The Rhythm of Language and Speech: Constraining Factors, Models, Metrics and Applications

Habilitationsschrift
an der
Philosophischen Fakultät
der
Rheinischen Friedrich-Wilhelms-Universität Bonn

vorgelegt von
Dr. phil. Petra Susanne Wagner

Bonn, 2008

Where utter patternlessness or randomness prevails, nothing is predictable.
DANIEL DENNETT, 1991, p. 30

All poetry is a reproduction of the tones of actual speech.
ROBERT FROST

Preface

This is a book on speech and language rhythm — of course, this should not imply that rhythmical phenomena are restricted to spoken language. In the past, very interesting studies have been carried out on multimodal aspects, e.g. in the coordination of spoken language and other gestures, e.g. deictic hand movements, emphatic hand movements, eyebrow movements to strengthen prosodic focus, foot and finger tappings etc. All of these fields are closely related to speech rhythm and likewise present interesting research topics, but since the phenomenon of rhythm is difficult to trace, measure, even to define, I believe a restriction to the domain of language and speech is necessary. However, we will see that even this limitation does not spare us the look to other, related fields of study and disciplines.

Why is rhythm difficult to describe and to model? The most problematic point seems to be that most people already have an intutive notion of what rhythm is, while being unable to describe its essence. This has lead to many methodological unclarities and vague definitions causing difficulties in the interpretation and comparison of research results.

I do not expect this book to be much wiser - but more to the point. I hope to have accumulated the essence of what we know today about different aspects of speech and language rhythm from different perspectives, linguistics, phonetics, psychology, psychoacoustics and speech signal processing. During the research for this book, I soon realized that the interdisciplinarity of the topic was enormous, even larger than in many other fields I had tackled so far. I then regarded it as my fate, coming from such an inherently interdisciplinary field as phonetics, to gather the different points of view and tried to organize them in a fashion as systematic as possible. I am sure that I failed in many respects and that a different approach to the topic could have been more successful or logical than the one I happened to choose. However, this was the pattern that slowly emerged in my head and I hope it is comprehensive to most readers. Mainly for the point of readability and understandability, it was sometimes necessary to leave out certain aspects of detail.

Intended audiences

I have been trying to write a book for a wide audience. Of course, I had in mind researchers and students from various disciplines who happen to have an interest in rhythm. I hope this book will be (at least partly) useful for people with a background in phonetics, applied linguistics, cognitive science, psycholinguistics, computational linguistics, speech technology or clinical linguistics. If one intends to write for a wide audience, it is necessary to chose a writing style much to the point and the avoidance of too much technical "mumbojumbo". I really hope I achieved this only half as much as I wanted to. However, I do expect a solid background in linguistics and phonetics. This means, that I did not introduce every technical term such as 'prosody', 'phonology' , 'phonetics', 'fundamental frequency' etc.

Overview

The next chapter 1.1 will accustom you first of all with the question, why research on speech and language rhythm is a relevant field at all. Maybe, it is slightly unusual to spend a whole chapter on the motivation. If the motivation for writing a book is not self-evident, this might cause criticism in first place. I did so because many linguists would argue that rhythm is not phonologically distinctive, thus it is only of marginal interest. I will try to show that rhythm plays a key role in language processing and is therefore an important research topic in all fields of cognitive science, including linguistics.

In chapter 2, much of rhythm related psychological and linguistic literature is reviewed trying to illuminate the fundamental cognitive processes involved in rhythm perception and production. This chapter closes with an overview of the various rhythm related influences that need to be taken into account when working in the field. In the following chapter 3 several models and metrics of speech rhythm are critically reviewed. I chose this order because in my view, is necessary to comprehend the basic cognitive procedures involved in rhythm perception and production in order to evaluate and assess the rationale behind the different models. In chapter 4 I make a suggestion for a new approach to visualize and quantify rhythmic patterns by integrating different prosodic levels of rhythmic structure. I believe that this approach helps to understand rhythmic phenomena in speech and language which have so far not been treated convincingly.

Chapter 5 will be mostly of interest to people who approach scientific topics with the pressing background question *What is all this good for?* I will illustrate various fields where insights of rhythm research can be of use, e.g. the classification of rhythm class within a typological approach, the characterization of L2 speech and the description of various speaking styles such as fast vs. slow speech and poetry vs. prose. Various applications of the model in speech technology will be discussed as well.

Contents

Chapter 1

Motivation and Prolegomena

1.1 Motivation

Why should linguistics or phonetics deal with rhythm at all? Unlike many other prosodic phenomena, rhythm cannot be straightforwardly connected to linguistic functions: I.e. it is difficult to find evidence that the rhythmic shape of an utterance changes its meaning. This is easier in other fields of prosodic research: The location of a a sentence accent may indicate prosodic focus and change an utterance's meaning (e.g. Altmann et al. (1989); Rooth (1996)). The shape of intonation contours may signal the pragmatic status of a piece of information as contextually given or new (Brown (1983); Baumann (2006)) and may provide information about the sentence mood (e.g. Altmann et al. (1989)). The linguistic function of a particular rhythm is — if there is any — less clear. In this chapter, we will describe the functions of rhythm for linguistic organization and the cognitive processing of speech and language, thus clarifying its status as an interesting research object within the field of cognitive linguistics and phonetics.

1.1.1 Rhythmic structure is an independent property of languages

It can be argued that rhythm is a byproduct of spoken language rather than an independent property of it. Indeed, the acoustic properties constituting the rhythmical pattern of an utterance are to a large extent determined by linguistic structure. In languages so different as Japanese (Kaiki et al. (1992)), Polish (e.g. Breuer et al.

(2006)), English (e.g. Klatt (1979); van Santen (1994)) or German (e.g. Möbius and van Santen (1996); Portele and Heuft (1997)) and French (e.g. Keller and Zellner (1995)) duration patterns can be predicted quite precisely on the basis of linguistic structure such as segment type, syllable structure, lexical stress, sentence position etc. Since most people share the intuition that rhythmic structure is predominantly expressed as timing patterns, the existence of rhythm being something independent of linguistic structure, is questionable. Furthermore, it has been shown that there is a strong correlation between syllable durations and the number and type of segments contained in them (Campbell (1992)): It would therefore be nor surprise at all if languages with complex syllable structures have an inherently different rhythm than those with more simple ones. This expectation goes hand in hand with the findings of Dauer (1983), who identified strong correlations between phonological rules, phonotactic constraints and rhythmic properties across many languages. Of course, rhythm is not only defined by acoustic duration or the number of phones contained in an utterance. It can also be shaped by placing accents on particular places in the utterance, thus forming a rhythmical pattern of accented and unaccented syllables. However, in this placement of accents, speakers are not completely free either. There exists an extensive body of research showing that syntax, semantics, pragmatics and lexical status and lexical stress location strongly influence placement, shape and strength of accents in an utterance (among many others, cf. Selkirk (1984, 1995); Wagner (1999); Baumann (2006); Schmitz and Wagner (2006)). Taken together, these insights lead some researchers to believe that the perceptual impressions of language specific rhythmical preferences are solely based on properties of the languages involved. Indeed, the phonological structure puts heavy constraints on the rhythmical pattern of a language. Thus, the rhythm might be regarded as being inherent in the linguistic structure rather than being the language independent product of rhythmical preferences within a linguistic community. Thus, unlike a rhythm in music, where a musician or a composer may impose just about any rhythm onto a sequence sequence of musical notes, the phonological structure forces much of the properties making up the rhythmical shape of an utterance onto a linguistic community. If such a view were true without exception, we would be "slaves to our native rhythm" rather than shaping it independently. The study of speech and language rhythm would become obsolete as an independent topic of research. Instead, we could concentrate on the production and perception constraints for stresses, accents and segmental durations.

So why has this book been written? I believe, that despite the heavily weighing linguistic constraints, a speaker has still some freedom in shaping the rhythmic structure of an utterance. E.g., a speaker may chose a stakkato rhythm, creating the impression of slight anger and impatience. Also, speakers show a significant amount of variation concerning accent placement, probably due to rhythmic rather than semantic-pragmatic constraints (Henke (1993)), since such rhythms are often speaking style related. Thus, speakers have some degrees of freedom when shaping the rhythmic structure of an utterance, independent of the text. According to Hirschfeld and Stock (2004), this rhythmic potential is used more systematically by trained speakers. Thus, we conclude with the point of view expressed by Jassem et al. (1984):

> *rhythm correlates with linguistic structure, but is not fully determined by it*

. Therefore, the rhythmic structure of language and speech qualifies as an independent topic of research.

1.1.2 Rhythmic patterning simplifies cognitive and linguistic processes

Very early in psychological research, the strong link between rhythmical structure and memory has been examined (Ebbinghaus (1964)). Gestalt psychologist even regarded rhythm as subdomain of research on human memory (Koffka (1909, 1935) quoted after Spitznagel (2000)). The insight that rhythmic strucuturing helps memorization has certainly been known for thousands of years and has become an integrative part of folk psychology. It is the reason for the existence of mnemonic rhymes, e.g. the one developed for German learners of Latin which chunks the different Latin prepositions that are followed by the ablative case into two groups carrying the same number of accents (marked in boldface). The third line contains a sentence explaining the grammatical rule. It follows the rhythmic pattern of the second line:

a **ab** **ex** **e**
de cum **si**ne **pro** und **prae**
stehen **mit** dem **A**bla **tiv**!

Recent findings in experimental phonetics show that rhythm indeed helps memorization, especially when it is expressed in duration patterns Gilbert and Boucher (2006, 2007), but it has long been unclear why such rhythmical grouping helps our memory. Psychological research explains this memory aiding effect by a chunking technique that can extend the capacities of short term memory to a certain extent (Neisser (1974)). This process can be explained using a metaphor we know from modern computer technology. If we want to send a number of data files per mail we may run into capacity problems because the files are too large. In these cases, we archive the different files and "zip" them, thus forming neat packages that need less memory space. The crucial thing of enhancing memory capacities is to know the order of things, i.e. "what comes next". If I use rhythmical structure to "zip" what I heard, I can also use my rhythmical knowledge in order to "unzip" things I stored. E.g. when memorizing Latin prepositions followed by the ablative case, rhythm tells us, that there are four prepositions per line. Thus, even if I cannot remember the full set of prepositions, at least I know how many prepositions I *should* retrieve. The mnemonic rhyme is furthermore structured into pairs of two phonologically similar prepositions, i.e. they start with the same phone (*a, ab; ex, e; pro, prae*), thus guiding us to retrieve another, similarly sounding preposition stored in the mental lexicon. The remaining rhythmical pair is not constituted by two but three prepositions, namely **de** *cum,* **si***ne*. Since these could not be paired based on their beginning sound, a trochaic pattern is used in order to form the structural similarity. Thus, rhythm provides us with an archiving structure minimizing cognitive memory load. Since it is clear that it helps to convey a linguistic message if listeners remember what been said, a clear rhythmic structure automatically aids human communication.

Listeners obviously feel the need to superimpose a rhythmic structure on sequences of auditory events. In early rhythm research, Dietze (1885); Bolton (1894) and Wundt (1911) described the following phenomenon: If we listen to a sequence of acoustically identical, isochronous acoustic stimuli, we perceive them as groups, each group starting with a stronger stimulus than the subsequent ones. E.g. a sequence of six identical stimuli can be perceived as three groups of two or two groups of three events each (cf. Section 2.3). Obviously, we feel the need to structure what we hear, thus establishing *rhythmic groups*. Rhythm provides us with a method to perform this process of structuring. Allen (1975) provides a good answer, why this structuring is useful:

> "Speech rhythm functions mainly to organize the information bearing elements of the utterance into a coherent package, thus permitting speech communication to proceed efficiently. Rhythm therefore does not carry much linguistic information, other than helping to signal the language of the speaker; without rhythmic organization, however, the linguistic message would be difficult to transfer."

According to Dennet (1991), a key function of all cognitive patterns is to constrain the number of possible future events. This is certainly true for rhythmic patterns as well. If I know that in my native language it is likely that a stressed syllable is followed by one or two unstressed ones and that the stressed syllables tend to contain the most important semantic content, e.g. a word stem or a focussed word, I can make excellent hypotheses concerning the next point of time it may be useful to pay attention. I can then use my cognitive capacities more economically.

In fact, there exists wide agreement concerning the function of rhythm as a means of simplifying cognitive processing, e.g. it helps segmental perception (Martin (1979)) and lexical perception, because by evaluating the rhythmical status of a perceived syllable it constrains the production alternatives of forthcoming events (van Donselaar et al. (2005)). Furthermore, rhythmic patterns improve the fluency and speed of articulatory gestures (Kalveram (2000)). Based on our rhythmical expectancies, we focus our attention at particular points in time, when something important is likely to be said (Quené and Port (2005)). It comes as no surprise that in general, rhythm speeds up cognitive processing (Buxton (1983)).

Besides providing cues concerning the prominent and important stretches of time in course of an utterance, rhythmical patterns provide us with information concerning important linguistic boundaries. In languages with a fixed word accent such as Polish (penultimate syllable) or Finnish (first syllable), listeners can use rhythm for word segmentation. It is also common knowledge that speakers use rhythm in order to structure their speech. According to German standard pronounciation and in line with rhetorical guides, groups of words belonging together semantically, so-called *sense groups* may not be interrupted by a pause (e.g. von Essen (1956); Wachtel (2000)). Such a point of view usually rules out pauses within syntactic phrases as well, which is in accordance with psycholinguistic results (Goldmann-Eisler (1968)). Thus, rhythmical grouping helps both to parse and interpret an utterance. Another classical example of rhythmical structuring are "spoken parentheses", where speakers tend to switch to a fast, monotonous rhythm (cf. Example 1) indicating the fact that the content of the parenthesis is not a part of the surrounding utterance.

Example 1
We will meet tomorrow — as I already explained earlier — in front of the hotel.

Speech rate phenomena even seem to play a role in signalling the semantic-pragmatic structure of an utterance: For Polish, Demenko (2003) found a tendency of deceleration before and a tendency of acceleration after a word in focus. Thus, the global timing across an utterance provides the listener with clues concerning the locus of the most important information.

When regarding the consequences of an impairment of rhythm related processes, the major impact of rhythm on the communicative chain becomes even more evident. Communicative processes can be seriously disrupted if rhythmic organization is disturbed. Many researchers regard stuttering to be a consequence of timing problems in speech production and perception (e.g. Riper (1986)). Similarly, Kalveram (2000) regards stuttering as the result of a problem in rhythmical processing: According to him the subdivision of the speech signal into consecutive syllables (= a process of rhythmical analysis) accelerates and automatizes speech production without the need of auditory feedback. He believes this mechanism to be impaired in people who suffer from stuttering. Another fact that indicates the link between rhythmical processing is the circumstance that in performance styles where the timing is controlled indepedently, like singing, stutterers tend to have fewer problems, independent of speech rate (Glover et al. (1996)). Methods of timing control, like an artificial reduction of articulation rate, accompanying tapping or the use of a metronome are often used in stuttering therapy (Ptok (2006)). Of course, stuttering has multiple causes and it may be oversimplistic to reduce it to a problem of rhythmical structure (Kaufmann (2006); Ptok (2006)). Therefore, very different therapeutical approaches may be similarly successful in its treatment (e.g. Kotby et al. (2003)). Even if the reasons for stuttering cannot be reduced to a problem of rhythm processing, its therapeutic effect does not come as a surprise taking into account its enhancement of linguistic planning, perception and performance. If the rhythmical abilities are impaired, linguistic performance can be seriously impaired, too. Stuttering is certainly the most well-known fluency disorder related to speech rhythm. Besides, rhythmical disablities may lead to problems in the language acquisition process. Lea (1980) showed the impact rhythmical disabilities had on the processing of syntactic structures. Weinert (2000) further specifies this connection. She finds that deficits in making use of the rhythm-related simplification strategies for language and speech processing correlate with developmental disphasia. Children suffering from it obviously have more difficulties using rhythmical cues such as prosodic grouping to derive syntactic phrases and memorize what has been said. Weinert even found a covariation between the strength of rhythmical impairment and the associated disphasia. The ability to make use of rhythmical structures in speech even influences linguistic skills that are only indirectly related to articulation, such as reading. Goswami et al. (2002) found a correlation between developmental dyslexia and the ability to detect rhythmical beats, so-called *p-centers* in speech. They argue that the detection of p-centers is a cognitively low-level process and a prerequisite to perform syllable segmentation and structuring thus enabling us to build up a speech chain both in listening and (obviously) reading tasks.

Summing up, rhythmic structures provide us with helpful anchors to produce parse, interprete and memorize the content of an utterance. This becomes even more evident when looking at the performance problems of speakers or listeners whose rhythmical abilities are impaired.

1.2 Prolegomena to the study of language and speech rhythm

1.2.1 Rhythmic Impressions

It is very likely that any reader of this book has some intuition what the rhyhm of language and speech is, what it sounds like, and how it varies between individual speakers or languages. It is also very likely that occasionally, you listened to a foreign language of which you would say, that it differed rhythmically from your own. Maybe, you also had the impression that in particular situations, speakers tend to use a certain rhythm that apparently is judged as appropriate and differs from other situations. Just imagine a situation where you listen to a religious sermon in the context of a wedding ceremony and later, you meet the preacher talking to the newly wed couple. Probably, the rhythmical style in both situations used by the preacher would be different. Also, if you imagine reading a poem as a poem or else, as a text passage. It is very likely, that the realisation of both versions differs rhythmically (Kruckenberg and Fant (1993), (Bröggelwirth, 2007, 29)) within certain degrees of freedom. One assumes that in the most metrical and mechanical manner of reading poetry, an abstract rhythm is emphasized which may deviate considerably from normal prose ((Kiparsky, 1975, 585), (Nespor and Vogel, 1986, 278)). This style of poetry reading we all know from school where memorized poems were recited in an automatic fashion, disregarding the meaning of the poem. Also, prose reading — or spontaneous speech — may disregard any rhythmic constraints, but it is by no means likely that it always will. In Example 2, the entry lines of a poem is once presented in its original form and also layouted as ordinary text. The poem is written in *blank verse*, an unrhymed iambic pentameter, thus, roughly every second syllable ought to carry a stress and each line ought to contain five stresses. In order to produce the blank verse, it is necessary to stress function words like the preposition “with” in the first line or the conjunction “and” in the second line. Most native speakers would probably not always stick strictly to the iambic pentameter when reciting the poem. But it can be expected that when reading the poem without the line breaks and with “normal” punctuation marks, the abstract *meter* would be obeyed even less. It can furthermore be expected, that readers would chunk the text into prosodic phrases different from the way indicated by Wordsworth’s line breaks and punctuations.

Example 2

William Wordsworth (1770 — 1850)
Original Text Layout:
FIVE years have past; five summers, with the length
Of five long winters! and again I hear

These waters, rolling from their mountain-springs
Layout without Poetic Line Breaks and Punctuation Marks:
FIVE years have past; five summers with the length of five long winters!
And again I hear these waters rolling from their mountain-springs...

Until now, the measurements and models describing this phenomenon are far from clear. There is still considerable dispute concerning the phonetic nature or realisation of rhythm and how language specific rhythms can be explained empirically. Early phonetic suggestions make a distinction between “syllable timed” or “machine gun rhythm” languages such as French and “stress timed” or “morse code rhythm” languages such as English (James (1940); Pike (1945)). This impressionist approach to a rhythmic typology of languages built upon the belief that the level of linguistic rhythmic organisation specified a level of acoustic phonetic organisation as well: One had the impression that in so-called syllable-timed languages, the syllables had the tendency to be isochronous, or at least near-isochronous — as you would expect it from the shots by a machine gun. In stress-timed languages, one expected feet or interstress intervals to be near isochronous. A foot, however, may consist of a long, stressed syllable and one or several short, unstressed syllables. This pattern of longer and shorter rhythmic events has more in common with a morse code than with a machine gun. In a third linguistic rhythm type, the so-called mora timed languages, is was believed that morae were nearly isochronous. A *mora* is a measure of syllable weight[1]. In Japanese, a syllable can consist of one mora, thus being short, or two morae, thus being long — in a strict version of the *isochrony hypothesis*, this would mean that syllables containing two morae are twice as long as syllables containing only one — or at least near so. One version of the isochrony hypothesis proposes that each language can be assigned one specific rhythm type. This radical assumption became influential through (Abercrombie, 1967, 97). Although nowadays, such a categorical view of a rhythm typology is heavily criticised, his hypothesis was not far fetched: In poetry, it is obvious that languages classified as syllable timed, prefer different metres compared to languages classified as stress timed or mora timed (Lehiste (1990); Cutler (1994)). The *endecasillabo* builds each verse out

[1] In some languages, called quantity sensitive, syllable weight or the number of morae contained in a syllable, determines lexical stress. E.g., it has been a matter of dispute, whether German is quantity sensitive or not. In **?**, it is argued that it is.

of eleven syllables and has become extremely popular in syllable-timed languages such as Italian and Spanish. In German poetry, there was a shift in poetic metrical style, after the presumably mora counting Middle High German had developed into a stress timed language (Opitz (1624); Vennemann (1995)), the so-called "Opitzian Reform". Hence, a fixed number of stresses per poetic line in German is regarded as more important than a fixed number of syllables or morae. Alternatively, Japanese poetry concentrates on the number of morae, e.g. the most famous type of Japanese poetic form, the *haiku*, consists of one verse with three groups of words built out of 5, 7 and 5 morae respectively. It is likely, that poetic speech maximizes the language specific rhythmical constraints which are probably often violated in spontaneous speech. Thus, poetic speech can be called to be characterized by a maximal level of rhythmical harmony:

Definition 1
An utterance where the rhythmical structure strictly obey the language specific preferences of rhythmical structure are perceived as rhythmically harmonious.

Not only poets seem to be aware of rhythmical peculiarities characterizing languages. Japanese school children are taught that each mora in Japanese is identical in length (Port et al. (1995)). In the second language classroom the isochrony hypothesis is still used uncontroversially, e.g. to teach English rhythm ((Underhill, 1994, 71), British_Council and BBC (2002), anonymous (2007), (Eckert and Barry, 2002, 195f.)). It is very unlikely that teachers should persistently stick to a concept that is completely without empirical foundation. The concept of isochrony must therefore be taken seriously, at least on an impressionistic or perceptual level.

1.2.2 Problems in Measuring Rhythmic Impressions

However, phoneticians systematically failed to find isochrony of the expected type for stress timing (Classé (1939); Shen and Peterson (1962); Uldall (1971); Roach (1982); Hoequist (1983)), syllable timing (Pointon (1980); Wenk and Wioland (1982); Roach (1982); Vayra et al. (1983); de Manrique and Signorini (1983)) and mora timing (Port et al. (1995); Warner and Arai (2001a)) in languages. Obviously, the rhythmical impressions of isochrony cannot be mapped straightforwardly to the absolute duration of rhythmic events such as syllables, feet or morae. In order to explain this perceptual impression, many sources of influence such as time processing, phonological organisation and linguistic top-down expectations need to be taken into account (cf. chapter 2).

The failure to find isochrony is certainly not surprising at all (also see Benguerel and D'Arcy (1986)). Many, if not all phonologically relevant, i.e. meaning distinctive perceptual impressions do not have a clear counterpart on the level of phonetic realisation. E.g. the phonological feature $[+voice]$ can be realised phonetically by voicing, i.e. a quasiperiodic vibration of the vocal folds, but depending on language and position in a word or syllable, it can also be realized by a short voice onset time, lack of aspiration or less intensity. Naturally, we should not expect that rhythm is less complicated. However, many approaches towards explanations of rhythmic patterns seem to expect that there is a simple,one-dimensional acoustic correlate for a such a complex prosodic phenomenon. Also, depending on the native language, the acoustic cues responsible for rhythm perception may differ. We know from an extensive body of research, that the native language competence heavily influences the way listeners perceive rhythm related prosodic phenomena such as stress or accent ((Jones, 1976a, 245), Eriksson et al. (2002)). Native speakers of languages with a fixed lexical stress tend to show a kind of "stress deafness" (Allen (1975); Dupoux et al. (1997, 2001); Peperkamp et al. (1999); Peperkamp and Dupoux (2002)), but also languages without a fixed stress system use different acoustic cues in order signal rhythm related phenomena. This implies that an operationalization of rhythm related phenomena must be carried out very carefully and designed in a flexible way. In this book, *rhythm is regarded as a suprasegmental property of language and speech* that manifests itself in speech perception, acoustic transmission and production. Therefore, it cannot be enough to describe rhythm in a single dimension such as acoustic phonetics (Lehiste, 1970, 5). Any acoustic effect that cannot be perceived (or at least shows an impact on rhythmic processing) should be treated as an irrelevant detail.

1.2.3 Is Rhythm something Linguistic?

A key issue concerns the status of rhythm within a taxonomy of the different components of a language system. In some models, rhythm is regarded as part of the grammatical prosodic system of a language, implying it to be part of the *langue* in the sense of de Saussure (1916). However, there exists no consensus on rhythm's place. According to Crystal (1969); Couper-Kuhlen (1986) and Lehiste (1970), rhythm is a subcategory of linguistic prosody, separate of paralinguistics (emotions, attitudes etc.) or extralinguistics (speaker specific voice quality etc.). However, Möbius (1993) regards rhythm as a non-linguistic feature of an utterance, in line with paralinguistic categories such as voice quality and speaking style, probably having in mind tempo or speech rate. (Pompino-Marschall, 1995, 236—239) regards rhythmical structure, quantity effects on utterance level such as final lengthening, pausing and articulation rate as related phenomena of all which he sees as part of linguistic prosody. In Figure 1.1 the various suggestions for rhythm's place in a prosodix taxonomy are sketched. In phonological approaches, rhythm rules (cf. Section 2.3.3) are regarded as postlexical, i.e. they are applied *after* the phonological shape of an utterance has been built out of the different words. Phonological rhythm rules operate on the level of lexical and sentence stresses, thereby implying that rhythm is part of the prosodic stress system. Also, phonology tends to regard rhythm rules as optional (Liberman and Prince (1977)), creating a harmonic prosodic shape rather than making an utterance grammatically well-formed in a strict sense. Rhythmic well-formedness is sometimes regarded as a constraint set on a different, abstract level of cognitive organisation which is not obligatory in speech production, but is obeyed much more when reciting poetry (Kiparsky (1975), Nespor and Vogel (1986)). Obviously, phonologists regard rhythm as part of a language's *orthophony*[2], i.e. a linguistic community may agree on certain rhythmical preferences which are not meaning distinctive but do aid comprehension.

What is often forgotten is that rhythm clearly can be influenced by factors that lie within the domain of segmental phonology, such as segmental quantity or the phonotactic complexity of a syllable. Both determine syllable weight and may lead to the perception of a syllable as stressed or unstressed. Another segmental cue to rhythm related phenomena is syllable reduction. Languages, that tend to reduce he unstressed syllables within a foot, tend to be characterized as stress timed rather than syllable or mora timed. The influence of the phonological factors *reduction* and

[2] The term "orthophony" is used to describe phonetic well-formedness constraints analogously to 'orthography'. It refers to well-formedness constraints which are not meaning distinctive, albeit they are used by a linguistic community

phonotactic complexity on rhythm class has been thouroughly described by Dauer (1983); Auer and Uhmann (1988); Bertinetto (1989) and Auer (1993) (cf. Chapter 3). Another factor that has rarely been taken into account as being influential falls into the domain of extralinguistics, e.g. some speakers may have more rhythmical or musical "talent" than others.

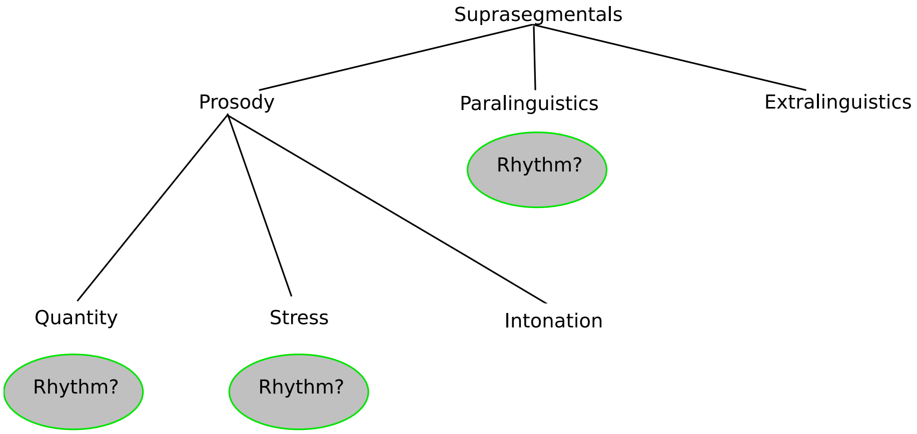


Figure 1.1: Various suggestions have been made in the literature concerning an appropriate level of suprasegmental features which describes the rhythm of language and speech.

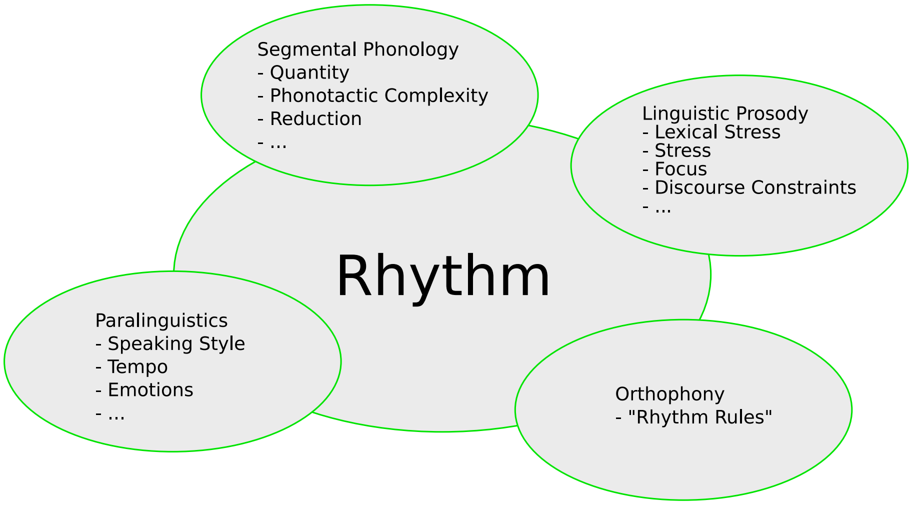


Figure 1.2: Various levels of influence on the rhythm of language and speech.

Clearly, the rhythmic structure of an utterance is determined by all of the levels of the prosodic taxonony mentioned so far (cd. Figure 1.2): It is influenced by linguistic prosody, i.e. placement of lexical and sentence stress, prosodic focus, syllable weight etc. Rhythm can also be influenced by — orthophonic — rhythmical preferences of a linguistic community such as the avoidance of stress clashes or a preference to stress the first or the last syllable within a phrase or utterance. Furthermore, the rhythmical shape can be influenced by speaking style (e.g. prose vs. poetry vs. sermon vs. informal speech), speech tempo (e.g. fast speech has less accents and more words are grouped into prosodic phrases (Trouvain (2004)), paralinguistic influences such as emotion or attitude (e.g. Anger is often connected with a staccato rhythm (Kehrein (2002)) and extralinguistic factors such as rhythmical talent. Thus, is can be said that the rhythm of an utterance is determined by linguistics, orthophony, paralinguistics and extralinguistics and does not genuinely belong to any level within a suprasegmental taxonomy.

1.2.4 Rhythmic Phenomena and their Interactions

Since it is obviously difficult to place rhythm unambiguously within the suprasegmental taxonomy, many researchers are approaching it on the phenomenal level and search for appropriate operationalizations of rhyhmic phenomena. There exists wide agreement that rhythm is the metrical organisation of language/speech, e.g. the structure of strong and weak events as described by *metrical grids* or *trees* in Metrical Phonology or by *prominence patterns* in perceptual phonetics (cf. Section 2.3.5.3). Also, it is usually regarded to be independent of sentence melody, *intonation* or *pitch* in the perceptual domain. This means that the same rhythmical pattern can be produced with a falling or a rising boundary tone—creating a statement or a question — or with a completely flat contour. However, one has to keep in mind that on the level of phonetics, intonational phenomena and intensity (stress) related phenomena are entangled. *Perceptual prominence* expressing stress is also linked to the acoustic fundamental frequency, which is the key correlate of intonation patterns[3]. Thus, a pitch accent which shapes the intonation contour will also be linked to a rhythmically relevant event. The degree to which a language uses intonation rather than intensity and duration as an indicator of prominence is variable (Tamburini (2006)). The influential isochrony hypothesis and the fact that rhythm interacts with tempo has lead to a wide belief that rhythm manifests itself predom-

[3] There is a detailed discussion on the acoustic correlates of perceptual prominence in section 2.3.5.2

inantly in the domain of *duration* on an acoustic phonetic level. E.g. Lehiste (1970) allocates rhythm to the prosodic dimension of quantity, because rhythm can only be described as a pattern in time. Nooteboom (1998) also regards rhythm patterns as something phonetically expressed as duration. There has been a discussion whether it is more appropriate to regard rhythm as a pattern of time or a recurring sequence of events varying in salience or perceptual prominence (Couper-Kuhlen, 1986, 51ff.). A complete description of rhythm clearly needs both of these dimensions: Rhythm consists or structured sequences of variable prominence taking place in time. For the measurement of rhythm in the phonetic domain, this means that several candidates need to be taken into account, all of which interacting with other levels of suprasegmental properties of language and speech. Figure 1.3 shows the different phenomenal levels of rhythm measurements in the domain of perceptual and acoustic phonetics. Rhythmic production is missing in this figure, mostly because it has been rarely studied directly in speech. The probable reason for this is that the articulatory production correlates of rhythm are not as clear as in other domains of rhythmical movement, e.g. finger tapping, gestural 'beats' or the movements of musicians which have been studied more extensively (cf. Wachsmuth (1999) for an overview).

1.2.5 A Working Definition of Rhythm—and some Consequences

The first working definition of rhythm follows Woodrow (1951) and Rammsmayer (2000), but unlike theirs, is not limited to the perceptual domain. This point of view is in line with widespread views on rhythm coming from various disciplines such as musicology, phonetics, cognitive psychology and psychoacoustics (e.g. Fraisse (1982); Lerdahl and Jackendoff (1983); McAngus Todd and Brown (1996); Benguerel and D'Arcy (1986)).

Definition 2
Linguistic rhythm is regarded as the structure of similar suprasegmental events in time. In order to describe this structure, it is necessary to (1) perform a segmentation of the linguistic chain into rhythmically relevant consecutive events, e.g. syllables. Furthermore, (2) these events must be grouped in such a way that consecutive groups become similar in structure, e.g. consist of similar patterns of strong and

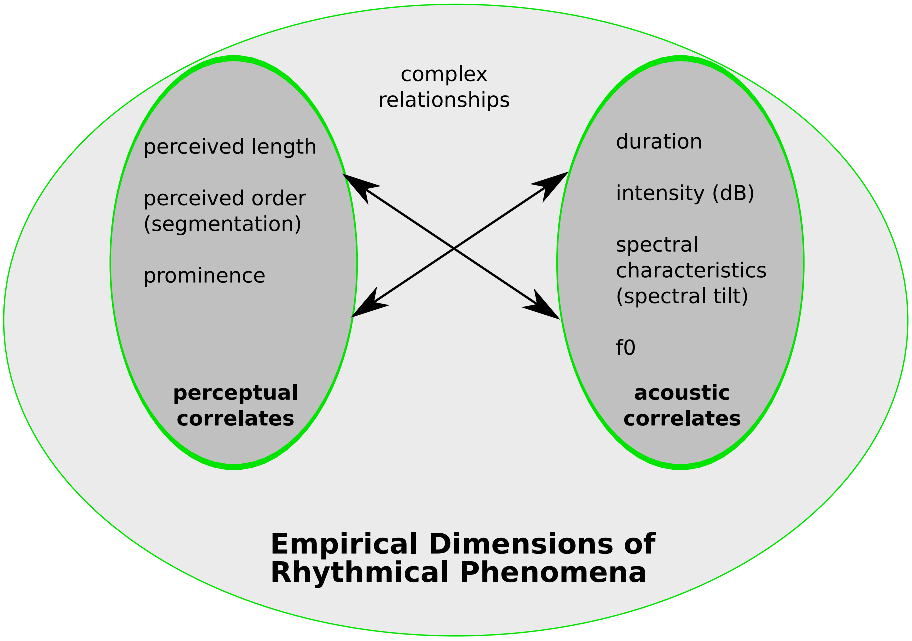


Figure 1.3: The various correlates of the abstract phenomenon rhythm in the perceptual and acoustic domain. The connections between the different descriptive levels are complex, e.g. prominence can be expressed as a complex function of duration, pitch excursion, characteristics of spectral and overall intensity.

weak events and/or contain a similar number of rhythmical events.

