cutoffs are somewhat arbitrary but have been determined through parametrically varied analyses of envelope-based metrics across different datasets.

The subsequent step involves low-pass filtering the magnitude of the vocalic energy. A fourth-order Butterworth filter with a cutoff of 10 Hz is employed, although the precise value of the cutoff is somewhat arbitrary. The purpose of the low-pass filter is to generate an envelope that varies on the time-scale of alternation between vocalic nuclei and consonantal margins, corresponding to the syllable-time-scale. The resulting output of the low-pass filtering process, termed the vocalic energy amplitude envelope or simply the "envelope," is depicted in Figure 3.1. The envelopes have been rescaled and the speech signals are divided in syllables. Figures 3.1a and 3.1b illustrate the envelope corresponding to the

same part of s sentence uttered by two different speakers, one male and one female respectively. Even though we hypothesize that each drop in the envelope corresponds to a syllable, Figure 3.1 demonstrates that this assumption does not always hold. For instance, in the last waveform, the final syllable consists of two drops, and represents the first syllable of the word "status".

It is recommended to also apply additional processing steps to make the signal more suitable for further analysis. First, the amplitude envelope is normalized by subtracting its mean and then rescaling it by its maximum absolute value, ensuring that it has an extremum at 1 or -1. Subsequently, the envelope is downsampled by a factor of 100 to expedite further computations. Finally, the envelope is windowed using a Tukey window with a parameter of 0.2. This choice of a moderate parameter value is aimed at minimizing the influence on a large time span of the signal and avoiding abrupt fluctuations near the edges. The resulting signal obtained after all these processing steps is referred to as the processed vocalic energy amplitude envelope.

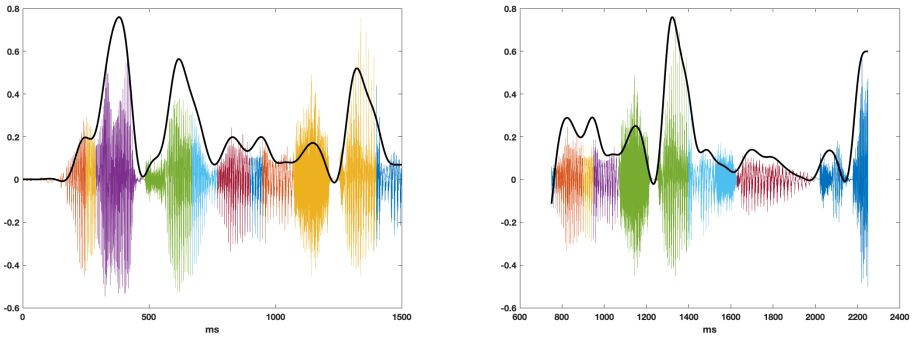
After obtaining the envelopes for all the chunks, they were used to extract a set of features that capture the rhythmicity. This was achieved by decomposing the envelopes using EMD and AM-FM techniques. Additionally, spectral analysis was performed on the envelopes to further analyze their characteristics and extract relevant information.

3.2.2 Empirical Mode Decomposition

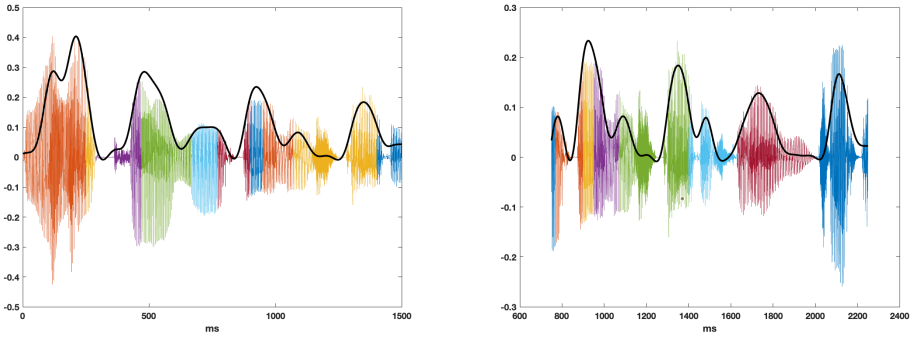
The EMD procedure is employed to decompose the envelope into a series of intrinsic mode functions (IMFs). These IMFs serve as a set of non-orthogonal basis functions. To ensure a well-defined instantaneous frequency, each IMF must satisfy certain properties. Firstly, the number of zero-crossings should be equal to the number of local extrema. This implies that all peaks and troughs are separated by a zero-crossing, indicating the oscillatory nature of the IMF. Secondly, for a function to be considered an IMF, the average of the negative and positive envelopes should be zero at every point. This condition ensures that the IMF possesses a zero local mean and captures the intrinsic oscillations present in the signal.

Each IMF represents an oscillation in the input signal at different time-scales. The first IMF reflects the highest frequency oscillations of the envelope, and thus we assume that it carries information about the "syllabic-driven fluctuations". Similarly the second IMF represents the second fastest oscillations of the envelope, and we consider that it corresponds to the "stress-driven fluctuations". As the order of the IMF increases the oscillations it reflects become slower. After a number of IMFs the power of the oscillations is reduced , and hence the IMFs go to zero. The number of the necessary basis functions can be determined with various criteria. Here, following the steps of [89], we use only the two first IMFs for the envelope decomposition, in order to extract the rhythmic features.

In EMD, as mentioned above, the basis function are non-orthogonal. They are



(a) Male speaker



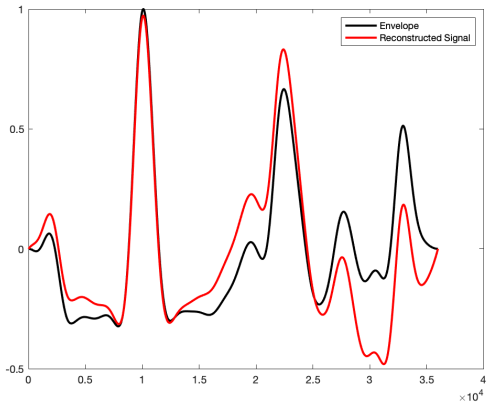
(b) Female speaker

Figure 3.1: Example of two vocalic energy amplitude envelopes

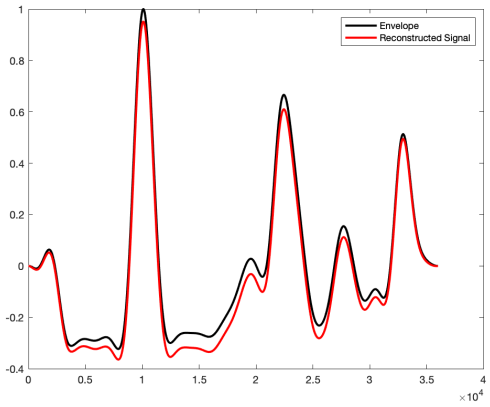
determined empirically through a sifting process. The sifting process is a step-by-step procedure used to extract intrinsic mode functions (IMFs) from the input signal. In this process, an initial sift is generated, which serves as a starting point for obtaining the IMFs. To obtain the first IMF the original signal is used as the initial sift. The next step consists of the calculation of the upper and lower envelope of the sift. This can be done by interpolating the local maxima and minima respectively. Using the upper and lower envelope we obtain the mean envelope, simply by taking their mean. A new sift is extracted by subtracting the mean envelope from the current sift. The new sift is then evaluated based on two specific criteria. The first criterion, known as the IMF criterion, assesses whether the sift satisfies the characteristics of an IMF. An IMF should have a zero-crossing between each pair of local extrema, ensuring that all maxima and minima are separated by zero-crossings. The second criterion, called the standard deviation criterion, is used to determine the stability and quality of the sift. It involves calculating the standard deviation of the sift and its preceding, namely the sum of the squared difference of the new sift and its preceding divided by the sum of the preceding sift squared. The standard deviation is then compared to a predefined threshold value. If the standard deviation exceeds the threshold, the sift is considered unstable and not suitable for an IMF. Here, we choose 0.01 as threshold. There are other methods that determine the termination of the sifting process, such as having a maximum number of iterations. If the sift meets both criteria, it is identified as an IMF. Otherwise, a new sift is generated, following the same method. This process is repeated iteratively until an IMF is obtained. Each iteration refines the sift by reducing the mean component, thereby enhancing the IMF characteristics.

After obtaining the first IMF, the residue is computed by subtracting the IMF from the original input signal. The residue contains the remaining components of longer time scales. The sifting process is then applied to the residue, aiming to extract additional IMFs.

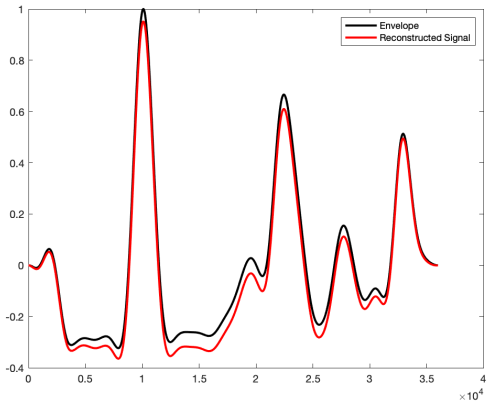
The sifting process continues as long as the residue exhibits two local extrema of sufficient amplitude. The resulting set of IMFs represents oscillatory components at different time scales present in the input signal. To reconstruct completely the original signal's envelope, the IMFs and the final residue are added together. Figure 3.2 illustrates the comparison between an amplitude envelope and its reconstruction with the use of different numbers of components: two, five, and ten. The reconstucted signal is obtained as the sum of the components. It is evident that utilizing only two IMFs results in a less accurate representation of the signal. However, the reconstruction becomes indistinguishable whether we employ five or ten IMFs. This observation suggests that higher-order IMFs do not contribute significant information and can be disregarded in subsequent analyses. The nature of the envelope indicates that it is a narrowband signal, thus a limited number of IMFs is expected to adequately capture its characteristics.This reconstruction process ensures that the energy and temporal structure of the original signal are preserved.



(a) 2 IMFs



(b) 5 IMFs



(c) 10 IMFs

Figure 3.2: Example of reconstructed envelopes after EMD, using 2, 5 and 10 IMF components

Algorithm 1 Empirical Mode Decomposition

Input:

S, the original signal

n, number of desired IMFs

Output:

$IMFs$, set of the n first IMFs

$IMFs \leftarrow empty\_set$

$residue \leftarrow S$

$i \leftarrow 1$

while $i \leq n$ **do**

  $IMFi \leftarrow sifting\_process(residue)$

  $residue \leftarrow residue - IMFi$

  $IMFs \leftarrow IMFs.add(IMFi)$

  $i \leftarrow i + 1$

end while

This observation is very crucial since it entails the IMFs contain energy from the envelope on a variety of time-scales inherent to the envelope itself. The authors in [89] use this to introduce a novel rhythmicity measure, which we also employ in this thesis. The idea is based on the fact that the first IMF predominantly encompasses oscillations at the syllable time scale, while the second IMF represents stress or foot-related oscillations. The measure quantifies the relative strength of these oscillations and is computed as the power of the second IMF (sum of squared values) divided by the power of the first IMF. This metric reflects the contribution of lower-frequency, stress-related energy compared to higher-frequency syllable-related energy. The authors call this metric $IMFR_{21}$.