This definition implies, that rhythm, like speech, consists of events in time. What is fundamental to a rhythmic structure now, is the definition of the units of speech which make up rhythmic events. It is very likely, that these are syllables, but language specifically, alternative or additional units like the mora, the foot or the phrase can be important as well (Cutler (1994)). This segmentation process cuts up speech into chunks of rhythmic events. Furthermore, we need to describe the rhythmical structure of the rhythm chunks, in order to assemble them into groups of similar successive events. During structuring, we might find strong (or stressed) syllables followed by one or several weak (or unstressed) ones. This implies that structure considers the number of events and their relative strength. In a poem with a regular meter, e.g. a dactyl, the groups will be quite uniform, consisting of one stressed syllable which is followed by two unstressed syllables. But in normal conversation, these groups will most likely be of variable size. Which kind of groups are regarded as structurally similar, is not contained in the definition, it is matter of empirical research and will probably vary across languages, speakers and speaking styles. This means, that structure is not necessarily generated by what is usually called *stress* and its position, e.g. at the end (iamb) or the beginning (trochee) of a foot. Structure can also be created by *counting* events at different levels of the prosodic hierarchy, as a dactyl consists of three syllables, a trochee of two, a haiku line of five or seven morae etc. Rhythmical patterns emerge through segmentation and grouping processes at different levels of rhythmical organization. In our approach, all grouping automatically implies a segmentation on the next higher level in our rhythmical hierarchy. E.g. if a syllable is perceived as stressed (structure), it also marks the beginning or the end of a foot on the next level of a rhythmical hierarchy. This assumption is in accordance with the *strict layer hypothesis* which states that prosodic boundaries must coincide with boundaries at lower levels of the prosodic hierarchy[4]. However, the definition does not make any statement whether every established level of the prosodic hierarchy is used during rhythmical structuring by a language. It is possible that some levels, such as the mora or the foot are simply ignored by some languages. Figure 1.4 exemplifies the process of segmentation and grouping for the beginning of the German children song *Alle meine Entchen*[5].

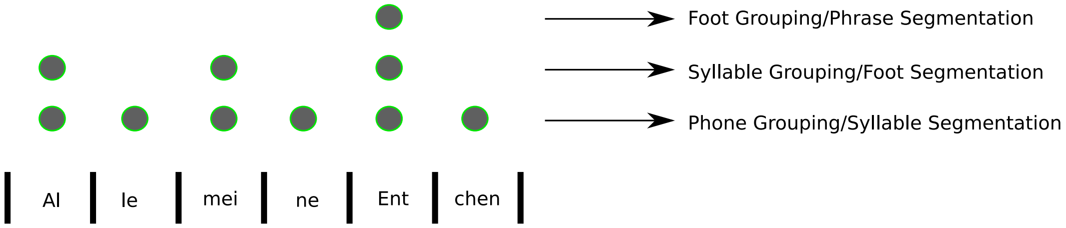

Foot Grouping/Phrase Segmentation
Syllable Grouping/Foot Segmentation
Phone Grouping/Syllable Segmentation
Al
le
mei
ne
Ent
chen

Figure 1.4: A string of phones is segmented into rhythmically similar events, e.g. syllables. Syllables are then grouped into higher level rhythmical events such as feet, phrases etc. based on rhythmical similarity, e.g. feet may start with a stressed syllable. phrases end with a stressed foot etc. Grouping constraints and their phonetic implementation as well as the levels of metrical organisation are language specific, i.e. it depends on the language whether it chooses to use morae as level of rhythmical organization below the syllable and whether speakers understand lengthening as a phonetic means of indicating phrase boundaries etc.

The following list subsumes the different ingredients a rhythmical pattern needs in order to exist:

[4] For one version of the prosodic hierarchy, see Nespor and Vogel (1986) and cf. Figure 2.26)

[5] All my little ducks

1. Segmentation - Where are the rhythm related events? Where are their boundaries?

2. Structure - Which event is weak, which is strong? Is it shaped in a particular way? How many events are there?

3. Grouping - Which event chains are similar in structure and thus belong together?

If we are interested in detecting the language specific rhythmical differences of rhythm on the phenomenal level, we need to find out appropriate acoustic and perceptual measurements telling about language specific mechanisms of segmenting and structuring an utterance. This is the key question an answer is sought after in the upcoming chapters.

Chapter 2

Rhythm Production and Perception

In the previous chapter, we introduced a working definition of rhythm which stated that rhythm can only be explained as an interaction of segmenting an acoustic stream into rhythmically relevant units and grouping these units into patterns based on their structural properties. Based on this definition, the different sources influencing rhythm perception and production will be investigated. The focus lies on the perception side but rhythm related production mechanisms will be discussed as well. When evaluating the various rhythm related processes, a complex picture emerges that disentangles the different sources of influence: peripheral influences of rhythm processing as well as linguistic and phonetic ones will be taken into account. Each subsection will be concerned with answering part of the following key question:

> Which (acoustic, phonetic or phonological) properties provide the relevant cues to rhythmical segmentation and grouping?

At the end of this section, it should become clearer, how the complex relationship between rhythmical experience and acoustic phonetic or phonological properties can be more adequately described. The end of this chapter then provide the building blocks for a model of the rhythmical processing of language and speech.

2.1 General Aspects of Processing Auditory Events in Time

2.1.1 Objective vs. subjective time

Rhythmic patterns are events in time. Thus, before regarding rhythmical events, it will be discussed how time is cognitively processed. It is important to realize that objective time spans, e.g. as we measure them with the help of clocks, can be very different from subjectively perceived time spans. A well known example for this is the effect that "time flies" in some situations while a minute can be perceived as almost neverending in other situations.

Experimental psychology has provided experimental data that these effects are indeed true: Past time spans can be judged as having been of very different length, e.g. when the attention of listeners is captured by a click sound, the subsequent auditory event will be experienced as longer in duration (Weardon (2004)). Wearden explains this effect by an acceleration of a *mental pacemaker* and shows that the effect of the preceding click is multiplicative, i.e. an objectively longer duration is stretched more than an objectively short duration. While the same amount of time is passing by objectively, our mental pacemaker notices more "ticks" because it is running faster. Thus, the amount of mental activity obviously plays a major role in how long or short we perceive events taking place in time. This explains the effect that with increasing age, many of our daily tasks are routine and need less attention, thus,"time flies faster", because our mental pacemaker is ticking with a lower frequency during routine tasks. The key role of attention in time perception also provides an explanation for the well-known effect that "a watched kettle never boils". Also, the level of stress or emotional involvement can play a role in changing our "internal timekeeper." Rao et al. (2001) could show in an fMRI study of the neural activities during a duration estimation task that the basal ganglia and dopamine level plays a role in the perception of time. This explains why people in situations with raised or lowered levels of dopamine, e.g. caused by stress, diseases or certain drugs, experience time faster or slower than in other situations. Thus, it is likely that there exists something like an internal timekeeper that can be influenced by the level of neural activity in the basal ganglia caused by neurotransmitters but also by the level of cognitive attention. Recent research on language specific duration perception and child directed speech revealed furthermore that our temporal resolution while listening to speech may be influenced by our mother tongue. Listeners whose mother language uses subtle differences in duration as a cue to meaning distinction show a better performance at rating duration differences than speakers of languages where duration has less functional load (Krull et al. (2003)). Krull and her colleagues draw the conclusion that the pace of listeners' inner clocks may depend on their native language.

From these results of (neuro)psychological and phonetic research, we can deduce that objectively measured durations cannot straightforwardly mapped to perceived durations.

The *mental pacemaker* or *internal clock* is not a new idea but has been an integrative part of psychological research in time perception for many years. Such an internal clock has been claimed to regulate much of human's and animal's life. In this study, we will not engage in the debate whether such an internal clock really exists and how it manifests itself on the level of neurophysiology, but we will take into consideration the experimental results suggesting the existence of it. There are different assumptions concerning the inner clock, some believe it to be standing in a linear relation to real time (Weardon (2004)), but such a simple model is challenged by many findings in the literature. E.g. we know that short intervals $< 0.6s$ tend to be overestimated and longer ones tend to be underestimated (Fraisse (1963, 1982)).

2.1.2 Processing the Continuous Time Stream

Another important rhythm related question concerning time perception is the somewhat philosophical but also psychological question, whether time is perceived as a continuous stream or as a sequence of individual percepts whithin each everything that is perceived is interpreted as being simultaneous. Since we are usually able to remember the order of the words (or phones or syllables) it is obvious that at some level we are able to make a decision concerning the order of auditory speech events. But it has been a matter of investigation how precise we are in this respect. Imagine the individual singers of a choir starting to sing simultaneously — it is very likely that the individual singers did not start precisely at the same time, due to differences in reaction times or different levels of attention with regards to the choirmaster's instructions. It has been a matter of experimental research how much to auditory events need to differ in their starting time in order to be perceived as nonsimultaneous. If we can answer that question, we know the limits of our temporal resolution.

Experimental data show that humans often perceive stimuli that are produced physically in distinct time frames as being simultaneous. Thus, human "temporal resolution" is certainly not comparable to, e.g., a tape recorder which is able to playback everything in the same order as the original physical signal. This will not even be the case if a human had a perfect memory, because her peripheral processing will lead her to perceive certain auditory events as simultaneous or overlapping even if they are (physically) not. Thus, unlike tape recorders, humans will, in a "playback situation" certainly deviate from the original physical stimulus. However, temporal resolution differs much between "speech", "non speech" and the type of stimulus. It is obvious that the perception of auditory events as simultaneous or consecutive has an impact on rhythm perception. If auditory events cannot be ordered, no rhythmical impression can emerge. A number of psychoacoustic investigations showed that isteners are able to judge the order of two different auditory events that differ roughly 20-40ms in their onset (e.g. Hirsh (1959); Rosen and Howen (1987); Wittmann and Pöppel (2000)) but Broadbent and Ladefoged (1958) report much higher thresholds which only dropped after rehearsal. This effect can be explained with the fact that temporal resolution is higher when we pay attention (cf. 2.1.1). It is very likely that subjects learnt to focus their attention at the relevant points in time in course of the experiment. However, a temporal resolution of 20-40ms would be enough to perceive the segmental order of most speech sequences. Much shorter differences are necessary for listeners to perceive that two stimuli are different, but the kind of difference does not enter conscious experience (see the discussion in (Plank, 2005, 2-3)).

When the auditory stimuli are not short clicks, or similar sounds used in psychoacoustic experiments, performance apparently varies: Listeners still perceive the different players in an orchestra as beginning in unison when the average deviation of each musician is between 31 and 51ms (Rasch (1988)) and Warren (1999) reports that the increasing complexity of a task causes this increase in threshold.

Pöppel (1997) states that a threshold of 30ms is typical for the detection of temporal order in different sense modalities, not only auditory perception. Below that threshold, auditory percepts merge into one complex sound impression. With regards to speech, temporal order is easier to determine when our linguistic experince provides us with hypotheses concerning the most likely order, e.g. based on our phonotactic knowledge and our experience concerning the probability of certain phone sequences, we derive hypotheses concerning segmental order (Fay (1966)). The circumstance that the order of some sequences is perceived better than that of others is explicable with the masking effects between a gradual tuning curve of a beginning and the fading of a previous auditory event. Especially comparatively loud speech events like vowels can heavily influence the perception of subsequent consonantal speech events (Zwicker, 1982, 93ff.). Segmentation of speech events is heavily influenced by these masking effects.

Recent models (cf. (Plank, 2005, 6-13)) that explain how the auditory system decodes acoustic information agree that the continuous flow of acoustic input is first processed in short temporal units consisting of few milliseconds, so-called *multiple looks*. These *multiple looks* are then integrated into larger units which can be compared with memory templates that are *optimal* representations of earlier experiences. Neurophysiological and psychoacoustic studies furthermore imply the existence of a so-called *sliding window of temporal integration* (Bregman (1990); Näätänen (1992)) where the individual *multiple looks* are accumulated and combined to an auditory impression. This window has approximately the size of 170ms (Yabe et al. (1998)). The size of such windows has been studied by measuring the brain's electric mismatch negativity (MMN) which is elicited by an abrupt sound change in a repetetive homogenous sound. Therefore, it is regarded to reflect automatic change detetion in the autitory cortex. The detection of such changes is vital in speech perception, e.g. in order to detect syllable boundaries which tend to be characterised by intensity changes. It is interesting that the size of this window has a similar duration as typical syllables in many of the world's languages.

It is safe to say that order estimation of auditory speech events is fuzzy and varies with the type of auditory stimuli involved as well as the listener's attention and learning experience.

2.1.3 Duration Estimation

Much research in psychoacoustics has studied the subjective duration of auditory events. In particular, researchers have studied the so-called *Just Noticable Differences* (JNDs) wich indicate the necessary difference in duration between two auditory percepts to be perceived as being of different length. An important psychoacoustic relationship concerns the relation between the JND (ΔT) and the absolute magnitude (here: duration) of the stimulus (T) where JND is the stimulus difference which results in 75% correct identification across subjects. For many sensory experiences, this relationship can be expressed as a constant ratio (*Weber's Law*) expressed in equation 2.1.

$$W = \frac{\Delta T}{T} \tag{2.1}$$

If Weber's law holds, then the ration between JND and the reference time span remains constant across different durations. Furthermore, it would strengthen the theory of Weardon (2004) that subjective and objective time stand in a linear relationship (cf. section 2.1.1). JNDs are often expressed as percentage values that indicate the necessary deviation from a reference duration T (cf. equation 2.2).

$$JND = \frac{\Delta T}{T} \times 100 \tag{2.2}$$

Psychoacoustic data provides some evidence to support the rationale behind Weber's Law for temporal processing, e.g. that longer intervals have higher JNDs than short intervals: When perceiving short clicklike sounds, temporal resolution is much more precise and may even perceive differences of 1 or 2 ms, while long musical sounds are difficult to discriminate in length (Bruhn (2000b)). Thus, the longer two sounds, the more difficult it is to discriminate their duration. However, Weber's Ratio is oversimplistic in order to estimate perceived duration differences. A large body of research implies that the ratio between ΔT and T does not remain constant. For stimuli between 50ms and 2000ms it can be better described as a nonlinear function indicating different processing mechanisms for short and long intervals. (McAuley, 1995, 25-27)) suggests to represent JNDs with a stepwise function referring to a time quantum of 50ms which is either halfed or doubled, depending on the the base interval. This would mean that for syllable sized units, we can expect a JND of roughly 25ms, for foot sized units of roughly 50ms. Friberg and Sundberg (1995) however found that below a threshold of 250ms JNDs remain quite constant around 6ms, while above 250ms, Weber's Law holds and the JND remains rather constant around 2,5-3%. From these findings we can deduce a higher temporal resolution for short intervals up to approximately 250ms. It is important to keep in mind, though, that due to their complexity and continuous nature, speech events are processed differently than non-speech events and JND measurements have to be interpreted with caution from a phonetic point of view (Lehiste, 1970, 13).

Another interesting fact that is possibly important for the processing of speech rhythm is the fact that filled intervals are processed more precisely than unfilled intervals (Rammsmayer (2000)), i.e. the JNDs for pauses are higher than those for auditory stimuli consisting of a sound. With regards to our assumptions concerning the relationship between attention and temporal resolution (cf. Section 2.1.2) this would imply that listeners pay less attention during pauses — and this certainly does not come as a surprise from a listener's point of view.

2.1.4 Temporal Instants and Time Spans

In order to explain the contradictory and diffuse results concerning duration estimation, some researchers have argued for two separate psychological mechanisms in the perception of time. I.e. they claim that relatively short and relatively long durations are processed very differently, e.g. (James (1890); Woodrow (1951); Fraisse (1963, 1982); Hellström and Rammsayer (2004)). Most researchers regard temporal intervals up to 400-600ms to be perceived as *temporal instants* and longer intervals as *time spans* where temporal or rhythmical structures are perceived. Temporal instants are perceived subcortically and their processing lies beyond conscious control. With regards to speech, we would thus expect a sequence of roughly two syllables (a typical "foot") to be a temporal instant, while a whole prosodic or intonational phrase would count as a time span which can be perceived as belonging together and forming a rhythmical pattern. The small window of 400-600ms is often referred to as the *psychological present*. This notion implies that even if a listener is able to perceive the relative order and durational structure within this interval, the auditory events are experienced as belonging together, creating a singular event in time that is perceived as *presence*. There exist different assumptions concerning the duration of the longer timing window comprising time spans. Assumptions concerning its length range between 2 and 8 seconds (Bruhn (2000b)). In his highly influential work, Pöppel (1990, 1994) proposes a temporal window of approximately 3 seconds as a neural segmentation unit of mental life. Such a window would typically cover an entire *intonation phrase* or *utterance* within which the different temporal instants are integrated into perceptual gestalts thus creating a rhythmical pattern. Within this unit, we are able to make estimations concerning the timing of upcoming events, e.g. regularly occurring auditory stimuli. It seems that the perception of rhythmical patterns is also closely linked to the capacity of our working memory (Pöppel (1997)). If one refers to the highly influential model of working memory proposed by Baddeley and Hitch (1974); Baddeley (2000) (cf. Figure 2.1), it is very likely that rhythmic patterns are stored in the *phonological loop* rather than the *visuo-spatial sketchpad*, because distracting articulatory tasks disrupt rhythm memorization much more than spatial tasks (Saito and Ishio (1998)).

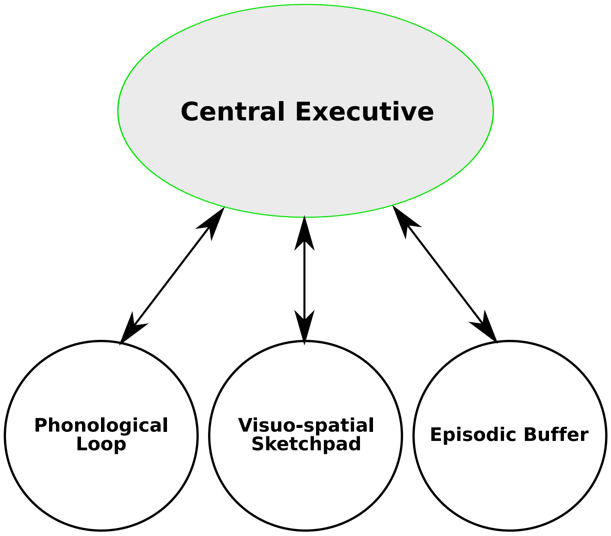


Figure 2.1: Baddeley's model of the working memory consists of the central executive controling information flow, the phonological loop acting as an "inner ear" and "inner voice", the visuo-spatial sketchpad storing and controlling visual and spatial information (but not reading!) and the episodic buffer, which integrates and stores visual, verbal and spatial information in chronological order.

The link between working memory and time processing also explain the effect that based on previous experience, humans make certain predictions concerning the timing of upcoming events, i.e. rhythmic patterns can also draw our attention to certain points in the near future, where something important, e.g. a stressed syllable, is likely to occur and listeners show a better performance in detecting fundamental frequency differences (Jones (1976b)).

2.1.5 Perception of Tempo

One of the main perceptual effects when integrating temporal auditory events across time is the perception of *tempo*. When successive auditory events become longer in duration, we should perceive this as deceleration or in musical terminology *ritardando*, while the opposite, with successive auditory events becoming shorter, should be perceived as an acceleration or *accelerando*[1]. The relationship between absolute duration and the perception of tempo has been examined mostly within cognitive psychology and music perception (e.g. the influential model by Povel and Essens (1985)), but also in speech (Pfitzinger (2001); Quené (2007)).

In speech, there have been several proposals concerning the most appropriate relevant unit in which to measure speech tempo. Suggestions have been to measure the tempo of speech as a function of *syllables per second*, *phones per second* or *consonantal plus vocalic intervals per second* (Dellwo (2008a)). Another point concerns the question whether these metrics should include pauses or not. The speech tempo where the pauses are included has been referred to as *speech rate*, while the measurement exclusive pauses has been referred to as *articulation rate* (e.g. Trouvain (2004)). This distinction relates to the fact that pauses constitute important events in spontaneous speech related to speech planning (Goldmann-Eisler (1968)) and they certainly play an important role for dialogue interactions, e.g. they may function as turn taking signals (Peters (2006)). If a tempo estimation aims to express the amount of *linguistic information transmitted to a listener*, it may therefore be useful to measure tempo as speech rate rather than articulation rate. However, it would be a misconception to believe that by omitting pauses, we get a metric that is independent of the amount of information transmitted in the speech signal. Since duration per-

[1] In musical notation, there exist several related tempo related terms, such as *rallentando*, *stringendo*, *ritenuto* which provide information concerning the time frame in which the musician should slow down or accelerate her interpretation. The most widely used *ritardando* and *accelerando* leave this to the musician's interpretation.

ception is too a large extent determined by attentional factors (cf. Section 2.1.2), it certainly will make a difference whether a syllable transmits much or few linguistic content. Since syllables can not only have different durations but also contain different amounts of phones, we can at least roughly conclude that the linguistic information contained in a syllable may vary considerably. In his model of *perceptual local speech rate*, Pfitzinger (1999, 2001) tries to combine linguistic content (phone rate) and articulatory tempo (syllable rate) in order to take care of this combinatory effect. His prediction model (cf. equation 2.3, where s is a coefficient multiplied with syllable rate (sr), p is a constant multiplied with phone rate (pr) and c is a constant minimizing the prediction error.)[2] is derived from a multiple linear regression analysis and reaches very high correlations between perceived and measured speech rate. With the help of this speech rate estimate, local perceived speech rate can be calculated for small windows across the entire utterance and a *speech rate contour* can be drawn. Pfitzinger's results covary nicely with findings by Dellwo (2008a) who could show that in languages usually regarded as syllable timed, more cv-intervals are produced per second than in so-called stress-timed languages. Since syllable-timed languages such as French tend to have less complex syllables (cf. Chapter 3), according to Pfitzinger's argumentation less linguistic content is transmitted per syllable. Consequently, a speaker of French has to produce more syllables per second in order to transmit the same amount of linguistic information. It is very likely that with regards to phone rate, French speakers do not speak much faster than those of a stress-timed language. It is very interesting, that an almost identical model has been described by Bruhn (2000a) to explain the tempo perception in music. He claims that the perceived tempo of a musical piece can be described as a function of the absolute tempo, called *horizontal flow* (e.g. the tempo measured by a metronome) and the musical complexity (called *vertical pressure*). Bruhn builds his model on suggestions of the former composer and conductor of the Munich philharmonic orchestra, Sergiu Celibidache. If "horizontal flow" is equated with syllable rate and "vertical pressure" with phone rate, then both models become very similar.

$$PLSR = s \times sr + p \times pr + c \tag{2.3}$$

[2] Pfitzinger's model works reliably for German data using the following constants: $s = 7.38$, $p = 4.06$, $c = -1.41$

Pfitzinger's model is aimed to detect smooth transitions in speech rate Abrupt changes are not treated adequately, as Pfitzinger notes himself (Pfitzinger, 2001, 220), since the local speech rate is always calculated across a larger window.

The question of a durational threshold leading to the perception of acceleration or deceleration has been treated by Quené (2007). He aimed to identify the *Just Noticable Difference* (JND) for the perception of tempo changes in speech. He claims that it is necessary to stretch or compress speech by 10% in order to achieve a noticable perceptual effect.

While information load leads to an increase in perceived tempo, we already learned in section 2.1.2 that attention leads to an increase in perceived durations. This also has an effect in tempo perception: Psychoacoustic studies show that frequent pitch changes or unexpected accentuations probably increasing the listener's attention, lead to the impression of a decreased tempo (Boltz (1989)), thus, attention has the opposite effect to information load.

Summing up, it is likely that there exists a tradeoff between information load, cognitive attention, objective tempo, local and global effects that need to be taken into account for a precise estimation of tempo perception.

2.1.6 Implications for the Rhythmical Processing of Speech

In this section, rhythm related research results from psychoacoustics, auditory phonetics and psychology have been presented. Some of these have direct implications on the perception of speech rhythm. A first issue concerns the cognitive resolution when processing auditory events: It seems obvious, that listeners cannot perceive the sequential order of events shorter than roughly 25 or 30ms. This resolution should be sufficient to process the temporal order of most phonetic segments, e.g. the order of vowels and consonants in sequences of speech. However, this resolution can vary, since it is influenced by a second major factor, namely attention: Paying attention to an auditory event causes its appearance of being longer because it increases the frequency of time resolution. Our mental pacemaker "ticks more often" in these cases, causing the impression of a longer duration. In speech, a listener's attention may be caught by sounds where a lot of dynamical changes are involved and no gradual tuning curves are present, i.e. plosives. This enhances the ability to identify such a segment within the rather continuous auditory stream. Plosives are thus ideal indicators for human speech segmentation (Zwicker and Fastl (1999)). Furthermore, plosives typically occur at syllable boundaries so that their identification provides useful cues for syllable segmentation. If our attention is caught and increased by a syllable intial plosive, the upcoming rest of the syllable would be perceived as longer than it objectively is. This explains the well-known effect that the prominence or metrical strength of syllables is mainly determined by its rhyme, i.e. the nucleus vowel and the coda consonants of a syllable. The onset consonants of a syllable have hardly any impact on a syllables subjective prominence (e.g. for German Mengel (2000);Wagner (2002)). Also, our attention may be drawn to places in an utterance where something important is likely to be said: Syntax may reserve particular slots for focussed constituents, therefore words or syllables uttered within these designated time frames of increased attention may appear as longer. This implies that a speaker may focus a word only by chosing a plausible spot — no special accentuation will be absolutely necessary in order to mark the word as perceptually prominent.[3] From child-directed speech we know that points of attention are rehearsed and duration plays a major role during rehearsal: Swanson and Leonard (1991) report that mothers lengthen content words when talking to their infants, but not function words, probably in order to direct their children's attention to those points in speech that are most important. After a child has mastered its native language's syntax, it should have learned the typical anchor points of important content in utterances, i.e. it knows when paying attention is most useful.

From neurophsyiological and psychoacoustic research (see 2.1.2), we furthermore learnt that there exist "sliding windows of temporal integration" of approximately the size of 150-200ms. These timing windows play a crucial role in auditory recognition: Listeners segment the continuous auditory stream into chunks of this window size and then compare each chunk with auditory templates stored in long term memory. It is difficult to imagine it a mere coindicence that an important rhythmical unit of speech, the syllable, typically has durations falling into this window

[3] It is certainly likely that a speaker may do something anyhow to accentuate the focussed word further - We know from much research on prosodic foci that speakers tend to produced focussed words with considerably more effort. However, the role of attention might explain why no focus marking is necessary in written speech.

of temporal integration (cf. Table 2.1.6 and Figure 2.2). Furthermore, the syllable is regarded as an important unit in human speech processing. A common belief is that syllables are stored as templates in in a mental *syllabary* (cf. Levelt (1994)). The similarities between general models of auditory processing and psycholinguistic ideas of speech processing are obvious. We therefore assume a close link between the temporal integration of auditory information and the syllable as a fundamental unit of human speech processing.

Language	Syll duration (mean)	Standard Deviation
English	178ms	92ms
French	155ms	58ms
German	184ms	75ms
Italian	142ms	68ms
Polish	172ms	67ms

Table 2.1: The average syllable durations and standard deviations for various languages in the BonnTempo database (Dellwo et al. (2004)). The database contains speech read at different speaker specific articulation rates so that the average durations reported here should factor our speech rate influences to some extent—it may also be responsible for the relatively large standard deviations. It is obvious that there are language specific variations of syllable duration but the mean durations are surprisingly close to the time span identified as *sliding window of temporal integration*. Thus, a syllable seems to be a good candidate for the identification of minimal gestalt units of temporal integration in speech.

The connection between duration, perception, points of attention, time stream segmentation and linguistic structure is certainly no coincidence. But the way that linguistic knowledge regulates our inner clock certainly depends on the way our native language employs duration in order to encode the linguistic message. If is is used on the segmental level, e.g. in Estonian with three levels of vowel length, listeners have to concentrate differently than listeners of English, were duration is predominantly used on the suprasegmental level, in order to signal lexical stress or phrase boundaries[4].

[4] It is clear, that listeners of English also need to pay attention to durational cues on the segmental level, e.g. to distinguish voiced and voiceless final stops, which are indicated by the preceeding vowel length, but duration in English certainly has less functional load on the segmental level.

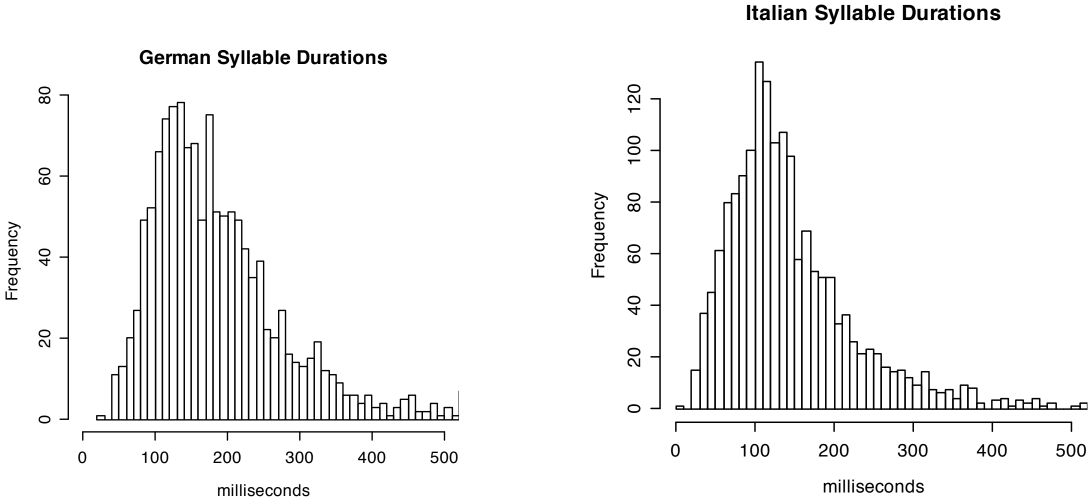


Figure 2.2: These histogrammes show the similar syllable duration distributions across three native speakers of German, a stress timed language and Italian, a syllable timed language. Both languages prefer feet of similar durations disregarding their supposed rhythmical difference. Speakers were selected randomly from the *Bonn Tempo* database.

The threshold of 25ms also seems to be relevant for noticing duration changes. Syllable sized entities need to differ roughly 25ms in order to be perceived of different length. For longer intervals around 500ms, the difference needs to be roughly 50ms. I.e. in order for a duration change to be of perceptual relevance, it needs to be of this size. Smaller duration changes can be safely ignored from a perceptual point of view.

Another important timing window seems to be a window of roughly 400-600ms which can be regarded as *temporal present*. Interestingly, the time span of the temporal present co-occurs with typical foot length across a variety of languages (cf. Table 2.1.6 and Figure 2.3). A foot sized unit is the minimal unit which can be seen as relevant for perceiving local rhythmical patterns, e.g. the unit in which we are able to perceive a minimal rhythmical structure like *weak—strong* or *long—short*. Also Fant and Kruckenberg (1996) suggested a 500ms window as in important reference unit of speech planning. Their investigations were carried out on Swedish data, but since the psychological data cited above also suggests such an interval as important for temporal processing, it might well be that their suggestion is more universal. The local rhythmical patterns comprising windows of temporal present are integrated into larger rhythmic-perceptual *gestalts* within time spans of 2-3 seconds. In speech, such time spans can comprise roughly 1-2 prosodic phrases. Often, they will constitute a whole utterance. Again, it is certainly no coincidence that such time spans are strongly connected to the capacity of our working memory. In these longer temporal units, we may obviously look for the global rhythmical structure of speech. It is unclear, whether it is useful to search for rhythmical patterns beyond the scope of utterances. Altogether, it seems that the timing and structure of rhythmically relevant speech units has been "taylored" to the human capacities of time processing.

Language	Foot duration (mean)	Standard Deviation
English	439ms	130ms
French	416ms	129ms
German	446ms	128ms
Italian	397ms	113ms
Polish	480ms	162ms

Table 2.2: The average foot durations and standard deviations for various languages in the BonnTempo database Dellwo et al. (2004). The database contains speech read at different speaker specific articulation rates so that the average durations reported here are not rate dependent but should factor our such influences. It is obvious that there are language specific variations of foot duration but the mean durations are surprisingly close to the time span identified as *psychological presence* — even taking into account the standard deviations most feet should fall into this time span. Thus, a foot seems to be a good candidate for identification of minimal local rhythmical structures.

2.2 Dectecting the Fundamental Entities of Rhythm: Beats and Perceptual Centers

In the last section, the psychophysiological mechanisms and processing limitations of temporal resolution, duration perception and tempo estimation have been discussed. The knowledge of these mechanisms is necessary to decide later, which temporal windows and durational changes that can be objectively measured in speech

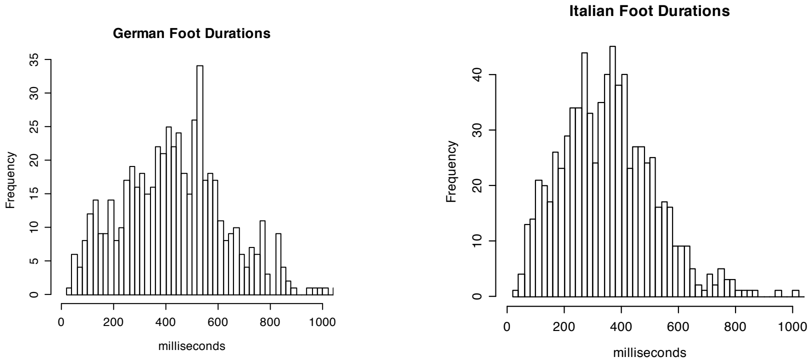


Figure 2.3: These histogrammes show the similar foot duration distributions across three native speakers of German, a stress timed language and Italian, a syllable timed language. Both languages prefer feet falling into a range between 200 and 500ms. The speakers were selected randomly from the *Bonn Tempo* database (Dellwo et al. (2004))

are relevant to detect rhythmical patterns in first place. The mere processing of an auditory stream as temporal events does not provide information, how the listener chunks this stream into rhythmically relevant units, e.g. the fundamental *beats*. Related to music, these *beats* would represent the tappings when imitating a sequence of musical notes. It is important to notice that these movements are not identical to the kind of tappings typical for accompanying a musical tune — in those cases, we will usually just tap in certain accented positions. Accentuation is a matter of *rhythmical grouping* (cf. Section 2.3.2). Before accenting selected beats, we need to identify the sequence of events some of which are accented, some of which are not. Thus, the rhythmical beats referred to here make up the fundamental rhythmic units which can be accented or grouped based into higher order units (see Figure 1.4 on page 17). These groupings are based on principles which make up the individual rhythmic pattern and may be characteristic for a certain language or speaking style. According to Klapuri et al. (2006), one can postulate an additional level *below* the level of fundamental beats, called the *tatum* which is the minimal duration in between two fundamental beats[5]. In this section, it will first be shown how an auditory stream can be segmented into chunks that may be perceived as fundamental beats, then different models of beat detection will be described that originate in

[5] Its linguistic equivalent is the mora, as will be explained in Section 2.2.4

psychoacoustics (cf. 2.2.1), in music technology (cf. 2.2.2), phonetics (cf. 2.2.3) and phonology (cf. 2.2.4) will be described and discussed.

2.2.1 The Perception of Fundamental Beats

The task of identifying fundamental beats in the acoustic stream is to detect the beginnings and ends of rhythmically relevant units[6]. The often continuously changing nature of auditory (speech) events shows that this mechanism is not self-evident, e.g. the acoustic representation of a sequence “lalalala”, it is difficult to detect the clear boundaries between various possible individual beats (cf. Figure 2.4). However, the high frequency of usage of “lalala” in singing provides evidence for its usefulness to transmit a rhythmic impression to a listener. Thus, the listener is obviously able to estimate the fundamental beats contained in the sequence and group these into a rhythmic pattern. In order to perceive auditory events as separate units, it is essential to hear when the separate events begin and when they end—even while they possibly partly overlap physically and acoustically. E.g. a vowel that is followed by a nasal consonant may carry traces of nasality without being perceived as part of the nasal — a normal phonetic phenomenon known as coarticulation. Besides, in reality we often have to segment auditory events of much higher complexity, e.g. when different speakers talk simultaneously or an orchestra plays.