After performing EMD and obtaining the IMFs, further analysis is conducted using the Hilbert transform. The Hilbert Transform is a mathematical operation that is used to analyze the time-varying properties of a signal. It provides a way to extract the instantaneous frequency and amplitude information from a given signal. The transform is based on the concept of analytic signals, which have a complex representation consisting of both real and imaginary components. By applying the Hilbert Transform to a signal, we obtain its corresponding analytic signal, which can be further analyzed to extract valuable information about its frequency content and phase characteristics. The analytic signal is given by:

$$H[x(t)] = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau} d\tau \tag{3.1}$$

Algorithm 2 Sifting Process

Input:

S, the initial sift

Output:

IMF, the ith IMF

$residue \leftarrow S$

$ZC \leftarrow False$

$std \leftarrow \infty$

while $std \geq 0.01$ **OR** $ZC = False$ **do**

$up\_envelope \leftarrow$ upper envelope of residue

$low\_envelope \leftarrow$ lower envelope of residue

$mean\_envelope \leftarrow mean(up\_envelope, low\_envelope)$

$new\_residue \leftarrow residue - mean\_envelope$

$std \leftarrow \sum (new\_residue - residue)^2 / \sum (residue)^2$

if $\#Zero\_Crossings(newresidue) = \#Local\_Extrema(new\_residue)$ **then**

$ZC \leftarrow True$

end if

end while

$IMF \leftarrow residue$

Return IMF

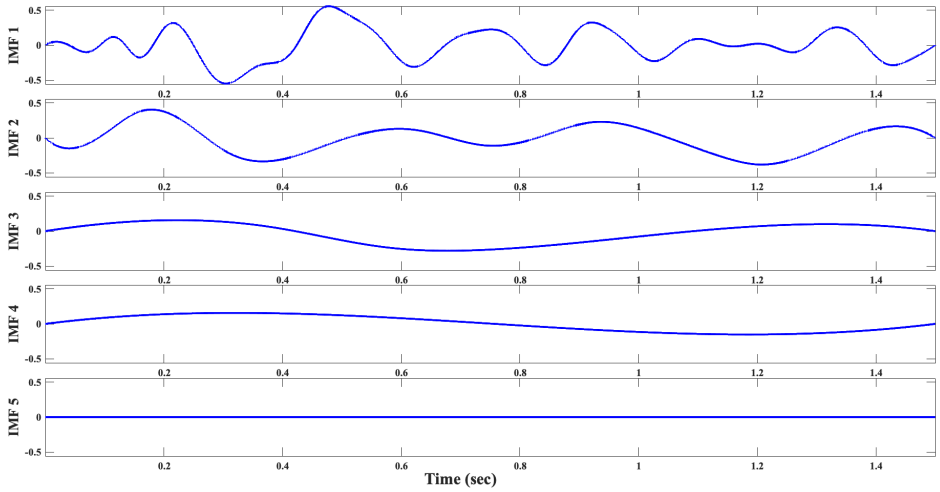


Figure 3.3: Example of the first four IMFs of an amplitude envelope.

where $H[x(t)]$ represents the analytic signal obtained from $x(t)$.

The instantaneous phase derived from the Hilbert transform allows us to define the instantaneous frequency as the time derivative of the phase [42]. Before obtaining the frequency phase unwrapping is performed to handle rapid changes, and each data point is smoothed by averaging with its nearest neighbors. Additionally, frequency values that deviate more than 3 standard deviations from the mean are excluded from subsequent analyses. To mitigate window-related edge effects, the first and last 100ms of instantaneous frequencies are also excluded.

The instantaneous frequency of the first IMF shows considerable variability. This variability is reduced as the order of the IMFs is increased. The variability in instantaneous frequency over time serves as an indicator of oscillation stationarity. In this thesis, following [89], we utilize the variance of the instantaneous frequency from each IMF as rhythmicity metrics. This variance reflects the rhythmic stability in the corresponding time scale. Lower values of variance indicate more stability and hence higher rhythmicity. Here only the first two IMFs, that resulted from the envelope decomposition, were used. Therefore, for every chunk of speech we obtain two rhythm metrics with this method.

3.2.3 AM-FM Decomposition

In this work, we propose extracting the same rhythm stability metrics, but with the utilization of an AM-FM envelope decomposition instead. The AM-FM algorithm was proposed by Pantazis et al. in [69] [67]. Since the envelope is a narrowband signal, the use of only one AM-FM component is sufficient for its representation. The instantaneous frequency of this component is obtained and its variance is used as a rhythmicity metric.

The decomposition algorithm is based on the adaptive Quasi-Harmonic modelling of speech. The procedure we followed was based on [69], where the authors use the algorithm to extract vocal tremor characteristics from speech signals. Firstly the envelope was downsampled to a 1000 Hz. sampling frequency. The downsampling doesn't cause any loss of information since the envelope does not contain high ($> 20Hz$) frequency modulations. As a next step the envelope is filtered with the use of a fourth-order Savinzky-Golay smoothing filter, to remove all the slow modulations, as we consider that they don't carry important information about the rhythmic content. This filter preserves important features of the distribution while reducing noise. Finally, the remaining modulations are modeled as an amplitude-modulated and frequency-modulated signal.

Following the notation of [69] and since we consider that the envelope is a mono-component signal, it can be represented as $x(t) = m(t)cos(\psi(t))$, where $m(t)$ represents the instantaneous amplitude (related to the modulation level), and $\psi(t)$ represents the instantaneous phase. The instantaneous frequency $\zeta(t)$ is calculated as the first derivative of the instantaneous phase. To estimate the AM-FM parameters, the aQHM algorithm is applied. The signal is divided into frames. A small hop-size of 15ms is used for better frequency resolution, and a Hamming window with a duration of 900ms is applied. For each frame the algorithm returns one value for the instantaneous amplitude and one for the instantaneous frequency, that correspond to the center of the frame. These values are determined, by solving analytically a least mean squares optimization problem and taking into consideration the values of the previous frame. Thus the algorithm requires an initial frequency estimation for the first frame. It was observed that different initial values resulted to different fits. One way to initialize the frequency of the first frame by making it equal to frequency that corresponds to the largest peak in the FFT of the frame. Here, in order to achieve the best fit for the component we search iteratively for the frequency value -in the interval between 5 and 10 Hz- that results to the highest Signal-to-Reconstruction Error Ratio. At the end the algorithm returns a set of values for the instantaneous amplitude and frequency, as many as the number of frames. The values are interpolated and we end up with the final curves of instantaneous amplitude and frequency. As the first value we have corresponds to the center of the first frame we cannot estimate the previous 450 (half the window size) values via interpolation. Similarly we can't estimate the last 450 values. To avoid losing information at the beginning and end of the envelope, we introduce Gaussian noise to the envelope's start and end. This ensures that the first sample of the envelope aligns with the center of the first frame, and the last sample aligns with the center of the last frame. By employing this approach, we preserve the integrity of the envelope and prevent information loss.

After obtaining the instantaneous frequency of the component, the next step is to calculate its variance. In this analysis, we adopt the assumption made by the authors in [89] that the variance of the instantaneous frequency reflects rhythmic stability. According to this assumption, as the variance of the instantaneous frequency decreases, rhythmicity increases. In other words, there is an inverse

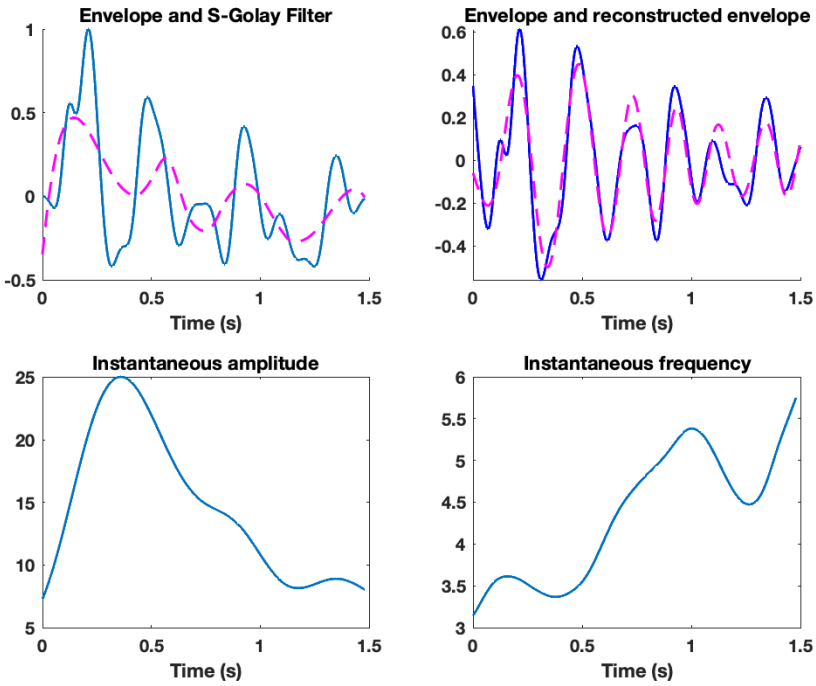


Figure 3.4: Example of an envelope, the AM-FM component and its instantaneous amplitude and frequency

relationship between rhythmicity and the variance of the instantaneous frequency.

3.2.4 Spectral Analysis

Another method to process the amplitude envelope and extract the information about rhythm it carries is by analysing its spectrum, as described in [90]. The envelope is first zero-padded and following the spectrum is computed by squaring the magnitude of the fast Fourier transform. The authors suggest dividing the spectrum by the length of the input and multiplying it by a factor of two in order to ensure smoothness and normalization. Then we perform smoothing across positive and negative frequencies by averaging within a 1 Hz frequency band centered on each frequency bin [13]. This smoothing procedure results in non-zero spectral power at 0 Hz. However, since the proposed spectral-based analysis does not incorporate spectral information below 1 Hz, this does not pose any significant issues.

Two different methods are employed to capture rhythm based on the envelope spectrum. The first approach involves separating the spectrum into low and high frequency bands and calculating the ratio of power within those bands. This

approach is similar to the previously described rhythm metric $IMFr_{21}$ as it is based on the same assumption, namely that lower-frequency periodicity in the envelope corresponds to supra-syllabic influence (such as stress or feet) on rhythm, while higher-frequency periodicity corresponds to syllable-level influence. This ratio metric is known as the spectral band power ratio (SBPr). Following our reference journal paper, the cutoff between the bands is chosen to be 3.25 Hz, representing a period of approximately 300 ms. The choice of cutoff is somewhat arbitrary but is generally set below the duration of typical syllables in fluent spontaneous speech.

The second approach is introduced because the definition of spectral bands is inherently arbitrary. To address this, the spectral centroid is computed over a range of 1.5-10 Hz. The centroid represents the weighted mean of frequencies, calculated by summing the frequencies within the range multiplied by their associated spectral power, and then dividing by the sum of all spectral power within the range. Unlike the ratio metric, the spectral centroid is not sensitive to the specific division between low and high frequencies but remains sensitive to the chosen frequency cutoffs.

3.2.5 Summary

The analysis of the envelope has given in total six rhythmicity metrics, which measures include $IMFr_{21}$, the variance of the instantaneous frequency of the first and the second IMFs, which will be symbolized as $varIMF_1$ and $varIMF_2$, the variance of the instantaneous frequency of the AM-FM component, denoted with $varAMFM$, ratio metric $SBPr$ and the spectral centroid, $CNTR$.

Before applying the metrics and analyzing the results, it was crucial to ensure that they measure the intended aspects. Validating these measures presents a challenge due to the absence of a definitive ground truth for comparison. However, we can establish their consistency by examining the pairwise correlation between them. This approach provides valuable insights into the extent to which our measures capture the same underlying phenomenon.

Pearson's Correlation	$varIMF_1$	$varIMF_2$	$varAMFM$	$IMFR_{21}$	$SBPr$	$CNTR$
$varIMF_1$	1.000	0.1517	-0.2411	0.3242	0.2409	-0.1794
$varIMF_2$	0.1517	1.000	0.5780	-0.5131	-0.63	0.7126
$varAMFM$	-0.2411	0.5780	1.000	-0.6247	-0.6905	0.7136
$IMFR_{21}$	0.3242	-0.5131	-0.6247	1.000	0.9222	-0.8457
$SBPr$	0.2409	-0.63	-0.6905	0.9222	1.000	-0.9607
$CNTR$	-0.1794	0.7126	0.7136	-0.8457	-0.9607	1.000

Table 3.1: Summary of rhythmic metrics and their pairwise correlation

To achieve this validation, we extracted all measures from the dataset with the 43 speakers. We used all speech files, regardless of speech rate. Every utterance was segmented into 1500ms chunks, and after extracting the measures for every chunk, the mean value was computed. Thus, every utterance was represented by one value per measure. Table 3.1 summarizes the pairwise correlation between each two of the metrics. The number of observations was equal to the number of total utterances.