Bregman (1990) calls this cognitive process *stream segregation*. According to him, at an early stage of neural processing different auditory streams are segregated on the basis of the following principles:

- Sounds standing in no relation to each other, do not start or end simultaneously.

- Sounds originating in a single source tend to change their frequency characteristics continuously rather than abruptly.

- Regular vibrations of a single source result in an acoustic pattern where the frequency components are integer multiples of the fundamental frequency.

[6] In Section 2.1.6 it has been discussed that plosives provide useful acoustic cues for segmenting a continuous speech signal into syllables. However, this is an insufficient strategy because syllable boundaries are not always marked by plosives.

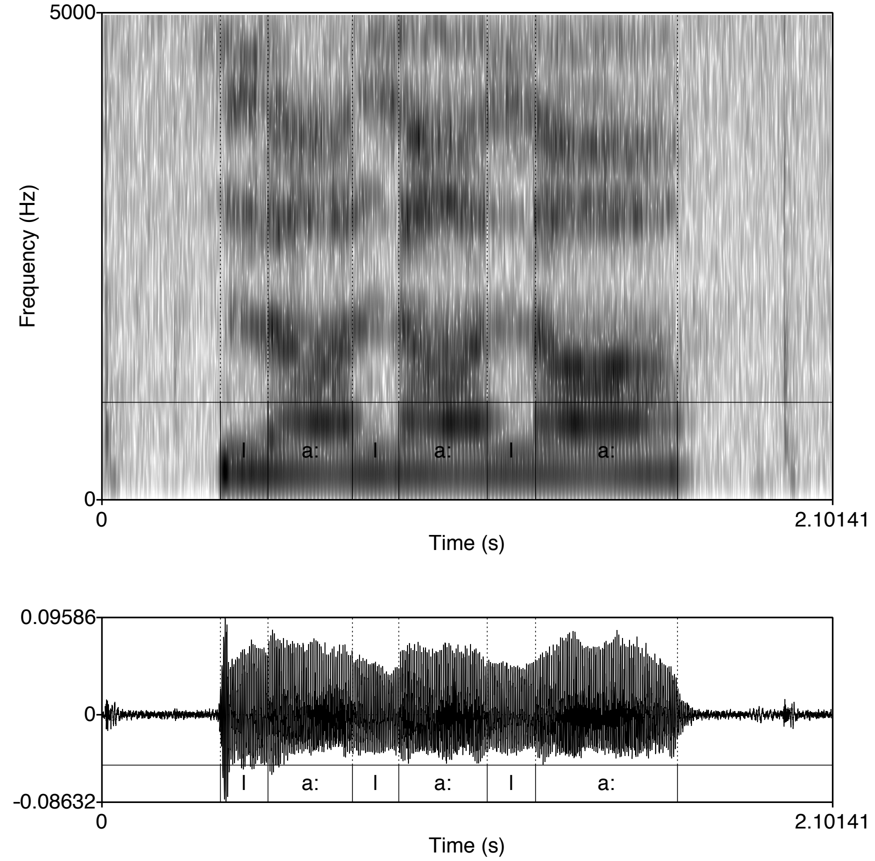


Figure 2.4: This figure illustrates the difficulty of segmenting the continuous acoustic speech signal into meaningful chunks that make up the fundamental beats necessary for a rhythmic pattern. The top signal shows the annotated spectrogramme of the sequence "lalala", the bottom signal shows the corresponding oscillogramme.

- Many changes occurring in an acoustic event affect all of its subcomponents of the resulting sound in the same manner and at the same time.

These effects combined lead to the phenomenon that in orchestral music or other types of polyphonic music, certain sounds coming from different instruments blend into a novel auditory image while others are perceived as different melodies, originating from different voices or instruments. When a composer obeys the laws of contrapuntal composition, the different melodies will not blend, if she prefers a homophonous style of composition, the auditory effect will be of one leading melody which is accompanied by the other instruments or voices. Similar effects, that something originates in a different source, or makes up a different auditory/melodic stream, also govern the way that listeners segment the continuous time stream into chunks of quasi-similar events, e.g. fundamental beats. Although *stream segregation* is mostly concerned with the task of identifying different sound sources, e.g. speakers talking simultaneously, similar cues may be used by listeners to segregate auditory events in time. For the segmentation of fundamental beats, the second principle seems to be fundamental, namely that "Sounds originating in a single source tend to change their frequency characteristics continuously rather than abruptly" (Darwin (1997)). In many sequences of speech, the places showing abrupt spectral change tend to be syllable boundaries, where the spectral characteristics of a voiced vocalic event with clear harmonics is often suddenly followed by a voiceless obstruent, containing noise rather than harmonics. However, our example above shows that such an abrupt change is not necessary for the detection of rhythmic events — here the spectral changes are smooth rather than abrupt, especially in the lower frequency regions.

Zwicker and Fastl (1999) propose a psychoacoustic model aimed to identify fundamental rhythmic beats based on changes in loudness. When looking at the oscillogramme of example 2.4, we see a clear change in the intensity parameter. Thus, a chunking based on both intensity and spectral change may be quite promising. Zwicker and Fastl (1999), however, suggest an identification of rhythm events based on the following simple rules:

- Only events of more than 0.43 relative to the the maximum loudness level are relevant

- The number 0.43 is used relative to the loudness level within a meaningful analysis window, e.g. a musical phrase
- The maxima detected with this analysis method must be separated by more than 120ms to be distinct

Their model indicates the number and location of beats, but does not calculate their beginnings and ends. This implies that beats are points in time rather than events with an inherent duration. Such a view is implausible, however, concerning the fact that in musical notation, beats are described as having duration. Alternatively, a point based view of beats implies their duration extension are limited by the point of occurence of a subsequent beat. When performing a thought experiment where an event identified as a loudness maximum is followed by a pause of an hour before the next beat occurs. It is highly unlikely that in this case, we would perceive the beat as lasting one hour.

In Figure 2.4 we can see that an loudness based approach to beat detection is promising as it identifies syllables in the acoustic stream, but it is clear that the dynamical intensity related aspects which can be regarded as more or less abrupt changes in the sense of Bregman (1990) play a role as well in the perception of beats. Zwicker and Fastl (1999) point out that the perception of an event beginning may differ from its acoustic beginning as a function of its initial tuning curve. The flatter and gradual the tuning curve's slope, the earlier is the beginning of the related event perceived. A preceding click may also influence the point of time, where the beginning of an event is perceived. We will see in the next section, that this phenomenon is also relevant for the detection of rhythmical event onsets in speech.

2.2.2 Detecting Fundamental Beats in Music

Speech is not the only field where researchers have been trying to identify beats based on computational or cognitive models. Musical reasearch has been busy with the development of *tempo* or *beat trackers* for many years. From an application oriented point of view, the usefulness of having good beat trackers is obvious in computational music processing, e.g. in order to adjust tempo changes or automatic musical transcription. It is important to realize that many beat detectors are actually searching for musical *downbeats*, i.e. stressed musical events, like the first quarter note out of three in a Waltz measure, which stands out and usually corresponds to foot tapping. A dancing couple will usually make a larger move on a downbeat and initiate a turn in the ballroom. Downbeats are not the only rhythmically relevant musical events. Other musical events may also correspond to (minor) rhythmical movements. Thus, downbeats are not the fundamental rhythmic units in music. Klapuri et al. (2006) formulate a hierarchy of metrical events that need to be taken into account in (automatic) beat detection (cf. Figure 2.5). They call the most fundamental level of metrical organization the *tatum* or "temporal atom" (Bilmes (1993)), which is the period of the shortest duration in a piece of music that can be encountered incidentally. All other durations

> are integer multiples of the tatum period and the onsets of musical events occur approximately at a tatum beat. (Klapuri et al. (2006))

The metrical level above the tatum is called the *tactus*, it is the level of fundamental rhythmic beats in music. The rate of the tactus defines musical tempo and its metrical level probably roughly corresponds to the syllable or location of *perceptual centers* in speech (cf. Section 2.2.3). The top level of metrical organization in musical rhythm is the *measure* which indicates *downbeats* in music. Measure detection is based on the identification of musical accent and belongs to rhythmical grouping. Other theoretical models for the description of musical rhythm postulate even higher levels of rhythmical grouping such as the musical phrase, e.g. Lerdahl and Jackendoff (1983). But since all of these issues are part of rhythmical grouping, they will be covered elsewhere (cf. Section 2.3).

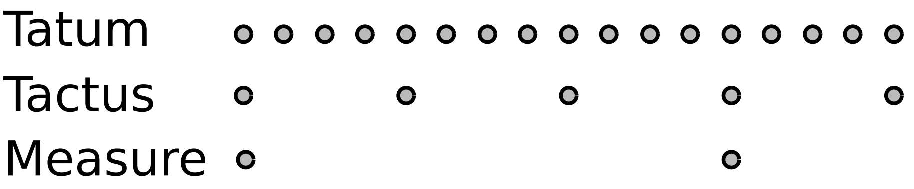


Figure 2.5: The three levels of metrical organization in music.The picture implies graphically, that a tactum, the fundamental beat, always consists of an identical number of tatums. In fact, this is not the case, each tactum beat may correspond to the duration of one tatum or be an integer multiple of it.

Dixon (2001) differentiates three different types of beat detectors: Models that are tracking beats on audio signals directly and those working on symbolic musical data, e.g. MIDI-like information containing information about the onsets of notes. Usually, the systems extracting beats from audio data directly, perform a first step where the symbolic data is extracted. Another distinction can be made between beat trackers working in real time and those performing the task offline. Various computational techniques have been used to detect beats, ranging from neural networks, adaptive oscillators, Bayesian Networks, probabilistic generative models etc. The earliest meter analysis working directly on audio data was developed by Goto and Muraoka (1995) based on mulitple agents tracking competing meter hypotheses. Davies and Plumbley (2005) compare several approaches of tactus detection and come to the conclusion along with Bregman (1990) (cf. Section 2.2.1) that spectral changes are the best indicators of fundamental rhythmic events. Klapuri et al. (2006) criticize models that are based on a symbolic representation level such as MIDI, because they tend to run into problems when finding beats across different musical genres, e.g. classical music vs. rock music. This is because they cannot use the full range of — genre dependent — acoustic cues used to indicate beats. Summing up, a new musical beat correlates with a point of time where the acoustic shape of the sound "changes significantly". Whether or not the amount of change is sufficient to be regarded as the beginning of another beat, is the task of the beat detector. In the following section 2.2.3 it will become evident that this general rule is applicable to speech as well.

A state-of-the-art beat detector model based on acoustic data is presented in Klapuri et al. (2006) (cf. Figure 2.6). As in this book we are mostly interested in speech rhythm, it is more important to know the acoustic cues to fundamental beats — the relationship between a musical score and rhythm is of less interest since it is unclear how such a representation would have to be mapped to speech signals or symbolic phonetic or phonological representations. It can be assumed that cognitive mechanisms detecting beats in music do not rely on completely different cues than those detecting beats in speech. But it is known that musical experience also have an impact on the way that the human brain decoded musical beats (Oerter and Bruhn (1998)). E.g., a person who grew up with Western classical and pop/rock music may have considerable problems detecting the rhythmical variety contained in classical jazz or oriental folk music. Thus, a genre independent beat detector may be somewhat too ambitious to be based only on acoustic features and not taking into account different genre specific types of beat structuring which are engraved in our musical memory and may generate genre specific expectations concerning the acoustic nature and distribution of musical beats. It therefore comes as no surprise that even the successful beat and meter detector described in 2.6 has problems finding beats in Western classical music and jazz reliably.

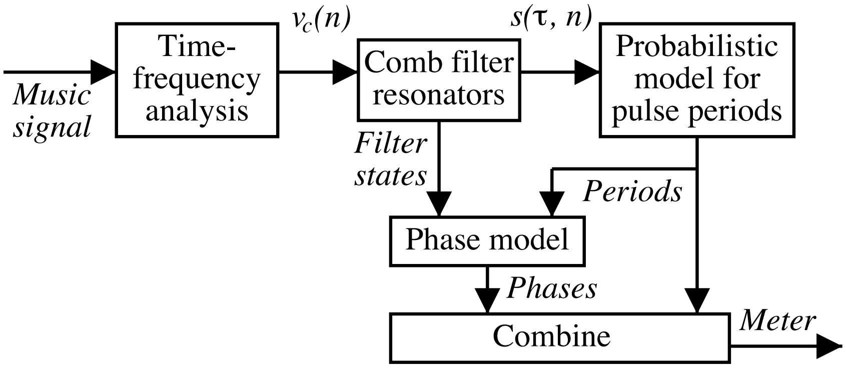


Figure 2.6: The musical beat/meter detector proposed by Klapuri et al. (2006).

2.2.3 Detecting Fundamental Beats in Speech

When trying to connect *beats* in natural speech, most people instantly think of the syllable as the entity that encodes it. In German primary schools, children are often asked to clap their hands in order to determine syllable boundaries, indicating a close relationship between rhythmical beats and syllables. This connection has been verified: Köhlmann (1984) found out that syllables and "clapped" beats indeed correspond to each other. But one has to be cautious equating syllables and perceptual beats, especially with regards to their duration. An extensive body of research reveals, that the physical onset of syllables does not coincide with their perceptual onset, which indicates a nonlinear temporal processing of the syllabic material.

In his early experimental phonetics studies, Meyer (1898, 1903) (cited after Janker (1993)) pointed out that when asked to accompany poetic speech with finger taps, taps are synchronized with the beginnings of vowels rather than with the (consonantal) onsets of the corresponding syllable. His studies were later confirmed by

Allen (1972) who located perceived beats shortly *before* vowel onsets in an experiment where subjects had to synchronize finger taps and clicks with the location of perceived beats in speech. He furthermore detected a connection between the duration of the syllable onset and the location of the tap/click before the start of the syllable rhyme. The general tapping paradigm is illustrated in Figure 2.7.

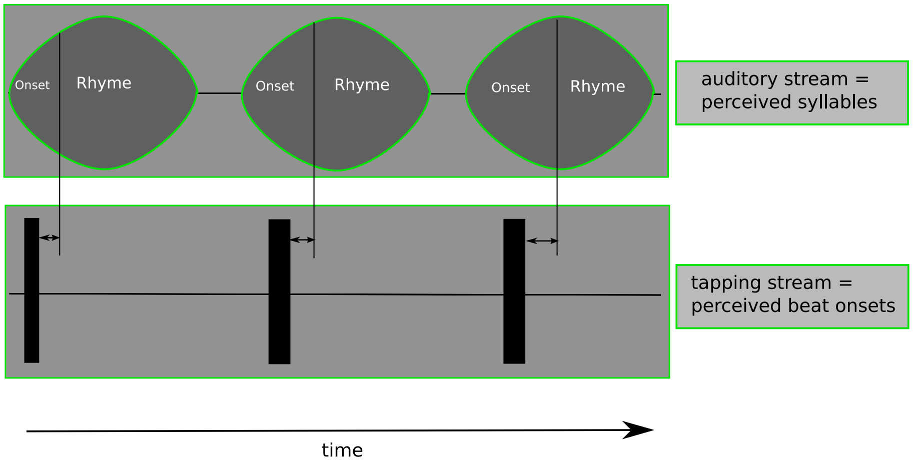

Onset
Rhyme
Onset
Rhyme
Onset
Rhyme
auditory stream = perceived syllables
tapping stream = perceived beat onsets
time

Figure 2.7: The figure illustrates the procedure to use finger taps as an indicator of perceived beat onsets in speech. The subjects are asked to listen to speech or produce speech overtly or silently and perform a finger tap whenever they feel that a beat event is taking place. The tappings are usually performed before the syllable rhyme but do not fall together with the beginning of the syllable. Taps are further away from the syllable rhyme if the syllable onset is longer. In a related experimental paradigm subjects are asked to synchronize clicks with the location of perceptual beats.

In an alternative experimental paradigm aimed to detect beat onsets in speech, subjects are asked to produce or to arrange speech in such a way that the beats are "regular" (cf. Figure 2.8).

Rapp (1971) found that when asked to synchronize the beginning of pseudo words with equally spaced auditive pulses, subjects tend to locate the pulse around the beginning of the vowel. With increasing duration and complexity of the consonantal syllable onset, the "beat onset" moves further away from the vowel beginning into the consonantal onset of the syllable. Later, Morton et al. (1976); Marcus (1981) detected the related effect that subjects perceive words as irregular, when they are presented with interstimulus intervals of identical length. They let the subjects synchronize the words until they were produced at perceptually regular intervals. That way, they determined, what they called the *moment of a word's occurrence* or

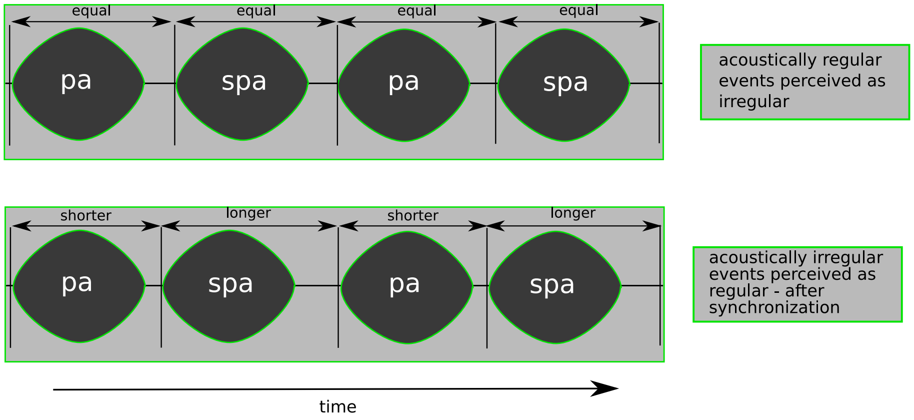

equal
equal
equal
equal
pa
spa
pa
spa
acoustically regular events perceived as irregular
shorter
longer
shorter
longer
pa
spa
pa
spa
acoustically irregular events perceived as regular - after synchronization
time

Figure 2.8: In order to detect the timing regularities of perceived isochrony, subjects are asked to produce words or syllables in such a way that they are perceived regularly. This experimental paradigm is based on the detection that subjects perceive syllables presented at isochronous interstimulus intervals as being irregular. There is a systematic deviation from the objectively measurable regularity to achieve perceived regularity.

p(erceptual) center.

> This P-center is equidistant from the surrounding P-centers of other words. For example, if words are set in time to a metronome, their P-centers would align with the metronome beat since the beats are equidistant from each other. (Pérez, 1997, 12)

According to the model defined in Marcus (1981) (cf. equation 2.4), a beginning of a p-center beginning relative to the syllable onset can be determined as a linear function of the syllable onset duration (C), rhyme duration (VC) and a constant ($const$).

$$P = 0.65 \times C + 0.25 \times VC + const. \tag{2.4}$$

In his alternative approach to p-center location, Howell (1984, 1988) postulates that the perceptual onset of events correlates with the point of maximal variation in the amplitude envelope of an aoustic event thus claiming that *change* causes perceptual beat detection and introduces the term *center of gravity* rather than p-center. His approach shows a resemblance to the psychoacoustic approach by Zwicker and Fastl (1999) that relied on loudness variation. In the psychoacoustically more refined model developed by Pompino-Marschall (1989, 1990), p-centers are calculated based on the onset and offset events for the rising and falling flanks of the corresponding syllable's loudness contour within single critical bands[7]. In related studies, Cooper et al. (1986) showed that the location of p-centers is not influenced by the segmental structure and Janker and Pompino-Marschall (1991) showed its independence of the fundamental frequency contour.

In an extensive evaluation of the various approaches to p-center location, Janker (1993) correlates finger tappings and p-center estimates and concludes, that the best approximation to p-center location is the vowel onset which also coincides with the point of maximal variation in the speech signal. However, due to the different processing of motor planning and auditory processing (cf. Section 98), finger tappings are no valid indicators of locating p-centers. To overcome the shortcomings of his earlier study, Janker (1996) therefore normalized his material for the motor planning and found that all models lead to a high correlation between tappings and predicted p-center location, with Pompino-Marschall's model being most precise. The models by Morton et al. (1976); Marcus (1981) and Howell (1984) appeared to be somewhat biassed in the direction of the vowel onset. It is unclear, though, whether this is explicable as a shortcoming of the models themselves or a simple result of the normalization method used to account for the distortions introduced by the tapping paradigm. Overall, the high correlations indicate that the various models are all trying to grasp similar things which happen to be predictable by different acoustic or phonetic parameters such as loudness, amplitude variation and consonantal-vocalic transitions.

One problematic issue of the entire p-center debate appears to be the question whether p-centers are regarded as determining the onset of a beat or whether they should aid to determine the beat's temporal extension as well. In the tapping paradigm, the answer is clear: The tapping serves as an indicator of the beginning of the perceptual event and the temporal extension of the corresponding beat may remain unspecific as in Allen (1972), who suggests the beat to be within a window of 200ms around the tapping location. On the contrary, in the synchronization task where words have to be arranged in an order of perceived or produced regularity,

[7] A critical band is a frequency selective channel of psychoacoustic processing. Noise that falls within the critical bandwidth can mask a narrow band signal. Our auditory systems contain a series of critical bands, each filtering out a specific part of the acoustic signal

it is very likely that the task is influenced by the duration of the entire word or syllable that needs to be arranged. Therefore, we feel that the criticism by Scott (1993) claiming that a location of p-centers based on the amplitude, loudness or segmental characteristics of the entire syllable rather than its beginning are inadequate, is only true for one of these paradigms. Scott proposed to detect p-centers with the help of determining the maximally rapid amplitude rise in frequency band of the first formant. An alternative suggestion concerning the regularity of p-centers has been made by Tuller and Fowler (1980); Fowler (1983) who claimed that the perceived regularity is not necessarily to be found in the acoustic signal but may be detected in the kinematic signal instead. Fowler (1983) suggests that the beginning of the vocalic gestures are timed at regular intervals when speakers are asked to produce speech at regular intervals.

In a more recent study by Patel et al. (1999), a database was recorded and examined in order to test the hypotheses by Scott and Fowler. Their recorded material consisted of productions of nonsense syllables with variable phonotactic complexity which had to be produced by subjects as regularly as possible — after a rehearsal phase with a metronome. The syllable streams consisted of alternating sequences of the syllable *ba* followed by syllables of different phonotactic complexity, i.e. the subject produced sequences like *"ba la ba la..."*, *"ba spa ba spa..."* etc. plus a reference consisting of identical syllables *"ba ba ba ba..."*. In addition to acoustic recordings, they recorded kinematic data with the help of an electromagnetic midsagittal articulometer. Their results confirm once more that timing is influenced by the syllabic onset complexity and that speakers tend to start their utterances earlier, the more complex the target syllable is going to be in order to uphold the impression of regularity. At first glance, their kinematic material shows a tendency towards regularity of underlying gestures which cannot be statistically confirmed. The amplitude slope maxima of the first formant and fundamental frequency do not show a clear tendency to be timed at regular intervals.

However, the search for objective "regularity" with respect to p-centers is apparently an approach that is doomed to fail — it is very unlikely to find a 1:1 correlate for a phenomenon measured in a different domain of phonetic realization. The aim of the former models was to define a function that maps an acoustic realisation onto the perceptual regularities — and from Janker's results we know that all models are more or less successful, especially when taking into account psychoacoustic processes. Concerning the question of the point of time where our brain starts perceiving a beat, it is of course not useful to base a model on the realisation of the entire syllable. But even the simple prediction model by Marcus (1981) could show the main effect that an increase in onset duration leads to a perception of the "beat" before the vowel. From a point of view that takes into account attention and expectancy, we know that a significant change in the acoustic input signal increases our attention and may start a new window of temporal integration (cf. page 23). An abrupt change in the acoustic signal also introduces an event perceived as autonomous according to stream segretation theory. Such an event is the beginning of the consonant, before the vowel has started. From an expectancy point of view, the syllable onset marks the point of time where a listener starts to prepare for the "beat event". While starting a new window of temporal integration, the listener expects a "beat", which may be usually identified with the syllabic rhyme. With time passing, we simply expect the beat to take place and our expectation overrides the reality of the rhyme event that has not yet begun. We therefore assume a rhyme to take place after enough time has elapsed after the beginning of a new window of temporal integration. That way, we are not forced to experience only vowel-like events as beats, but also noise events can be rhythmic — a phenomenon known from various percussion instruments relying on noise, e.g. rattles or from languages allowing fricatives as syllable rhymes, e.g Czech.

It remains unclear how the p-center phenomenon can be applied to running speech without pauses in between syllables, since our auditory system is inhibited by the backward masking effects of previous syllables, i.e. loud auditory events persist in the auditory perception *after* they ended physically, thus masking subsequent auditory events (cf. (Zwicker, 1982, 151f.)). Thus, a following event may appear shorter than it objectively is. It is unclear how we have to integrate such effects in a model of p-center detection.

Summing up, despite the difficulty of its proper detection, we provide the following definition:

Definition 3
Fundamental beats in speech are cover time spans. Their beginnings are located at the perceptually estimated or expected beginning of a syllable rhyme (a perceptual

center) after a new syllable event has started. Their end is limited by the syllable boundary.

One of the shortcomings of the p-center model is the question how the phenomenon is linked to known levels of prosodic organization other than the syllable, e.g. the prosodic word, the foot or the mora? The original observations have been made in relation to stressed monosyllabic words or nonsense syllables and few studies are concerned with the relationship between p-centers and the prosodic hierarchy. Most works found a strong relationship between p-center organization and the phonotactic structure of the syllable, thus implying that the syllable is the level of prosodic organization that creates the impression of *beats*. However, it seems possible that in a bisyllabic word of a stress timed language such as English, not every syllable, e.g. when it is reduced, qualifies as a beat. I.e. the clapping based "beat detection rule" used by primary school teachers relies on the fact that in German, every syllable *may be* but does not *have to be* stressed.[8] It is therefore unclear, in what way the p-center model needs to incorporate reference to stress or accent. So far, no interaction between f0 or accent type were found (Patel et al. (1999); Janker and Pompino-Marschall (1991)). Pérez (1997) argues for it being likely that p-center effects are influenced by other, language specific prosodic levels which have an impact on timing organization. She examined the influence of metrical foot structure on p-centers in disyllabic words of American English and found that stressed syllables in trochees are nearer to the previous syllable than stressed syllables in iambs. The relation between morae and p-centers is somewhat difficult to judge. However, looking back at the discussion on beat perception in music, we know that a rhythmical level of organization below the beat has been introduced as well, called the *tatum* (cf. section 2.2.2). The musical tatum corresponds to the minimal duration span between two rhythmical beats. Likewise, the mora corresponds to the minimal phonological duration span in between two perceptual centers in speech, since a syllable must consist of at least one mora. The exact interrelations between both levels of organization cannot be solved here, but theoretical parallels in musical and linguistic theory are most likely not coincidental. The connection between mora, syllable and beat strength has been analysed in detail by phonologists and the main

[8] Clearly, some syllables in German are rarely stressed, e.g. those containing ə or a syllabic consonant. However, even these syllables may be stressed, e.g. in a correction.

insights will be reviewed in the following section 2.2.4.

2.2.4 Defining Fundamental Beats in Linguistic Models

In the two previous sections, the detection of *beats* in music and speech perception was examined. Research of both fields has implications — or rather parallels — in phonological descriptions of language rhythm. For music, the issue of introducing a hierarchy became apparent, since it was felt to differentiate between beats of different strength or weight. This lead to the introduction of different descriptive levels, the *tatum* and the *tactum*. While the tactum corresponds to the musical *beat*, the tatum describes the minimal duration or "strength" in a musical piece. Each tactum must be an integer multiplicative of the basic tatum. In phonological descriptions, a parallel distinction can be found between the notion of the syllable and the mora, the former providing the structural frame for a linguistic beat, the latter providing its strength or — in linguistic terminology — its *weight*. As the tatum defines the minimal strength or duration of a musical beat, the mora defines a syllable's minimal *phonological weight* (Hyman (1985)). Depending on the language and the syllable structure, a syllable may contain of one, two or three morae (cf. Figure 2.9), while in music, there exists much more freedom concerning the number of tatums contained in a beat. Within the mora-based approach to language rhythm, each mora is assumed to contribute equally to a syllable's weight or strength. Therefore, just like the musical tactus duration being a multiplicative integer of the basic tatum, a syllable's strength is an integer multiplicative of the basic mora. In so-called mora-timing languages like Japanese, this relationship is phonetically expressed in the domain of perceived duration, i.e. two morae are roughly perceived as twice as long as one mora. In a language that is not mora-timed, the strength may be phonetically expressed in a different manner.

A further parallel between phonological descriptions and findings in perceptual phonetics concerns the relationship between syllable structure and the notion of the p-center. In section 2.2.3 it was described that the perceptual onset of a beat in speech is closely linked to the onset of the syllable's vowel. Phonological analyses came to the conclusion that a syllable's *weight* is largely determined by its rhyme, i.e. the syllable nucleus, usually consisting of the vowels plus the syllable coda,

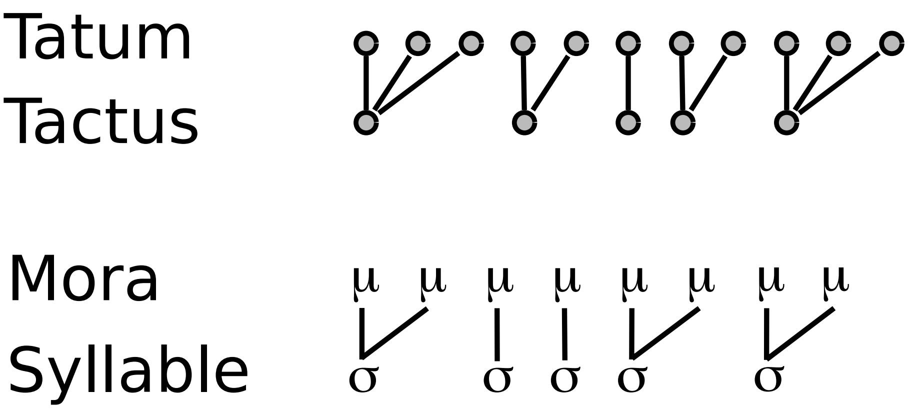


Figure 2.9: The distinction between tatum and tactus in musical rhythm is analogous to the one between mora and syllable in language rhythm. Mora and tatum define the minimum quantity units possible, while the tactus and syllable define the beat. The number of tatums or mora contained in a beat defines the beat's strength or weight. Naturally, music has more degrees of freedom concerning the number of tatums contained in a beat, while in language, the number of morae contained in a syllable is —language dependently— restricted to one, two or three.

consisting of the final consonant(s) (cf. Figure 2.10, van der Hulst (1984); Selkirk (1982)). The notion of *syllable weight* was introduced into phonological models for several reasons. . In so-called quantity-sensitive languages, syllables having comparatively much weight attract phonological accents. Besides, syllable structure constrains type, number and order of the phonological segments contained in it, e.g. the syllable nucleus consists of the segment(s) with the highest sonority, i.e. vowels or syllabified consonants like nasals or laterals. Clements (1990) defends the position that a segment's level of sonority goes hand in hand with its perceptual prominence. This connection between phonological sonority and perceptual prominence fits in nicely with the p-center concept introduced earlier, since it is useful to assume that the perceptual onset of a beat is perceptually salient. Due to the fact that the syllable rhyme determines the perceptual weight and marks the beginning of its most salient part, phonologists introduced a hierarchical structure to a syllable with a dominant syllable rhyme and nucleus (cf. Figure 2.10). This concept again goes hand in hand with another phonological prosodic unit introduced earlier, the mora, since the number of morae in a syllable is defined by the structure of the syllable rhyme (cf. Figure 2.11). Usually, an open syllable with a short vowel is regarded as a *light syllable* while a syllable with a long vowel is regarded as heavy.

Closed syllables can be either light or heavy, depending on language specific constraints, i.e. in some languages, closed syllables may be heavy, in other languages, their weight may depend only on vowel quantity and the coda may be irrelevant for syllable weight. In some languages, even "superheavy" syllables do occur. A light syllable corresponds to one mora, a heavy one to two, a superheavy one to three morae. However, the lacking impact of the onset on syllable weight appears to be a universal phenomenon[9]. The segment's sonority seems to have an impact on the number of morae contained in a syllable as well: Zec (1995) showed that in some languages, short-vowel syllables closed by sonorants are heavy, while short-vowel syllables closed by obstruents are light. It is obvious that all of the syllable related effects mentioned above have an impact on the respective languages' rhythm.

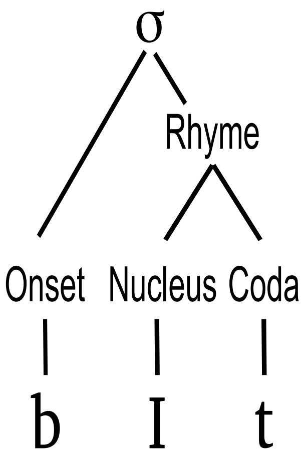


Figure 2.10: In standard phonological approaches, a syllable is hierachically structured, consisting of the *onset*, the initial consonant(s), a *nucleus*, consisting of the vowel or most sonorous segment and the *coda*, consisting of the final consonant(s). Nucleus and coda comprise the syllable *rhyme* (or *rime*) which determines the syllable weight.

The insight that the syllable makes up a phonologically relevant and independent level of phonological descriptions has —among other reasons— lead to the development of the so-called Nonlinear Phonologies in the 1980s. The most relevant phonological theory for the description of rhythmic phenomena has become the theory of (Autosegmental-)Metrical Phonology, e.g. Liberman and Prince (1977),

[9] However, see Mengel (2000) for a phonetic study that found a small effect of onset complexity on perceptual prominence in German.

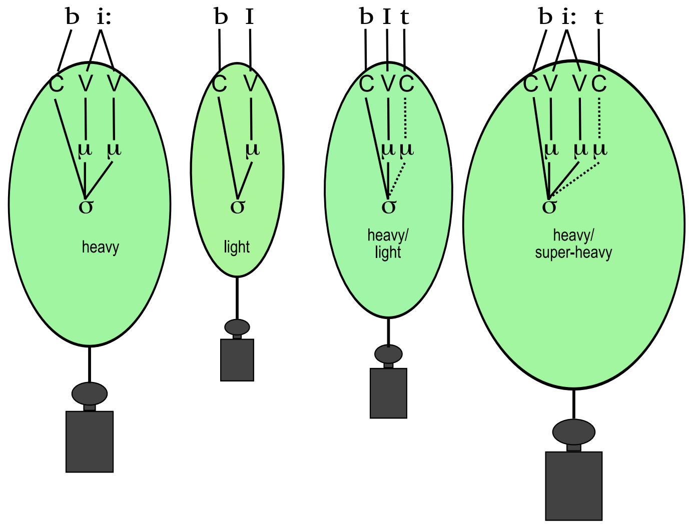


Figure 2.11: Syllable structure defines whether a syllable is light, heavy or even superheavy. The connections between syllable structure and syllable weight are language dependent, but syllable weight and number of morae contained in a syllable stand in a linear relationship. I.e. a light syllable comprises one mora, a heavy one two and a superheavy one three morae.

Selkirk (1984), Hayes (1994). In these phonological theories, or sometimes rather, descriptive models, the various descriptive levels, such as syllable, segments, tones etc. are notated on more or less independent *tiers*. Between those tiers exist *association lines* indicating temporal overlap in the phonetic realization, e.g. segments that are associated with a syllable temporally co-occur. If association lines are deleted or changed, phonological processes like reduction, elision etc. can be modelled. In nonlinear descriptions, the mora plays an important role as well. Moraic sequences form the lowest tier within the *prosodic hierarchy*. There exists no one-to-one-mapping between different tiers, i.e. a mora may comprise one or more segments and one segment may be even associated with two morae (Cohn (2003)). It is important to understand that the moraic structure is independent of the so-called *skeletal tier*, which organizes the linear order of segments. We have discussed above (cf. section 2.1.2), that the perception of linear order does not imply a cognitive estimation of the pertaining durations. Thus, the cognitive processing of segmental order — illustrated by the skeletal tier in phonological models — is largely independent of the estimation of phonological durations or *phonological weight*. The phonological models imply that the mora is the smallest meaningful quantitative unit in speech.

In line with this reasoning, Goedemans (1998) argues that syllable weight (or number of morae) is a better phonological equivalent of *perceptual* syllable duration than the number of segments, based on findings that the JND is much higher in onset segments than in rhyme segments.

One could argue that the assumption of the mora as minimal timing unit does challenge the idea of equating p-centers and syllable rhymes as fundamental psycholinguistic rhythmical entitites in speech. Such a view receives further support from so-called mora timing languages (c.f. section 1.1). We know from these languages (e.g. Japanese) that their entire poetry is organized based on morae. However, since the p-center is defined relative to the syllable, the mora does not have a 1:1-correspondance to our previously defined fundamental beats: One syllable may consist of one or two morae. Obviously, a speaker of Japanese may "count" two fundamental rhythmical entities (=beats?) when listening to a bimoraic syllable.

This superficial contradiction can be solved if we take into account the notion of the *tatum*. Language specifically, listeners inner clocks "tick" more fine-grained in listeners of mora timed languages (also see Sagisaka (1999), quoted after Warner and Arai (2001a)). Thus, their windows of temporal resolution may be more fine-grained as well and react more sensitive if a syllable is significantly longer than normally the case[10]. Thus, a listener of a mora-timed language may start a new window of temporal integration during a bimoraic syllable, thus counting one more fundamental event without this being a beat in the traditional sense. Findings by Kato et al. (1997) suggest that the psychoacoustic boundary between vocalic and consonantal portions across morae play the most important role for opening such a new window of temporal integration, e.g. the area of strongest syllable internal spectral change. Thus, an alternative approach towards explaining the relationship between morae and beats/syllables might be that in mora timed languages, beats are not restricted to syllable rhymes, but can also be realized as consonants. The syllable itself may simply play a less central role as a perceptual unit, but remains to have an impact on segmental durations (Campbell (1999); Warner and Arai (2001b)).

[10] For further evidence concerning this point, refer to section 2.3.1. Here, it will be shown that listener's temporal resolution becomed more fine grained the longer they listen to a sequence of stimuli of equal length. Thus, a language which has less variability due to less phonotactic complexity make listeners more sensitive of temporal distinctions.

It is likely that the perception of morae as fundamental rhythmical units are explicable by a combinatory effect of the higher perceptual temporal resolution and the classification of syllable final consonants as additional minimal rhythmical units or additional rhymes.

Contrary to mora timed languages, listener of a syllable or stress timed language may have more flexible windows of temporal integration. Thus, listeners of such languages will perceive the bimoraic syllable as being longer, more prominent and having more phonological weight, but will not distribute the bimoraic syllable across two windows of temporal integration. The listener perceives its weight, but does not count this weight as two separate temporal entities. In musical terminology, we might say that a listener of a stress timed language perceived a quarter note where the listener of the mora timed language perceived two eighth notes. Summing up, we conclude that in mora timed languages, the windows of temporal integration are less flexible. This can be explained with the relatively uniform syllable structure exposing listeners to more similar and overall shorter syllable durations (Dauer (1983)).

Still, speakers of stress timed languages may use syllable weight for rhythmical structuring, thus making an indirect use of the mora. In these languages, moraic weight may be a good indicator to locate rhythmical accents forming rhythmical groups (cf. section 2.3.2).

It is likely that the connection between perceptual processing and the different levels of phonological structure behaves differently, depending on the respective language.

2.2.5 Implications for Rhythm Research

It seems that the detection of fundamental beats in music and speech has a lot in common and is mainly driven by abrupt spectral and intensity changes in the signal. In speech, the beginning of a beat co-occurs with the onset of the vowel or syllable nucleus. The onset consonants do not contribute to the perception of a beat and its pertinent strength. This is in accordance with phonological models that also regard the syllable rhyme as the determiner of a syllable's weight, a measure which expresses the perceptual strength or duration, i.e. the strength of a linguistic beat.

Syllable weight, however, is not derived from the segmental level but is better expressed in terms of the number of morae contained in a syllable. The mora describes the minimal unit of perceived rhythmically meaningful duration. The number of morae contained in a syllable is language dependent and determined by vocalic and consonantal quantity and sonority. Altogether, it seems that the interface between moraic and p-center descriptions are a good starting point to determine a rhythmical interface between phonetic and phonological descriptions of fundamental rhythmical beats. There are indications that the detection of p-centers is not solely determined by syllabic structure but also — at least to some extent — by higher level organization, e.g. the question whether a beat is embedded in an iambic or trochaic pattern. Thus, the p-center perception may be also influenced by top down processing based on linguistic expectancies.

2.3 Grouping Beats

In our initial working definition of rhythm given in section 1.2.5, it was stated that a rhythmic pattern emerges via the segmentation of a continuous string into fundamental rhythmic events which are then grouped in such a way that the groups are in some way similarly structured. It was claimed that a similarity in structure may be derived from the number of rhythmical segments contained in a group or the pattern of rhythmically stronger and weaker events. In this section, several approaches are discussed how rhythmical groups are identified by listeners and can be characterized in different descriptive domains such as psychology, acoustics, phonology and music. A pressing question is the one why listeners aim to group events at all. We follow the idea (in line with Povel and Essens (1985); Handel (1992)) that listeners are trying to match their perceptions with their expectations based on their previous listening experiences, e.g. they are trying to maximize rhythmical harmony. This strategy is useful because it helps the listener to predict which kind of rhythmical event is going to follow, e.g. after having listened to several unstressed syllables, the probability rises that something stressed (and possibly important) is going to be uttered. Handel (1993) supposes two generally different ways of perceiving rhythmical groups: Either based on "global metrical templates" for fast rates or by accurate timing expectancies based on previous experience (for slow rates with onset-to-onset intervals >250ms). A maximisation of rhythmical harmony, where metrical expectation and perception are in perfect agreement, is often found in musical pieces and poetry. The templates or local expectations are based on relatively small sequences — or groups — of subsequent events, e.g. sequences of two or three fundamental beats and it is likely that these are constrained by the so-called *window of temporal presence* introduced in Section 2.1.2.

2.3.1 Psychological Principles of Grouping

A phenomenon already noticed in early listening tests (Bolton (1894); Wundt (1911)) is that listeners impose a rhythmical grouping pattern on sequences of acoustic stimuli even if these are — acoustically — identical, i.e. they have the same duration, frequency, intensity and interval duration between successive events. Such stimulus sequences are referred to as *isochronous pulse trains*. In section 1.2.5 we defined rhythmical structuring as a process of grouping sucessive rhythmical events into similar sequences. In this section, we propose that similarity can be achieved either by grouping *an equal number of auditory events* or by *creating similar patterns of stronger and weaker events* in each sucessive group.

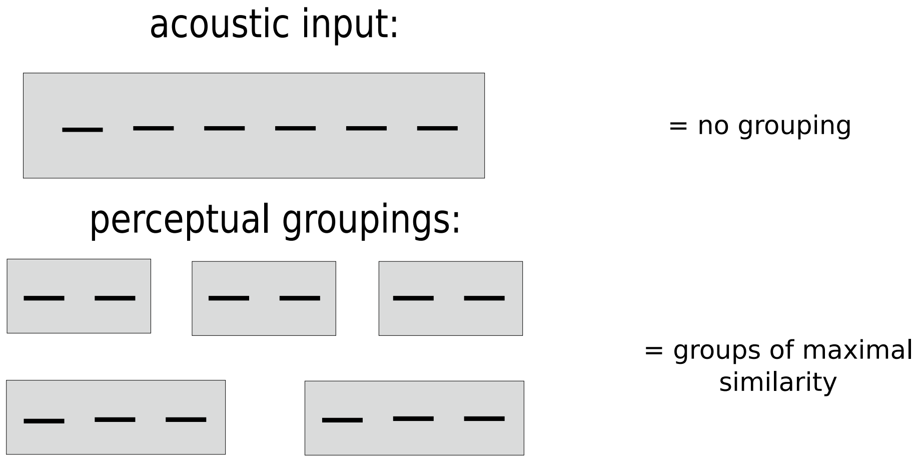


Figure 2.12: Perceptual groupings of listeners when perceiving sequences of identical stimuli, so-called *isochronous pulse trains*.

When listening to a sequence of isochronous intervals, listeners tend to group them into pairs or triples, thus perceiving each second or third stimulus as more prominent (Handel (1989); see Figure 2.12). This undermines the fact that the perception of rhythm is very much guided by top-down expectancies or perceptual conditions that may override the physical reality. Listeners obviously feel a need to impose a structure on what they hear. The reason for this can be manifold, likely explanations are to be found in the cognitive processing of auditory events, e.g. the temporal integration of auditory events into windows of a temporal present (cf. section 2.1.2). This means, that a chain of events may be perceived as one group due to the fact that the events contained in it belong to the same window of temporal presence. Schreuder (2006) replicated this effect for Dutch showing a different rhythmical structuring of listeners at different speaking rates — in fast speech, more syllables were grouped together than in slow speech, where listeners perceived more stressed beats. She explained this phenomenon with the limitations of the window of temporal presence. The idea that rhythmical structuring is the effect of grouping chunks of auditory percepts into windows of temporal presence receives further support by the findings that this grouping does not take place if the interval sequences are too long, too short or the pauses between them are very long (e.g. Pöppel (1994); Fraisse (1982)). Therefore, this type of grouping process is constrained by absolute time rather than the relative durations of the stimuli themselves. Gestalt psychology refers to this effect as being the result of the Gestalt law of *proximity*. We call this effect the *principle of temporal presence* (see Figure 2.13). We

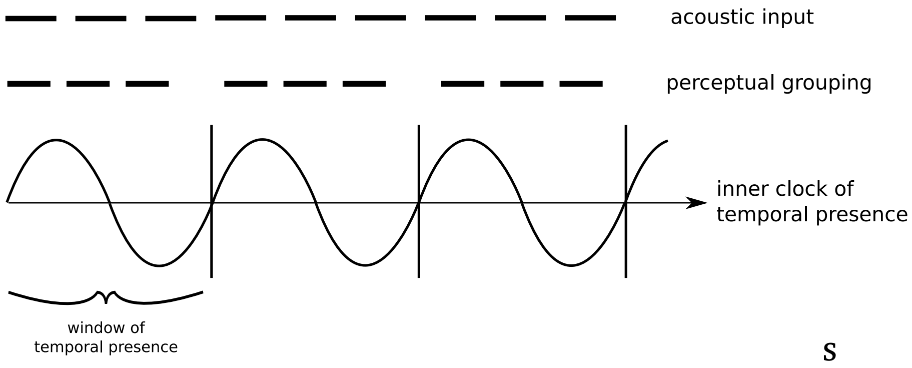


Figure 2.13: An illustration of grouping principles 1 and 2. Listeners tend to collect groups within single windows of temporal presence. In the ideal case, successive groups contain an identical number of elements to which the inner clock is adjusted. In the example above, each window contains three rhythmical beats.

therefore conclude that grouping can emerge through cognitive processing strategies and by the result of integrating events into windows of temporal presence. There exists a growing body of evidence that such windows of temporal presence are approximately 400-600ms long and coincide with typical durations of linguistic organisation in various languages and speaking styles. Fant and Kruckenberg (1996) found for Swedish rhythmical organization a reference quantum of 500ms and Schreuder (2006) showed for Dutch fast speech, that rhythmical groups contain more syllables due to window length. Thus, there seems to be a tendency to preceive a grouping of temporal events within a window of similar length and the temporal organization of speech processing tends to adjust to this window of human temporal processing.

Besides integrating auditory events into groups of similar absolute duration, listeners obviously feel the need to make the group internal structure as similar as possible, since listeners tend to group an identical number of events, e.g. either two or three events into successive groups. Thus, one strategy of creating rhythmical harmony is called the *principle of similarity in number* (see Figure 2.13). This principle explains the poetic technique of keeping the number of fundamental rhythmical events — beats — equal in each sucessive foot or verse. The number of fundamental beats contained in each upcoming window matches our rhythmical expectations.

However, rhythmical harmony is not only governed by grouping an identical number of events, but also by grouping events which show other types of structural similarity. This may be important if the duration of subsequent events are too dissimilar to group identical numbers of them into windows of temporal integration, so that other types of harmony need to be found. It has been shown that listeners tend to perceive successive events as equal in length, which leads to the phenomenon that objectively long events are perceived as shorter if preceded by short events, thus they appear more isochronous in perception. The most interesting aspect of this psychoacoustic finding known as *time shrinking* probably is, that is can propagate across several events, e.g. an antepenultimate event can influence the perceived duration of a final event in a sequence of three (Sasaki et al. (2002), cf. Figure 2.14). Time-shrinking can be blocked, however, in the presence of strict alternation of intervals. Thus, listeners not only feel the need to impose structure on equal successive events but also try to equalize the characteristics of events within a group, if there is a tendency towards deceleration. We calls this perceptual mechanism the *principle of local similarity* (see Figure 2.15).

However, learning may change this effect, i.e. listeners are becoming more sensitive to anisochrony after listening to sequences of several isochronous stimuli (Schulze (1989)). We can conclude from this that there exists a tendency to perceive sucessive events as being rather similar unless we learnt to expect a different length based on recent experiences. In such cases we are very sensitive of anisochrony (*principle of local dissimilarity*, see Figure 2.15).

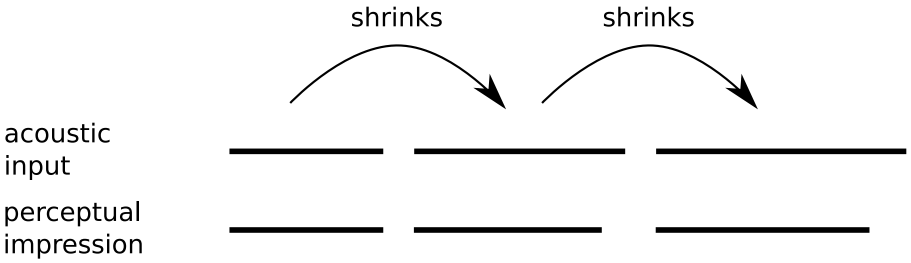


Figure 2.14: An illustration of the psychoacoustic phenomenon known as time-shrinking. A short preceding event can make a subsequent event appear as shorter. Thus, both events appear as more similar in temporal extension. The effect can propagate across several intervals, i.e. a perceptually shrunken interval may itself cause time shrinking.

Local anisochronies are probably the most well-know method of rhythmical grouping: A group is created by patterns of longer and shorter intervals, e.g. *(long—short—short) (long—short—short)* or *(long—short)(long—short)*. Rhythmical harmony is then constructed by structural similarities of long and short events across groups. The rhythmical patterns may lead to the perception of rhythmical groups even if the group internal timing relations are not objectively identical. Handel (1992, 1993) found that listeners have problems hearing differences between groups if their relative timing patterns were identical on an ordinal scale, i.e. subsequent groups consisted of identical patterns of longer and shorter events but varying absolute relative durations. However, grouping seemed to work differently for different tempos: In slow rhythms, groups were identified based on similarities of the initial element, while in fast rhythms, groups were built based on a similarity of final elements. If a rhythm is perceived because the different groups show strong local anisochrony but a similar structural pattern, the rhythmical group is built based on the *principle*

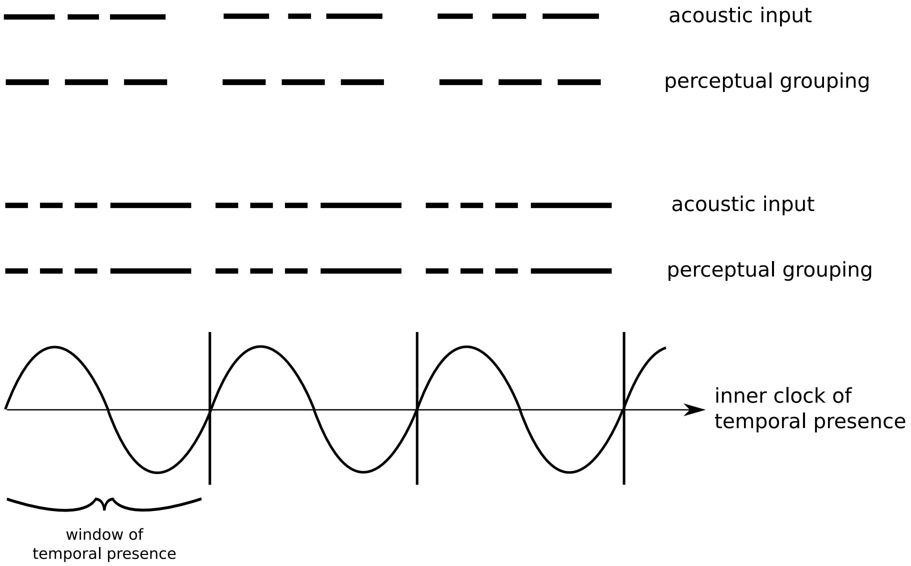


Figure 2.15: An illustration of grouping principles 3 and 4. Listeners tend to equalize the perceived duration of rhythmical beats within each group, unless they adjusted to a particular duration in sequences of isochronous events. In such cases, they are more sensitive towards differences from the trained reference duration and perceive it as anisochrony .

of global similarity (see Figure 2.16). It is likely that our inner clock defining the duration of temporal presence is continually adjusted with the help of such indices of group boundaries.

Rhythmical grouping involves several mechanisms:

1. **The principle of temporal presence:** Groups tend to be built within identical windows of temporal integration or temporal presence. Thus, the inner clock defines the upper limit of such groups and will usually operate around 400-600ms.

2. **The principle of similarity in number:** Listeners prefer to have groups containing identical number of fundamental rhythmical events.

3. **The principle of local similarity:** Listeners tend to equalize the perceived duration of sucessive events, given constant variability, they are less sensitive to it.

4. **The principle of local dissimilarity:** If listeners are trained to a particular isochronous duration in successive events, they become very sensitive towards

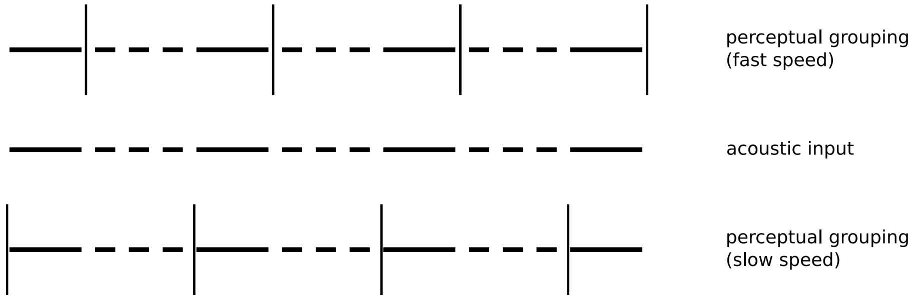


Figure 2.16: An illustration of the grouping principle 5. Based on the perception of anisochronies, groups are built showing the same distribution of anisochronous beats. At rather slow speeds, the anisochronous beat tends to be aligned with the beginning of a group, at fast speeds with its end.

anisochronies. This means, that the more similar sequences of fundamental events are, the better is any upcoming dissimilarity perceived.

5. **The principle of global similarity:** If anisochronies are perceived they are a major cue to indicate the beginnings or ends of a group. Tempo defines wether anisochronies are perceived as beginnings (slow tempo) or endings (fast tempo). Global similarity is created by patterns of stronger (longer) and weaker (shorter) events in identical positions across adjacent groups.

Thus, we have two fundamentally different strategies to structure rhythmical groups: Rhythmical structure is built by the number of fundamental beats contained in them and by their pertinent pattern of relatively isochronous and anisochronous events. If a relatively harmonious sequence of groups has been processed, a listener may formulate hypotheses concerning the rhythmical structure of upcoming events. These hypotheses are abstract generalizations and can be called *meter*. In speech, a perfectly regular meter is rare and only likely in stylized speech such as poetry, sermons or rhetorical figures. In music, however, meter is usually essential and normally obeyed by the performed rhythm within the limits of a musician's interpretative freedom. We close this section with the following definition:

Definition 4

Meter is the abstract rhythmical generalization based on previous rhythmical experience. Meter leads to hypotheses concerning internal structure of upcoming rhythmical groups. Depending on context, metrical hypotheses can be very precise (e.g.

when listening to a march showing a high degree of rhythmical uniformity) or relatively weak (e.g. when listening to spontaneous speech containing a high degree of rhythmical variability).

2.3.2 Creating Groups via Accentuation

It is clear that in groups where anisochrony is involved, the longer events stand out in a particular way. These events are often called *accented* or *prominent*. In much psychological literature, the only dimension of accentuation examined is duration. However, the effect of accentuation can be achieved in different ways. Clarke (1999)[11]differentiates three different types of (musical) accents:

1. **Phenomenal accents**, caused by a local physical intensification, e.g. via an increase of duration, intensity, pitch excursion or a significant change of timbre.

2. **Structural accents**, indicating the endings of a rhythmical event sequence, typically involving a deceleration and a characteristic decline of pitch. In music, structural accents are called *cadence*, in phonetics and phonology, their equivalents would be *boundary tones and final lengthening phenomena*

3. **Metrical accents**, caused by their position in a group of rhythmical events. Metrical accents are governed by listeners expectancies based on previous experience. After having listened to a waltz for a while, we may experience the first note of a tactus as being accented by its position rather than by an objective lengthening or intensification.

It is a common claim (e.g. Cooper and Meyer (1960); Allen (1975)) that different acoustic means of accentuation have an effect on grouping, e.g. if in a sequence of acoustic stimuli every second stimulus is accentuated by an increase in fundamental frequency or intensity, the accented beats are perceived as the beginnings of a group. However, if the accents are created by an increase of duration, the thus accented beats are perceived as being the endings of each group . Thus, grouping is not only determined by tempo (cf. Section 2.3.1) but also by the individual accent type.

[11] Clarke's definition is heavily influenced by Lerdahl and Jackendoff (1983).

It is unclear, however, how these strategies can be related to speech, since here duration signals both beginnings of a group, e.g. by lengthening the first syllable in a foot, or ends of a group, e.g. by final lengthening[12] Final lengthening is often stronger than accentual lengthening. An increase in spectral intensity usually correlates with perceived accentuation and is thus in accordance with the previously mentioned claims, while f0-movements can indicate both indicate group beginnings (as pitch accents) and ends (as boundary tones), but tend to be accompanied by an increase in duration as well. From these facts several hypotheses were derived which were subsequently tested in a small pilot study:

- A moderate increase in duration is interpreted as the beginning of a group, a strong increase in duration is interpreted as the end of a group.
- An increase in intensity is interpreted as the beginning of a group.
- An increase in fundamental frequency does not have a clear impact on grouping.

For the perception task, 4 beat patterns were created all of which were alternating in one acoustic property. The basic event was a 220 Hz, 200ms, 60dB sine wave. The pause between successive stimuli was kept constant at 50ms. One stimulus (intended trochaic) varied the duration between sucessive stimuli in a ratio 1.5:1, one stimulus (intended iambic) varied the duration between sucessive stimuli at a ratio 2:1, a third stimulus (intended trochaic) varied the intensity between sucessive intervals at 10dB, thus every second beat was half as loud as the next one, a last stimulus chain alternated in f0 (440Hz:220Hz = 1 octave). 10 musically trained and 2 phonetically trained native speakers of German were asked to listen to each pattern as long as necessary in order to identify it as either iambic or trochaic. They were allowed to say that the rhythmic pattern was unidentifyable. The results (cf. Figure 2.17) clearly confirm the findings that relative durations can differentiate between an iambic and a trochaic grouping. They also confirm the interpretation of intensity increases as trochaic, while subjects had many difficulties to identify a rhythm on

[12] These grouping principles were further confirmed in poetic speech by Bröggelwirth (2007), who found that German iambs are characterized by a stronger increase in duration on the stressed syllable than trochees or dactyls.

the basis of fundamental frequency. Anecdotal responses from the musicians can be interpreted in such a way that they are mainly guided by whatever instrument they play: Bass players tend to align low fundamental frequency with group beginnings, while sopranos, tenors and violin players tend to do the opposite. This feedback indicates that fundamental frequency certainly can intensify an accent indicating grouping, but its role in grouping itself may be less important than often assumed. I conclude, that fundamental frequency "accentuates" grouping accents but is not the key signal to initiate the grouping process. It should be noted, however, that subjects' intuitions became less reliable when being given "false starts", e.g. an intended iamb starting with a long beat. It thus seems to be crucial to know when to start perceiving a particular rhythmic pattern.

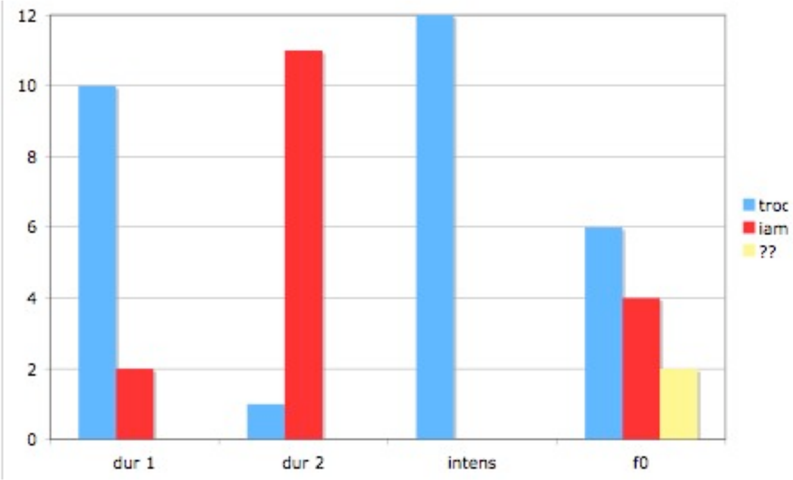


Figure 2.17: Results of the pilot study in order to verify ealier claims concerning the correlates of rhythmical grouping. *dur 1* denotes an intended trochaic interpretation caused by a moderate increase in duration, *dur 2* indicates an intended iambic interpretation, caused by a strong increase in duration, *intens* denotes an intended trochaic interpreation caused by an increase in intensity and *f0* denotes an intended trochaic interpretation caused by an increase in f0.

Normally, the different strategies of accentuation will covary: In music performance, metrical accents (e.g. the first note in a waltz measure) will be produced louder by a musician and typically the composer makes that note stand out using the different possibilities of phenomenal accentuation. Thus, normally the musical performance and composition will be in accordance with our expectations. In speech and language perception, listeners are also quite good at predicting phenomenal accents simply based on their linguistic knowledge: Syntax, semantics and pragmatics provide us with cues to predict the location of (important) content words and phonology may give us plenty of hints where to expect a lexical stress (also cf. Allen (1975)). In many cases, a speaker of the corresponding language will also mark such a predicted (metrical) accent by producing the repective syllable longer, more accurately articulated and with a significant excursion of fundamental frequency. However, in many cases such a strategy may not be necessary: If listeners guide their attention towards an upcoming prominent event, their inner clock may go at a slower pace (cf. Section 2.1.1), thus creating the impression of an accent based on their metrical expectancy alone.

The *structural accents* described above are not a phenomenon known from musical theory alone: The endings of intonational phrases in speech tend to be characterized by typical rises or falls of pitch and lengthening effects (cf. Section 2.3.5). Thus the listener can use similar cues to structure a sequence of auditory events in both music and speech.

2.3.3 Hierarchical Grouping

Grouping is not constrained to one level. Basic rhythmical groups within windows of temporal presence can also be adjoined to form larger groups. Again, such groups are formed based on the principles of maximising global similarity, e.g. listeners tend to perceive larger groups of identical or at least similar rhythmical patterns. Martin (1972) proposed that event sequences are always hierachically structured — on a fundamental level we perceive a sequence of beats, on a higher level, beats are grouped into sequences of two or three, and on an even higher level these groups can be assembled into similar structures in order to form a complex pattern (see Figure 2.18) that is maximally harmonious, i.e. in line with our *rhythmical expectancies based on previous experience*. The higher the hierarchical level of any beat, the more prominent this beat is perceived. Beats reaching the highest levels mark the beginnings or endings of larger grouping patterns. The higher the hierarchical level, the higher is the probability that this beat is realized by a phenomenal accent (Lerdahl and Jackendoff (1983)). It is certainly no coindicence that similar rhythmical hierarchies have been developed in theoretical models describing musical Lerdahl and

Jackendoff (1983) and language rhythm (Liberman and Prince (1977), Nespor and Vogel (1986)). One of the most influential points made by early work on rhythm perception (e.g. Martin (1972)) concerns the issue that we can use our expectancies to perceive and structure upcoming rhythmical events. E.g. using our linguistic experience we know that the probability of an accented and important content word to occur increases dramatically after two or even more function words. It is likely that we have certain "expectancy profiles"[13] after having listened to a lot of speech of various styles or languages. On each level of our rhythmical hierarchy, we can estimate the occurence of the next rhythmically prominent event based on our expectancy profile that matches best to the hitherto processed event sequence. Thus, our expectancy can also draw our attention to certain points of increased attention.

In more recent times, dynamic theories modeling the location of points of increased attention along the timeline have been created. Attention points at different hierachical levels can be predicted by the periodicity of internal, often coupled, oscillators (e.g. Large and Kolen (1994); Gasser and Eck (1996), also cf. Section 3.3.1). These models assume that on each level of our rhythmical hierarchy there exists an individual oscillator with a pertinent eigenfrequency. A harmonious model now expects these different oscillators to have aligned periodic cycles, e.g. they start a new cycle simultaneously. This effect can be easily illustrated by the well known phenomenon known from tapping: It is easy to tap a fast rhythm with the right hand and a slower rhythm with the left hand, as long as these are synchronized, e.g. the taps are co-occur at regular intervals. We can now assume that each hand represents a clock modelled by an oscillator and each tap corresponds to the beginning of an oscillator's period (von Holst (1937)). The term *coupling strength* is often used to refer to the alignment strength of two oscillators. In a perfectly harmonious rhythm, the period of each oscillator on a higher hierachical level comprises an integer multiple of periods of the oscillator periods of the oscillators on lower hierarchical levels. Usually, empirical data shows a less strong coupling of two oscillators — it is often claimed that the oscillators periods are readjusted dynamically during perception in order to better predict the temporal structure of upcoming rhythmical groups. Such readjustements may be necessary if the tempo of the fundamental beats changes.

An alternative dynamic approach to model rhythmical expectancy has been pro-

[13] For the original use of this terminology, see Jongsma et al. (2004)

posed by Desain (1992). Here, complex expected rhythmical patterns emerge with a maximum of attention on the dominant time interval and inferior, additional expectancies occur on multiples or subdivisions of this interval. The predictions made by this model are not much different from the oscillator based view, though (Jongsma et al. (2004)).

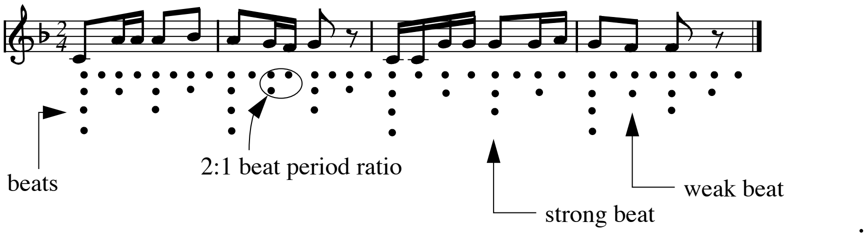


Figure 2.18: An illustration of rhythmical hierarchical organization taken from Lerdahl and Jackendoff (1983). The dots representing fundamental beats can be stronger or weaker, thus being beats at different levels of the hierarchy. Each row indicates a different level of the metrical hierarchy. Beats reaching the highest levels tend to be realized as phenomenal accents according to the definition given in 2.3.2

Some researchers claim that there exists larger temporal windows constraining the temporal extension within which such hierarchically organized groups are formed, e.g. Kien and Kemp (1994) suggested a universal window of timing performance of 2-3 seconds. It is possible that such a window attracts rhythmical grouping at higher levels of rhythmical organization, e.g. musical or phonological phrases. Some empirical findings support this claim, e.g. Bröggelwirth (2007) found out that the large majority of German poetic verses are indeed produced within this timing unit.

2.3.4 Summary: The Fundamental Processes of Rhythmical Grouping

As outlined above, the perception of rhythmical groups depends on a variety of independent but entangled influences of cognitive *top-down* and *bottom-up* processes. The different basic principles of grouping and pattern building outlined in Sections 2.3.1 to 2.3.3 can be summed up to the following fundamental processes:

1. Listeners try to assemble a similar number of lower level rhythmical events

into rhythmical groups. Rhythmical groups or events are constrained by the size of the different windows of cognitive temporal processing. A fundamental window involved in grouping has the size of 400-600ms, a larger window, also regarded as window of universal timing performance, ranges between 2-3 seconds.

2. Listeners tend to equalize the perceptual durations of subsequent fundamental events, unless they are adjusted to a specific duration range. In the latter case, listeners are more sensitive to perceive deviations from the expected (learned) duration.

3. Different tempos lead to different perceptual groupings. In slower tempos (onset to onset interval of fundamental beats $= 167ms$) listeners tend to perceive longer beats as group beginnings, in fast tempos (onset to onset interval $= 88ms$) listeners tend to perceive longer beats as group endings.

4. The kind of accentuation has different grouping effects. Durational accents are usually perceived as endings of a group, accents caused by an increase in intensity or a moderate increase in duration are perceived as beginnings of a group, f0 intensifies an accentuation, strong increases in duration are interpreted as the end of a group.

5. Listeners can experience an accent simply by an increase of attention based on previous experience. After having adjusted to a certain reiterating rhythm, a listener increases her attention within a limited range of time where a new rhythmic event is likely to occur. Such an increase of attention may lead to a perception of increased duration without any objective lengthening.

6. Rhythmical groups are organized hierarchically, i.e. beats may form groups, these groups may form larger groups. The kind of grouping often falls together with windows of cognitive timing. Fundamental beats falling together with boundaries of rhythmical groups at high levels of the rhythmical hierarchy tend to be strongly accentuated - thus, rhythmical expectancies and phenomenal accentuation usually correlate.

Summing up, we see that rhythmical grouping is determined by both relative and absolute timing timing of events, since grouping interacts with temporal windows of auditory processing. It will become clear in the next section, that these fundamental processes of perceiving rhythmical structures are essential in explaining rhythmical grouping in music, language and speech.

2.3.5 Grouping Linguistic Structures

The last sections focused on the cognitive processing of rhythmical structures. Necessarily, most of the reseach results presented were based on perception experiments using highly controlled stimuli sets, e.g. sine tones only deviating with respect to their duration. Therefore, one has to be cautious when trying to apply these results on the processing of language and speech phenomena. In speech processing, the listener necessarily concentrates more on linguistic content than form and the fact that speech can be rhythmic is just one parameter out of many a listener may pay attention to. In chapter 1.1 we reflected on the question whether rhythm is of linguistic interest at all and came to the conclusion that it is not inherently linguistic but may aid the communication process because it helps to organize linguistic content and provides us with timing hypotheses concerning relevant upcoming linguistic events. If this is so, we should be able to apply at least some of the insights known from the processing of rhythmic events on the processing of speech — the mechanisms underlying the helpful effect of rhythm for speech and language processing should not be much different from the ones that come into play when processing non-linguistic stimuli. However, we need to take into account the fact that during speech production, a speaker has to encode content. This content may be subject to some constraints aiming at an optimal rhythmical structure in order to enhance the listener's ability to process what has been said. It would not come as a surprise if the kind of enhancement goes hand in hand with the rhythmical restrictions placed by the phonological structure of the language spoken. Indeed, in the case of rhythm perception many of the insights seem to be directly applicable to the processing of rhythm in speech.

The most obvious connection between language-specific grouping and language-specific rhythm can be found when looking at the almost classical distinction between *stress timed*, *syllable timed* and *mora timed* languages. The distinction between these different rhythmical classes is usually connected to the work by Pike

(1945) and Abercrombie (1967). The traditional view built on a perceptual metaphor capturing the difference between a "machine gun" rhythm of syllable timed languages and the "morse code" impression of stress timed languages (James (1940)) with the so-called *isochrony hypothesis*, i.e. a tendency to make syllables, morae or interstress intervals similarly long, at least on a perceptual level (see Chapters 1.1 and 3). Furthermore, many researchers have found a rhythm type specific preference for grouping in poetry (e.g. Lehiste (1990); Cutler (1994); Noel (2006), also see Chapter 1.1) and linguistic constraints correlating with language specific rhythm (Bertinetto (1989); Dauer (1983), Hoequist (1983), Auer and Uhmann (1988),Auer (1993)). When comparing some of these findings with the previously stated psychological insights on the processing of rhythm, we will understand, *why* certain linguistic structures correlate with language specific rhythmical preferences — not only, *that* they do.

2.3.5.1 Evidence for the Application of Fundamental Grouping Principles in Language and Speech

Here, we will examine whether the grouping principles proposed on page 55ff. appear to have an impact on the rhythmical typology of languages, especially the distinction between *stress timing*, *mora timing* and *syllable timing*.