These results provide valuable insights into the metrics used. Firstly, it is evident that $varIMF_1$ exhibits the lowest correlation with the other metrics. As we have previously established, IMF_1 represents rhythmicity at the smallest timescale, specifically at the syllabic level. In contrast, we assume that IMF_2 reflects rhythmicity at a supra-syllabic level. The relatively low correlation with IMF_1 and the relatively high correlation with IMF_2 suggest that the other metrics capture rhythm on a larger time scale beyond the syllabic level.

Furthermore, the relatively high correlations of $varAMFM$ with the other metrics validate that our proposed measure can indeed be used as a rhythmic metric. Another crucial observation is that, excluding $varIMF_1$, we can categorize the metrics into two groups. Metrics within the same category positively correlate with each other and negatively correlate with metrics from the other category. The first category includes $varIMF_2$, $varAMFM$, and $CNTR$. Based on our assumption, inspired by [89], these measures increase inversely to rhythmicity. In other words, lower values of these metrics indicate higher rhythmicity. Drawing from this assumption and the highly negative correlations observed, we can conclude that the metrics in the second category, namely $SBPr$ and $IMFR_{21}$, increase alongside rhythmicity.

Lastly, the metrics $IMFR_{21}$, $CNTR$, and $SBPr$ exhibit the highest correlations among each other. This can be attributed to the methods used to construct these measures. All three metrics quantify the ratio of power between lower and higher frequencies. Although they show strong correlations, suggesting the possibility of retaining only one of them, we choose not to exclude any from further analysis. This decision is based on the understanding that each metric is sensitive to different aspects of rhythmicity, and thus they provide complementary information that enriches our analysis.

3.3 Rhythm Extraction from the Speech Signal

Despite strong evidence indicating that the amplitude envelope carries rhythmic information in a speech signal, there are limitations associated with this approach. We recognize that extracting the envelope results in a significant loss of information, potentially excluding crucial aspects of rhythmic content, such as pitch. To

address this concern, we propose an alternative method to capture rhythmic characteristics directly from the speech signal, aiming to obtain a more comprehensive representation of the rhythm. Building upon the assumptions made in the previous section, our method focuses on extracting features that enable a fair comparison between speakers. By leveraging these features, we aim to capture and quantify rhythmic characteristics in a manner that allows for meaningful comparisons across different individuals.

When it comes to comparing speakers based on their speech rhythm using the metrics, we encounter additional limitations with the previous method. This comparison is crucial as we aim to investigate whether rhythmicity has an impact on preference. However, it has been observed that speech rate can introduce a bias to our metrics and the variations among speakers might primarily reflect their individual rates rather than their rhythmicity.

To address this issue, we require a method that enables a fair comparison of speakers. This method should provide a means to evaluate speakers on an equal basis, ensuring that the comparison is not influenced by variations in speech rate or other factors that could introduce bias to the analysis. Moreover, it is crucial to acknowledge that rhythm manifests itself at the syllabic or supra-syllabic level. While our previous approach involves approximating syllables based on the envelope and its oscillations, it is important to recognize the inherent limitations of this approximation.

In summary, our proposed method serves three purposes.

1. **Extract rhythmic features directly from the speech**: Instead of relying solely on the envelope extraction approach, we aim to extract rhythmic features directly from the speech signal. By doing so, we can avoid potential information loss and capture relevant rhythmic characteristics that may contribute to preference.

2. **Develop a fair pairwise comparison**: In order to conduct a fair comparison between speakers, it is essential to have a fair and unbiased pairwise comparison between speakers. This approach will enable us to compare speakers based on specific rhythmic features, thereby eliminating potential biases introduced by variations in speech rate or other confounding factors.

3. **Accurately capture rhythmicity related to syllable transitions**: We recognize that rhythmicity is influenced by the transitions between syllables. To capture this aspect accurately, we will focus on refining our methodology to specifically target and analyze rhythmic patterns associated with syllable transitions.

In order to achieve all this, we used the text-speech pair from the dataset and convert the text/transcript into phoneme sequences using Apple internal graphene to phoneme conversion tool. Afterwards, we performed forced alignment on the speech signal to identify the boundaries of each phoneme sequence. Subsequently,

we use this information to separate syllables. It is important to note here, that in our dataset the utterances were common among speakers and rates. In order to study the transitions between syllables, we propose performing EMD separately on the part of the signal that contains two consecutive syllables, to extract the rhythmic metrics.

3.3.1 Forced Alignment

In order to compute the phoneme boundaries for all the phoneme sequences in our speech dataset, we used a forced alignment tool called Montreal Forced Aligner (MFA). [55]

The MFA provides automatic alignment of speech data with corresponding text transcriptions at a high level of accuracy. This alignment process is crucial for various applications, such as speech recognition, speaker diarization, and language processing tasks.

MFA operates by deploying a combination of acoustic and language models to align the speech and text data. Initially, it performs a phonetic segmentation of the speech signal, dividing it into smaller units corresponding to individual phonemes. It then utilizes a statistical model to estimate the most likely temporal boundaries for each phoneme, aligning them with the corresponding words in the text transcription.

The MFA algorithm uses the following procedure. The alignment process begins with data preparation. The speech recordings and their corresponding transcriptions are pre-processed and formatted. This includes cleaning the data, removing non-speech sounds, and normalizing the audio. Once the data is prepared, an acoustic model is trained. The acoustic model is a statistical model that captures the acoustic characteristics of different phonemes and their variations. The acoustic model is trained using a large amount of labeled speech data. After the model training, a lexicon is created. The lexicon maps each word in the transcriptions to its corresponding phonemes. This ensures accurate alignment between the words and phonemes in the speech signal.

The alignment process itself is performed in two steps: forced alignment and refinement. Forced alignment breaks the speech signal into small, overlapping windows. The acoustic model is then used to predict the likelihood of each phoneme occurring in each window. The alignment is performed by selecting the phoneme with the highest likelihood at each window. Refinement involves adjusting the boundaries of phoneme segments to improve their alignment with the speech signal. This is done using a Hidden Markov Model (HMM) framework.

The final output of the algorithm is a phoneme-aligned transcription, where each phoneme is temporally aligned with its corresponding segment in the speech signal. This output was used to separate syllables. The syllable separation was done manually.

3.3.2 Extraction of Rhythmic Metrics from Syllables

The main idea behind the syllable separation was to be able to study the speech signal during the transition between syllables. To achieve that, we utilized EMD for a second time. This time the decomposition was performed on the speech signal, instead of the envelope. The EMD algorithm was applied for every two consecutive syllables, aiming to examine the signal's oscillations for every syllable transition. Since speech is a much more complicated signal than its envelope, more IMF components were needed to get a good representation. Figure 3.5 depicts the five first IMF components of a speech signal. It is evident that even the fifth IMF carries information about the signal, and has relatively high power, as opposed to the envelope's higher order components shown in Figure 3.3.

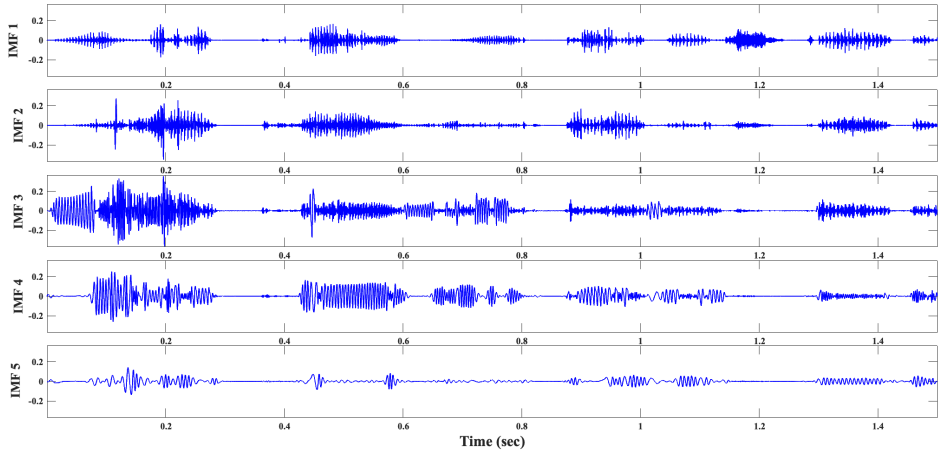
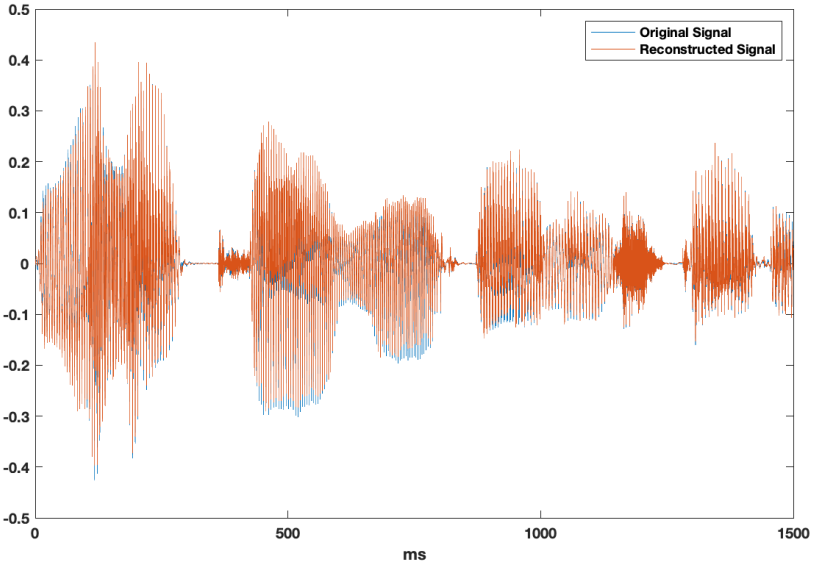


Figure 3.5: Example of five first IMF components of a speech signal

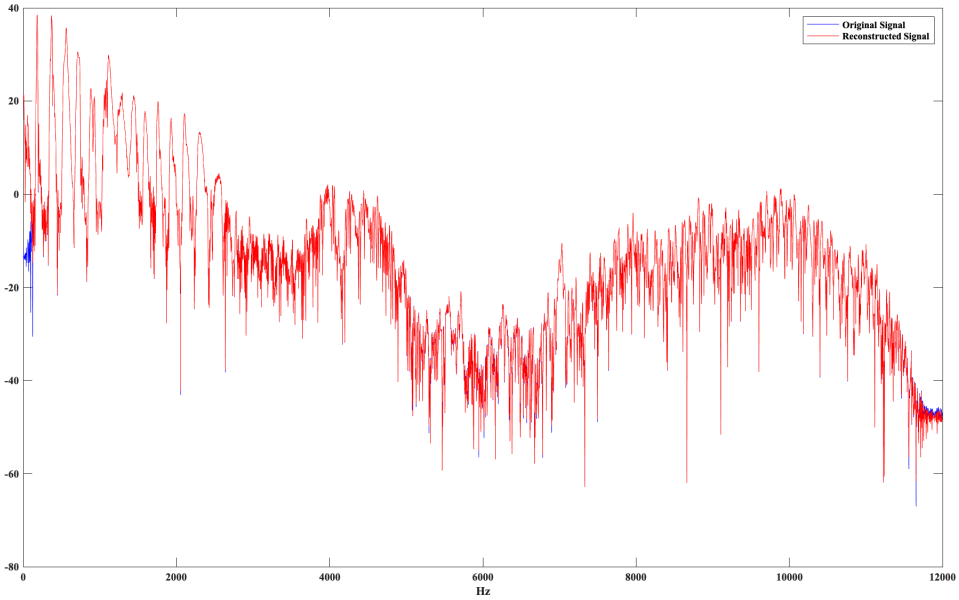
We selected the first five Intrinsic Mode Functions (IMFs) for this analysis. Figure 3.6 illustrates that the original signal and its reconstruction using these five components exhibit a high degree of overlap, indicating that the important signal information is captured within these IMFs. The spectral comparison in Figure 4.6d demonstrates a close resemblance between the spectrum of the reconstructed signal and the original signal, with minor discrepancies observed primarily in the lower frequency range. This discrepancy can be attributed to the exclusion of higher-order IMFs, which correspond to slower oscillations or lower frequencies. Each segment of the signal associated with a syllable transition was decomposed into five IMF components, and the Hilbert transform was applied to extract the instantaneous frequency. The variance of the instantaneous frequencies for each IMF was then utilized as rhythmic metrics.