Grouping principle 1: The Principle of Temporal Presence Grouping principle 1 claimed that low level grouping is restricted by the window of temporal presence, which has been defined as being approximately 400-600ms long. This principle receives support by our findings that foot durations tend to be within this range and are rarely found to be longer (cf. section 1.1). Furthermore, it follows from this constraint that longer feet would then be covering two or more instances of temporal presence. It can be expected that in such a case the foot would have to be split up into two groups. Such an effect would have to lead to a higher number of perceived accents in slow speech, because group endings or beginnings tend to be perceived as more prominent. Also, it follows from this principle that in cases where many unaccented syllables occur in a chain, extra feet and extra prominences must be introduced because otherwise, the group would extend its temporal frame. These hypotheses were confirmed by investigations on speech timing showing similar windows of speech organization and showing that slow speech tends to contain more perceptual (!) accents than fast speech (Fant and Kruckenberg (1996); Schreuder (2006)). Also, it was found that languages with a comparably simple syllable structure tend to comprise more syllables in a foot than languages showing more complex phonotactic patterns, probably because the simple syllable structure makes more syllables fit into one window of temporal integration: Dellwo (2008a) shows that perceived syllable timing exhibits higher articulation rates than stress timing. van Dommelen (2006) reports speech rate to be the most influential factor for an identification of native language in a discrimination study of L2 Norwegian. Dauer (1983) reports that — phonotactically complex — stress-timed languages have less interstress syllables than syllable-timed ones. In her English data, she found a majority of one to four interstress syllables, for syllable timed Spanish, a majority of two to six. For English, she claims that stress is moved for rhythmic purposes or extra stresses are included to break up interstress intervals that are becoming too long. This strategy has also entered phonological models as the well formedness constraint of *beat insertion* (Selkirk (1984), cf. Section 2.3.3). It may be a simple consequence of the fact that the window of temporal presence is not large enough to encompass as many interstress syllables. Therefore, a new window is "opened" hence creating the perceptual impression of a new foot.

Grouping principle 2: The principle of similarity in number In her seminal study, Dauer (1983) observed comparatively stable peaks for the number of interstress syllables. Both stress timed and syllable timed languages seem to favor a relatively small number of syllables allowed in a foot, preferring 2-3 interstress syllables. Thus, her data supports the view that languages constrain this number. However, this finding does not provide convincing support for the hypothesis that this number is kept stable across an utterance. In fact the principle of similarity in number seems to be obeyed most strictly in poetic speech. In his typology of poetic metrical systems, Wagenknecht (1999) states that a language's poetry may be based simply on a restriction concerning the number of syllables[14] per verse. When looking at the way that different languages encode their poetry, one gets the impression that the principle 2 is more important to syllable and mora timed languages, who apparently like to fix the number of fundamental rhythmical entities in their poetry,

[14] Wagenknecht refers to produced beats or sonority peaks rather than phonological syllables.

e.g. the Japanese *haiku* (5 or 7 morae per line), the Italian *endecasillabo* (11 syllables per line, split into two groups of 6 and 5 syllables) the French *hexameter* consisting of 12 syllables per line which are split into two symmetric groups , the Finnish *kalevala* (8 syllables per line), the Tagalog *tananga* (7 syllables per line). Of course, the metrical structure of most mentioned meters is more complex and may demand a certain pattern of long and short syllables, particular accent patterns or rhymes. We can therefore conclude that the application of principle 2 tends to be insufficient in order to create poetry. Still, it plays a major role in the creation of rhythmical harmony in the respective languages. Certainly, the same patterns are used in stress timed languages as well, e.g. at all times, poets adopted meters from languages other than their own in order to create rhythmical harmony. Thus, modern literature knows many istances of English haiku[15] or German endecasillabi[16]. However, often these approaches make certain prosodic adaptations necessary, such as the usage of syllables instead of morae for English haiku. Often, the attempts at transplanting non-native meters have been regarded as clumsy or artificial[17] and lead to rhythmic reinterpretations, such as modern German hexameters allowing for less or more than 12 syllables per line (Noel (2006)). In stress timed meters usually the number of accents is regarded more important than the number of syllables, but also here, a similar number of fundamental rhythmical events enhances the poetic effect, as can be seen in the high popularity of iambic (e.g. Shakespeare's blankverse) or trochaic meters with a fixed number of syllables per foot. Noel (2006) argues that a transplantation of particular meters into another language are usually restricted to sophisticated, or rather learned varieties — in more "down to earth" types of poetic speech, such as folk song or hip hop, they tend to be rare or accidental. It is likely, that these provide better examples for the study of language specific, native rhythmical phenomena.

Grouping principle 3: The principle of local similarity The principle of local similarity captures the phenomenon that within stimulus groups showing durational variability, these variabilities are perceived less. It apparently delivers a good ex-

[15] e.g. by Jack Kerouac

[16] e.g by Johann Wolfgang von Goethe

[17] For example, the poor reception of Longfellow's adaptation of the Finnish kalevala in his *Song Of Hiawatha*

planation of the well-known effect of perceptual isochrony of syllables. Many researchers were puzzled by the phenomenon when measuring syllable durations in syllable timed languages, since these apparently show a significant amount of variability — in fact their durational variation seems to be very similar or sometimes even higher than those of traditional stress timed languages (Roach (1982); Bertrán (1999)). Still, intuitively and perceptually, many languages have been repeatedly argued to be *syllable timed*, i.e. their syllables are perceived to be more or less isochronous. The principle of local similarity may be the reason why the objective variability present in these languages is smoothed out effectively by the listener's auditory processing. However, the principle of local similarity leaves open the question why the auditory perception apparently is guided differently when listening to stress timed languages, since according to the literature, these show a similar amount of variability. In his often quoted study, Roach (1982) measured the standard deviations of syllables in different languages. For each representative language, he examined one speaker. While some language showed variability according to the predictions of the isochrony hypothesis (stress-timed English the highest, syllable-timed Telugu the lowest), French, Yoruba, Russian and Arabic exhibited similar variability. It is possible that his measurements are blurred by the fact that in the standard approach, the standard deviation is not independent of the mean value and increases along with it[18]. Since some languages or individual speakers may systematically produce shorter or longer syllables, this circumstance might influence the measurements. Therefore, more controlled data needs to be evaluated in order to verify Roach's results. His simple investigation was repeated with the help of the BonnTempo database (Dellwo et al. (2004)). As representative candidates for syllable timed languages, French and Italian, as stress timed languages, English and German were chosen from the database. For each speaker, the data contains the same text read by each speaker in five different speaking rates. Thus, variations introduced through individual tempo preferences are to some extent factored out. For each language, three speakers were randomly chosen and examined according

[18] The standard deviation σ_x of a population is usually estimated as $s_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}$, where s_x is the estimate, N is the sample size, x_i is the characteristic value of the ith element of the sample and $\bar{x}$ is the mean value.

to the syllabic standard deviations and the variation coefficient[19], to get more results independent from language specific speaking rate (see Table 2.3).

	syllable timed		stress timed	
	English	German	French	Italian
mean dur.	163	176	168	137
standard dev.	83	78	56	61
variation coeff.	51	42	35	44

Table 2.3: Mean syllable durations (milliseconds), standard deviations (milliseconds) and variation coefficients (%) across three speakers and five speaking rates for each language

When looking at our data, the durational variation indicated by the standard deviation seems to support the isochrony hypothesis as French and Italian show much lower values of standard deviation — contrary to Roach (1982). However, the variation coefficient reveals that at least Italian exhibits even more relative variation as stress timed German, after factoring out the Italian speakers' faster articulation rate. Also, with respect to relative variability French is closer to German than German to English, even though all three languages hardly differ in average syllable durations. To conclude, it is striking that while all four languages show significant durational variations — French showing slightly less and English slightly more — the syllable timed languages deviate less from the average duration in absolute milliseconds. This indicates that the principle of local similarity may be constrained by an absolute duration threshold in speech, i.e. variation is perceived less if the corresponding unit, e.g. the syllable is relatively short. This assumption goes hand in hand with findings by (Dellwo, 2008a, 92ff.) who convincingly shows that perceived regularity is correlated with an increase in speech rate. He postulates a threshold between perceived syllable and stress timing at 11 consonantal plus vocalic intervals per second[20]. A faster rate will be perceived as syllable, a slower rate as stress timed. Such a threshold definitely lies above the JND for duration differences in non-speech stimuli (see Section 2.1.3). In order to verify Dellwo's somewhat bold claim, evidently

[19] The variation coefficient is the normalized standard deviation. It is calculated by dividing the standard deviation by the mean value. It is usually given as a percentage value.

[20] Dellwo counts every c *or* v-interval, thus rate of 11 cv-intervals comprises roughly 5-6 syllables/second as a threshold of variability smoothing.

further research is needed but rate should definitely taken into account more closely when regarding speech rhythm. For the moment, we will refer to this threshold as the *variability smoothing threshold*.

Another issue not revealed by global measures of dispersion lies in the possibility that the perceptional effect of similarity in syllable timed languages is the result of distributional effects. E.g. even if a lexical or phrasal stress will have the effect of a durational increase larger than the variability smoothing threshold in both English and Italian (both languages have lexical stress), it will come into play more frequently in English than in Italian, because lexemes tend to be shorter in English. Thus, the smoothing effect will be disturbed more often in English thus creating even more perceptual variation. In French, where lexical stress is expressed less in terms of duration, the durational peaks may even be restricted to the end of "stress groups". Furthermore, stress timed languages show more alternation of very short reduced followed by very long stressed syllables. Both of these tendencies combined lead to rather smooth duration profiles in syllable timed languages and rather harsh duration profiles in stress timed languages. Figure 2.19 shows duration profiles of the different languages indicating longer sequences of syllables with relatively similar durations in French and (somewhat less) in Italian, but a completely different style in English and (somewhat less) German, where very short syllables are often followed by very long ones. Thus, less frequent *strong* deviations from an average or neighboring duration plus less strong absolute variation (very short and very long syllables) may in combination enhance the perceptual effect of relative syllable isochrony in syllable timed languages.

Grouping principle 4: The principle of local dissimilarity This grouping principle can be regarded as the complementary principle to the previous one. While principle 3 explained the effect of perceptual smoothing of variabilities, principle 4 enhances durational differences by increasing the listener's sensitivity towards them. A necessary prerequisite for this grouping principle is the listener's adaptation to a particular, usually short, duration. After getting used to such a duration, any deviance will be perceived better. Thus, the perception of lengthening ought to increase after being exposed to a sequence of identical or at least very similar short (!) durations. Such an effect explains several duration related phenomena in

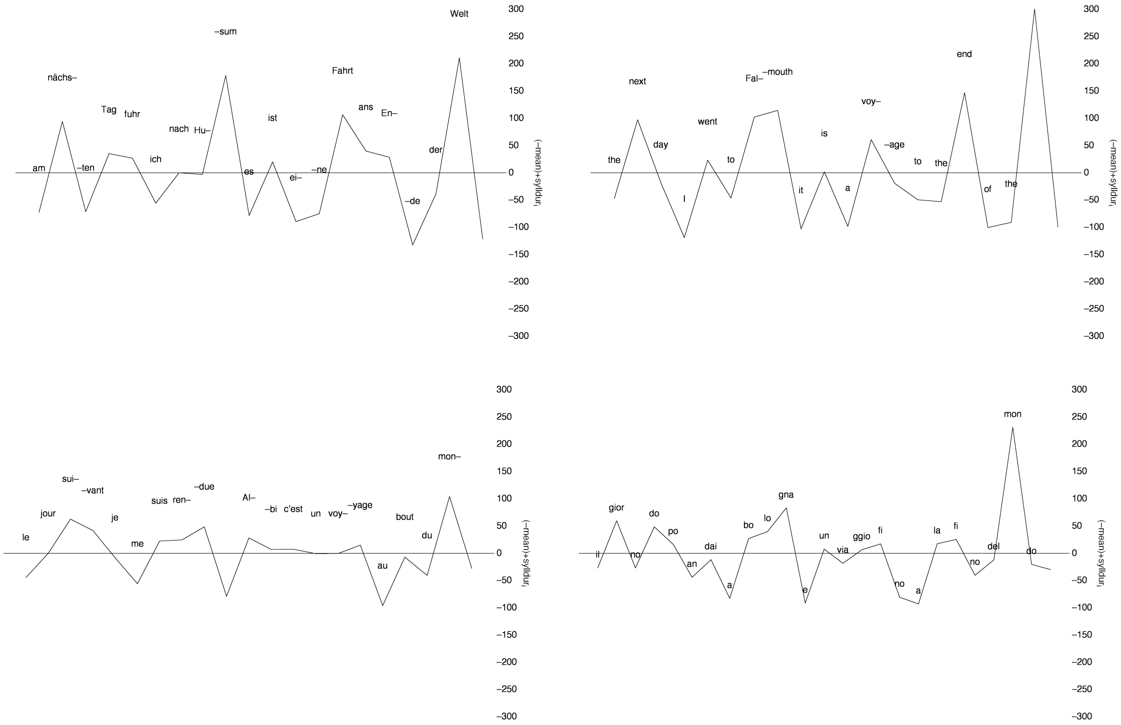


Figure 2.19: An illustration of rather steep duration transitions in stress timed German and English (top) and more smooth transitions of syllable timed French and Italian (bottom). Each example shows two sentences with semantically and syntactically similar content read at normal tempo. Each example lies within a range of 4.5-5.5 seconds. Syllable durations are expressed as deviations from the mean duration (horizontal line) in milliseconds. By adding each syllable duration to the negative mean duration, short durations are shown as valleys and long durations as peaks. Our data definitely is in favour of a weak isochrony hypothesis showing much more pronounced and more frequent peaks and valleys in German and English.

speech: In a syllable timed language such as French (see Figure 2.19), the listener is frequently exposed to sequences with relatively little duration variation — possibly this increases the listener's sensibility towards final lengthening effects which has sometimes lead to French being perceived as *iambic* or *trailer timed* (Wenk and Wioland (1982)). Studies in Language Acquisition confirm that L1 learners of French reach such an iambic bias (Paradis and Deshaies (1990)) within the first months of life (Höhle et al. (2007)), although such a pattern may be phonologically marked in the sense that most languages prefer a trochaic pattern (e.g. Allen and Hawkins (1979); Turk et al. (1995)). In fact, the ability to perceive subtle durational changes when occurring systematically may play an important role in French, since it marks word boundaries with a small increase in duration, albeit not in fast speech, where lengthening is solely used in order to signal phrase finality (Keller et al. (1993); Keller and Zellner (1996)). Thus, grouping principle 4 ought to lead to an effect of minimizing durational variations in French before word boundaries and phrase boundaries. Even though Figure 2.19 seems to support this hypothesis impressionistically, it definitely needs further research to be confirmed.

A similar phenomenon which has already been explored empirically concerns a stress timed language, English. An early suggestion by Jassem (1952) is that rhythm in English is not explained by lengthening of lexically stressed syllables but rather by shortening *anacrusis* syllables, i.e. the unstressed syllables before the stressed ones that do not belong to a *narrow rhythm unit*, i.e. a prosodic unit starting with a lexically stressed syllable and ending with the word boundary (cf. Figure 2.20). This prosodic segmentation differs from the more widespread view that English rhythm can be explained relative to feet, which span from one lexically stressed syllable to the next one. In Jassem et al. (1984), evidence was presented that syllables within the anacrusis are produced in a very quick, uniform manner but are not subject to any *compensatory shortening* (cf. Section 2.4.1.2), i.e. they are not compressed with an increasing number of syllables in the anacrusis. They also detected that the phones within narrow rhythm units are compressed with an increasing number of phones but are also produced less compressed if time allows. A more recent reinvestigation of Jassem's model carried out on a large corpus of British English confirmed these results (Bouzon and Hirst (2004)) and further made the detection that stressed syllables hardly show any significant lengthening effect - the effect of an increased duration on stressed syllables is a result of coinciding with the beginning of a narrow rhythm unit (Hirst and Bouzon (2005) thus being less compressed. According to them, subjective lengthening is described better as a lack of compression.

Grouping principle 4 now gives a perceptual explanation for this phenomenon: Listening to the short, relatively uniform anacrusis syllable(s), the deviating duration indicating the beginning of a narrow rhythm unit increases the perceptual prominence of the stressed syllable because the listener is more sensitive to it.

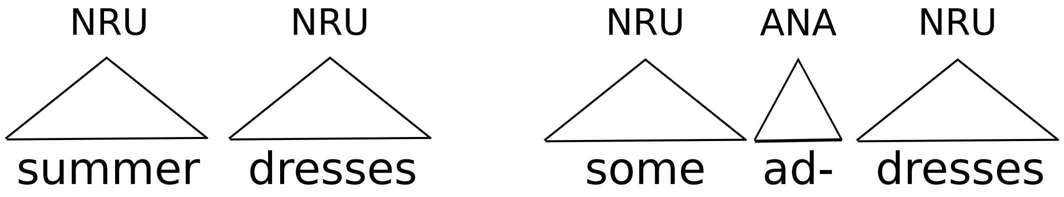


Figure 2.20: An illustration of the principle of the narrow rhythm unit (NRU) and the anacrusis (ANA) in English. The two sequences *summer dresses* and *some addresses* are identical on the segmental level and share the location of lexical stresses. Jassem's model predicts a shorter production of the anacrusis syllable thus creating a prosodic minimal pair. Within the NRU, any shortening effect will be only due to *compensatory shortening* with an increasing number of syllables contained in the NRU. The ANA duration increases as a function of the number of syllables contained in it.

Grouping principle 5: The principle of global similarity This grouping principle explains the perception of similarity in structure across different rhythmical groups. Two fundamental mechanisms are apparently responsible for the phenomenon of perceiving neighboring groups of acoustic stimuli or beats as more or less similar.

First, the group internal relative durations apparently are more important for subsequent groups to be perceived as identical than absolute durational similarity (Handel (1992)). I.e. for two groups to be perceived as identical, maybe even isochronous, it is sufficient for them to share the same global structure of longer or shorter fundamental rhythmical events, i.e. beats. This fundamental mechanism may explain — at least to some extent — the phenomenon of perceptual foot isochrony in stress timed languages, provided that consecutive feet are built of similar numbers of fundamental beats. For this phenomenon, evidence so far has been non-existent. However, this mechanism still does not explain why in stress timed languages adjacent feet are still perceived as more or less equal in duration if the feet are built of different numbers of syllables.

Second, there exists a mechanism to have listeners group differently depending on the tempo of the stimuli they are listening to. With an increase in rate, listeners interprete an increase in duration as the end of a rhythmical group, at slower tempos the same increase is interpreted as the beginning of a rhythmical group. This mechanism covaries with the circumstance that some syllable timed language such as French have been described as *trailer timed* or *iambic* whereas stress timed English has been described as *leading timed* or *trochaic* (Allen and Hawkins (1979); Wenk (1982); Wenk and Wioland (1982); Paradis and Deshaies (1990); Turk et al. (1995)). Since it has been shown that syllable timed languages tend to be produced at somewhat higher articulation rates (Dellwo and Wagner (2003); Dellwo (2008a)) this may be more than an accidental phenomenon. These language specific grouping preferences may be a direct consequence of the languages tempo preferences. One should be more careful, though, to generalize from this too much, since iambic grouping — as reported for French — apparently is the phonologically marked case across the languages of the world. If one equates syllable timing with a fast tempo and tempo with an iambic grouping preference, it would follow that all fast syllable timed languages are iambic. This is certainly oversimplistic since grouping is not constrained to durational variation but can be achieved by a variety of acoustic cues, as will be explained in the following section. However, if duration is the prevalent acoustic cue to rhythmical patterning — as it may be the case in French — an iambic impression may be the consequence of the language's preferred tempo.

2.3.5.2 Grouping through Accentuation in Language and Speech

Rhythmical grouping emerges as the result of perceiving some fundamental beats as more prominent than others. That way, a pattern of stronger and weaker beats is perceived and we group the perceived chain of beats into sequences of a strong beat possibly followed by one more more weak beats. In the paragraphs above, we mainly discussed duration cues to rhythmical grouping, e.g. the way that variability or similarity in duration may influence the way we perceive a sequence of speech events as a chain of similarly or differently long. One underlying assumption here is that duration or perceived length positively correlates with perceived strength of a fundamental beat. In phonetic research this perceived strength of a speech event is usually referred to as *perceptual prominence*.

While there is general agreement that an increase in duration also increases perceptual prominence across many, probably most languages, other acoustic cues to prominence have been under consideration in phonetic research as well. Among these, pitch excursions or f0-variations have been the most intensively studied. In fact, prominence has received much attention in the phonetic literature at all times. However, it has often been explored under different names, such as lexical or sentence stress, lexical or sentence accent, prosodic focus etc. In such studies, researchers have examined the acoustic cues to words or syllables which as perceived and produced prominent because they fulfill a specific linguistic function. Nowadays, perceptual prominence tends to be regarded as the result of a complex interaction of various prominence lending acoustic parameters, among them[21] being

- duration (e.g Fry (1955); Fant and Kruckenberg (1989); Dogil (1995); Jessen et al. (1995); Eriksson et al. (2001); Streefkerk (2002); Mixdorff and Widera (2001); Batliner et al. (2007); Andreeva et al. (2007a,b); Tamburini and Wagner (2007)

- intensity, loudness (e.g. Isačenko and Schädlich (1966); Silipo and Greenberg (1999, 2000); Kochanski et al. (2005))

- presence of a significant fundamental frequency excursion (= a pitch accent) (e.g. Fry (1958); Dogil (1995); Jessen et al. (1995); Heuft (1999); Wagner et al. (2000); Eriksson et al. (2001); Streefkerk (2002); Andreeva et al. (2007b)

- shape of fundamental frequency excursion (late vs. early peak, L+H* vs. H* or H*+L) (e.g. Kohler (1991); Reyelt et al. (1996); Heuft (1999))

- spectral emphasis, signalling *vocal effort* or laryngeal setting (e.g. Campbell (1995); Sluijter and van Heuven (1996); Claßen et al. (1998); Heldner (2001); Eriksson et al. (2001); Tamburini and Wagner (2007))

- formant frequency (e.g. Dogil (1995); Andreeva et al. (2007a))

[21] It should be noted that the list below is almost entirely concerned with correlates of prominence in Germanic languages such as German, English, Swedish and Dutch and far from complete, but see Andreeva et al. (2007a) for an interlanguage study.

- linguistic or rhythmical expectancy (e.g. Eriksson et al. (2002); Wagner (2005); Quené and Port (2005); Tamburini and Wagner (2007); Arnold (2008); Arnold and Wagner (2008))

Thus, accentuation in speech works very similar to accentuation in music as it has been described by Clarke (1999) (cf. Section 2.3.2). Clarke's *phenomenal accents* are caused by acoustic cues such as duration, intensity, pitch excursions or changes in timbre, i.e. he also regarded very similar cues as being responsible for making a musical note appear as prominent. His *structural accents* are marking musical boundaries, i.e. they signal the ending of a musical theme, line, phrase or piece. In both music and speech, endings involve lengthening effects (= ritardando) and a decline in pitch. The distinction between structural accents and phenomenal accents in musicology parallels the phonological distinction between *prominence lending* and *non prominence lending* pitch phenomena, the latter marking boundaries or the end of intonational phrases (see Grice et al. (2000) for an overview). This distinction already indicates that equalizing prominence and rhythmical structuring via accentuation cuts too short: Non-prominence lending phenomena certainly add to rhythmical strucuture. It simply is unclear whether such phenomena ought to be called accents, as Clarke (1999) suggested for music, since a phonological accent typically implies an increase of prominence. Still, structural accents play an important role in the rhythmical structure of speech. The phenomenon of *final lengthening* apparently is a universal phenomenon across all languages hitherto examined (Crystal and House (1988); Beckman and Edwards (1990)). It explains the fact that the ending of a prosodic phrase, typically an utterance, is indicated to the listener with the help of a pronounced lengthening of the final syllable(s). The amount of lengthening is positively correlated with the depth of the boundary, i.e. boundaries within an utterance are marked more strongly by final lengthening than those at the end of utterances (Wightman et al. (1992); Gussenhoven and Rietveld (1992)). Also, endings are often marked by typical declination. These have also been studied extensively, but unlike the rather universal final lengthening effect, there seem to be language specific meanings conveyed by the type of fundamental frequency contour occurring at boundaries. While the lengthening itself indicates the boundary as such, the contour signals whether the speaker is asking a question, has doubts, has still something to say etc. Thus, the *boundary tones* are not confined to the expression of sentence mode, i.e. differentiating a declarative from an interrogative utterance. The semantics and pragmatics of boundary tones have been encoded in phonological approaches to intonation, most prominently the Autosemental-Metrical framework for various languages (e.g. Pierrehumbert and Hirschberg (1990); Ladd (1996); Grice and Baumann (2002)).

With regards to musicology's third type of accent, the *metrical accent*, there exist growing evidence that similar mechanisms are present in speech perception. Sometimes, accents are perceived due to their location rather than their acoustic realisation, i.e. it is possible to perceive a content word as prominent simply because it is a content word and is produced at a time where it is likely to be realized prominent, e.g. towards the end of an utterance. However, some well-known positional effects, e.g. the bias of perceiving the first syllable of a word as stressed, can be eliminated by manipulating the acoustic signal (van Heuven and Menert (1996)). Thus, we can assume that a rhythmical bias is learned.

Extensive studies on language acquisition for German have shown that these complex acoustic cues are used from an early age on (Lintfert and Schneider (2005)) and are made use of consistently in German child-directed speech (Schneider and Möbius (2007)). Even with regards to one language or language familiy, there exists considerable disagreement concerning the relative contribution of the various acoustic cues to prominence. While most studies do find that overall intensity plays an utmost minor role in the signalling of prominence (see for example Fry (1958); Nöth et al. (1991), perceptual loudness has claimed to be a much more important cue than fundamental frequency by Kochanski et al. (2005). Lately, loudness or intensity parameters tend to be regarded as indicators of vocal effort which is encoded in a different phonation mode. An increase in vocal effort enhances the intensity in the higher frequency spectral regions due to an abrupt closing of the vocal folds. This phenomenon has been described and quantified as *spectral tilt* or *spectral slope*. Some studies have had difficulties in finding convincing evidence for this type of prominence correlate (e.g. Streefkerk (2002); Mooshammer and Harrington (2005); Wagner (2005)) but it is possible that these results are mainly due to measurement inconsistencies, lack of a robust metric for the measurement of spectral slope and confounding segmental influences (Eriksson et al. (2001)).

Especially with regards to larger databases, fundamental frequency has been rejected as a good predictor of prominence (Silipo and Greenberg (1999, 2000)). However, traditionally, fundamental frequency excursion has been claimed as *the* prototypical indicator for perceptual prominence. While most researchers still agree that in presence of a pitch accent, fundamental frequency is a major and very reliable indicator of perceptual prominence, given its absence, one cannot conclude that a syllable is always completely deaccented or has no prominence whatsoever. Studies which found fundamental frequency to be the best predictor of prominence tend to examine lexical stress in content words or sentence stress only. It is probable that these types of linguistically relevant prominences indeed are typically pronounced with a pitch accent. Batliner et al. (2007) observed that given the circumstance that a pitch accent in most cases correlates with a significant increase in duration, the latter remains to be a reliable predictor of prominence under any circumstance.

With regards to rhythmical grouping and accentuation, duration is involved in both by signalling boundaries and prominence, or are mainly responsible for signalling *structural* and *phenomenal* accents. Furthermore, even with regards to the signalling of boundaries, we learnt in section 2.3.5.1 that an increase in duration can be interpreted both as the beginning or the ending of a group, depending on the underlying tempo. This double function of durational grouping may easily lead to a certain amount of ambiguity, since every durational increase may indicate either the beginning, the end of a group or an accentuation within it. Thus, it is possible that fundamental frequency and intensity related parameters as rather unambiguous prominence lending parameters may fulfill the function of marking the beginning of a group (rather than its end) and indicate phenomenal accents different from boundary signals (Cooper and Meyer (1960)). The role of non durational cues in rhythmical grouping are a well-known phemenon in music as well: In Reggae rhythm, a 4/4 meter, the typically less accentuated *downbeats*[22] are pronounced by high frequency instruments like the rhythm guitar. In oriental rhythms, the distribution of high and low frequency beats, often referred to as “tak” and “doum”,

[22] A 4/4 meter is comprised of 4 quarter notes, the first and third are usually regarded as the downbeats, where one would clap with a hand or tap a foot while listening. The offbeats are the second and fourth quarter notes, which are typically produced less prominent. Thus, the accentuation of the offbeat certainly is a deviation from *prototypical* listening experiences. While offbeat accentuation is typically found in Reggae, Ska and New Wave music, a related stylistic device used in Western classical music is the *syncope* which completely alters the predominant meter temporarily.

respectively, play a crucial role.[23] These are further indicators for the way that in rhythm perception, duration is not the only contributor to rhythmical structure. Further evidence for the assistant role of fundamental frequency in the detection of rhythm class comes from listening tests with adults and newborns (Ramus and Mehler (1999); Ramus (2002)). When delexicalized speech was presented with a flat intonation, all listener groups were able to distinguish rhythm patterns characteristic for stress timing or syllable timing. This indicates, that duration is a sufficent cue to rhythm perception. However, discrimination improved when intonation could be used as an additional cue. Keeping this in mind, the rather controversial role that intonation may play in the signalling of rhythmic structure can be seen in a novel light. As in oriental rhythm, frequency and intensity may add information to the complex pattern in order to disentangle speech rhythm from the — potentially — ambiguous signals we receive from duration. Given the well known phenomenon of fundamental frequency decliation or downstep across an utterance, the beginning of a rhythmical phrase will usually correlate with a high fundamental frequency, also known as *reset*. Since rhythmical groups often start with pitch accented syllables, it is possible that altogether, fundamental frequency is interpreted as the beginning of a new rhythmical group. Intensity related phenomena obviously are necessary to create the impression of phenomenal accents that are no boundaries or create groups at a lower level of the rhythmical hierarchy. It is therefore concluded that duration, fundamental frequency and spectral intensity parameters all contribute to rhythmical grouping in different ways, where their combination delivers valuable cues for a rhythmical interpretation of an utterance:

- End markers: A strong durational increase is interpreted as the end of a rhythmic group, unless the tempo is very slow and it is not also combined with additional cues to phenomenal accentuation, i.e. its tonal realisation is not equal to a pitch accent. If it is combined with a rising fundamental frequency, it seems to be ambiguous because it marks both the end of a group and the beginning of a subsequent group.[24]

[23] The different frequency characteristics are typically created by playing either the center or the rim of a percussion instrument.

[24] This also explains that this type of prosodic event is usually interpreted as a pitch accent when presented to listeners in isolation (Wagner and Paulson (2006)).

- Beginning markers: An increase in fundamental frequency tends to be interpreted as the beginning of a rhythmic group and simultaneously as a phenomenal accent. This prosodic event is equivalent to a pitch accent in the sense of Ladd (1996); Kohler (2006) and others.

- Group internal accents/Beginning markers at lower levels: An increase in prominence lending intensity related parameters tend to be interpreted as phenomenal accents. They tend to be combined with an increase in duration as well. This prosodic event is equivalent to a force accent according to Kohler (2003, 2005). It is possible that these are also interpreted as beginning markers of smaller groups, as a suborganization of the entire rhythmic group. With regards to music, this would be the third quarter note in a 4/4 meter which is usually produced with a slight accent and could be interpreted as initiating the second half of the entire measure.

The extent, to which the various cues contribute to prominence and boundary signalling has been described as being language specific. Also, each language may make use of the different boundary markers on various levels of the prosodic hierarchy to a stronger or lesser extent. This is not surprising, since marking both the beginnings (pitch accent) and ends (final lengthening) of each rhythmical unit would show a significant amount of redundancy. Such a rhythmically fully specified group and various less redundant varieties are illustrated in Figure 2.21 and 2.22. It is furthermore possible that listeners are less sensitive to acoustic cues not used in their native language to signal prominence and boundaries, e.g. native speakers of French have been shown to be rather unable to perceive prominent syllables in Spanish (Peperkamp et al. (1999); Dupoux et al. (2001); Peperkamp and Dupoux (2002)). This phenomenon has been called “stress deafness”. However, since many researchers claim French to *have* stress or accent (e.g. Wenk and Wioland (1982); Martin (2002), it may simply happen to be phonetically realized differently from other languages. E.g. while many languages are characterized by lexical stress which correlates with the perceptual prominence of the stressed syllable, French can be described better by stressing the first and the last syllables of a “stress group” (e.g. di Cristo (1998); Wenk and Wioland (1982); Kohler (2006)). Thus, it may well be, that French listeners are not “stress deaf” but they are paying attention to different

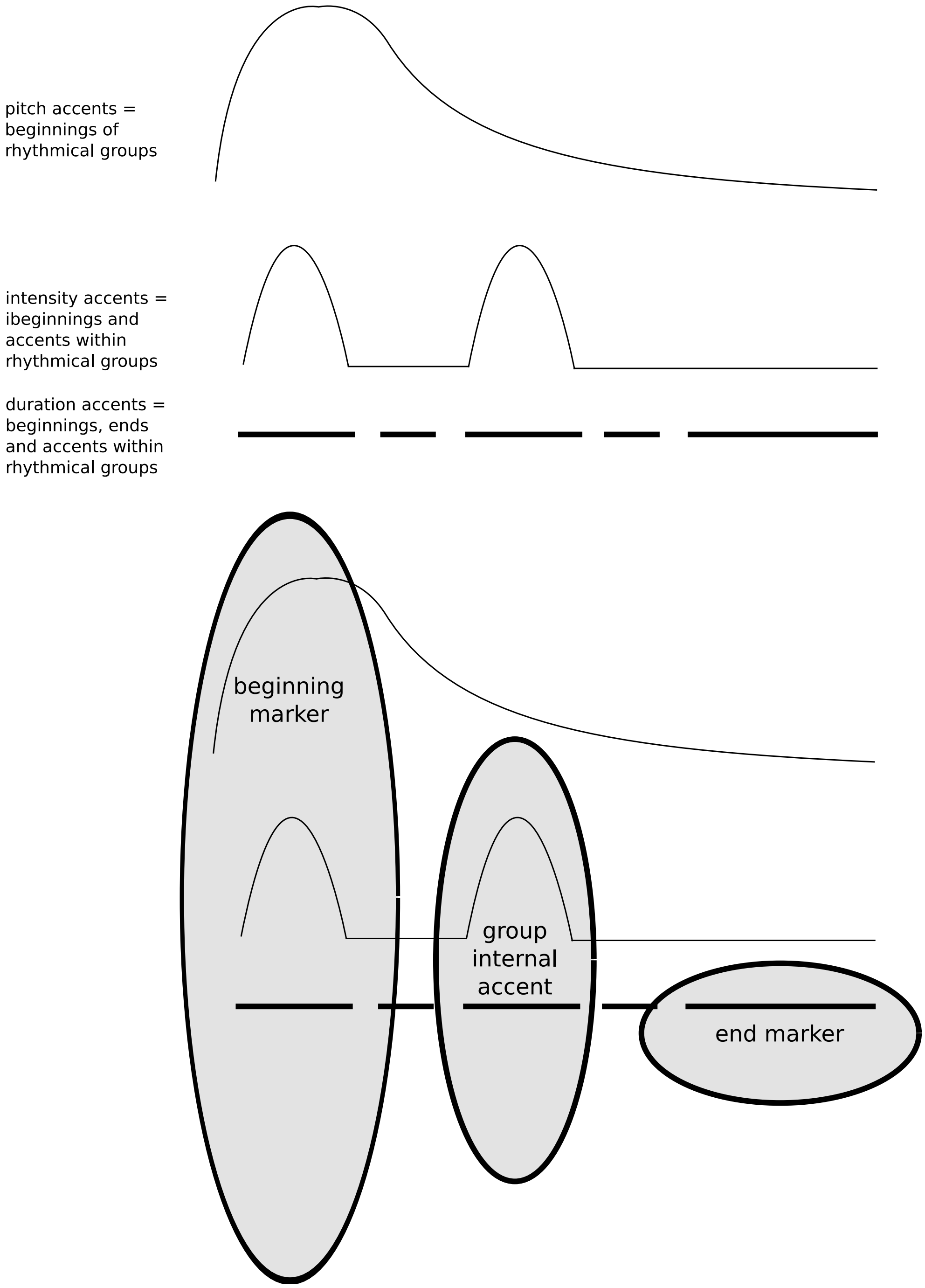


Figure 2.21: The combination of the different acoustic parameters causing the impression of accentuation lead to different rhythmical structures. While increased duration on a beat is perceived as the beginning of a group, its end or a group internal accent, pitch usually indicates the beginning of a rhythmical group unambiguously. Intensity related parameters aid the interpretation of a beginning but may also mark a group internal accent, typically in combination with a durational increase. Pitch and intensity help to disambiguate the various rhythm related functions of duration.