After completing the aforementioned procedure, we have acquired five features corresponding to each syllable transition. As all speakers utter the same sentences, we can perform a one-to-one comparison of these features. This enables us

to directly compare the rhythmic characteristics among speakers and analyze any variations or patterns that may exist. By examining these features, we can gain insights into the individual rhythmic profiles of each speaker and explore potential relationships between rhythmicity and other factors of interest. The ability to conduct a direct and meaningful comparison of these features facilitates our investigation into the role of rhythm in speech and its potential effects on various aspects, such as preference or perception.



(a) Comparison in time domain



(b) Comparison in frequency domain

Figure 3.6: Comparison of original speech signal and reconstructed with five IMF components

Chapter 4

Results

This chapter presents the findings of the research study, which sought to investigate the measurement of speech rhythm and its influence on listeners' preferences. Initially, we present the outcomes and conclusions derived from employing the envelope-based metrics. We analyze their limitations and shortcomings and subsequently turn to the utilization of the second proposed method that circumvents the use of the envelope. By adopting this alternative approach, we anticipate achieving more promising results and addressing the shortcomings encountered earlier.

4.1 Experimental Evaluation of Envelope Based Methods

Through the analysis of measures derived from the decomposition of the amplitude envelope, we have gained valuable insights into the efficacy of these measures in capturing rhythmicity as well as their inherent limitations. Additionally, our investigation has shed light on the influence of speech rate on these metrics. The conducted experiments, utilizing two distinct datasets, have allowed us to assess the performance of the proposed rhythmic metrics in various tasks, providing valuable empirical evidence.

4.1.1 Effect of Speech Rate

One of the initial experiments conducted aimed to assess the sensitivity of our metrics to variations in speech rate. These results hold significant importance for two main reasons. Firstly, they serve as a crucial validation step for our metrics, as rhythm is inherently influenced by speech rate and our measures should accurately capture this relationship. Secondly, it is essential to ensure that our metrics do not solely reflect speech rate as that would render them ineffective in capturing other aspects of speech rhythm. By examining the impact of speech rate on our metrics, we not only validate their sensitivity but also assess their ability to capture broader rhythmic characteristics beyond rate variations. These findings provide valuable

insights into the robustness and reliability of our proposed metrics in quantifying speech rhythm.

As an initial step, we implemented our envelope-based methods on the first dataset which comprised 42 distinct speakers delivering speech at four different rates. The speech files were segmented into chunks of 1500 ms duration with a 750 ms overlap. For each speech chunk, we extracted the measures mentioned in Table 3.1. Based on the rate of speech, the files were categorized into four distinct groups. To assess the variations in the distributions of the measures across these rate groups, we conducted a one-way ANOVA test. The results of this analysis are visually depicted in Figure 4.1, providing insights into the differences observed among the measures' distributions across the different speech rates.

Note that, the null hypothesis of the test assumes that all groups are the same while the alternative hypothesis suggests that at least two of the groups are different. The significance of the results is reflected by the F-statistics and the p-value. The F-statistic is a measure of the ratio of the variability between the groups to the variability within the groups. A larger F-statistic indicates stronger evidence of differences among the groups. The corresponding p-value associated with the F-statistic provides the probability of obtaining such a result purely by chance. The small p-values obtained in our analysis reject the hypothesis that all groups are from the same distribution, indicating significant differences among the groups. However, this overall conclusion does not provide specific information about the pairwise statistical differences among the groups. Therefore, it is necessary to conduct additional pairwise comparison tests to determine which specific group comparisons show significant differences. For the pairwise comparison tests, Tukey's Honestly Significant Difference (HSD) [94] procedure was used to determine which specific group means differ significantly from each other. It calculates a critical value called the Honestly Significant Difference which represents the minimum difference between group means that is considered statistically significant. The Tukey HSD procedure takes into account the overall significance level (we used 95%) and the number of groups being compared. It adjusts the significance level for each pairwise comparison to maintain an overall family-wise error rate, which helps control for multiple comparisons.

It is evident from the analysis that all measures, except for $varIMF_1$, exhibit clear sensitivity to speech rate. By combining the findings from Table 3.1 with the results depicted in Figure 4.1, it becomes evident that there is a linear decrease in rhythm as the speech rate increases. This finding aligns with our intuition as maintaining rhythmicity becomes more challenging in faster speech, resulting in a "flatter" speech pattern. Pairwise comparisons were conducted among the rate groups, revealing statistically significant differences in all cases except for two. Specifically, for $varIMF_1$, the means of the Normal and Fast groups as well as the Fast and Fastest groups were statistically equal. Additionally, for $IMFr_{21}$, there was no significant difference between the Fast and Fastest groups.

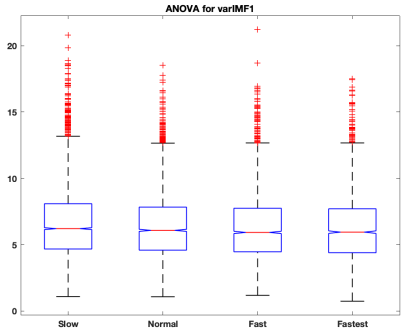
Among the metrics analyzed, the most sensitive ones to speech rate were be

$CNTR$ and $SBPr$. These measures demonstrated notable variations corresponding to different speech rates, indicating that potentially small differences in rate might cause bias to their estimation of rhythm. On the other hand, the inability of $varIMF_1$ to capture the change in rhythmicity resulting from the increase in speech rate raises doubts about its suitability as an indicator of rhythm. This observation is further supported by its lack of correlation with the other metrics examined in the study. Further investigation and refinement of this metric may be necessary to enhance its relevance and reliability as a rhythmic indicator in future studies.

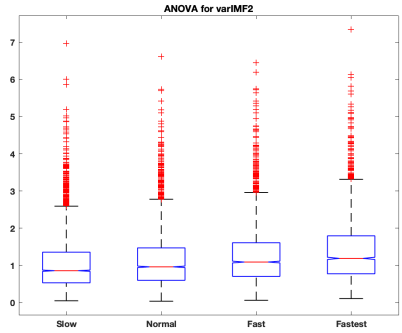
The presence of numerous outliers in the data is not surprising considering that each group comprises different speakers and each speaker has their unique rhythm. Consequently, the data points within each group do not necessarily belong to the same distribution. Moreover, due to the unavoidable individual differences in speech rate among speakers, the groups are not only heterogeneous in terms of rhythm but also in terms of rate. Since the recordings were done by the speakers' personal devices, there was also variability in the recording conditions. Despite these inherent limitations, the outcomes of this experiment offer a certain level of validation for the efficacy of our metrics. While the heterogeneity of the data introduces variability, the overall trends observed in the relationship between speech rate and rhythmic measures support the notion that our metrics can capture and quantify rhythmicity to some extent.

To mitigate the variability and heterogeneity observed in the previous dataset, a similar test was conducted using a dataset consisting exclusively of studio recordings from the four initially selected speakers. This approach aimed to achieve more consistent and uniform results by minimizing the influence of different speakers and environmental factors. By focusing on a smaller group of speakers with controlled recording conditions, we expected to obtain more reliable and comparable data for further analysis. The results from this experiment were more clear and lead us to the same conclusion. The results of the ANOVA test consistently demonstrated a similar pattern across all cases, albeit with minor variations. Notably, all metrics exhibited consistent trends of increasing or decreasing values as shown in Figure 4.1. Of particular significance, the metrics of $SBPr$ and $CNTR$ remained consistently the most responsive to changes in speech rate, showcasing their heightened sensitivity in capturing this type of variations. At the same time, the number of outliers was significantly reduced as expected.

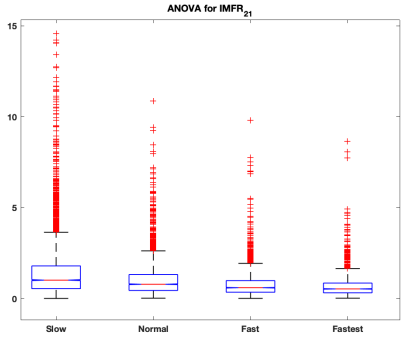
One noteworthy observation concerns the behavior of $varIMF_1$. The corresponding results are depicted in Figure 4.2. We observe that the finally selected speakers (Male Speaker 1 and Female Speaker 2) exhibit a steady decrease of this metric as rate increases, as opposed to the other two. This pattern may indicate a higher degree of rhythmic stability in the selected speakers. It could suggest that they maintain a consistent rhythm or exhibit more consistency in the way they adjust their speech speed.



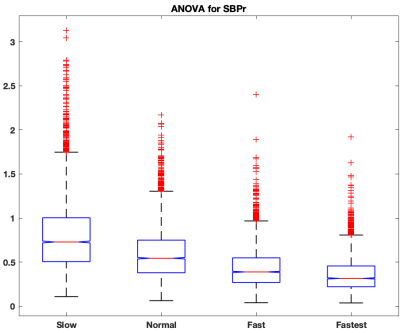
(a) $F(3, 16061) = 14.76, p < 0.001$



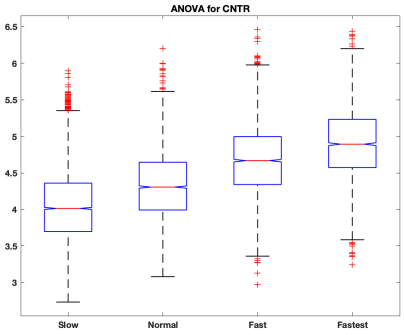
(b) $F(3, 16061) = 146.58, p < 0.001$



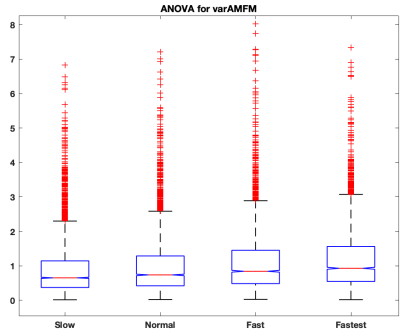
(c) $F(3, 16061) = 202.75, p < 0.001$



(d) $F(3, 16061) = 1607.2, p < 0.001$

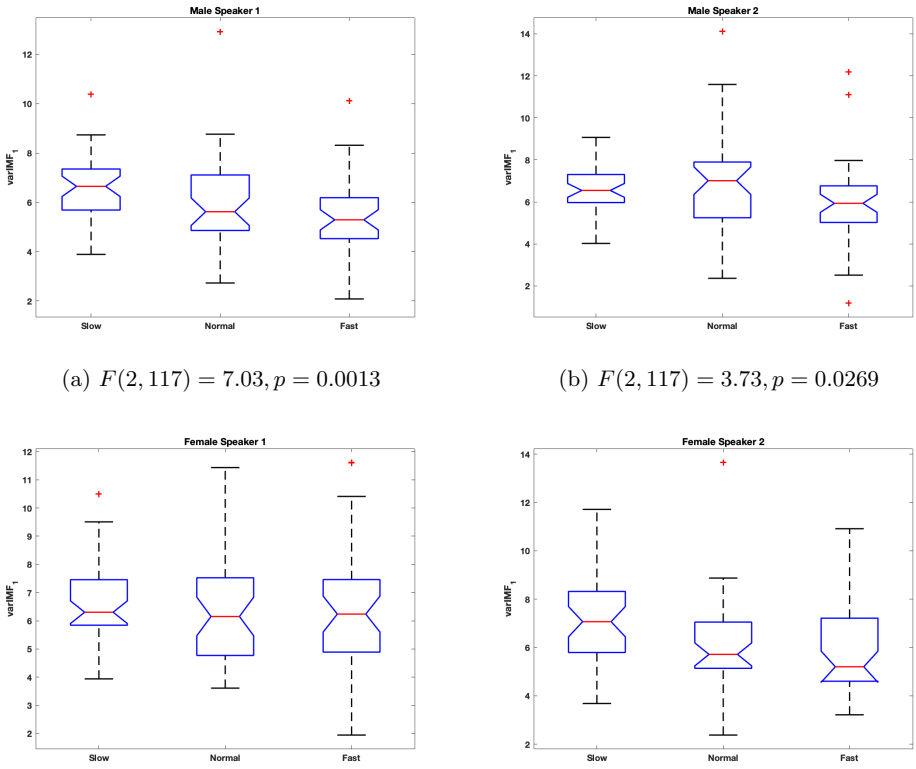


(e) $F(3, 16061) = 2345, p < 0.001$



(f) $F(3, 16061) = 90.75, p < 0.001$

Figure 4.1: Results of ANOVA tests between different speech rate groups, that include utterances of all 42 speakers. In $F(3, 16061)$, 3 represents the between-groups degrees of freedom (number of the groups minus one) and 16061 reflects the within-groups degrees of freedom. The within-groups degrees of freedom is equal to the total degrees of freedom (number of total observations minus one) minus the between-groups degrees.