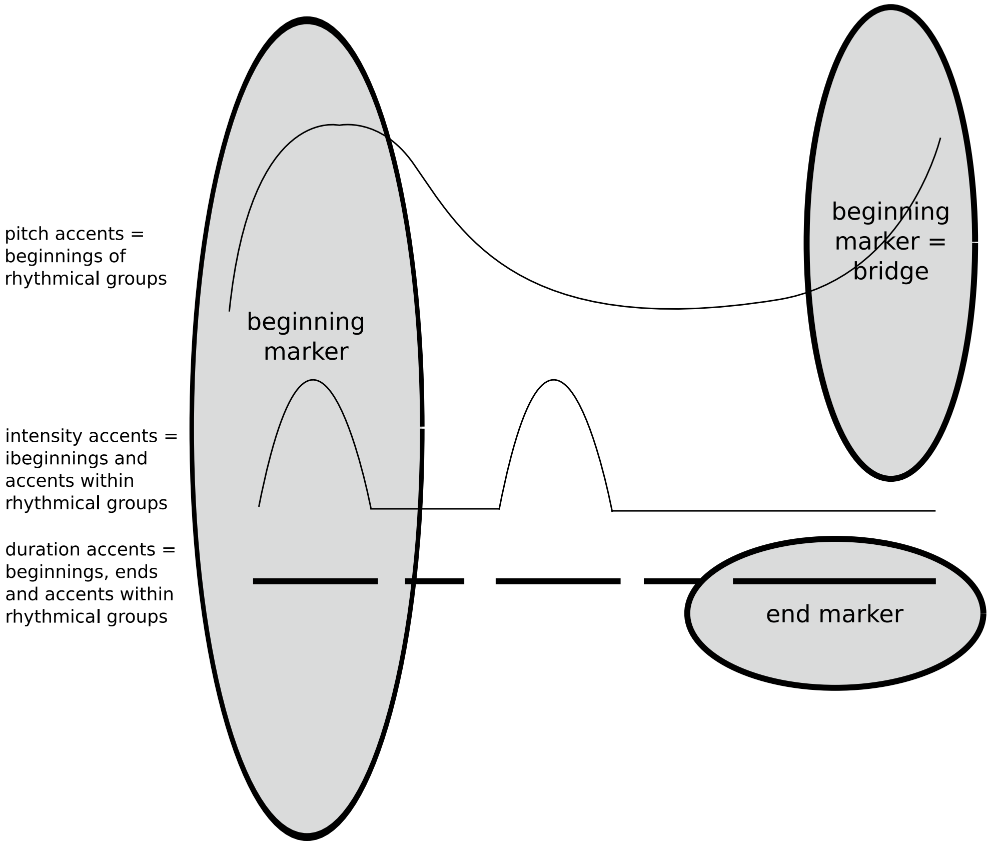


Figure 2.22: If increasing pitch and final lengthening are combined in boundary tones, this may cause a rhythmical ambiguity. Here, final lengthening may be even more pronounced, but the ambiguity builds a rhythmical bridge to the upcoming speech event, e.g. an answer (in question intonation) or to the continuation of the utterance, as a so-called *continuation rise*.

rhythmical cues.[25] French rhythm seems be be particularly shaped by signalling the endings of rhythmical groups, while English or German tend to be characterized by marking the beginnings of rhythmical groups and the endings of intonational phrases. With regards to language specific rhythm perception many more factors may play a dominant role whether a language is quantity sensitive, e.g. whether its heavy syllables attract lexical stress, whether quantity is distinctive or whether it has tone, which means that fundamental frequency fulfills phonological functions other than signalling accent. However, it is likely that if a particular acoustic parameter already carries high functional load, e.g. as fundamental frequency in tone language, it is likely it will be used to a somewhat lesser degree in another domain such as the signalling of rhythm.

Summing up, duration appears to be the basic structural parameter of speech rhythm. It is used to indicate rhythmical boundaries and rhythmical accents within a rhythmical group. The typology of accentuation in both rhythm and speech implies that there exist accents indicating rhythmical groups and those that do not. Since duration apparently is used as an acoustic cue to both types of accents, this may create rhythmical ambiguity in many cases. Thus, further cues may come into play which help to disambiguate the rhythmic structure in both music and speech. While *pitch accents* are obviously used to signal the beginnings of groups, so-called *force accents*[26] indicate group internal accents. An alternative, hierarchical point of view would be that force accents mark the beginnings of rhythmical groups at a lower level of rhythmic-hierarchical organization. Where boundaries and pitch increase fall together, a boundary also marks the beginning of a new group, as in an utterance internal phrase boundary, a so-called progredient phrase, or in a decision question, where the speaker already hints at an answer. Thus, the expectation of an answer is prosodically marked by placing a beginning of a new rhythmical group. This can be interpreted as a *rhythmical bridge* for the listener who is expected to answer the question.

It is not necessarily the case that each rhythmical group marks both the beginning and its end. Theoretically, it should be sufficient for a listener to have one clear cue of either beginning or end in order to interpret the grouping structure.

[25] Volker Dellwo, personal communication

[26] I.e. accents that are produced by an increase in spectral intensity and duration.

Thus, a language may chose to concentrate on marking endings rather than beginnings or vice versa. This option comes in nicely when remembering that rhythmical structure is organized hierarchically, i.e. grouping and segmentation take place on different prosodic levels such as the syllable, prosodic foot, prosodic phrase and intonation phrase (cf. 1.2.5). If a speaker now employs different cues to indicate grouping at various levels, this may simplify the decoding of the hierarchical structure of rhythm enormously. E.g. a speaker uses intensity related cues at a rather low level of prosodic organization to indicate the beginnings of rhythmical feet, pitch accents at the beginnings of prosodic phrases and duration at the end of intonation phrases. Alternatively, both endings or beginnings could be marked by pitch. A language specific rhythmical analysis thus needs to specify whether a language marks beginnings and/or ends at the various levels of the prosodic hierarchy and which prosodic cues are predominant/desambiguating on each level.

We conclude that as a basic approach, a duration oriented approach to rhythm is sufficient, but in order to fully understand any language specific pattern, all cues to rhythmical structure need to be taken into account. A listener will be able to interpret the acoustic cues to rhythm selectively in her language and disentangle them from other, purely phonological cues. She will be able to infer from a given rhythmical sequence whether the upcoming beat will be rather strong, weak or whether the ending of a phrase is about to occur etc. Such an abstract, language and speaking style dependent knowledge can be called *meter*.

2.3.5.3 Hierarchical Grouping in Language and Speech — Metrical Grids, Metrical Trees and Acoustic Correlates

In sections 2.3.5.2 and 1.2.5 hierarchical grouping of rhythm has already been introduced: A rhythmical group may consist of several subgroups or constitute a subgroup of a higher order rhythmical group. In order to facilitate the listener the perception of the kind of grouping, it is possible to strengthen either the beginning or ending of each group in various ways introduced above. The level within the hierarchy positively correlates with the amount of perceived prominence in the entire stretch of speech under examination: Let us presume we have an utterance which also constitutes a rhythmical group. This rhythmical group is identical with one intonational phrase in the prosodic hierarchy. The most prominent beat in this intonational phrase would be equated with the prominence of the entire utterance. At lower levels, the intonation phrase may consist of several stress groups or prosodic phrases, each of them beginning with an accent. The prominence of each accent determines the prominence of each prosodic phrase. Thus, each rhythmical group can be said to have a rhythmical *head* determining its rhythmical prominence or strength. The head is always the most prominent element within the group. Smaller rhythmical groups, e.g. feet, can be combined into larger groups which are dominated by a more prominent head. Thus, a head typically is coexistent with a prominence-lending phenomenal or metrical accent. Structural accents have the predominant function to indicate boundaries, they are less important in order to signal prominence across larger rhythmical groups (= phrases). However, given the circumstance that they also may indicate the depth of phrase boundary, they should at least indirectly serve as a marker of rhythmical prominence of the pertinent rhythmical group. The interaction of phenomenal, metrical and structural accents in the signalling of boundaries and prominence are illustrated in Figure 2.23.

The location of the prominence lending accent within each group is usually defined in a language's phonology. The framework of Metrical Phonology (Liberman and Prince (1977)) postulates that there are language specific preferences of placing a stress either towards the right or the left edge of a prosodic unit. This belief has also been adopted in constraint-based phonological frameworks such as Optimality Theory (Prince and Smolensky (1993)), e.g. by Kager (1999). Since these left or right-dominating preferences can differ for each level of prosodic organization, it helps to explain many complex phenomena of stress distributions, e.g. the fact that in German and English, lexical and phrasal stress tends to orientate itself towards the right edge, while in compounds, typically the left constituent is rhythmically dominant. The dominance is expressed with the help of metrical trees. In a *metrical tree*, on each branching level one branch is labelled *s(trong)*, while the others are labelled *w(eak)*. The most prominent prosodic unit within a larger rhythmical group is dominated purely by strong branches (see Figure 2.24). The prominence relations expressed in metrical trees are better illustrated in *metrical grids*. In a metrical grid, prominences are expressed as columns, where each column represents a time slot of a fundamental beat which can be produced more or less prominent. Prominence level is indicated by the column height, illustrated by more or less stars. Metrical

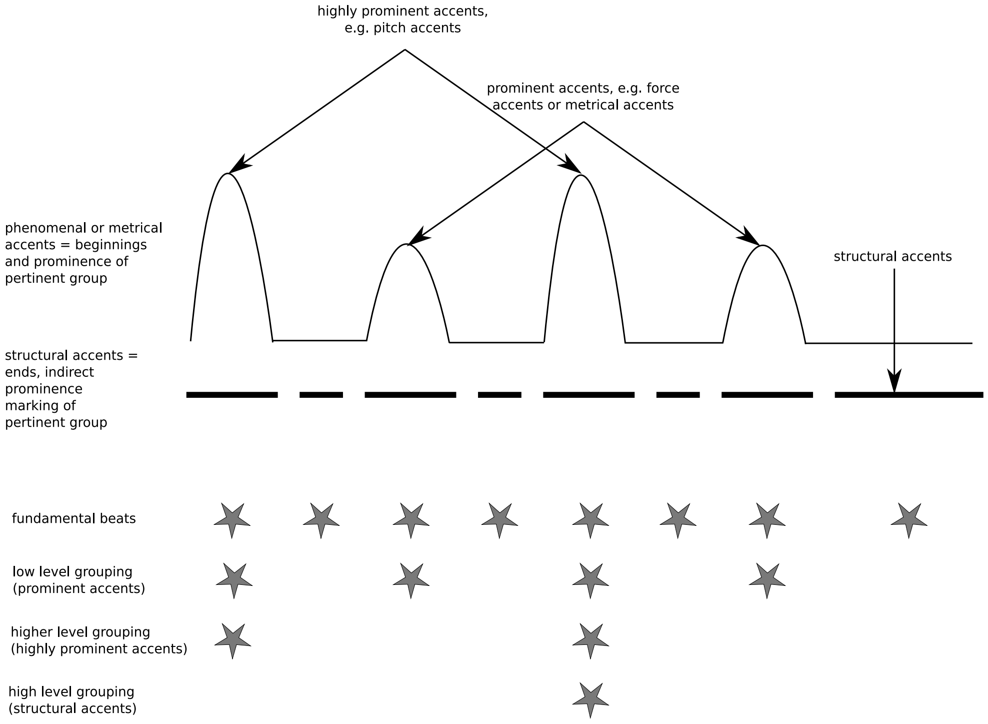

highly prominent accents, e.g. pitch accents
prominent accents, e.g. force accents or metrical accents
phenomenal or metrical accents = beginnings and prominence of pertinent group
structural accents
structural accents = ends, indirect prominence marking of pertinent group
fundamental beats
low level grouping (prominent accents)
higher level grouping (highly prominent accents)
high level grouping (structural accents)

Figure 2.23: Pure structural accents are mostly involved in marking grouping while phenomenal and metrical accents also signal prominence of the pertinent group. However, since structural accents such as boundary tones also provide information about the placement of the group within the rhythmical hierarchy, an indirect prominence lending effect of structural accents is possible.

grids can be extracted out of metrical trees in the following way: First, each fundamental beat is indicated by a single prominence expressing star. Then each column which is dominated by a branch labelled *s(trong)*, receives one more star than its neighbors labelled *(w)eak*. The column which is only dominated by strong branches must have at least one more star than the remaining columns. There is no logical upper limit of branching complexity but we can assume that in natural language, the highest meaningful rhythmical structure probably can be called a paragraph, marking the ending of a text passage or a sequence of coherent utterances. With regards to prominence perception, though, there seems to be an upper limit around 4-5 levels of prominence that listeners are able to distinguish (Marbe (1904); Wagner (2002); Jensen and Tøndering (2005)).

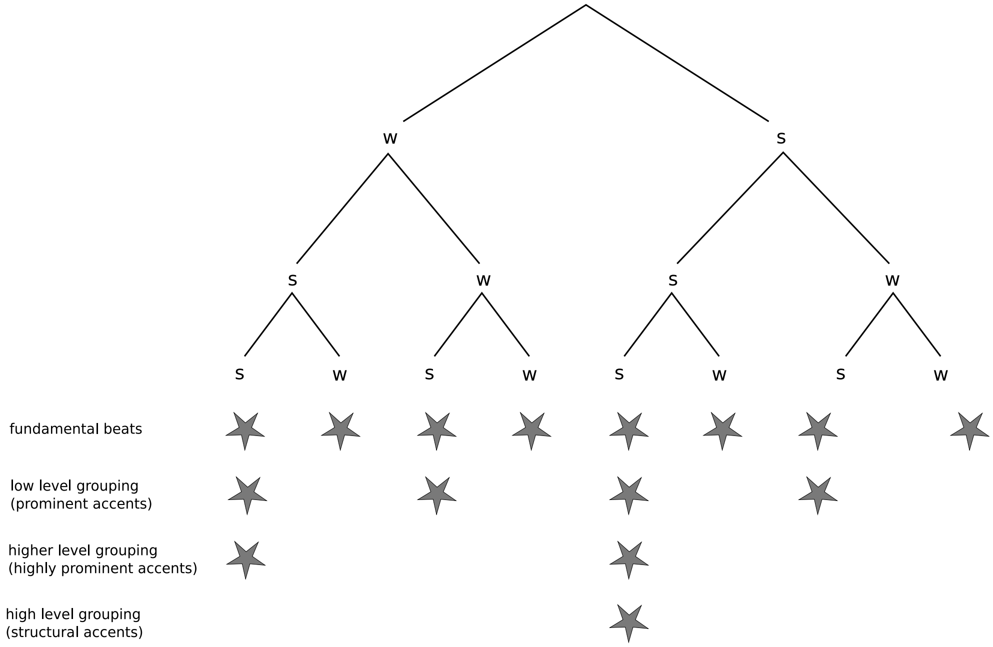


Figure 2.24: In metrical trees, the most prominent unit is entirely dominated by strong branches. The given tree illustrates the left dominant metrical pattern of lower level rhythmical groups and the right dominant metrical pattern of the high level rhythmical group. The former correspond to rhythmical feet and stress groups, the latter correspond to prosodic or intonation phrases. The pertinent metrical grid with the corresponding prominence pattern is illustrated below the metrical tree. The presented phrase may represent the metrical structure of language or speech, e.g. it would correctly describe the rhythm of the German phrase *Bie.le.fel.der Fuss.ball.spie.ler* (= Bielefeld football player).

Metrical Phonology has formulated well-formedness criteria for metrical trees and grids, thus not any rhythmical structure is allowed. Especially in the framework of Optimality Theory, universal constraints expressing general preferences of the world's languages. The most important constraints subsume the following phenomena:

- prosodic units are marked prominent at the left and right edge
- syllables are parsed into rhythmical feet (= anacrusis and other "extrametrical" syllables are avoided)
- rhythmical clashes (adjoining accents) and lapses (sequences of nonprominent beats) are avoided (cf. Figure 2.25)

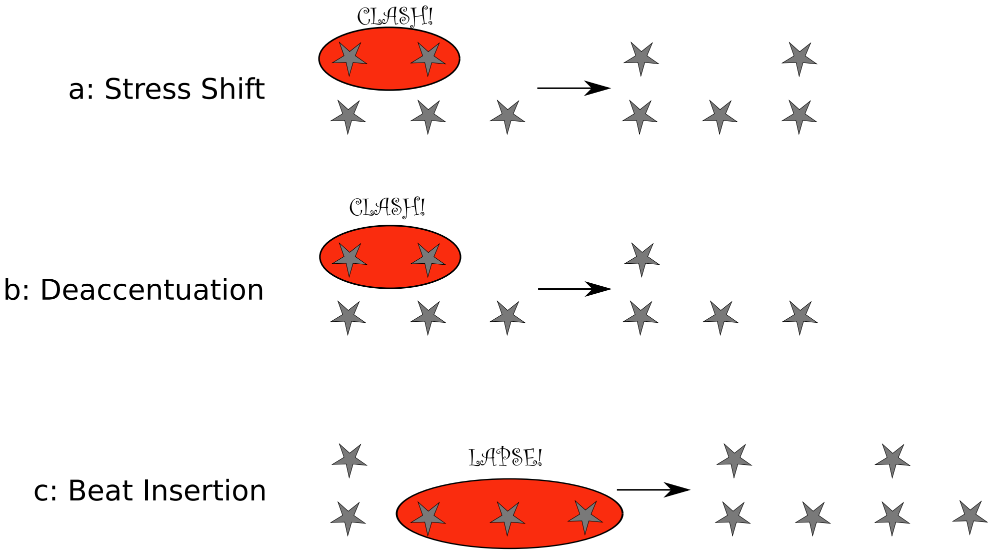


Figure 2.25: An illustration of the fundamental euphonic principles known in Metrical Phonology. Rhythmical clashes are avoided either by a *stress shift* onto a neighboring syllable (a) or by deaccentuation of the clashing syllable (b). Rhythmical lapses are avoided by inserting an additional stress on a syllable (c) that would have been nonprominent under different circumstances.

The constraints (or rules in Generative Phonology) are formulated for each level of the language specific *prosodic hierarchy* (Selkirk (1984); Nespor and Vogel (1986)) (cf. Figure 2.26), i.e. different preferences may be formulated for the level of the prosodic word, the prosodic phrase, the prosodic foot etc. The theory of a prosodic hierarchy has been widely accepted in phonological theory, although some of its standard assumptions have been subject to considerable debate, most prominently the non-recursivity of prosodic constituents which has been manifested in the *strict layer hypothesis* (Nespor and Vogel (1986)). For example, this hypothesis states, that an intonation phrase cannot be split up into one or more intonation phrases but only into one or more prosodic phrases which are at a directly subordinate level. Also, it forbids that certain parts of a structure "leave out" a level, e.g. an anacrusis would not be allowed in this point of view. This strict interpretation certainly poses problem for many analyses.

In	Pa	ki	stan	Tues	day	is	a	ho	li	day	*syllable*
(In)	(Pa	ki)	(stan)	(Tues	day)	(is)	(a)	(ho	li)	(day)	*foot*
(In)	(Pa	ki	stan)	(Tues	day)	(is)	(a)	(ho	li	day)	*pros. word*
(In	Pa	ki	stan)	(Tues	day)	(is	a	ho	li	day)	*pros. phrase*
(In	Pa	ki	stan)	(Tues	day	is	a	ho	li	day)	*int. phrase*
(In	Pa	ki	stan	Tues	day	is	a	ho	li	day)	*utterance*

Figure 2.26: An illustration of the different levels of the prosodic hierarchy as suggested by Selkirk (1984). This view of a prosodic foot differs from a more phonetic approach, e.g. as suggested by Abercrombie (1967), where a foot covers the range from one stressed syllable to the beginning of the next stressed syllable. In this book, we follow Abercrombie's definition.

Depending on a language specific constraint hierarchy, some constraints are obeyed and others are violated. In a language where right edge marking dominates left edge marking, may in effect sound more iambic than a language where left edge marking is typically obeyed. A language which strictly obeys the constraints concerning rhythmical clashes and lapses, the rhythmical structure will be strictly alternating between prominent and non-prominent fundamental beats. The lack of parsing syllables into rhythmical feet may lead to phenomena such as *anacrusis* where parts of the speech chain do not belong to rhythmical entities and are thus produced in a rhythmically uniform, unstructured manner (cf. section 2.3.5.1). Also, in case a phenomenal accent is missing on the phonetic surface, a native listener may still perceive the alternating sequence because it is very dominant in her native phonology. In general, one can say that the constraints formulated by phonological theories provide further support for the grouping principles formulated in the previous sections.

These phonological approaches to rhythm unfold furthermore the key function of rhythm that was already put forward in chapter 1.1: Rhythmical preference rules or constraints are developed by a linguistic community in such a way as to support the listener in the cognitive processing of an utterance by two fundamental processes:

- Boundary marking at the relevant levels of the prosodic hierarchy or other relevant linguistic levels (morae, syllables, prosodic words, morpheme boundaries, syntactic phrases, "sense units", utterances), thus simplifying parsing.
- Prominence marking in order to guide the listener's attention to those units most relevant for the communicative goal, thus making the perception and production more economical.

Not only for speech, well formedness criteria for rhythmical groups and structures have been formulated. In musicology, the highly influential book by Lerdahl and Jackendoff (1983) also postulates rhythmical structures typical for Western classical music which show a lot of resemblance to the metrical patterns described by phonologists. Despite the concentration on Western classical music it is likely that similar structures can be found in other musical genres as well - rhythm certainly is a mechanism used in all cultures. They build up a set of *well-formedness rules* describing possible rhythmical structures and a set of perceptual *preference rules* operating on well-formed but still rhythmically ambiguous structures. Schreuder (2006) points out that Lehrdahl and Jackendoff's *preference rules* could be called predecessors of the basic idea of the Optimality Theory-framework developed much later. Both operate on several possible output forms (here: several possible interpretations of rhythmical groupings and structures) and in both frameworks preferences rules or constraints are violable - however, Optimality Theory claims that constraints can only be violated if a higher ranked constraint is thus satisfied. Lehrdahl and Jackendoff's basic rules for analyzing the hierarchical structure of musical rhythms resemble many rules for grouping and structure postulated earlier. The most important distinction probably is that they try to differentiate between rules for grouping, metrical structure, and so-called *time-span reduction* which introduces a hierarchy on the entire musical piece. However, grouping and metrical structure assignment are difficult to seperate practically, since metrical structure usually indicates the beginning or end of a group, so that metrical and grouping structure are developed hand in hand. The time-span reduction process serves to identify a single, structurally most important event within each group on the various hierarchical levels. This event is then called the *head* while the rest of the group is called the *elaboration*. The result of this process is the

> listener's organization of pitch events into a single coherent structure, heard as a "hierarchy of relative importance".(Lerdahl and Jackendoff, 1983, 106)

This idea goes hand in hand with the metrical tree concept formulated by Metrical Phonology where prosodic events split into a strong "head" and a weak part. Indeed, the tree structures produced by the different approaches to illustrate a metrical hierarchy look similar. However, Lerdahl and Jackendoff's trees resemble prominence more straight-forwardly by using right- and left-branching structures. The longer branch always marks the (strong) prosodic head (see Figure 2.27).

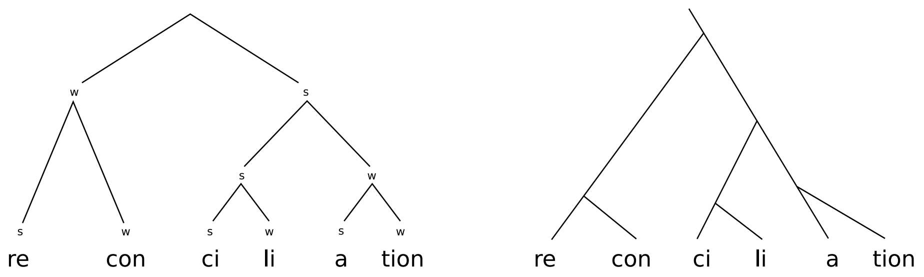


Figure 2.27: The above example illustrates the difference between metrical trees constructed according to Metrical Phonology (Liberman and Prince (1977)) and according to Lerdahl and Jackendoff (1983) for the word *reconciliation*.

Due to this direct encoding of perceived prominence it does not come as a surprise, that McAngus Todd (1994); McAngus Todd and Brown (1996) claim that Lehrdahl and Jackendoff's tree structure illustrations give quite an adequate illustration of rhythm perception — they base their line of argumentation on their own model of auditory perception which pays special attention to highly prominent parts in the auditory input, e.g. vowels. In their model, these events leave longer traces in the auditory memory which leads to a prominence colouring of subsequent events. When plotting their data, the resulting structures show resemblances to Lerdahl and Jackendoff's model (cf. Figure 2.28).

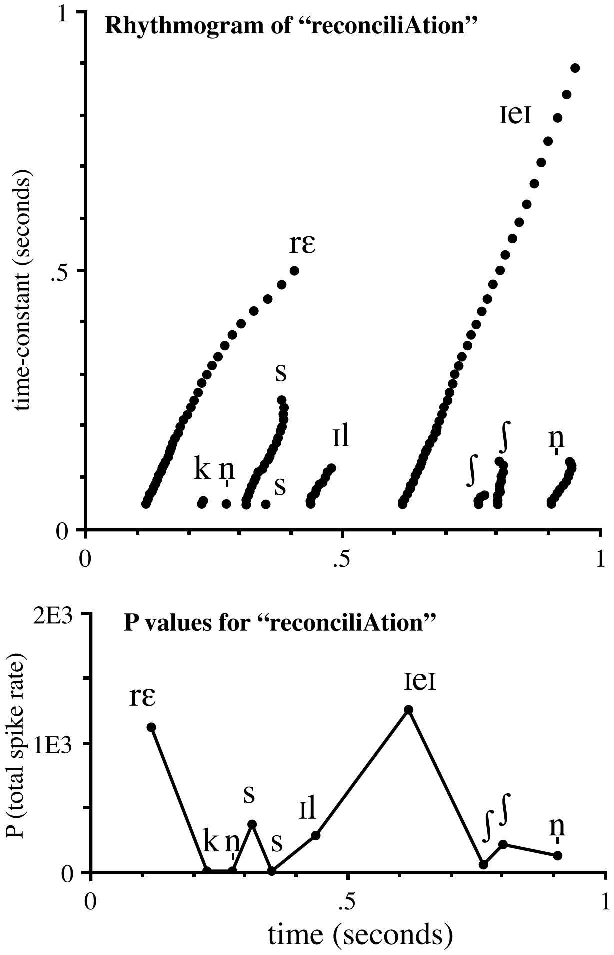


Figure 2.28: The above example of Lee and McAngus-Todd (2004) illustrates the auditory prominence pattern of the word *reconciliation*. The prominent vowels leave longer traces in the auditory memory than the less prominent consonants thus reflecting their relative rhythmical weight. The main stress clearly sticks out in terms of prominence. The "skewness" resulting of the memory traces of highly prominent auditory events leads to a resemblance of the metrical trees introduced in the work by Lerdahl and Jackendoff (1983).

For the moment, we leave it up to further research whether the hierarchical structure of rhythm is better described with classical metrical trees or the "skew" structures put forward in musicology. Obviously, both musical and linguistic rhythms show similar structures of hierarchical patterns. For the purpose of further analyses of rhythm it remains a striking coincidence that very similar theoretical models have been developed both for phonological and musicological purposes. Both approaches closely resemble the fundamental principles of rhythmical organization.

2.4 Rhythmic Speech Production

It seems intuitively clear that the phenomenon of perceptual rhythm is closely related to motor skills and motor behaviour or motion in general. Many of the metaphors used in everyday language suggest this: E.g. we describe rhythms as "stamping", "hand-chopping" or "driving", musical genres are called "swing", "beat" or "rock", singers tap their feet or clap their hands in order to keep up with the rhythm. The relationship between music and dancing has always been a close one: There is no self-evident reason why one should not just dance or watch a ballet without simultaneously listening to music unless one accepts the close connection between both performing arts. Thus, trying to establish a close connection between motor behaviour and perceptual rhythms in language, speech and music has always been a key concern in rhythm research — especially with Gestalt psychology that stresses the link between rhythm perception, time perception and motor behaviour, e.g. (Katz, 1948, 129) reports that subjects' writing becomes faster when they are asked to increase the size of their written letters. Recent fMRI-studies in speech rhythm perception (Geiser et al. (2006)) revealed that subjects who are asked to pay attention to an utterance's rhythm show activation in the supplementory motor area (SMA), a neural region which is not a classic language perception area. However, it is known for being involved in motor preparation and attention tasks. Unpredictable rhythmical patterns show supplementary high activation in the planum temporale, which has been identified as playing a major role in the temporal processing of auditory signals. For the timing of motor tasks, there is consensus that this these are somehow associated with the cerebellum (Max and Yudman (2003)) which has also been assigned a role in the rhythm-related speech disorder of stuttering

(Howell (2004)). Summing up, the connection between — predictable — rhythms and motor activity receives more than just anecdotal support. While is has become clear that a 1:1 identification between rhythm and rhythmic production in speech is oversimplistic and that perception has major impact of rhythmic structures as well, this section tries to establish to what extent speech and language rhythm and speech production can still be linked to each other.

2.4.1 Fundamentals of Rhythmic Production, Predictability and Variation

The vast amount of studies on rhythm production have been based on finger tapping experiments. This section points out some of the most important findings of these studies. Later, their potential relevance for rhythmic speech production will be discussed.

A key issue of rhythmic motor behaviour is its predictability. It is easy for people to tap to regular beats when these can be estimated in advance and many production studies (mostly tapping studies) confirmed the effect of anticipation, i.e. subjects who are asked to tap to a particular rhythm systematically tap "too early", i.e. the subjects' motor commands were initiated long before they can hear the acoustic stimulus, thus they must have anticipated it (see Aschersleben (2000) for an overview). This anticipitory effect is less strong but remains stable for musically trained subjects or even professional musicians. Thus, production experiments deliver further evidence for our rhythmical guidance by an abstract time keeper controlled by previous experiences.

However, rhythm production is highly constrained, i.e. not any complex but repeating "rhythmical" pattern can be produced. First, it seems to be difficult to produce two interacting rhythms when these are not related to each other in whole numbers, i.e. sharing beginnings: Imagine you are tapping a rhythm with your right hand and an accompagnying rhythm with your left hand. Typically, one hand will tap faster than the other, but sometimes both fingers will tap simultaneously, i.e. the faster rhythm fits four times into one interval of the slower rhythm. One could argue, that such polyrhythms need to be *in phase* for us to produce them. This is explicable by the findings that during rhythm production, humans are unable to produce two different rhythms independently. Rather, they relate each rhythmic production, e.g. a finger tap, to the previous one — always using the short interval as fundamental "counting" interval (Krampe et al. (2000)). Naturally, this relationship only works if the interval of the slower rhythm is an integer multiple of the shorter interval. A second fundamental finding is that tempo and produced rhythm interact. Rhythms become less complex with an increasing tempo (Peper et al. (1995)), e.g. in tapping experiments with two hands a 3:8-rhythm may become a 1:2 one with increasing tempo. However, the overall temporal variability of rhythmical intervals is higher at slower tempos (Ivry and Hazeltine (1995); Krampe et al. (2000)). A third finding is that the duration of any interval, e.g. finger taps, is very much influenced by the duration of the previous interval. This phenomenon has been described first in the highly influential model by Wing and Kristofferson (1973). In their experiment they had subjects tap to a given rhythm. After a rehearsal phase, subjects had to continue tapping in the trained rhythm. In order to account for the variability in the tapping data, Wing and Kristofferson (1973) propose a model constisting of an external timekeeper and motor behaviour. Both parts contribute to the variability, but they do so in different ways: It is assumed that rhythmical mistakes caused by the motor timing component are compensated for in the upcoming interval, while rhythmical variability caused by the external timekeeper shows more random behaviour. Thus, motor timing variability causes a negative lag1 autocorrelation[27]. The remaining variance is caused by the external timekeeper. The model has later been extended to several timekeepers which are ordered hierarchically (Vorberg and Wing (1994)). An alternative view of motor timing is based on coupled oscillator models, also known as the Haken-Kelso-Bunz-model (Haken et al. (1985)). While it is clear that rhythms can be complex and are correctly described as the interaction of two independent rhythms, humans are unable to produce any two rhythms independently. In a study by Krampe et al. (2000), cf. Figure 2.29, professional pianists were asked to tap a complex rhythm that can be described as consisting of two independent rhythms. Each of these showed isochronous intervals and was easy to tap independently. The subjects performed one rhythm with their left and the other rhythm with their right hand and mastered the task sucessfully. During performance, the subjects could be shown to relate the timing of each individual

[27] For an explanation of autocorrelation analyses in rhythm research, cf. Sections 2.4.1.2 and 3.2.1

tapping event relative to the previous one, even though that was performed with the other hand. No subject showed a performance that indicated an independent rhythm production by each hand. Thus, the interaction constraints of hierarchically ordered rhythms must be taken into account by rhythm models.

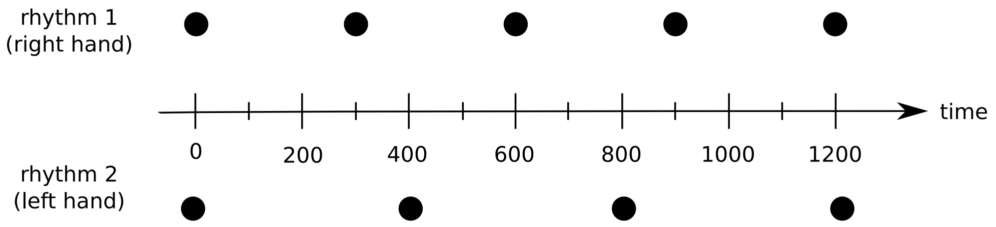


Figure 2.29: Illustration of the production experiment by Krampe et al. (2000). Professional pianists were asked to produce two independent rhythms, one with each hand, constituting a complex rhythm. The rhythms are characterized by being coupled at some points, i.e. in some cases, both hands have to perform a tap simultaneously.

Altogether, it still seems to be a matter of ongoing research whether the relative timing model by Wing and Kristofferson (1973) or the Haken-Kelso-Bunz model is better suited for the description and prediction of rhythmic motor timing events. Also, it appears to be yet unresolved whether a cognitive timekeeper — if it exists — is shared by the production and perception system. It is uncontroversial, however, that the motor timing heavily relies on some kind of proprioceptive or auditory feedback, since the timing of motor commands is heavily impaired if the auditory feedback is delayed, e.g. also people without any obvious speech production impairments start to stutter, a phenomenon known as the Lee-Effect.

Summing up, we can conclude that:

- The standard assumption in rhythm production models is that humans orientate their rhythmic behaviour on an internal timekeeper and attempt to adjust their motor productions to its prescribed intervals.

- In tapping experiments, these intervals are rehearsed and tend to be near isochronous or — in polyphonic rhythms — regular in the sense that the shorter intervals are related to the longer invervals by whole numbers.

- The deviations from these isochronous patterns are easily predictable. Our rhythmic timing in production shows a negative lag1 autocorrelation, thus, humans tend to produce an alternating timing pattern.

2.4.1.1 Fundamental Timing Constraints on Rhythmic Speech Production

With regards to the duration of production constraints of fundamental beats, Lenneberg (1967) suggested a neurally determined fundamental timing frequency of 4-6 Hz. Such a window corresponds to a duration of 160-250 ms. This corresponds highly with the average syllable duration across many languages (see Figure 2.1.6) and is furthermore strikingly similar to the average windows of temporal integration in auditory perception (see section 2.1.2).

Another interesting time frame is the typical duration of interstress intervals which — in the sense of Abercrombie (1967) — is coexistent with a prosodic foot. For this sequence, many researchers have found a reference value of approximately 400-600 ms (see Figure 2.1.6 and the discussion in (Eriksson, 1991, 10-11)). It has already been mentioned in section 2.1.4 that it seems to be more than just a coincidence that this timing window is so close to the duration that has been identified as *temporal instant* in auditory perception research. Linguistic entities constisting of something stressed are likely to convey important content while syllables that are less stressed are likely to carry less semantic weight, e.g. they may be affixes, function words etc. Thus, a speaker ties a neat package for the listener by squeezing one important item into a window of one temporal instant each. The interval of 400-600ms also plays a major role in human's life of motor behaviour, as Fraisse (1982) regards it as both the average personal tempo of rhythmic behaviour and preferred tempo in rhythmical perception. It may be more than just a coincidence that it also relates to basic human movements such as walking and is known as the standard tempos (*andante*) in Western classical music.