(a) $F(2, 117) = 7.03, p = 0.0013$

(b) $F(2, 117) = 3.73, p = 0.0269$

(c) $F(2, 117) = 0.36, p = 0.69$

(d) $F(2, 117) = 4.84, p = 0.009$

Figure 4.2: Results of ANOVA tests between different speech rate groups for each one of the preferred speakers.

Having established the sensitivity of our metrics to variations in speech rate, it is imperative to verify that this sensitivity is not solely driven by rate differences. Consequently, we need to delve into whether the observed sensitivity to rate is indeed rooted in the variations of rhythm associated with different speaking rates. To tackle this challenge, we conducted research on the connection between our metrics and listeners' preference. This approach was driven by our initial assumption that rhythm plays a significant role in shaping preference, namely that higher rhythmicity renders a speaker more preferable.

4.1.2 Metrics and preference

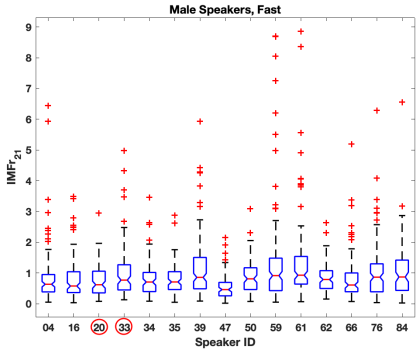
To explore the relationship between our metrics and preference, we used the first dataset. The utterances within the dataset were categorized into four groups based on their speech rate: slow, normal, fast, and fastest. Additionally, we conducted separate analyses for male and female speakers considering that the comparison of

preference was within the same gender.

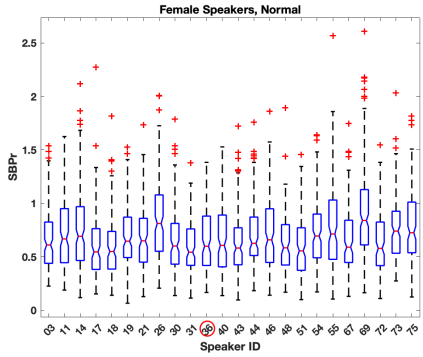
All speakers in the dataset have an assigned ID number. Among the male speakers, the two initially preferred speakers were 033 and 020. Ultimately, speaker 033 was selected as the final choice. As for the female speakers, initially 036 and 049 were preferred, with 036 being the final selection. Speaker 049 is absent in the analysis within the normal speech rate because they did not provide a speech file suitable for our study (it contained only one out of 25 sentences). It should be noted that the difference between these speakers was minimal and the selection process was not solely based on voice quality. Other factors, such as the ability to perform specific vocal tasks, were also taken into consideration during the selection process. We expect that the metrics that correspond to these speakers should be distinguishable and reflect higher rhythmicity. Furthermore if our assumption holds true that these metrics effectively quantify rhythm, they should be capable of differentiating between speakers as each individual possesses a unique rhythm in their speech patterns.

To examine the separation between different speakers, we conducted a one-way ANOVA test. This test allowed us to assess the statistical significance of the differences among the metrics within each gender and rate group. In total, we performed 48 tests, considering the combination of 2 genders (male and female), 4 rate groups (slow, normal, fast, fastest), and 6 metrics. In each ANOVA test, the groups were defined based on the individual speakers, and the observations within each group were determined by the number of 1500 ms chunks derived from each speaker's utterances.

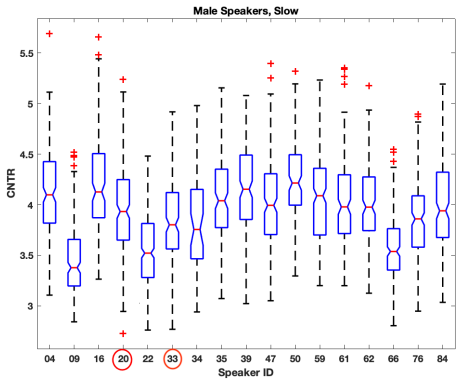
The findings from our analysis revealed that the desired performance of our metrics was not achieved as anticipated. Specifically, concerning the metrics $IMFR_{21}$, $SBPr$ and $CNTR$, while they were successful in distinguishing speakers with statistically significant differences, the selected speakers did not exhibit extreme values in these metrics. Figure 4.3 depicts indicatively some of the ANOVA test results that correspond to these three metrics. It includes plots from both genders and all rates. This observation raises concerns regarding the potential bias introduced by rate differences among the speakers within the same rate group. The similarity observed in the rankings of speakers and the high sensitivity of these metrics to rate variations further suggest that their outcomes might be influenced by the rate disparities present in the dataset. While it is true that different speakers may vary in their natural speech rate, these differences alone do not determine preference or intelligibility. Therefore, it is important to recognize that any bias arising from variations in speech rate can potentially lead to misleading conclusions.



(a) $F(14, 1574) = 5.95, p < 0.0001$



(b) $F(23, 3329) = 10.02, p < 0.0001$



(c) $F(16, 2837) = 60.51, p < 0.0001$

Figure 4.3: Results of ANOVA tests between different speakers among the same rate and gender group. The circled ID numbers corresponds to the preferred speakers.

Correlation between speech rate and metrics							
Gender	**Rate Group**	**varIMF$_1$**	**varIMF$_2$**	**varAMFM**	**IMFR$_{21}$**	**SBPr**	**CNTR**
Male Speakers	Slow	-0.51	0.37	0.22	-0.83	-0.75	0.70
	Normal	-0.61	0.02	0.12	-0.70	-0.59	0.52
	Fast	-0.24	0.53	0.59	-0.77	-0.89	0.85
	Fastest	-0.53	0.46	0.68	-0.77	-0.90	0.91
Female Speakers	Slow	-0.28	0.61	0.43	-0.84	-0.89	0.87
	Normal	-0.37	-0.1	0.07	-0.87	-0.81	0.74
	Fast	-0.42	0.46	0.36	-0.88	-0.91	0.89
	Fastest	-0.02	0.58	0.5	-0.62	-0.72	0.77

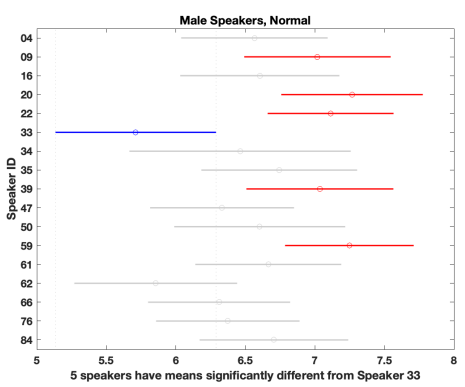
Table 4.1: Summary of correlation between speakers' speech rate and the corresponding metrics. The red color denotes correlation values above 0.7 which is considered high.

To validate this assumption, we calculated the approximate speaking rate for each speaker within the same speech rate group. It is essential to acknowledge that these rates were computed automatically, introducing the possibility of some error in the calculations. Additionally, it is important to note that some speakers did not strictly follow the provided transcript, resulting in slight misalignment between the text and their actual speech. Despite this limitation, we believe that these results can provide insights on how rate effects our metrics.

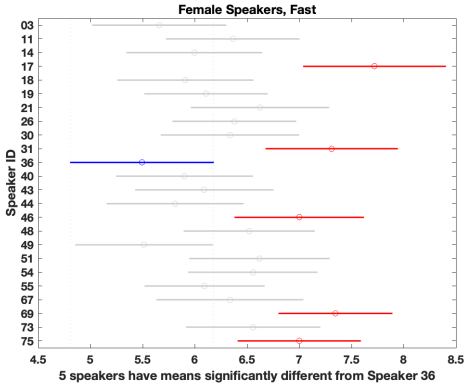
For each rate group and metric, we computed the correlation between the speech rates of all speakers within that group and the corresponding mean values of the metric (average of all chunks). Table 4.1 provides an overview of our results. The findings confirm our initial assumption, as the metrics $IMFR_{21}$, $SBPr$ and $CNTR$ demonstrate a significantly high correlation with the speech rate. The observed high correlation between these metrics and speech rate suggests that these metrics are heavily influenced by rate variations and may not accurately capture rhythmicity. Consequently, they may not be reliable predictors of preference. Therefore, alternative metrics or approaches need to be explored to capture the desired aspects of rhythmicity and its impact on listener preference. It is worth noting that the high variance among the correlation values can be attributed to several factors. Firstly, the estimation of speech rates may not be perfect, introducing some degree of error in the calculations. Additionally, the variable number of speakers within each gender and rate group can also contribute to this variance. The smaller sample sizes in certain groups may lead to more variability in the correlation values.

In contrast to the metrics discussed earlier, the metrics $varIMF_1, varIMF_2$, and $varAMFM$ show smaller correlation values with speech rate. This suggests that while there may still be some influence of rate on these metrics, it is not as definitive or strong. The smaller correlation values indicate that these metrics have the potential to capture aspects of rhythmicity beyond just rate variations. It is encouraging to see that these metrics exhibit a weaker association with rate, as it suggests they may offer a better opportunity to quantify rhythmicity and potentially predict preference.

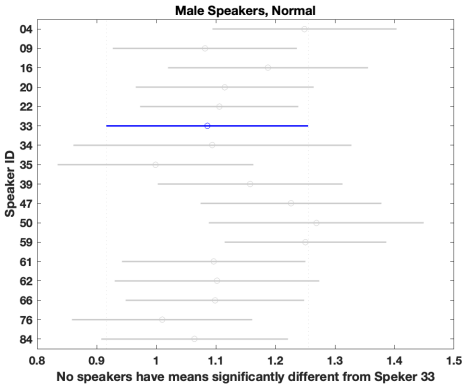
However, despite their weaker correlation with speech rate, the metrics $varIMF_1$, $varIMF_2$ and $varAMFM$ were unable to effectively separate the means of the distribution that corresponded to each speaker. Figure 4.4 illustrates the comparison of the means of the metric values for the selected speaker (denoted by the blue line) and the other speakers. The figure only includes the normal rate for male speakers and fast rate for female speakers, but it is representative of the overall trend. From the figure, it is evident that there is no statistically significant separation between speakers, including the selected speakers and the rest of the group. There is a partial separation observed in the case of $varIMF_1$ and the selected speakers have obtained smaller values which indicates higher rhythmicity. However, it is not sufficient to draw definitive conclusions.



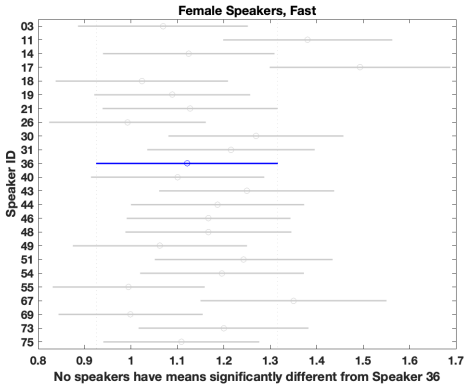
(a) $varIMF_1$



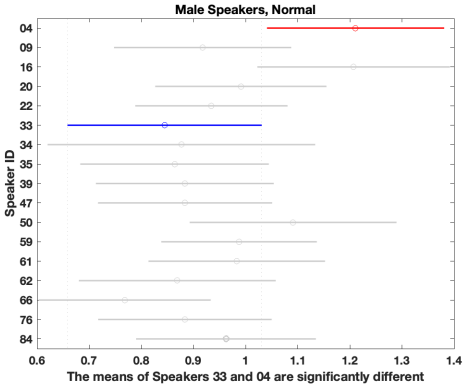
(b) $varIMF_1$



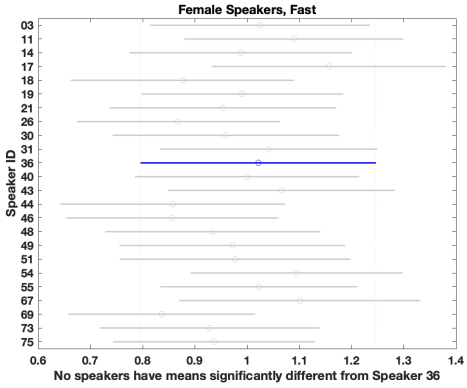
(c) $varIMF_2$



(d) $varIMF_2$



(e) $varAMFM$



(f) $varAMFM$

Figure 4.4: Results of pairwise comparison of means. The blue line corresponds to the selected speaker, the red lines to the speakers with a significantly different mean and the gray lines to the rest of the speakers. The circle symbol denotes the mean value for every speaker (x-axis). The length of the line reflects the confidence interval (95%). In order for two groups to be significantly different, their lines should be non-overlapping.

Therefore, despite these metrics potentially reflect rhythmicity to some extent, they are not suitable for the current task. They do not effectively distinguish among speakers and therefore do not provide meaningful information about listener preference. These limitations have motivated us to explore an alternative approach that shows more promise and potential for accurately capturing rhythmicity and predicting preference.

4.2 Experimental Results for Syllable Transition Method

Motivated by the failure of the previous metrics, we used the method described in 3.3 to extract rhythmic features directly from the speech signal. This way we obtained five measures, one for every utilized IMF that corresponded to each syllable transition in every speaker's utterances.

Initially, it was crucial to identify the IMF that yielded the most effective separation ability among speakers. Given our focus on differentiating the speakers we know were preferred, we conducted a comparative analysis of the results for each metric between the selected speaker and each of the remaining speakers. We utilized again the dataset with the 42 speakers. It is important to acknowledge that in order to ensure a consistent comparison between speakers' utterances, certain criteria were applied, leading to the exclusion of some speakers from the analysis. Specifically, speakers who did not adhere to the given script were excluded. This included speakers who did not complete the script, those who stated the sentence number before each sentence or individuals who pronounced certain parts significantly differently (e.g., pronouncing "2021" as "two thousand twenty-one" instead of "twenty-twenty-one"). As a result of these criteria, the number of male speakers included in the analysis was reduced to 11 for the normal rate and 7 for the fast rate. Similarly, the number of female speakers included was reduced to 19 for the normal rate and 14 for the fast rate. The analysis did not cover the slowest and fastest rates; it focused solely on the normal and fast rates of speech.

As all speakers recorded the same set of utterances, including the same syllables, it allowed for a meaningful comparison of results on a one-to-one basis for each syllable transition. This comparison was implemented by obtaining the difference of the metrics' result of the selected speaker minus the results for each one of the other speakers. This resulted to a sequence of values equal to the number of syllable transitions in the dataset, namely 402. To examine how consistently the selected speakers were separated from the rest we studied the transition of the sign of this difference. In other words, we looked at whether the differences between the selected speaker and the other speakers were predominantly positive or negative. This experiment was conducted on the normal speech rate utterances.

The findings of this comparison were promising, although not conclusive. It was observed that the male speaker ultimately chosen, 033, did not demonstrate values that could provide a definitive conclusion. However, both the male speaker ranked "second" in the selection process (020) and the chosen female speaker (036)

consistently displayed higher values compared to the other speakers. Considering the difficulty in selecting between the two male finalists, indicating that both possess high-quality voices, we proceeded with our analysis based on this observation.

Table 4.2 summarizes the percentage of syllable transitions where these two speakers showed greater values than each one of the rest. Upon analyzing the data, it becomes evident that the metric $varIMF_1$ demonstrates better separation ability for the preferred speakers compared to the other metrics. Specifically, for male speakers, both $varIMF_3$ and $varIMF_4$ show promising results in terms of distinguishing power. However, among the female speakers, only $varIMF_1$ performs well in this regard. It is worth noting that female speakers 011 and 018 displayed larger values overall compared to the selected speaker, 036. This is a second indicator that while the metric appears to perform well, it may not provide definitive results. However, it is important to consider that both speakers 011 and 018 were initially included among the ten speakers chosen for the second round of auditions. This observation suggests that the metric's effectiveness in separating the preferred speakers is not absolute, and additional factors and considerations should be taken into account during the evaluation process.

These findings contradict our initial assumptions, which suggested that higher rhythmicity would be indicated by lower variance in the instantaneous frequency of the IMFs, and that higher rhythmicity would be preferred. However, there is a potential explanation for this outcome. When we obtained these measures from the envelope of a larger chunk of speech (1500 ms), a low variance in frequency represented greater rhythmic stability. Conversely, when we zoomed in on the syllable transition, a higher variance in frequency reflected a better distinction between syllables and, consequently, enhanced intelligibility. In essence, the contradiction in results could be attributed to the different levels of analysis. At the larger chunk level, rhythmic stability was prioritized, while at the syllable transition level, the ability to differentiate between syllables took precedence, even if it led to higher variance in frequency.

After determining that $varIMF_1$ exhibited the best separating performance among the metrics, we proceeded with further analysis using this metric. It is important to note that speakers may not always utter the same phrases or sentences. Therefore, in addition to individual comparisons, it is crucial to consider the overall distribution of syllable transitions determined by the values of $varIMF_1$. Considering the distribution of syllable transitions provides a broader perspective on how the selected metric captures the differences between speakers. It helps us account for the variability in their utterances and ensures that our analysis is not solely based on specific phrases or sentences. By incorporating the overall distribution of syllable transitions, we obtain a more comprehensive evaluation of the metric's performance and its ability to accurately separate speakers.

Figures 4.5 and 4.6 provide clear evidence that, for both the normal and fast speech rates, the two speakers consistently exhibit statistically significantly higher mean and median values compared to the majority of the other speakers. The median value for male speaker 020 was equal to $8.42 \cdot 10^6$ for both normal and fast

Male Speakers, Normal Rate					
Speaker ID	$varIMF_1$	$varIMF_2$	$varIMF_3$	$varIMF_4$	$varIMF_5$
004	81%	48%	67%	73%	65%
016	91%	67%	69%	66%	61%
022	72%	54%	57%	66%	70%
033	77%	62%	77%	72%	55%
035	70%	52%	70%	78%	70%
047	94%	85%	81%	80%	68%
050	93%	83%	92%	90%	81%
061	58%	48%	62%	64%	54%
076	62%	46%	60%	73%	62%
084	67%	26%	43%	56%	57%

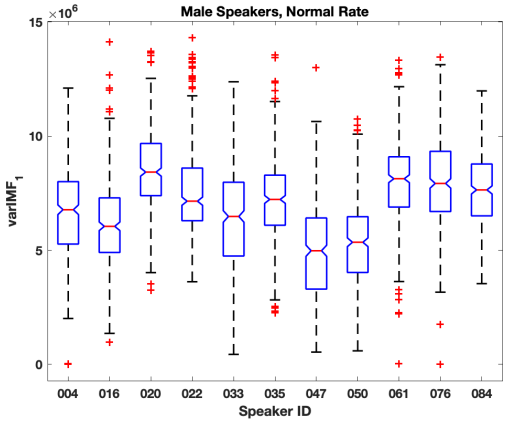
Female Speakers, Normal Rate					
Speaker ID	$varIMF_1$	$varIMF_2$	$varIMF_3$	$varIMF_4$	$varIMF_5$
003	81%	75%	64%	52%	51%
011	37%	36%	43%	42%	54%
014	76%	61%	54%	45%	46%
018	35%	49%	54%	33%	42%
021	60%	42%	34%	28%	36%
026	71%	55%	53%	42%	45%
030	86%	49%	37%	38%	48%
031	88%	71%	61%	56%	60%
040	52%	47%	53%	48%	50%
043	75%	63%	62%	53%	56%
044	85%	51%	57%	60%	67%
046	73%	72%	60%	50%	54%
051	87%	51%	60%	57%	51%
055	86%	84%	78%	66%	63%
067	66%	55%	57%	44%	48%
072	75%	67%	58%	48%	54%
073	82%	63%	57%	53%	57%
075	78%	65%	53%	55%	54%

Table 4.2: Percentage of syllable transitions where the selected speakers exhibited larger metrics values. The red values represent percentages below 50%, indicating that in the majority of transitions, the non-selected speaker displayed higher values compared to the selected speaker.

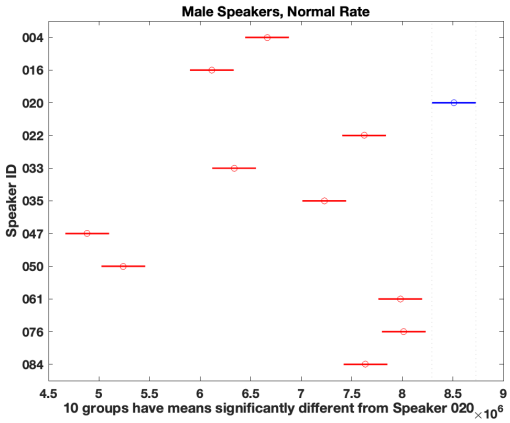
rate, and largest than all other speakers. The second largest median value was $8.13 \cdot 10^6$ for normal rate and $8.22 \cdot 10^6$ for fast rate. The smallest values observed were equal to $4.97 \cdot 10^6$ and $5.85 \cdot 10^6$ for normal and fast rate respectively. The finally selected speaker, 033, had medium values, namely $6.42 \cdot 10^6$ for normal and $6.82 \cdot 10^6$ for fast rate. For the female speakers the median values were slightly larger in general. The selected speaker, 036, had a median value equal to $9.44 \cdot 10^6$ for normal and $9.16 \cdot 10^6$ for fast rate, and exhibited the third largest values. The largest values observed for female speakers were $9.99 \cdot 10^6$ for normal speech and $9.96 \cdot 10^6$ for fast. This finding supports the notion that the selected speakers are distinguishable from the rest not only in individual syllable comparisons but also in terms of the overall distribution of syllable transitions.

The discovery of the effective separation achieved by the metric $varIMF_1$, not only in individual comparisons but also in the overall distribution of syllable transitions, presents an opportunity to apply this metric in datasets where speakers do not necessarily have the same utterances. This flexibility allows for the analysis of diverse datasets that may contain different speech samples or varying sets of utterances for each speaker.

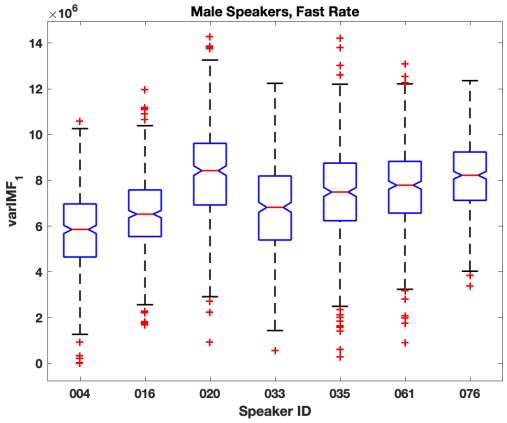
The proposed metric demonstrated improved performance in differentiating speakers based on listener preference, although it was not able to achieve complete accuracy. It is important to acknowledge that preference is a complex phenomenon influenced by various factors, not all of which were fully captured in the initial speech files provided by the speakers. Our metrics may identify certain characteristics that contribute to the appeal of a voice. Further research could explore additional measures and their combination to enhance the prediction of preference. The fact that the first choice for male speakers was not clearly distinguished could be interpreted in this context, since he might possess a different characteristic that makes him preferable.