2.4.1.2 Compensatory Shortening and Isochrony

Early models in speech rhythm research, most prominently the ones by Pike (1945); Abercrombie (1967) were also speech production oriented by claiming the difference between *stressed timed* and *syllable timed* languages to be explicable as

> a periodic recurrence of *chest pulses* acompanying each syllable in a syllable timed language, or of *reinforced chest pulses* accompanying each

> stressed syllable in a stress-timed language." (Dauer, 1983, 51)

For stress-timing, this notion has the consequence of the phenomenon of *compensatory shortening*. Compensatory shortening is the effect that with an increasing number of unstressed syllables between two stressed syllables, the unstressed syllables need to be more and more compressed in order to keep up foot isochrony (see Figure 2.30). Pike (1945) states that

> uniform spaces of stresses [...] can be achieved only by destroying any possibility of even time spacing of syllables. [...] The syllables of longer [rhythm units] are crushed together, and pronounced very rapidly [...]. This rhythmic crushing of syllables [...] is partly responsible for many abbreviations — in which syllables may even be omitted entirely — and the oscuring of vowels.

The idea is deeply rooted in traditional poetic metrics and still used in English textbooks for L2 learners (see section 1.1).

Tim **walks**.
Tim has **walked**.
Tim has been **wal**king.
Tim has not been **wal**king.

Figure 2.30: The sentences above illustrate the effect of compensatory shortening, i.e. the effect to shorten the syllables in between two stressed ones (here: boldface) as a function of the number of interstress syllables. Ideally, the speaker thus keeps up *foot isochrony*, e.g the duration between the two stressed syllables *Tim* and *walk*(s/ed/ing) will stay the same despite an increasing number of interstress syllables.

Compensatory Shortening on foot level has been regarded as typical feature of stress timed languages but not for syllable timed ones. After the notion of a more or less strict isochrony, which would be the result of a "perfect" isochrony was ruled out by various studies (see section 1.2.2) many researchers moved on to show that the effect still existed — the reasoning was that since speakers are forced to obey the phonetic form of the utterance, strict isochrony is almost impossible to uphold. Since the phonetic form of two adjacent feet is usually very different and since the articulatory movements cannot be accelerated beyond a certain limit, isochrony should be regarded as an ideal rather than a phonetic production reality. However, since speakers of stress timed languages strive for isochrony, it is believed that typical phenomena of such languages are phonetic reduction, especially vowel reduction and elisions as consequence of hypospeech (Lindblom (1990)) in the interstress intervals. Indeed, stress-timed languages are often characterized by vowel reduction in unstressed syllables and the occurrence of the neutral vowel [ə] while syllable timed languages often do not show such phenomena. As a consequence of this reasoning, researchers examined the level of duration increase as a function of interstress syllables. With a larger number of interstress syllables, the duration of that interval still ought to increase, but the syllables should become shorter. This indicates the speakers attempt to uphold isochrony. Such tendencies were indeed found for English (e.g. Fowler (1977); Jassem et al. (1984); Fourakis and Monahan (1988)) and German (e.g. Kohler (1982, 1983); Ikoma (1993)) but not for syllable timed Spanish (e.g. Font and Mestre (1991)). In a slightly different approach, Cutler (1980) collected elision and speech error phenomena as production evidence aiming at upholding the perceptual impression of isochrony in English. The degree of compensatory shortening has also been regarded as an indicator for L2 competence (Bond and Fokes (1985)). Compensatory shortening has been examined on various levels of the prosodic hierarchy like the mora in Japanese (e.g. Ota et al. (2003)) and is used as evidence for rhythmic prosodic organization. Thus, its study can be called a relaxed approach to the concept of isochrony where a native speaker attempts to keep the most relevant timing unit as isochronous as possible but is often forced to deviate from it, due to other production constraints.

However, alternative suggestions have been made to explain the effect of compensatory shortening. In his in-depth study of (stress timed) Swedish, Eriksson (1991) claims that compensatory shortening is rather explicable as an effect of language-specific behaviour of stressed vs. unstressed syllables than as the effect of compensatory shortening across all syllables contained in an interstress interval. That way, he takes up a much older approach by Delattre (1966), who found a higher ratio of stressed vs. unstressed syllable durations in stress timed languages. Based on the data described in Dauer (1983), he calculates linear regressions where the interstress interval durations are a function of the number of syllables. He finds that the slopes of the regression are very similar across the various stress-timed and syllable-timed languages and the regressions coefficients are very high (> 0.9 for all languages). From the almost identical slopes, he concludes that the amount of duration added to an interval with increasing syllables is almost the same, independent of the language. However, he finds a systematic difference between stress-timed and syllable-timed languages, namely the first coefficient of the regression equation, indicating longer stressed syllables in stress-timed languages. As a rough approximation for calculating the difference between both rhythm types, he gives equations[28], where I is the interstress interval duration and N is the number of syllables contained in the regarded interval:

$$stress\ timing : I = 232ms + 99.5ms * N \tag{2.5}$$

$$syllable\ timing : I = 112ms + 109.5ms * N \tag{2.6}$$

He also concluded that some phenomena that appear as compensatory shortening are mere artifacts of an influence of final lengthening — he finds final lengthening not limited to the end of intonation phrases[29]. Very similar data were reported by Nooteboom (1998) for Pseudo-Dutch nonsense productions. He reports that the domain of compensatory shortening is mostly the stressed syllable, while unstressed syllables are affected less. From Nooteboom's results, we might draw conclusions that fast speech may affect the stressed syllables in a similar way to an increase in the number of interstress syllables. If this were the case, fast Dutch would become more similar to a syllable timed language showing less differences between the stressed and unstressed syllable. We have to be cautious to draw too many conclusions from Eriksson's data, since this is based on mean durations which certainly are not representative for individual utterances, while Nooteboom used nonsense and reiterant speech. However, their results certainly point out that one should not jump to conclusions when measuring durations without taking into account rhythmical group internal relative durations. The instability of stressed syllable duration may be an important factor explaining the rather unexpected results by Bouzon and

[28] The equations presented here are calculated from the average values he gives for several stress and syllable timing languages.

[29] It has been reported for other languages such as French (Keller et al. (1993); Keller and Zellner (1996)) and Japanese (Ota et al. (2003)), that final lengthening is not limited to intonational phrases but may also occur on word or foot or stress group level.

Hirst (2004); Hirst and Bouzon (2005) who could not find evidence for a systematic durational increase in English stressed syllables. Instead, they explained perceived stress as the result of durational increase at the beginning of a foot relative to the preceding anacrusis syllable(s) (also cf section 2.3.5.1). Their results may have been influenced by a lot of stylistic variation resulting in a increased instability of a stress related lengthening effect. Further evidence for the rhythmical importance of the relationship between the stressed and the unstressed syllable can be found in a study by Bröggelwirth (2007) on the different durational properties in German poetic meter. The major indicator for the difference between an iamb, trochee and dactyls in German turned out to be the relative durations between the stressed and the unstressed syllables in a poetic foot.

Based on Pike's reasoning, one of the key evidence for a language being stress timed or syllable timed would be the presence or absence of reduction phenomena. Indeed, this coincides with Dauer's findings, since many, if not most languages that have phonetic reduction are classified as stress timed. However, some languages apparently do not satisfy this hypothesis, e.g. Catalan shows phonetic reduction but is traditionally regarded as a syllable timed language. Summing up, any study on compensatory shortening needs to be carried out very carefully and ought to take into account the durational properties within the interstress interval or foot very carefully.

While compensatory shortening remains a difficult concept and it is yet unsolved whether or not it creates the impression of isochrony in the listener, studies by Port et al. (1995); Cummins and Port (1998); Cummins (2002) have successfully shown, that speakers are indeed able to produce stresses in a more or less isochronous fashion. They examined native speakers of American English in an experimental paradigm called *speech cycling*. They asked speakers to align stressed syllables of simple phrases[30] with high and low beats given by a metronome. The high beats indicated the beginning of the first stressed syllable and the interval between the two high beats was modified in each experimental session. In another task, speakers were asked to produce word list speaking in synchrony. They could show that subjects were indeed able to align stressed beats in a more or less isochronous fashion.

[30] The phrases consisted of the pattern "X *for a* Y". The X and Y slots were filled with monosyllabic, content words of the structure CVC beginning with voiced consonants, e.g. *beg for a dime*.

The main effect they found out was that there were dominant anchor points in the interstress intervals that subjects aligned the perceptual centers of stressed syllables to. The main and most stable anchor point lies halfway, two more anchor points lie more or less halfway between the beginning or end and the middle of the prosodic phrase, i.e. the anchor points divide the phrase into simple integer intervals around 1/3, 1/2 and 2/3 of the prosodic phrase. The anchor points remained stable even if the phrase duration (the interval between two high metronome beats) was modified. Their finding is interpreted as evidence for attractors of basic oscillators governing simple motor processes similar to the proposal made in the Haken-Kelso-Bunz model described earlier (Haken et al. (1985)). According to their view the oscillator governing foot timing is coupled to the oscillator governing the timing of prosodic phrases. It is argued that the alignment becomes easier detectable under laboratory conditions, since their task involved a rehearsal phase during which subjects could tune their utterances to a given phrase duration. In Port (2003) it is furthermore argued that the identified attractors are tuned to points in neurocognitive attention thus aligning rhythmically salient events to points of the listener's attention. Such a strategy would appears to be even more useful given that an increase in attention leads to subjectively perceived longer durations (cf. Section 2.1.1). Within the same experimental paradigm, Tajima (1998); Tajima et al. (1999) examined whether the attractors found in speech cycling are different across rhythmically different languages and detected differences of temporal adjustment which can be explained by different phonological properties of the languages involved. In general, the argument is that the stability of an attractor provides evidence for the importance of the examined level of prosodic organization. Furthermore, the phonological entity that tends to be organized around the attractor is regarded as phonologically salient. These salient entities can also vary across languages.

In a follow-up study Cummins (2005) examined whether there are any order effects on interval durations. He used production data of earlier experiments where people had produced word lists in a more or less isochronous fashion. Finger tapping experiments, most importantly in Wing and Kristofferson (1973) had revealed systematic deviations from an isochronous leading metronome rhythm which were explicable as effect of immediately adjacent taps. The tapping data had revealed a negative autocorrelation at lag1[31]. Cummins word lists showed a strikingly similar effect: Intervals between adjacent words are made shorter if the preceding interval was rather long and vice versa. This effect has been explained by Wing and Kristofferson as evidence for compensation for production mistakes (see page 98). However, Cummins also found systematic positive (!) autocorrelations at longer distances (lag 7) which cannot be accounted for by the Wing and Kristofferson model. This long distance dependency appears to be related closely to hierarchical organization of rhythm, since it relates the first and the last items of the produced word list.

2.4.2 Implications for Speech Rhythm Research

Although there exists relatively little systematic research on rhythmical timing in speech (not taking into account duration analyses related to the isochrony hypothesis), the studies seem to hint at a close connection between the rhythm related processes of perception and those of production. The main results of these studies can be summarized as follows:

- The neural areas linked to rhythm perception are also responsible for motor coordination and planning tasks. Neural areas responsible for temporal auditory processing are less active if the rhythm is regular and thus highly predictable. This is one more indicator for the function of rhythm in minimizing the cognitive load.

- It is striking that the fundamental timing intervals related to rhythm production are almost identical to the intervals relevant for auditory temporal processing, i.e. fundamental production frequencies correspond closely to average durations of syllables and prosodic feet across many of the word's languages.

[31] Autocorrelation analysis is a method to examine time series. Time series are sequences of observations in the course of time, e.g. in a series consisting of events $N = 7$ can be called $X_t(t = 1, ..., 24)$. A typical feature of time series is that the observations are correlated, i.e. an event in a time series is influenced by preceding events. The correlation between an observation t and the immediately preceding observation $t-1$ is expressed by the autocorrelation coefficient at lag 1. A correlation between an observation t and an observation at distance k is expressed by the autocorrelation coefficient at lag k.

- Rhythmic production models show a systematic negative correlation between adjacent events, even if a subject attempts to produce an isochronous sequence. This is explained as a compensation for previous alignment mistakes, resulting in an alternating sequence. Thus, alternation is the normal way to produce an "isochronous" rhythm and the way that alternation is somewhat expected and taken as the norm in perception[32]

- The domain of compensatory shortening tends to be restricted to the stressed element within a foot.

- Subjects are able to produce more or less isochronous feet after rehearsal by adjusting stressed syllables to the beats of a metronome. However, they do make systematic "mistakes": when tuned to a specific phrase duration, their stresses fall at predefined anchors covering 1/3, 1/2 or 2/3 of the framing intonation prase.

2.5 Conclusion: Building Blocks for Rhythm Research

The many details assembled in this chapter can be taken together as indicators that rhythm is perceived the way it is perceived for a wide number of reasons. Such "conspiracies" are not new in rhythm research. They were proposed before by Dauer (1983); Bertinetto (1989) and Auer (1993). However, their previous proposals all tend to regard of a number of *per se* unrelated *phonological* properties *somehow* leading to the impression of a particular rhythmic impression. This chapter was an attempt to go beyond this in trying to explain *how* various factors might trigger the known effects. Although it has been found that phonology puts heavy constraints of the rhythmic pattern of a language, it should be possible in principle to somehow generate different rhythmic impressions in any language, although such attempts may be in conflict with the languages' orthophony. E.g. such a deviating rhythm might sound like a foreign accent.

We have shown in the present chapter that the perceptual phenomenon described as rhythm is the result of a complex of cognitive auditory perception strate-

[32] E.g. as it is encoded in phonological constraints avoiding "stress clashes" or sequences of unstressed syllables does not come as a surprise.

gies, production and perception constraints, top-down expectancies and language specific preferences. Keeping this in mind, it is possible to derive certain expectations for a language's rhythmical structure. It also helps us to derive hypotheses for empirical work on rhythm and may prevent us from oversimplifications, such as "rhythm is..." or the trap of equalizing physical durations with perceptual ones. In order to work on rhythm, much has to be taken into account to get hold of the underlying processing differences that emerge via the different languages or varieties spoken. Up to now, many fundamental questions, most importantly with regards to duration perception in speech have not been solved yet. This makes it difficult for us to interpret any durational data we get (see the discussion in Lehiste (1977)). With the remaining myriad of unclarities and vaguenesses, it is difficult to postulate a straightforward line of research for speech rhythm phenomena. However, a number of factors need to be taken into account that play a demonstrable role in rhythm perception and production.

- Since cognitive processing places constraints on the perception of rhythmical events, most strikingly in the relevance of existing timing windows for temporal integration (roughly syllable sized), temporal present (roughly foot sized) and rhythmical pattern integration (roughly utterance sized), it is important to find out more about the phonological properties of syllables, feet and intonation phrases of the language or variety under examination. These factors are mostly concerned with the question how rhythm is linked to the prosodic hierarchy in the given language. Evidence for this can be derived from phonological and phonetic analyses but also from poetry. This knowledge may make it easier to detect the language specific rhythmical characteristics, e.g the question whether the mora plays a role in a language's rhythmical duration pattern or not. Also, this knowledge can give insights into metrical expectancies based on phonological knowledge, i.e. where do listener's place their attention? The rhythm study itself then may indicate how the language implements these phonological properties phonetically.

- It should also be looked at the distribution of stresses or accents in the language in question. Does the language have fixed or free lexical stress? Does it have lexical stress at all? How is it distributed?

- Any examination of rhythm ought to take into account a language's phonotactic complexity which may lead to a certain increase in variation, since syllable structure poses the most heavy constraint on syllable duration. However, it is not sufficient to simply define whether or not a language possesses the possibility to produce complex syllables or not. Rhythm needs to take into account the relative frequency of these more or less complex syllables as well as their sequential distribution. Else, a phonology based rhythm description may lead to wrong results.

- With regards to acoustic correlates of rhythm, duration seems to be the fundamental indicator of rhythmical structure. It serves to indicate both rhythmical grouping — usually by final lengthening — and group internal structure — usually by accentual lengthening. Even though lengthening indicating finality (=iambic) is apparently stronger than lengthing indicating the beginning of a group (= trochaic), this may cause a certain amount of ambiguity. Thus, other acoustic parameters ought not to be forgotten when studying rhythmic phenomena, especially those indicating prominence: E.g. when examining a tone language, one might concentrate on intensity related correlates as additional cues to rhythmic structure.

- It is of essence to look at tempo related features, since tempo influences the number of fundamental rhythmic beats that can be grouped into feet, i.e. it indicates the absence of presence of pure *metrical accents*. Also, there exist many tempo related factors leading to a predominance of trailer vs. leading timing. Furthermore, we know that tempo is linked to information density which can be roughly measured in phones. In a phonotactically less complex syllables, more syllables (= beats) need to be produced in order to communicate the same amount of content, possibly leading to an increased speaking rate. Since variability is perceived less at high rates, it is a main factor in rhythm analyses.

- The relative timing at the boundaries between unstressed and stressed syllables apparently play a major role in order to account for effects such as compensatory shortening. Also relative timing may provide valuable insight into the metrical expectancies a listener may extrapolate based on what he has heard previously. Thus, a rhythm study should not only concentrate on global variability, since this may oscure many factors of relevance. Instead, relative timing patterns across an utterance should be paid attention to.

- Obviously, listeners pay attention to certain points in time where rhythmically important (= stressed) events are likely to occur. During production, speakers align stress with certain points within an intonation phrase. Thus, important speech events are aligned with points of time where the listener can guide his attention to. These attractors apparently are chosen relative to intonational phrases. The study of such alignment constraints may provide valuable insights into metrical expectancy and information structure related speech timing.

- Physical isochrony does not imply perceptual isochrony and vise versa. Instead, a high degree of variability rather causes listeners to perceive deviations less, while rehearsal of listening to isochronous events causes listeners to perceive an upcoming durational variation better and stronger. Thus, the amount of local durational variation as such is not sufficient to deduce its perceptual relevance. However, the amount of variance across a longer stretch of speech may indicate the perceptual relevance of a local durational variation. Thus, speech rhythm studies need to look at both local and global variation.

The various factors involved in rhythm perception and production lead to a complex picture of interactions and influences (cf. Figure 2.31) all of which determine the level of perceivable variability and structure of metrical expectancy or *meter*. We believe that both factors mingle, maybe by weighing both factors based on certain external sources of influence into a cognitive percept one might call *speech rhythm*.

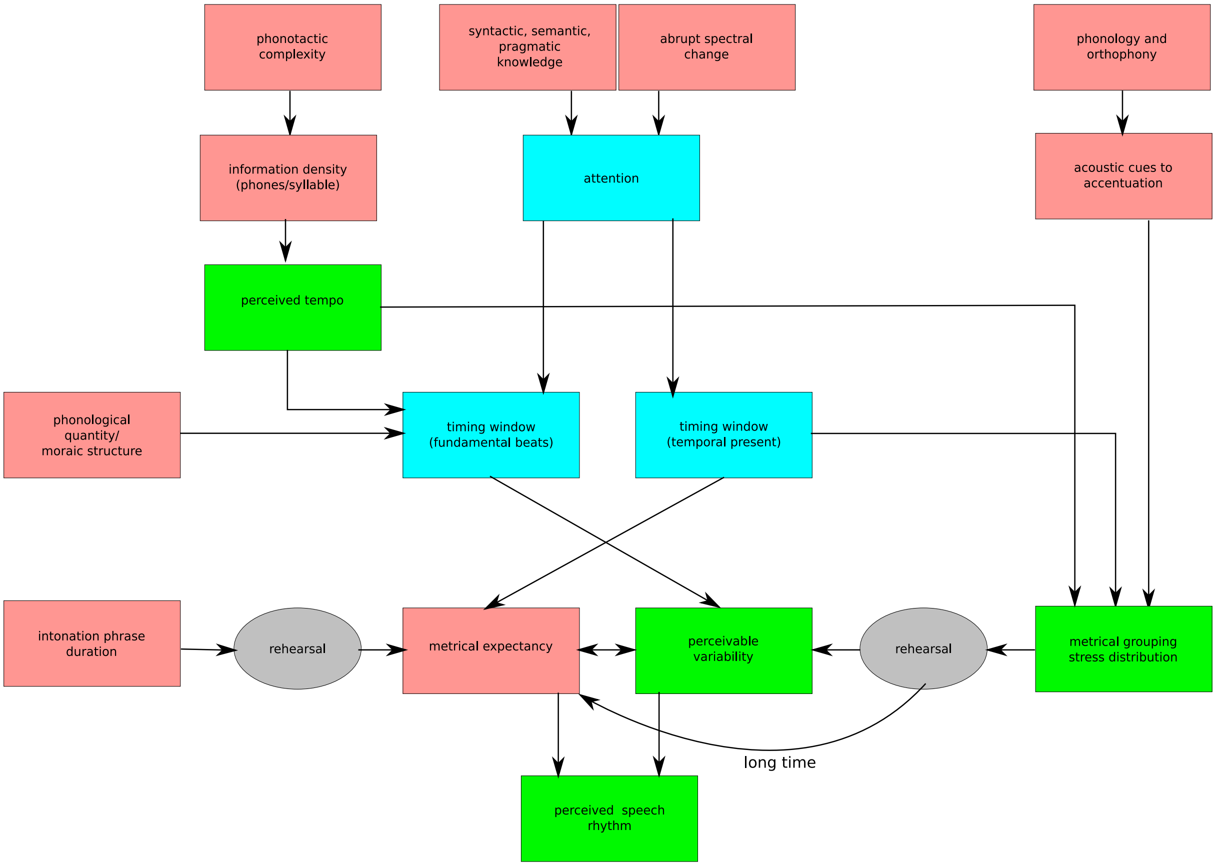

phonotactic complexity
syntactic, semantic, pragmatic knowledge
abrupt spectral change
phonology and orthophony
information density (phones/syllable)
attention
acoustic cues to accentuation
perceived tempo
phonological quantity/ moraic structure
timing window (fundamental beats)
timing window (temporal present)
intonation phrase duration
rehearsal
metrical expectancy
perceivable variability
rehearsal
metrical grouping stress distribution
long time
perceived speech rhythm

Figure 2.31: Rhythm perception is influenced by numerous factors. External factors, originating in phonological structure and the acoustic input signal, are indicated in red. The levels of perception are indicated in green. The blue boxed denote cognitive constraints of temporal processing, “our inner clocks”. The grey fields indicate that a particular influence is mostly reached after rehearsal, e.g. long-term or short-term tuning to a particular rhythmical structure. It is proposed that the finally perceived rhythmical structure is extrapolated by weighing metrical expectancy with perceived variability. The potential link between rhythm production and rhythm perception is not illustrated.

Chapter 3

Measuring Rhythm — A Critical Assessment

In this chapter, rhythm models and their pertinent ways to measure or characterize rhythms in language and speech will be reviewed. It is obvious and along model theoretic reasoning (Stachowiak (1973)) that any model will concentrate on those attributes of the modeled reality it assumes to be relevant. Thus, whatever a model builder believes to be the essence of speech and language rhythm will be taken into account. When assessing models of speech and language rhythm, we therefore need to pay attention to the attributes the researchers had in mind when building models and developing metrics quantifying the phenomenon of speech or language rhythm. For this reason, this chapter is organized around various major characteristics referred to in rhythm models.

One central phenomenon widely discussed in research has been the isochrony hypothesis already discussed in the previous chapters. Typological approaches have attempted to capture this phenomenon based on phonological features. Furthermore, various approaches to quantify such typological models have been suggested. These will be discussed in the beginning of this chapter. In the same section, various other models that are not explicitly built on typological grounds are also discussed, simply because they attempt to capture rhythmic phenomena by quantifying the degree of lacking isochrony, i.e. temporal variability. Thus, they can be regarded as being built on reasoning very similar to that of phonological models.

An alternative approach towards the quantification of rhythm has been attempted by researchers focusing on the distributional characteristics of rhythmical patterns. They tried to find evidence for the regularity of short-term and long-term alternations, rhythmical grouping, deceleration or acceleration across utterances.

Another method to quantify speech rhythm are hierarchical models. It is indeed the case that some of these models also take into account the notion of isochrony (e.g. O'Dell and Nieminen (1999); O'Dell et al. (2007)) and some of the typologically based models take into account various hierarchical levels of rhythmic description (e.g. Asu and Nolan (2005, 2006)). The fundamental difference between the two types of models is the following one: Models and metrics collected in Section 3.3 were largely built on the assumption that rhythm is best described in a hierarchical manner. Such a hierarchical approach *inherently* takes into account different levels of rhythmic-prosodic organization. The models described in Section 3.1 can certainly be used to investigate various levels of prosodic organization — but this aspect is *not* one of their built-in features.

3.1 Measuring Variability and Isochrony

As discussed in the previous chapter, no convincing evidence for what has been called the strict version of the isochrony hypothesis could be found so far, i.e. no researcher has shown evidence for morae, syllables or feet/stress groups being of equal length in any language which has been claimed to have mora-, syllable- or stress timed. Cauldwell (2002) pointed out that due to the many extra-rhythmic timing constraints each discourse poses on our speech production, the search for isochrony must be unsuccessful. He suggested that it may rather be the degree of irregularity that characterizes a particular rhythmical style of speech or language. In line with this thinking, the search for rhythm styles has rather gone into the direction of measuring durational variability: If pure isochrony cannot be found, then at least there should be less variability in syllable durations in syllable timed languages compared to stress timed languages. Some researchers have described variability on the level of phonology alone, probably having in mind the strong perceptual relevance of phonological patterns, most prominently Dauer (1983). Since such variability ought to have a phonetic realization as well, several proposals have been made with regards to the level of phonetic and phonological organization that this variability ought to be measured in. Another dispute regards the way the variability is best captured quantitatively. The various proposals will be described and critically reviewed in the following.

3.1.1 Identifying Typological Constraints

After the much quoted failure of phoneticians to find isochrony or stable variablity in speech data, Dauer (1983) took another approach in explaining the impressions of isochrony on syllable or foot/stress group level for different languages. By compiling phonological properties of various languages and comparing these with their pertinent rythm class, she was able to identify strong phonological constraints which she found sufficient to create the impression of a particular rhythm type (cf. Table 3.1).

syllable timing	*stress timing*
	Presence of accentual lengthening
Low phonotactic complexity	High phonotactic complexity
quantity distinction (vocalic, consonantal) possible everywhere	quantity distinction (if any) restricted to stressed syllables
presence of tone	
	Vocalic and consonantal reduction in unstressed syllables
fixed lexical stress	free lexical stress

Table 3.1: Overview of typical phonological features correlating with the impression of syllable or stress timing.

In his theoretical analysis of 34 languages, Auer (1993) suggests to complement Dauer's list with the following typical features of syllable timed languages:

- Few assimilations
- Vowel harmony possible
- If fixed lexical stress, then often realized weakly

- Geminates possible

Most of the proposed phonological features attributed to being responsible for rhythm class distinction are the ones directly connected to syllable duration. It is also apparent that the key distinction lies in the way, phonology strengthens the durational variability or isochrony. The following list discusses the various influential factors contributing to the impression of one or the other rhythm type:

- Accentual lengthening contributes to a larger distinction between stressed and unstressed syllables in the temporal domain, thus should highlight temporal variability at long distances.

- A low phonotactic complexity will necessarily lead to a relatively smooth durational variability of adjacent syllables. Since languages with a high phonotactic complexity usually allow for less complex syllables as well, a certain amount of variability is a natural consequence. If phonotactically complex languages attract stress, a widespread phonological phenomenon called *quantity sensitivity* this difference is enhanced even further through accentual lengthening.

- Interestingly, syllable timed languages often possess quantity distinctions in stressed and unstressed syllables and in consonants and vowels. This does create a significant amount of durational variability in syllables and at first glance, stands in contrast to the definition of rhythm class along the scale of more or less durational variability in syllables. This undermines both phonetic findings on lacking syllable isochrony and the finding of perceptual studies (cf. Chapter 2) that a constant amount of slight variability rather blurs the perception of these differences. While listeners of the respective languages definitely are able to perceive the quantity contrasts, the chain of beats realized as syllables is perceived as relatively uniform. If the quantity contrasts are restricted to the stressed syllables, as it is the case in stress timed languages, these further enhance the effect of contrasting stressed vs. unstressed syllables.

- The lack or presence of tone seems to enhance the lack or presence of reduction, since a tonal realization necessarily needs a certain amount of time. Thus, a language where each syllable is connected to a tone, may be less inclined to make a large durational contrast between stressed and unstressed syllables.

- Vocalic reduction makes a vowel draw nearer to the neutral vowel schwa. In consonantal reduction, the consonantal gestures do not reach their articulatory target, e.g. plosives become fricatives, fricatives become approximants. It is also possible that a segment is completely elided, because an articulatory gestures overlaps with other gestures in becomes inaudible or because it is deleted out of the articulatory plan altogether. Such reduction processes always aim at a high degree of hypospeech (Lindblom (1990)) which probably coincides with the larger information density in the syllables that are less affected by reduction, i.e. the stressed ones. Again, this will enhance the difference between stressed and unstressed syllables.

- A fixed lexical stress renders prosodic minimal pairs impossible. Thus, the position of lexical stress may play a completely different role in cognitive parsing in such languages. Obviously, this difference plays a role in the perception of durational variability of stressed and unstressed syllables, possibly based on the varying degrees of attention paid. Another factor may be that languages with a fixed lexical stress tend to indicate it less strongly (cf. Keller et al. (1993); Keller and Zellner (1996) for French and O'Dell et al. (2007) for Finnish).

The major distinction between the two rhythm types is along the two dimensions of phonotactic complexity/accentual lengthening and reduction/assimilation of unstressed syllables. Thus, the effect of a language belonging to either type lies mostly in the more or less strong distinction between stressed and unstressed syllables and can be illustrated by the surprisingly simple Table 3.2.

reduction of unstressed	*enhanced distinction stressed/unstressed*	
+	+	*stress timed*
-	-	*syllable timed*

Table 3.2: Simple discrimination matrix between stress timed and syllable timed languages. Both influences are positively correlated, i.e. a language favouring reduction of unstressed syllables automatically enhances the distinction between stressed and unstressed syllables.

It is obvious, that languages fitting perfectly into this classification scheme will be rare. Also, the third suggested type of rhythm class, namely mora timing, is missing. In fact, many languages are difficult to classify along these lines. Two famous examples pointed out by Nespor (1990) are Catalan (reduction, simple phonotactics, free lexical stress) and Polish (no reduction, complex phonotactics, fixed lexical stress). But another unclarity is what to do with languages having many but not all of the features which are typical for one rhythm class or the other. Italian fits nicely into syllable timing in some respects (geminates, no reduction, simple phonotactics) but is also characterized by strong accentual lengthening. Its lexical stress may be less free than in most languages but is not completely fixed, either. Thai, although being a tone language, has been classified as stress timed. It remains to be explained, now, to what extent a language may "fail" the prerequisites and still belong to one or the other rhythm class. This problem lead some researchers to the alternative working assumption that, instead of having distinct classes of stress timed, syllable timed (and mora timed) languages it may be more useful to think of a rhythm continuum (e.g. Dauer (1987)). This new hypothesis let several researchers design their rhythm metrics accordingly.

3.1.2 Metrics of Dispersion as Indicators of Timing Variability

Probably the most straightforward measure of variability has been used (among others) in the often cited study by Roach (1982), namely the standard deviation, which is the average durational deviation from a mean value (see footnote on page 72). The rationale behind this is that given a higher level of syllable isochrony, the durations should cluster more around the mean compared to the stress timed languages. One of the main problems with this metric is that the standard deviation obviously does not take into account any sequential distribution, i.e. a sequence of $\{100, 100, 100, 200, 200\}$ would of course be regarded as equally variable to the strictly alternating sequence of $\{100, 200, 100, 200, 100\}$, since the metric is calculated across a whole sample, without taking into account the relative order of the individual samples. An intuitive view would regard the first sequence as less variable, since it only contains one rhythmical change, while the second sample shows a strictly alternating sequence. A further problem with the standard deviation is its dependence from the mean value, i.e. it is very sensitive to outliers and calculates much higher variation for data samples containing larger numbers. With respect to speech data, both issues may be problematic. Outliers are a frequent phenomenon, especially if one considers less controlled data, e.g. spontaneous speech. The second issue is most problematic regarding different speaking rates, which even occur under laboratory conditions, unless it is highly controlled. Thus, this metric is problematic when comparing two data sets, even if they stem from the same speaker, since individual tempo might vary due to numerous factors even intra-individually. This inherent problem of standard deviation was shown to be a key problem in the robustness of rhythm metrics by White and Mattys (2007); Dellwo (2003).

A suggestion to overcome some of the problems related to the standard deviation has been made by Dellwo (2003). In order to factor out tempo variation in speech rhythm analyses, he used the variation coefficient instead. This enabled him to measure variability on a relative scale rather than an absolute one and indeed, in his sample the influence of speaking rate lost much of its influence. The variaction coefficient enables us to compare the variation of two data sets with different mean values, by normalizing the data, usually taking the mean value, represented as the expectancy value E in 3.1[1]:

$$VarCo(X) = \frac{\sqrt{\sigma^2(X)}}{E(X)} \tag{3.1}$$

It can be argued, though, that a factoring out of absolute durations may be problematic when taking into account some of the facts we know about rhythm perception discussed in chapter 2. When perceiving temporal patterns, there are several absolute duration thresholds playing a role, such as the window of temporal integration (approximately syllable size) or the window of temporal presence (approximately foot size). Now as long as our relative durations that are relevant for rhythm perception stay stable within the limits of such thresholds, the variation coefficient may remain a suitable metric for variation analysis across several speakers or speaking styles. However, once a speaker talks very fast or very slow, thus changing the relationship between what fits into one window of temporal presence may change thus creating a completely different rhythmical impression. Thus, even the more

[1] As usual, variance is indicated as σ^2 and standard deviation as $\sqrt{\sigma^2}$

stable metric may not be robust enough for the suggested purpose.

If looking for a more robust global dispersion metric, one could use the mean deviation from the median (cf. Equation 3.2)[2]. This metric has to my knowledge not been used before in rhythm analyses, though Bröggelwirth (2007) highly favours the median over the mean in his rhythmical analyses using large data sets. This metric has the advantage that it delivers robust values in a non-normalized fashion, i.e. it is useful to gain insight into variability on an absolute duration scale. In case one wants to compare data sets of different tempos, the metric could also be normalized, thus leading to a relative measure analogous to the variation coefficient.

$$MD = \frac{1}{n} \sum_{i} |x_1 - \tilde{x}| \tag{3.2}$$

It is not at all surprising that global dispersion metrics only are able to capture *some* characteristics of speech rhythm that are implied by the phonological models. The phonological models do not predict an overall higher variability of stress timed languages, with regards to unstressed syllables even the opposite is the case. On the other hand, syllable timing does not exclude accentual lengthening, it is just supposed to be less strong. As some analyses of large data sets imply, even the accentual lengthening in a prototypical stress-timed language such as English cannot be detected straightforwardly. Instead, relative timing of adjacent unstressed and stressed syllables play a major role in transporting a certain perceptual variation. Also, keeping in mind that slight variabilities are rather smoothended perceptually, while a sudden local increase in duration may cause an impression of strong variability, global dispersion metrics should always be interpreted with a certain amount of caution from a perceptual point of view.

3.1.3 Early Variability Metrics

Two reletively early metrics have been proposed by Donovan and Darwin (1979); Darwin and Donovan (1980a,b) and Scott et al. (1985). Both have been extensively discussed in (Eriksson, 1991, 64ff.) and I closely follow his analysis. The metric sug-

[2] Where n is the number of sample values, x_1 the ith sample value and $\tilde{x}$ is the median of the entire sample.

gested by Donovan and Darwin (1979) (cf. Equation 3.3)[3] has serious shortcomings. As it is based on ratios of adjacent durational values, decelerating sequences count as more irregular than accelerating sequences. E.g. the sequence $\{1, 2, 3\}$ would be ascribed the variability $\left|(1 - \frac{1}{2})\right| + \left|(1 - \frac{2}{3})\right| = 0.833$, while the variability of the sequence $\{3, 2, 1\}$ would add up to $\left|(1 - \frac{3}{2})\right| + \left|(1 - \frac{2}{1})\right| = 1.5$. Such an order sensitivity is of course highly questionable, despite the fact that it takes into account the durational deviation of neighboring events. Thus, it makes an attempt to integrate more rhythmical information than a purely global metric, albeit not in a very promising manner.

$$variability = \sum_{i=1}^{n-1} \left| 1 - \frac{a_i}{a_{i+1}} \right| \tag{3.3}$$

With regards to the proposed metric by Scott et al. (1985) (cf. Equation 3.4), Eriksson sees less problems, although he does notice that due to the usage of a logarithm, the metric overrates small and underrates large durational differences. However, he mentions that it is far from clear whether this constitutes a problem. With regards to psychoacoustic results on duration perception, this feature may even be advantageous since humans tend to behave in a similar way. The measure may be even relatively robust with regards to different speaking rates as it assigns identical scores to the sequences that are identical in relative durations, e.g. $\{1, 2, 3\}$ receives the same variability score as $\{1, 2, 3\}$. Also an absolutely isochronous sequence would come out as 0 as desirable. Still, the model needs to be tested with regards to its perceptual adequacy. Also, it is a global measure not taking into account local rhythmical phenomena. Thus, one of the main criticims is, that it does not adequately treat one of the main characteristics of rhythm that were identified by phonological models, namely the relative variation of adjacent syllables.

$$variability = \sum_{1 \leq i < j \leq n} \left| ln \frac{d_i}{d_j} \right| \tag{3.4}$$

[3] The equation is given with the amendments suggested by Eriksson (1991), since the original metric intended to compare two rhythms.