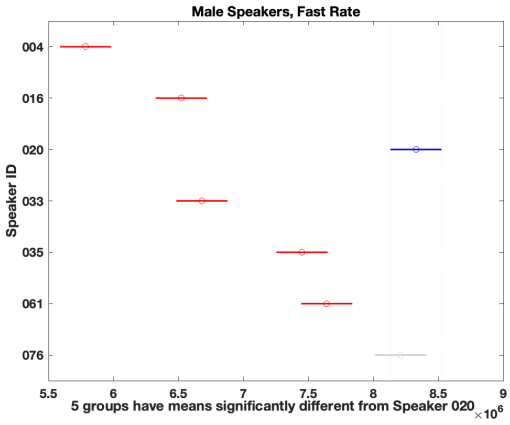
(a) $F(10, 4411) = 154.47, p < 0.0001$



(b) Comparison of means

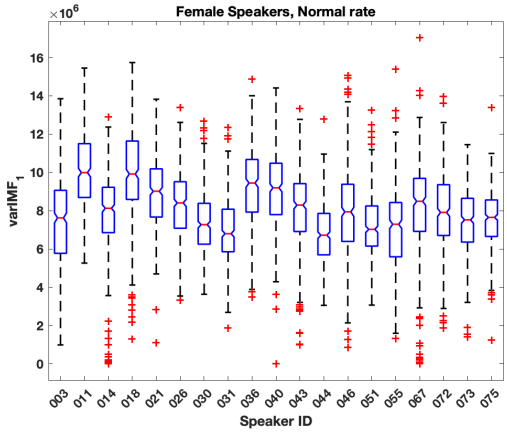


(c) $F(6, 2807) = 98.26, p < 0.0001$

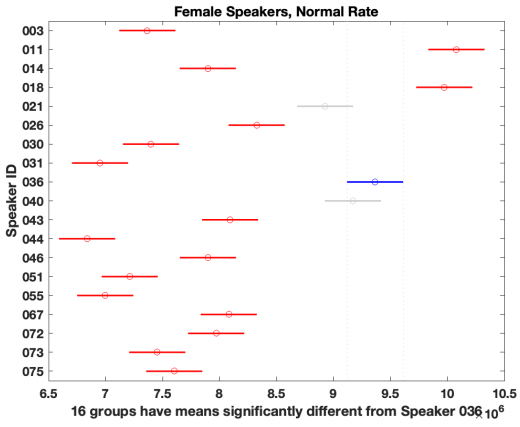


(d) Comparison of means

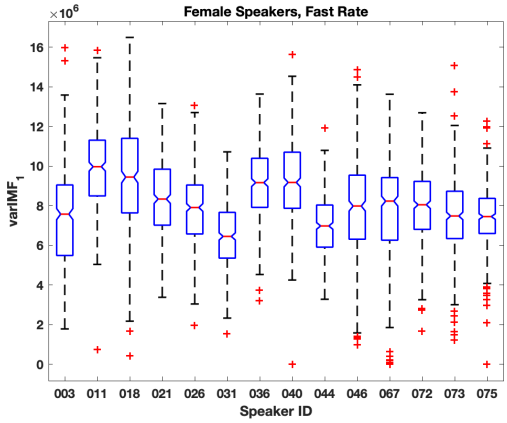
Figure 4.5: Results of ANOVA tests between males speakers in normal and fast rate. Selected speaker is 020.



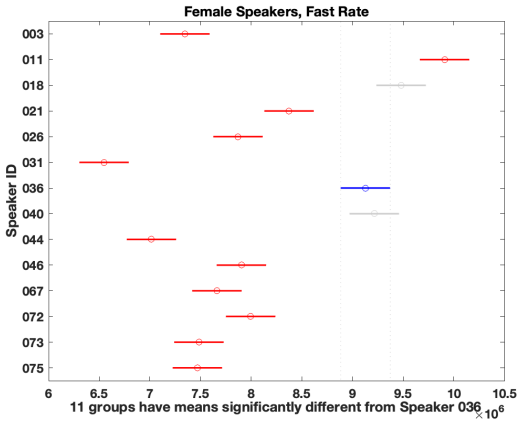
(a) $F(18, 7619) = 99.7, p < 0.0001$



(b) Comparison of means



(c) $F(13, 5614) = 92.79, p < 0.0001$



(d) Comparison of means

Figure 4.6: Results of ANOVA tests between female speakers in normal and fast rate. Selected speaker is 036.

Chapter 5

Discussion and Future Work

In this thesis the main focus was on manipulating speech characteristics in order to explain and predict listener preference. Through our investigation, we have demonstrated that an envelope-based approach is not well-suited for the task at hand. The metrics derived from this approach tend to be either heavily influenced by variations in speech rate, leading to biased results, or they fail to adequately capture the distinctions between different speakers.

Our proposed alternative method showed promising results in separating between the preferred and less preferred speakers. Even though our original goal was to link rhythmic metrics with preference, there is no clear evidence that our measure does indeed reflect rhythmicity. However, it is indeed the case that speech rhythm is not clearly defined. Nevertheless, even if our metric is not clearly connected to rhythmicity, it originated from the same assumptions and methods as the rhythmic metrics in the envelope analysis.

The thesis aimed to achieve a challenging goal of predicting subjective preference through objective measures. It was anticipated that obtaining definitive results in this endeavor would be difficult. Predicting preference automatically is a complex task, as it involves subjective factors that are not easily quantifiable.

This study paves the way for numerous future research possibilities. While our metric has demonstrated promising experimental performance, there remains an opportunity to develop a theoretical framework that can provide insights into the underlying principles driving this performance. By establishing such a framework, we can gain a deeper understanding of the factors influencing speaker differentiation and potentially enhance the effectiveness of our metric.

Additionally, the manual syllabification process used in this study presented limitations when applying the method to larger and more diverse datasets. To overcome this limitation and facilitate the generalization of our approach, it would be valuable to explore automatic methods for extracting syllable boundaries. By incorporating automated syllabification techniques, we can not only enhance the scalability of our method but also evaluate its performance across a wider range of datasets, thereby assessing its generalizability.

An intriguing avenue for further exploration involves applying the same method using the AM-FM decomposition instead of the EMD. The AM-FM algorithm offers a robust mathematical framework that can facilitate the development of theoretical explanations for the observed metric performance. By incorporating the AM-FM decomposition, we can potentially gain deeper insights into the underlying dynamics of speaker differentiation and strengthen the theoretical foundations of our metric.

Lastly, due to the multifactorial nature of preference, relying solely on a single metric may not be sufficient for accurate prediction, if such prediction is indeed possible. Therefore, it is crucial to explore and incorporate additional measures to enhance the representation of preference. By combining multiple metrics, a more comprehensive and robust approach can be developed, offering a more nuanced understanding of subjective preference. This avenue of exploration holds significant potential for improving the prediction of preference and opens up opportunities for future research in this domain.

Bibliography

[1] David Abercrombie. *Studies in phonetics and linguistics*. Oxford University Press, London, 1965.

[2] David Abercrombie. *Elements of general phonetics*. Edinburgh University Press, Edinburgh, 1967.

[3] George D Allen. The location of rhythmic stress beats in english: An experimental study i. *Language and speech*, 15(1):72–100, 1972.

[4] George D. Allen. The location of rhythmic stress beats in english: an experimental study i. *Language and Speech*, 15(1):72–100, January 1972.

[5] George D Allen. Speech rhythm: its relation to performance universals and articulatory timing. *Journal of phonetics*, 3(2):75–86, 1975.

[6] Yousef A. Alotaibi, Ali H. Meftah, Sid-Ahmed Selouani, and Yasser M. Seddiq. Speaker environment classification using rhythm metrics in levantine arabic dialect. In *2014 9th International Symposium on Communication Systems, Networks Digital Sign (CSNDSP)*, pages 706–709, 2014.

[7] Amalia Arvaniti. The usefulness of metrics in the quantification of speech rhythm. *Language and Speech*, 55(2):147–175, 2012.

[8] Peter F Assmann and Terrance M Nearey. Relationship between fundamental and formant frequencies in voice preference. *The Journal of the Acoustical Society of America*, 122(2):EL35–EL43, 2007.

[9] W. Barry, Bistra Andreeva, Michela Russo, Snezhina Dimitrova, and T. Kostadinova. Do rhythm measures tell us anything about language type. 01 2003.

[10] Pier Marco Bertinetto and Chiara Bertini. On modeling the rhythm of natural languages. In *Proceedings of the Fourth International Conference on Speech Prosody*, pages 427–430, 2008.

[11] Danielle Bragg, Cynthia Bennett, Katharina Reinecke, and Richard Ladner. A large inclusive study of human listening rates. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2018.

[12] Ferenc Bunta and David Ingram. The acquisition of speech rhythm by bilingual spanish- and english-speaking 4- and 5-year-old children. *Journal of Speech, Language, and Hearing Research*, 50(4):999–1014, aug 2007.

[13] C. Chatfield. *The Analysis of Time Series*. Chapman & Hall, 1975.

[14] Dasom Choi, Daehyun Kwak, Minji Cho, and Sangsu Lee. " nobody speaks that fast!" an empirical study of speech rate in conversational agents for people with vision impairments. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

[15] Luis Coelho, Daniela Braga, Miguel Sales-Dias, and Carmen Garcia-Mateo. An automatic voice pleasantness classification system based on prosodic and acoustic patterns of voice preference. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

[16] Fred Cummins and Robert Port. Rhythmic constraints on stress timing in english. *Journal of Phonetics*, 26(2):145–171, April 1998.

[17] Eric Deléchelle, Jacques Lemoine, and Oumar Niang. Empirical mode decomposition: an analytical approach for sifting process. *IEEE Signal Processing Letters*, 12(11):764–767, 2005.

[18] Volker Dellwo. Rhythm and speech rate: a variation coefficient for delta c. *Journal of the Acoustical Society of America*, 119(1):597–608, 2006.

[19] Volker Dellwo and Petra Wagner. Relationships between rhythm and speech rate. 01 2003.

[20] Francisco Gutiérrez Díez, Volker Dellwo, Núria Gavaldà, and Stuart Rosen. The development of measurable speech rhythm during second language acquisition. *The Journal of the Acoustical Society of America*, 123(5):3886–3886, May 2008.

[21] Hongwei Ding, Oliver Jokisch, and Rüdiger Hoffmann. The effect of glottalization on voice preference. In *Proceedings of speech prosody*, pages 851–854, 2006.

[22] Catharine H. Echols and Megan J. Crowhurst. The perception of rhythmic units in speech by infants and adults. *Infant Behavior and Development*, 22(4):675–694, 1999.

[23] R Rashidi Far and Saeed Gazor. Am-fm decomposition of speech signal using mwl criterion. In *Canadian Conference on Electrical and Computer Engineering 2004 (IEEE Cat. No. 04CH37513)*, volume 3, pages 1769–1772. IEEE, 2004.

[24] Anne Fernald and Patricia Kuhl. Acoustic determinants of infant preference for motherese speech. *Infant Behavior and Development*, 10(3):279–293, 1987.

[25] Patrick Flandrin, Gabriel Rilling, and Paulo Goncalves. Empirical mode decomposition as a filter bank. *IEEE signal processing letters*, 11(2):112–114, 2004.

[26] Elena Flaugnacco, Luisa Lopez, Chiara Terribili, Stefania Zoia, Sonia Buda, Sara Tilli, Lorenzo Monasta, Marcella Montico, Alessandra Sila, Luca Ronfani, and Daniele SchÃ¶n. Rhythm perception and production predict reading abilities in developmental dyslexia. *Frontiers in Human Neuroscience*, 8, June 2014.

[27] Carol A Fowler. b••perceptual centersb•• in speech production and perception. *Perception & Psychophysics*, 25(5):375–388, 1979.

[28] Soumaya Gharsellaoui, Sid Ahmed Selouani, Wladyslaw Cichocki, Yousef Alotaibi, and Adel Omar Dahmane. Application of the pairwise variability index of speech rhythm with particle swarm optimization to the classification of native and non-native accents. *Computer Speech &amp Language*, 48:67–79, March 2018.

[29] Asif A. Ghazanfar, Ryan J. Morrill, Annika Paukner, and Pier F. Ferrari. Monkey lip-smacking develops like the human speech rhythm. *Developmental Science*, 15(4):557–568, 2012.

[30] Francesco Gianfelici, Giorgio Biagetti, Paolo Crippa, and Claudio Turchetti. Am-fm decomposition of speech signals: an asymptotically exact approach based on the iterated hilbert transform. In *IEEE/SP 13th Workshop on Statistical Signal Processing, 2005*, pages 333–338. IEEE, 2005.

[31] Dafydd Gibbon and Ulrike Gut. Measuring speech rhythm. *Phonetica*, 56(1-2):86–115, 1999.

[32] Anne-Lise Giraud and David Poeppel. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience*, 15(4):511–517, 2012.

[33] Usha Goswami. Speech rhythm and language acquisition: an amplitude modulation phase hierarchy perspective. *Annals of the New York Academy of Sciences*, 1453(1):67–78, June 2019.

[34] Usha Goswami. Language acquisition and speech rhythm patterns: an auditory neuroscience perspective. *Royal Society Open Science*, 9(7), July 2022.

[35] Usha Goswami and Victoria Leong. Speech rhythm and temporal structure: Converging perspectives? *Laboratory Phonology*, 4(1), January 2013.

[36] Usha Goswami, Jennifer Thomson, Ulla Richardson, Rhona Stainthorp, Diana Hughes, Stuart Rosen, and Sophie K Scott. Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences*, 99(16):10911–10916, 2002.

[37] Frédéric Gougoux, Franco Lepore, Maryse Lassonde, Patrice Voss, Robert J. Zatorre, and Pascal Belin. Pitch discrimination in the early blind. *Nature*, 430(6997):309–309, July 2004.

[38] Esther Grabe and Ee Ling Low. Durational variability in speech and the rhythm class hypothesis. *Papers in laboratory phonology*, 7(1982):515–546, 2002.

[39] Ulrike Gut. Rhythm in l 2 speech. 2012.

[40] Charles Andrew Harsín and Kerry P Green. Perceptual centers as an index of speech rhythm. *The Journal of the Acoustical Society of America*, 96(5):3350–3350, 1994.

[41] Charles E Hoequist Jr. The perceptual center and rhythm categories. *Language and Speech*, 26(4):367–376, 1983.

[42] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, March 1998.

[43] George P Kafentzis, Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. An extension of the adaptive quasi-harmonic model. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4605–4608. IEEE, 2012.

[44] Naoko Kinoshita and Chris Sheppard. Validating acoustic measures of speech rhythm for second language acquisition. In *ICPhS*, volume 17, pages 1086–1089, 2011.

[45] Peter Ladefoged. *A course in phonetics*. Harcourt Brace Jovanovich, New York, NY, 1975.

[46] Jean Laroche. A new analysis/synthesis system of musical signals using prony's method-application to heavily damped percussive sounds. In *International Conference on Acoustics, Speech, and Signal Processing,*, pages 2053–2056. IEEE, 1989.

[47] Victoria Leong, Michael A. Stone, Richard E. Turner, and Usha Goswami. A role for amplitude modulation phase relationships in speech rhythm perception. *The Journal of the Acoustical Society of America*, 136(1):366–381, July 2014.

[48] Qin Li and Les E Atlas. Over-modulated am-fm decomposition. In *Advanced Signal Processing Algorithms, Architectures, and Implementations XIV*, volume 5559, pages 172–183. SPIE, 2004.

[49] Low Ee Ling, Esther Grabe, and Francis Nolan. Q uantitative characterizations of speech rhythm: Syllable-timing in singapore english. *Language and Speech*, 43(4):377–401, December 2000.

[50] Sha Liu and Kaye Takeda. Mora-timed, stress-timed, and syllable-timed rhythm classes: Clues in english speech production by bilingual speakers. *Acta Linguistica Academica*, September 2021.

[51] Anastassia Loukina, Greg Kochanski, Burton Rosner, Elinor Keane, and Chilin Shih. Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, 129(5):3258–3270, 2011.

[52] Stephen Michael Marcus. Acoustic determinants of perceptual center (p-center) location. *Perception &amp Psychophysics*, 30(3):247–256, May 1981.

[53] William Marslen-Wilson and Lorraine Komisarjevsky Tyler. The temporal structure of spoken language understanding. *Cognition*, 8(1):1–71, 1980.

[54] Neil P. McAngus Todd and Guy J. Brown. Visualization of rhythm, time and metre. *Artificial Intelligence Review*, 10(3):253–273, Aug 1996.

[55] Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal forced aligner: Trainable text-speech alignment using kaldi. In *Interspeech*, 2017.

[56] Ali H. Meftah, Mustafa Qamhan, Yousef Alotaibi, and Sid-Ahmed Selouani. Emotional speech recognition using rhythm metrics and a new arabic corpus. In *2020 16th IEEE International Colloquium on Signal Processing Its Applications (CSPA)*, pages 57–62, 2020.

[57] Pat Mirenda, Douglas Eicher, and David R. Beukelman. Synthetic and natural speech preferences of male and female listeners in four age groups. *Journal of Speech, Language, and Hearing Research*, 32(1):175–183, March 1989.

[58] Ryan J. Morrill, Annika Paukner, Pier F. Ferrari, and Asif A. Ghazanfar. Monkey lip-smacking develops like the human speech rhythm. *Biology Letters*, 16(4):20200232, 2020.

[59] John Morton, Steve Marcus, and Clive Frankish. Perceptual centers (p-centers). *Psychological review*, 83(5):405, 1976.

[60] Murray J. Munro and Tracey M. Derwing. The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning*, 48(2):159–182, June 1998.

[61] Thomas Murry, WS Brown Jr, and Howard Rothman. Judgments of voice quality and preference: Acoustic interpretations. *Journal of Voice*, 1(3):252–257, 1987.

[62] Thierry Nazzi and Franck Ramus. Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, 41(1):233–243, 2003.

[63] Mikhail Ordin and Leona Polyanskaya. Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *The Journal of the Acoustical Society of America*, 138(2):533–544, August 2015.

[64] Tobias Overath, Josh H McDermott, Jean Mary Zarate, and David Poeppel. The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature neuroscience*, 18(6):903–911, 2015.

[65] Michael Ob••Dell and Tommi Nieminen. Coupled oscillator model of speech rhythm. In *Proceedings of the XIVth international congress of phonetic sciences*, volume 2, pages 1075–1078. University of California Berkeley, 1999.

[66] Yannis Pantazis. *Decomposition of AM-FM signals with applications in speech processing*. PhD thesis, PhD thesis, University of Crete, Department of Computer Science, 2010.

[67] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. Am-fm estimation for speech based on a time-varying sinusoidal model. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[68] Yannis Pantazis, Olivier Rosec, and Yannis Stylianou. Adaptive am–fm signal decomposition with application to speech analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(2):290–300, 2010.

[69] Yannis Pantazis, Yannis Stylianou, and Maria Koutsogiannaki. A novel method for the extraction of vocal tremor. *A Novel Method for the Extraction of Vocal Tremor*, pages 1000–1004, 2009.

[70] Elinor Payne, Brechtje Post, LluG•sa Astruc, Pilar Prieto, and Maria del Mar Vanrell. Measuring child rhythm. *Language and Speech*, 55(2):203–229, sep 2011.

[71] Jonathan E Peelle and Matthew H Davis. Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology*, 3:320, 2012.

[72] Kenneth Lee Pike. *The intonation of American English*. University of Michigan Press, Ann Arbor, 2nd edition, 1946.

[73] David Poeppel. Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, 21(6):322–334, 2020.

[74] Leona Polyanskaya and Mikhail Ordin. Acquisition of speech rhythm in first language. *The Journal of the Acoustical Society of America*, 138(3):EL199–EL204, September 2015.

[75] Bernd Pompino-Marschall. On the psychoacoustic nature of the p-center phenomenon. *Journal of phonetics*, 17(3):175–192, 1989.

[76] Robert F Port. Meter and speech. *Journal of phonetics*, 31(3-4):599–611, 2003.

[77] Michael Proctor and Athanasios Katsamanis. Prosodic characterization of reading styles using audio book corpora. *The Journal of the Acoustical Society of America*, 130:2553, 10 2011.

[78] Franck Ramus, Marina Nespor, and Jacques Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292, 1999.

[79] Naveed Rehman and Danilo P Mandic. Multivariate empirical mode decomposition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 466(2117):1291–1302, 2010.

[80] Gabriel Rilling, Patrick Flandrin, Paulo Goncalves, et al. On empirical mode decomposition and its algorithms. In *IEEE-EURASIP workshop on nonlinear signal and image processing*, volume 3, pages 8–11. Grado: IEEE, 2003.

[81] Donia R Scott and S D Isard. Perceptual isochrons in english and in french. *Phonetica*, 52(3):131–147, 1995.

[82] Sophie K Scott. P-centers in speech. *Unpublished PhD dissertation). University College London, UK*, 1993.

[83] S Chandra Sekhar and Thippur V Sreenivas. Novel approach to am-fm decomposition with applications to speech and music analysis. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages ii–753. IEEE, 2004.

[84] Rajib Sharma, Leandro Vignolo, Gastón Schlotthauer, M.A. Colominas, H. Leonardo Rufiner, and S.R.M. Prasanna. Empirical mode decomposition for adaptive AM-FM analysis of speech: A review. *Speech Communication*, 88:39–64, April 2017.

[85] Olympia Simantiraki and Martin Cooke. Exploring listeners' speech rate preferences. In *INTERSPEECH*, pages 1346–1350, 2020.

[86] Kenneth N Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000.

[87] DG•vid SztahG•, MiklG•s Tulics, Klara Vicsi, and IstvG•n ValG•lik. Automatic estimation of severity of parkinson's disease based on speech rhythm related features. 09 2017.

[88] Adam Tierney, Jessica Cardona Gomez, Oliver Fedele, and Natasha Z. Kirkham. Reading ability in children relates to rhythm perception across

modalities. *Journal of Experimental Child Psychology*, 210:105196, October 2021.

[89] Sam Tilsen and Amalia Arvaniti. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, 134(1):628–639, July 2013.

[90] Sam Tilsen and Keith Johnson. Low-frequency fourier analysis of speech rhythm. *The Journal of the Acoustical Society of America*, 124(2):EL34–EL39, 2008.

[91] Ruth Tincoff, Marc Hauser, Fritz Tsao, Geertrui Spaepen, Franck Ramus, and Jacques Mehler. The role of speech rhythm in language discrimination: further tests with a non-human primate. *Developmental Science*, 8(1):26–35, 2005.

[92] Ingo R Titze and Daniel W Martin. Principles of voice production, 1998.

[93] Neil P McAngus Todd and Guy J Brown. A computational model of prosody perception. In *ICSLP*, volume 94, pages 127–130, 1994.

[94] John W. Tukey. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99, June 1949.

[95] Alice Turk and Stefanie Shattuck-Hufnagel. What is speech rhythm? a commentary on arvaniti and rodriquez, krivokapiD•, and goswami and leong. *Laboratory Phonology*, 4(1):93–118, 2013.

[96] Janet F. Werker and Richard C. Tees. Development of speech perception: Maturation, plasticity, and constraints. *Monographs of the Society for Research in Child Development*, 60(4):1–183, 1995.

[97] Nicole Whitworth. Speech rhythm production in three german–english bilingual families. *Leeds working papers in linguistics and phonetics*, 9:175–205, 2002.

[98] Clare Wood and Colin Terrell. Poor readers' ability to detect speech rhythm and perceive rapid speech. *British Journal of Developmental Psychology*, 16(3):397–413, September 1998.

[99] Brigitte Zellner. Pauses and the temporal structure of speech. In *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.*, pages 41–62. John Wiley, 1994.