3.1.4 Variability Metric by Ramus and Mehler, 1999.

In their strive for an adequate rhythm metric, Ramus and colleagues (Ramus et al. (1999, 2003)) followed a different methodology. Their study is based on two fundamental assumptions:

1. The ability to perceive rhythm classes is a fundamental cognitive process.
2. The perceptual differentiation between rhythm classes is based on language specific variations of vocalic and consonantal intervals.

Ramus' reasoning is built on various studies of speech perception based on delexicalized speech, especially those with newborns showing an ability to discriminate between languages of two different rhythm classes (e.g. Japanese and English), but not between languages belonging to identical rhythm classes (e.g. Dutch and English) (Nazzi et al. (1998)). Based on Dauer's work on phonological properties of rhythm class (cf. Section 3.1), he draws the conclusion that the two main influential parameters determining rhythm class are reduction and phonotactic complexity. He argues that these manifest themselves most clearly in the distribution of consonantal and vocalic intervals in a language. Further empirical support for such an account come from perception studies with newborns and adults who are presented delexicalized speech (Ramus and Mehler (1999); Ramus (2002)) where every vocalic interval is replaced by an /a/ and every consonantal interval by an /s/, the so-called "sasasa"-method. In alternative experimental paradigms, the delexicalization is performed taking into account the relative sonority of the individual consonants, e.g. each fricative is replaced by [s], each plosive by [t], each nasal by [n], each lateral by /[l]/ and each approximant by [j] (The so-called "saltanaj"-method). He also examines the impact of fundamental frequency on rhythm perception and finds it helpful, but not necessary for rhythm class perception. By simply listening to delexicalized speech mimicking the durational properties of consonantal and vocalic intervals, newborns are able to distinguish languages belonging to different rhythm classes, but not those belonging to the same rhythm class.

From these insights and based on the phonologically based rhythm class distinction, he derives a measure that is able to classify languages belonging to one of the three established rhythm classes (syllable, stress, mora timed). Furthermore, he aims to find a way of classifying languages as being rhythmically mixed. Thus, his metric is an attempt to quantify the rhythm continuum derived from the phonological typology introduced earlier. Using a relatively small corpus of annotated speech from eight languages, an ANOVA revealed that the best prediction of rhythm class can be made based on the two parameters ΔC (standard deviation of consonantal intervals) and $\%V$ (percentage of vocalic intervals in the data set). These two factors can be related to the two phonological properties *reduction* and *phonotactic complexity* as illustrated in Figure 3.1.

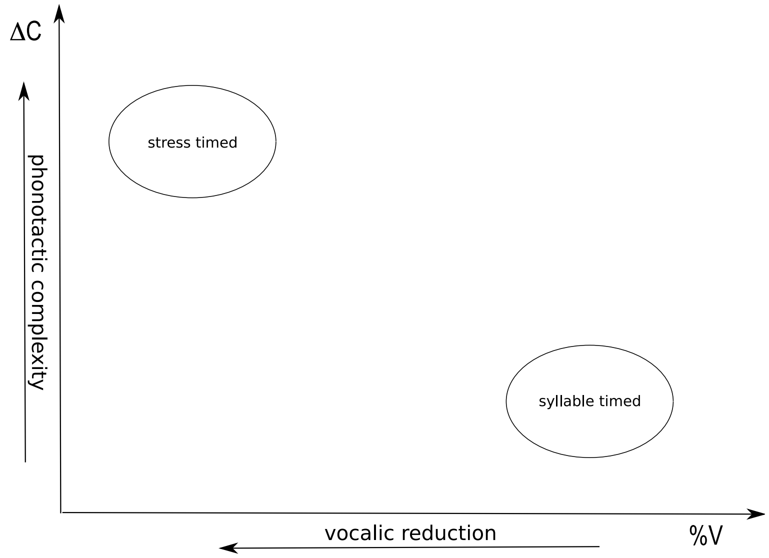


Figure 3.1: The relationship between phonological and phonetic predictors of rhythm class according to Ramus and Mehler (1999). (Illustration adopted from Wagner and Dellwo (2005)).

Ramus' metric has become very popular in the rhythmical categorization of languages hitherto unclassified (e.g. European and Brazilian Portuguese Frota et al. (2002) or Polish Ramus et al. (2003)) but also was also applied in the examination of non-native and bilingual speech rhythm (e.g. Bond et al. (2003)).

However, the model has also been faced with a lot of criticism. One argument the metric has been confronted with concerns the fact that it does not adequatly capture the phenomenological side of rhythm, rather an epiphenomenon by making a quantification of a qualitative classification based on phonological properties (e.g. see Cummins (2002); Gibbon (2003a,b)). The shortcomings of this approach become most evident in its application to Polish, a language that looked rhythmically "mixed" from a phonological point of view, e.g. Polish or Czech. When describing Polish speech data, the outcome provides further evidence that Polish is not adequately described as either stress timed or syllable timed, neither does it show a behaviour "in between" the endpoints of the continuum (Ramus et al. (2003); Dancovičova and Dellwo (2007)). Maybe an even more problematic point of the measure is that the metric does not reflect the rhythmical *patterns* that the languages consist of, e.g. in which fashion do speakers produce and perceive sequences of stressed and unstressed beats, how do they group beats into larger rhythmical entities and how do these rhythmical entities relate to phonological properties? The mere possibility to distinguish languages into predefined rhythm classes based on the rhythm metric is not very informative. Furthermore, it has been shown that such a distinction can also be made based on alternative phonetic properties having to do with syllabic structure, such as the durational distribution of sonorants, thus probably capturing rhythmically important phenomena such as reduction phenomena leading to consonantal syllable nuclei (Steiner (2003a,b)). E.g. in stress timed German, the word /hˈaːbən/ is usually pronounced as [hˈaːbm̩], thus possibly leading to a high proportion of lengthended nasals in the language. Dellwo et al. (2007) even could show that in order to get a stable distinction between stress timed and syllable timed languages, it is sufficient to annotate their voiced and voiceless parts.

By now, numerous studies have proved that the proposed rhythm metric is indeed a somewhat more detailed phonetic description of phonotactic complexity. Wagner and Dellwo (2004) showed that if one expresses duration of consonantal and vocalic intervals in terms of simple CV-sequences, counting geminates as CC (= 2), long vowels and diphthongs as VV (= 2) and all other consonants and vowels as C or V (= 1), one receives exactly the same distribution of stress timed and syllable timed languages (cf. Figure 3.2). In another study, Dancovičova and Dellwo (2007) could show the relationship between phonotactic complexity of a language, also taking into account the frequency of more or less phonotactically complex syllables, and Ramus metrics.

Thus, while the model distinguishes rhythm classes which have been phonologically predefined, it is not very informative. However, one could argue that it indeed captures rhythmic differences between languages that make more or less use of phonological quantity, thus having different rhythmic patterns. This should

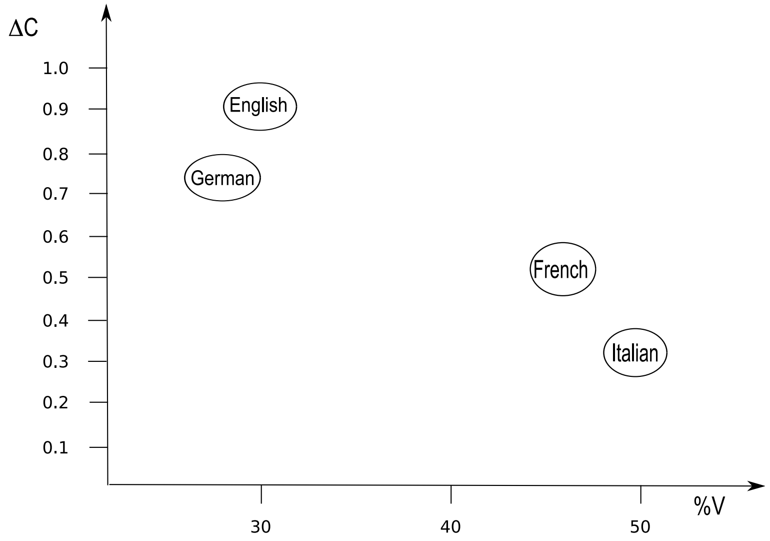


Figure 3.2: Ramus metric applied to simple CV-sequences derived from the texts of the BonnTempo database for stress timed German and English and syllable timed Italian and French, adapted from Wagner and Dellwo (2004).

be even more the case if the languages in question are quantity sensitive, i.e. phonological weight attracts stress. However, the vagueness of the phonological criteria in determining rhythm class cannot be solved convincingly with the proposed metric.

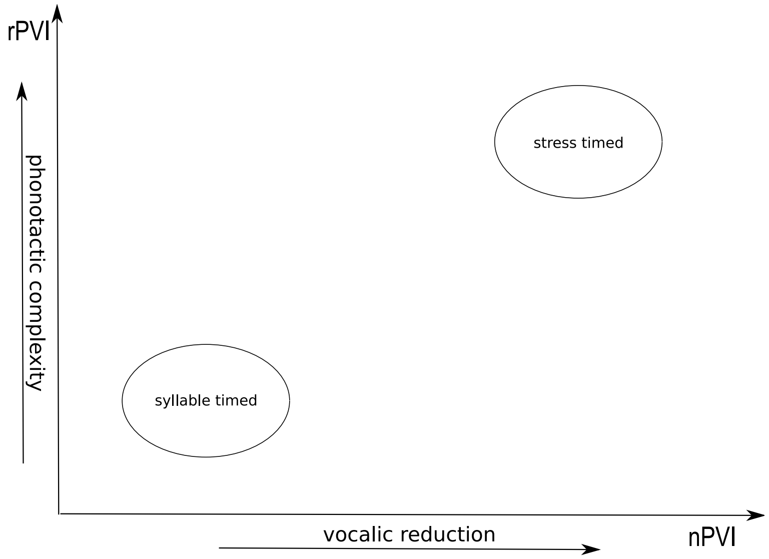
Another weakness of the metric concerns its robustness and has been pointed out by Barry et al. (2003); Barry and Russo (2003); Dellwo and Wagner (2003); Dellwo (2003); Wagner and Dellwo (2004). They show that the metric is strongly influenced by speech rate, especially the consonantal properties captured by ΔC, while the vocalic proportions remain more stable. Furthermore, while the classification works fine across a larger data set, a lot of inter- and intraindividual variation is evident for speakers of the same language (Dellwo and Wagner (2003)). Taking into account the points made in Section 3.1.2 it does not come as a surprise that the use of the standard deviation in ΔC runs into problems when regarding different speaking rates, speakers or speaking styles — as standard deviations depend on the mean, any variation in articulation rate — as can be expected across speakers or speaking styles — will influence it.

3.1.5 Measuring Rhythm with Pairwise Variability Indices

Based on the close link between phonotactic complexity, syllable reduction and rhythm class distinction, Low et al. (1999, 2000); Grabe and Low (2000) proposed another metric based on the relative distributions of vocalic and consonantal intervals across utterances. For each pair of sucessive consonantal or vocalic intervals across an utterance, the durational difference is calculated. The differences are summed up and the mean difference of neighboring consonantal and vocalic intervals is calculated. Overall little variation of adjacent intervals results in small PVI-values. These mean values are plotted in a two dimensional diagramme with the vocalic variation on the x-axis and consonantal variability on the y-axis. In order to factor out articulation rate related influences on the vocalic intervals, these are normalized locally by division with the mean of two adjacent vocalic intervals. The normalization achieves that the local difference between two successive intervals is expressed as the ratio of the absolute durational difference and an "ideal" isochronous value given that both intervals are identical in length. The ratio ranges between 1 for complete isochrony and 2 for maximum deviation. This vocalic variability metric is called the $nPVI$. The consonantal metric is referred to as the "raw" $rPVI$, since it does not employ any normalization. Equations 3.5 and 3.6 show the way the pairwise variability indices are calculated, with d indicating interval duration and i indicating the number of the interval within the sequence of intervals. The rationale behind that approach is that in stress timed languages, neighboring vocalic intervals ought to be highly different, due to reduction phenomena. Furthermore, the higher phonotactic complexity in stress timed languages ought to lead to an more frequent alternation of complex and less complex consonantal clusters. This is expected to show in a high degree of durational variation between neighhoring consonantal intervals. Thus, the larger amount of variability leading to stress timing compared to syllable timing is captured. The connection between phonologically and phonetically determined rhythm class in illustrated in Figure 3.3.

$$nPVI = 100 * \sum_{i=1}^{n-1} \left| \frac{d_i - d_{i+1}}{(d_i + d_{i+1})/2} \right| /(n-1) \tag{3.5}$$

$$rPVI = \sum_{i=1}^{n-1} \frac{|d_i - d_{i+1}|}{n-1} \tag{3.6}$$



).

Figure 3.3: The relationship between phonological and phonetic predictors of rhythm class as determined by the nPVI and rPVI. (Illustration adopted from Wagner and Dellwo (2005

It is intuitively clear that unlike the rhythm metrics described above, the PVI *does* take into account the rhythm related phenomenon of alternation, as it is calculated of adjacent phonetic events instead of measuring global utterance features. Thus, at first glance it appears to be more adequate than the metrics shown previously. Next to the study of typological and dialectal variation as in the original papers, the metric has been used to describe a wide variety of related phenomena such as L2 speech (e.g. Whitworth (2002); Bond et al. (2003); Stockmal et al. (2005); Lleó et al. (2007)), language acquisition studies (e.g. Grabe et al. (2000)) and even for the distinction of musical style (Patel and Daniele (2003a,b); Dalla-Bella and Peretz (2005); Patel et al. (2006)). Naturally, the measure itself is not restricted to the usage of vocalic and intervocalic intervals within an utterance. Alternative suggestions have been made to calculate variability based on the syllable (Deterding (2001); Wagner and Dellwo (2004)) or the foot or a combination of both (Asu and Nolan (2005, 2006)). Thus, the metric can also be used to state tendencies towards isochrony (or at least less variability) on any level of the prosodic hierarchy that may appear relevant.

Despite of its ability to capture alternation of successive rhythmical events better than other metrics, the PVI has been criticized due to various shortcomings: Gibbon (2003a,b) points out that it treats sequences alike which are rhythmically different, e.g. a linearly decelerating sequence $\{100ms, 200ms, 300ms, 400ms\}$ has the same PVI as a strictly alternating sequence $\{100ms, 200ms, 100ms, 200ms\}$, namely 100. This is because the PVI only measures local variation no matter whether it is decelerating or accelerating. Global deceleration or acceleration phenomena are of course not captured by the metric, since it only calculates the absolute durational difference between two intervals, independent of its direction. Also, the PVI-metric in both the normalized and raw version is very sensitive towards outliers. Thus, the metric might run into problems when it comes to styles other than laboratory speech. This might be responsible for the finding that despite the normalization of vocalic portions, the metric is influenced by speaking rate (Barry and Russo (2003); Wagner and Dellwo (2004)). Of course, this can be interpreted in such a way that the rhythmic style changes with tempo, but this relationship remains to be shown. The benefit of the normalization method becomes even less clear when taking into account global rhythmic characteristics of an utterance — due to the local normalization for each pair of adjoining intervals, a relative difference is calculated, independent of the two intervals being rather long or short. Thus, a spondeic sequence, as may occur in a final lengthening context, will probably receive a very similar local $nPVI$ value as a sequence of two very short, reduced syllables. In total, the profit of the normalization procedure seems to be outweighed by the problems. Another point of critique arises when calculating the vocalic and consonantal PVIs simply based on their segmental distribution. Similarly to the segmental Ramus' measure, raw PVIs were calculated for texts of the BonnTempo database for successive consonantal and vocalic intervals, respectively. Instead of measuring durations, each consonant was assigned the value 1 and each geminate $= 2$. In the vocalic intervals, each reduced vowel [ə] was assigned the value 1, each short vowel the value 2 and each long vowel or diphthong the value 3. After calculating PVIs based on this very rough procedure, we get the same distributions between syllable and stress timed languages as the PVI-metric (cf. Figure 3.4). Thus, a similar question as for the Ramus measure can be raised, namely, whether the PVI is really much more informative than the phonological approaches. There exist suggestions to modify the PVI metric, e.g by Gibbon and Gut (2001) who quantify the durational difference of two adjoining intervals in terms of an absolute ratio between the shorter interval d_i and the longer interval d_j (cf. Equation 3.7). Their so-called *Rhythm Ratio* approximates 100 in a perfectly isochronous sequence and becomes lower with an increase in local

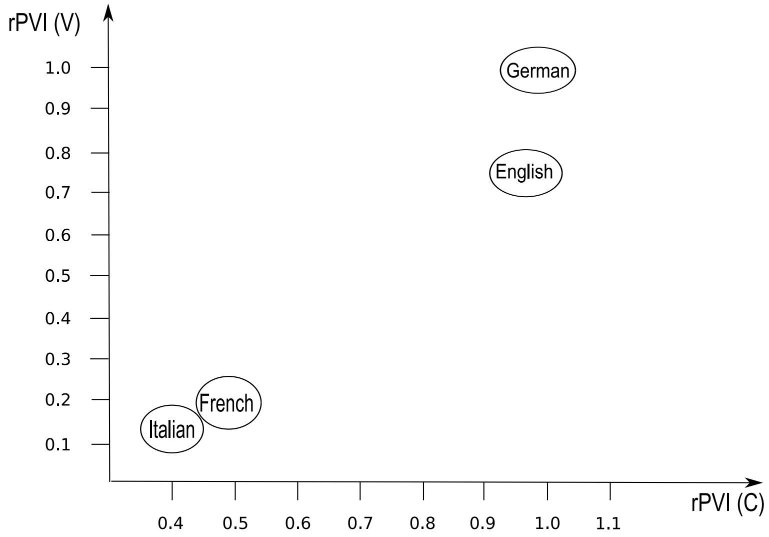


Figure 3.4: Raw PVI metric applied to simple CV-sequences derived from the texts of the BonnTempo database for stress timed German and English and syllable timed Italian and French.

variation. The metric has been applied to account for a number of typological phenomena such as the classification of African Tone Languages (Gut et al. (2001)), but it remains unclear in what way this approach is more or less advantageous over the original $nPVI$.

$$RR = 100 * \sum_{k=1}^{n-1} \frac{d_i}{d_j} / (n-1) \tag{3.7}$$

Summing up, the PVI apparently captures some important features intuitively related to speech rhythm, but still has a significant amount of problems concerning its robustness (also see Ferragne and Pellegrino (2004)) which are in need of clarification, e.g. with the help of perception experiments.

3.1.6 Metrics Based on Vocalic Variation

As to a large extent discussed above, rhythm metrics based on consonantal and vocalic intervals have been shown to be instable, especially with regards to the consonantal intervals. These were shown to be very much influenced by speech rate and speaking style, while metrics calculated on the vocalic parts of the utterance remained more or less stable (e.g. Barry and Russo (2003); Dellwo and Wagner (2003); Wagner and Dellwo (2004); Ferragne and Pellegrino (2004)) unless they are based on non-normalized standard deviation (White and Mattys (2007)). Thus, it has become more and more popular to use measures of vocalic variation only to account for rhythmical phenomena. Often, these metrics are normalized such as the $nPVI$ or the vocalic variation coefficient $VarCoV$ (Dellwo (2003)). Especially the former has become a widespread metric and has been suggested for being applied to the comparison of native language and musical style (Patel et al. (2006)). In combination with the percentage of vocalic intervals, Ramus' %V metric, it was very successful in showing the influence of L1 to L2 speech (White and Mattys (2007)) and the perceptual distinction of rhythmically different dialects (White et al. (2007)). When using global rhythm metrics, calculations based on the vocalic intervals are so far the most robust and reliable choice we have. It is likely that this circumstance is explicable by the close link between vowels and the perception of fundamental beats, the p-centers discussed in Section 2.2.3. The time span in between the vocalic onset and the end of the syllable rhyme, often consisting of a pure vowel, play an important role in the perception of rhythm prominence while consonants in the syllable onset do not. Furthermore, the vocalic regions are the ones showing the strongest amount of deliberate variation or "elasticity" (Campbell (1992)), as expressed in accentual or final lengthening and reduction phenomena. However, we have to be extremely cautious to assume a close link between a metric that has been shown to be stable and able to discriminate rhythm classes among a variety of speaking styles and languages and its perceptual relevance. Why is that so? Evidently, the ability to discriminate two categories, e.g. the colours blue and red, does not automatically imply we are able to identify them. Imagine, e.g. blue and red are being represented in different shades of grey. We would be still able to discriminate both colors without being able to make any useful statement concerning their true color. Thus, while we know that metrics of vocalic variation are closely linked to some features characterizing rhythm and that listeners use these cues in order to differentiate between rhythmic classes, we are still not sure how these metrics do capture rhythm in terms of its structure as patterns of stronger and weaker events. Until now, it is far from clear how variations in the dimensions of $nPVI$, $\%V$ or $VarCoV$ are interpreted perceptually.

3.2 Quantifying Rhythmical Distributional Patterns

Instead of searching for global evenness or lack of variability, some researchers have rather concentrated on the detection of rhythmical patterns across an utterance. However, such approaches are less frequent in rhythm analyses, despite them being the more promising approaches in the search for evidence of the hidden regularities underlying rhythmical grouping and structure in speech.

3.2.1 Detecting Predictable Alternations with Time-Series Analysis

While most measures discussed above are mostly concerned with the quantification of global or local rhythmical variation based on typological considerations, some alternative approaches attempt to capture the distribution of rhythmical strong-weak patterns across utterances. Instead of trying to measure global or local variations in timing, these approaches aim at discovering the serial dependencies in sequences of rhythmical events. E.g. a sequence of intervals $x_1, x_2, ...x_k$ is examined in order to predict the behaviour of the interval x_{k+1}. If such predictions are possible, the listener can use them to derive hypotheses for upcoming speech events. Thus, the empirical well-foundedness of possible metrical expectancies are examined. Within this field of research, an important examination and prediction method is autocorrelation analysis (also cf. Section 2.4.1 on production). Here, correlations between events within a sequence are calculated. However, the examined events are not necessarily immediately successive (as in the PVI metric). E.g. in weather forecast the knowledge that the current season is summer can be used to predict low temperatures in the time of the year after the next season = winter, but also average temperatures in seasons immediately approaching (= fall) and following winter (= spring). The proximity of two intervals is expressed in terms of lag, with $lag1$ referring to an immediately contiguous event, $lag2$ to an event with one other event in between and so forth. In a strictly binary alternating rhythm $\{50, 110, 40, 100, 60, ...\}$, each interval would correlate positively with intervals at even distances, $lag2, lag4, lag6, ...$ etc., but negatively with intervals at uneven distances $lag1, lag3, lag5, ...$. In autocorrelation analyis, usually correlations are calculated for a variety of lags and significant correlations are interpreted such that an event serves as predictor for an upcoming event.

Time series analyses are much less frequent compared to studies on speech rhythm using global rhythm metrics. Keller et al. (2000) examine autocorrelations in French and English for the purpose of improving the duration prediction in synthetic speech. Their starting point is that even though a timing model built on strict rhythmic alternations may appear oversimplistic, there might exist some tendency towards an alternating rhythm in speech. According to them, this general tendency is blurred by linguistic factors which are mainly determining duration patterns (also cf. the discussion in Section 1.1). They examine corpora of French and English and compute autocorrelation coefficients across utterances in two conditions, an *absolute time condition* and a *linguistic time condition*. In the absolute time condition, after each 500ms, the number of syllables contained in this timing window is calculated. E.g., if the interval contains 50% of the remaining part of one syllable and 100% of another syllable, the interval is assigned the value 1.5. For the linguistic time condition, they compute syllable durations and calculate their fit into the 500ms reference window. E.g. a syllable of 250ms would receive the value of $500/250 = 2$. They find small but significant anticorrelations in various speaking rates at $lag1$ for their French and English data. They also report an effect of "swingyness" in synthetic speech that had been manipulated based on these anticorrelations at $lag1$. This seems to indicate a small but quite stable effect of rhythmical alternation - such an effect had already been explained as being the result of compensation of production mistakes (cf. Section 2.4.1). From a perception point of view, however, such a systematic alternation effect can help the listener making predictions on the beginning and duration of an upcoming event. Thus, such systematic relationships would guide the listener's rhythmical expectancy.

Another study using time series analysis has already been reported in section 2.4.1. Cummins (2005) examined the sequential properties of spoken English word lists and also detected an anticorrelation at $lag1$, thus providing further evidence for the preference of alternations. Buder and Eriksson (1999) used time series analysis in order to find acoustic evidence rhythmical structuring of conversational speech, but their results are difficult to interpret due to a lack of description of their mathematical procedure.

The overall tendency to perceive a binary alternation has already received support by anecdotal observations of perceptual prominence patterns in German (Heuft (1999)). Thus, its presence in speech data can be seen as evidence for so-called euphony rules proposed by Metrical Phonology that are aiding to uphold a rhythmical alternation by avoiding *stress clashes*, i.e. sequences of two stressed beats and *lapses*, i.e. sequences of more than two unstressed syllables (cf. Figure 2.25 on page 92). In an evaluation of these rules, it has been verified that especially the avoidance of *lapses* helps improving the match between perceived and predicted prominence patterns (Wagner (2000)). The lack of such a bias of metrical alternation has also been shown to impair the performance of an automatic prominence detection algorithm for German (Tamburini and Wagner (2007)). However, the existence of the same tendencies in syllable timed French as reported by Keller et al. (2000) are somewhat surprising and may indicate a universal phenomenon that is independent of the language spoken.

3.2.2 Detecting General Deceleration Trends as Indicators for Rhythmical Grouping

Besides looking for rhythmical alternations in speech, a simple metric indicating rhythmical grouping was sought after by Gibbon and Williams (2007). In their study of Welsh rhythm, they chunked utterances based on the criteria of deceleration. That way, they attempted to automatically detect prosodic feet. This straightforward chunking process lead to fairly good results when matching it with grammatically determined feet based on the assumption that a foot ends on a stressable (=content) word. Their approach provides further evidence for the grouping principle saying that durational increase indicates boundaries. General deceleration trends across utterances have also been verified for Czech (Volín and Skarnitzl (2007)) and French (Benguerel and D'Arcy (1986)), the latter having also shown that deceleration is actually a prerequisite for a sequence to be perceived as even (or isochronous). Thus, the general search for temporal downtrends in order to detect grouping mechanisms appears to be fruitful. It has already reported in Section 2.3.5.2 that pure durational increases tend to be perceived as group endings. However, all studies reported above also show a strong influence of independent segmental, morphological and lexical phonological properties on this general deceleration trend. It is thus difficult to perform such studies based on the measurement of duration patterns alone.

3.3 Rhythmical Hierarchies

It has been thoroughly discussed in Chapter 2 that rhythm is most adequately described as grouped rhythmical entities which the individual groups having a particular structure. Such a grouping into a larger rhythmical entity automatically introduces a hierarchicy of larger groups emcompassing smaller ones. In Section 2.3.3, such hierarchies were presented with regards to phonological models and in Section 2.3.1 we described some insights into perceptual aspects involved in rhythmical grouping. The following sections will present various approaches towards modeling such hierarchies with respect to speech data.

3.3.1 Coupled Oscillator Models

In recent years, coupled oscillators have become popular as cognitively plausible models of rhythmic speech production and perception. Such models rely on the belief that speaking or hearing is an interplay of abstract representations and biomechanical systems which are modelled best as dynamical systems. Of course, the close connection between motor system control and rhythm production (cf. Section 2.4.1) on the one hand, and the notion of an oscillating inner clock that may even link our production and perception (2.1.1) renders these approaches attractive. Having this in mind, such models are very elegant solutions to the complex problem of rhythm perception and production. While Port, Cummins and colleagues (Port et al. (1995); Port (1990); Cummins and Port (1998); Cummins (2002); Port (2003)) mostly concentrate on collecting empirical support for such an approach and the description of an interface to discrete phonological descriptions, O'Dell and Nieminen (O'Dell and Nieminen (1999); O'Dell et al. (2007)) and Barbosa (Barbosa and Madureira (1999); Barbosa (2000, 2001, 2002)) developed hierarchical models aiming to explain the difference between stress timed and syllable-timed languages. Barbosa's apporach is integrated into a more general model of speech production with interfaces to prosodic and segmental planning (cf. Figure 3.5).

The general idea shared by all coupled oscillator models is the existence of in-

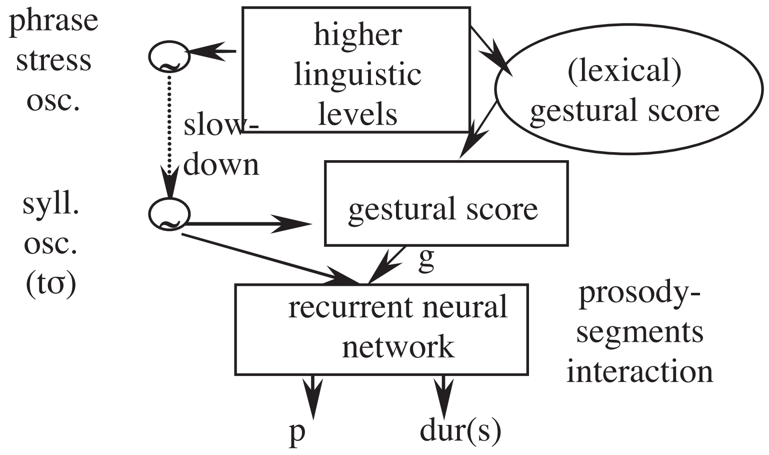


Figure 3.5: The dynamical production model suggested by Barbosa and Madureira (1999); Barbosa (2002).

dividual oscillators being responsible for timing at various levels in the prosodic hierarchy, E.g. it is proposed that there exists a syllable oscillator guiding timing at syllable level, another one at a superior level, like the foot or the stress group, another one at phrase level, a very low one at mora level etc. The various oscillators do not produce independent timing patterns, instead, they are connected with a specific strength, the *coupling strength*. The coupling strength determines the influence of one oscillator on the other. Coupling is achieved using the notion of *entrainment*, i.e. a modification of the period of another oscillator.

The production model by O'Dell and Nieminen restates the long-term durational data described for various stress timed and syllable timed languages in Dauer (1983) in terms of coupling strength between a stress group oscillator and a syllable oscillator. Their description is based on phase differences between the oscillators and coupling strength. The coupling strength between the stress group oscillator and the *entrained* syllable oscillator modifies the period of the latter. As empirical basis they make use of the analyses by Eriksson (1991) (cf. Section 2.4.1) where stress group durations were computed with the help of linear regression equations of the form given in Equation 3.8.

$$I = a + b * n \tag{3.8}$$

In Eriksson's formula, I is the stress group duration, a is the x-axis intercept , b determines the regression slope and n is the number of syllables contained in the stress group. Coupling strength is now identified as the ratio $r = a/b$ and the higher the coupling strength, the more stress-timed a language is supposed to be.

Their approach was developed further (O'Dell and Nieminen (2001); O'Dell et al. (2007)) to model the interaction of more than one oscillator. A linear function expresses the period of the slowest oscillator, e.g. the prosodic phrase based on the number of the hierarchically lower levels it includes. A five-oscillator model is thus expressed in Equation 3.9, where T_1 is the expected duration of the top level cycle, e.g. the prosodic phrase, n_k is the number of level k oscillator cycles synchronized within, e.g. the number of syllables, and c_k could be further expressed in terms of the oscillators eigenfrequency and the mutual coupling relations.

$$T1 = c_1 + c_2 n_2 + c_3 n_3 + c_4 n_4 + c_5 n_5 \tag{3.9}$$

While their approach is able to distinguish stress timed and syllable timed languages based on long term variability (cf. Figure 3.6), their model does not provide a "moment-to-moment" description of rhythmic phenomena.

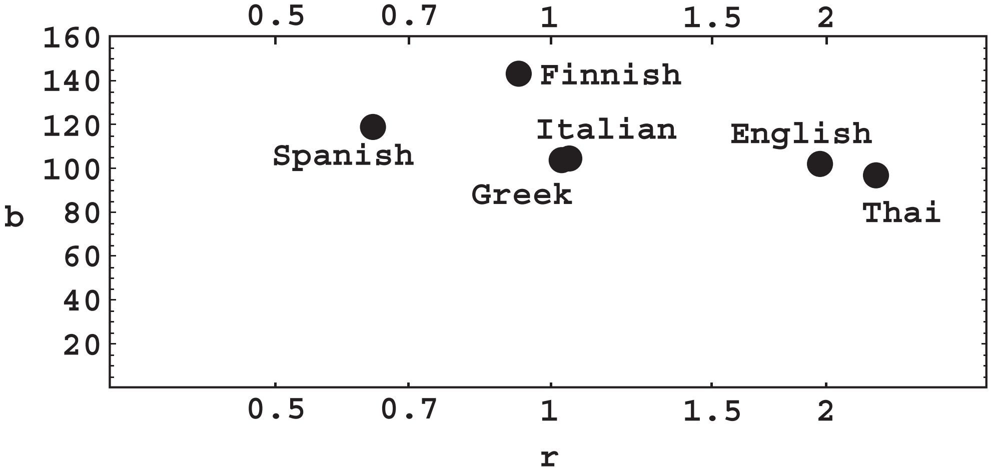


Figure 3.6: Relative coupling strengths r vs. regression slopes b for various languages computed in O'Dell and Nieminen (1999). Stress timed Thai and English have a high coupling strength compared to the other languages.

Barbosa's model (cf. Figure 3.5) aims to describe the interface between discrete phonological representations such as lexical stress and segmental production. It is able to model entrainment as a continuous process which can be followed from cycle to cycle of an oscillator. According to his view, the two oscillators representing phrasal and syllable durations are a universal property of human language, while the coupling strength between the two is language dependent — similar to the suggestions by O'Dell and Nieminen. One oscillator may entrain another one, e.g. the syllable oscillator may be entrained to the phrasal oscillator by a decelerating mechanism which increases the time span in between two vocalic onsets (= perceptual centers or beats, cf. Section 2.2.3) thus modeling the phenomenon of final lengthening.

His oscillator model is much inspired by the general rhythm perception model described in McAuley (1993, 1995), where a cosinoidal oscillator (here: syllables) is entrained by a train of rhythmical pulses (input signal). In Barbosa's model the to-be-entrained oscillator represents syllables and the pulse train input signal represents the stresses of a stress group oscillator. The timing of the input signal is generated by higher level linguistic representations such as semantics, syntax, lexicon, pragmatics, i.e. all linguistic levels that may influence stress location. The period coupling is achieved by a phase resetting mechanism that takes place at each syllable-sized oscillator maximum (cf. Figure 3.7). In addition to phase reset, the period is gradually adjusted to the period of the input pulse train, but with missing input by estimating the level of synchronicity between both oscillators. The amount of coupling strength guiding this syllable-by-syllable adjustment is determined by a language specific parameter $w_0 r$. The synchronicity function furthermore contains a decay rate which leads to a slow "recovery" of the syllable oscillator back to its intrinsic period. After each stress impuse, the period of the syllable oscillator is reset (cf. Figure 3.7). In Barbosa (2002) it was shown that the model can imitate some properties of "syllable" and "stress timing" only by variation of the coupling strength parameter, i.e. the phenomenon of compensatory shortening without stress group isochrony and the lack of compensatory shortening without syllable isochrony. When computing the different models' coupling strengths based on the generated durations according to O'Dell and Nieminen (1999), results for stress and syllable timing respectively were similar to the ones based on empirical data (cf.

Figure 3.6).

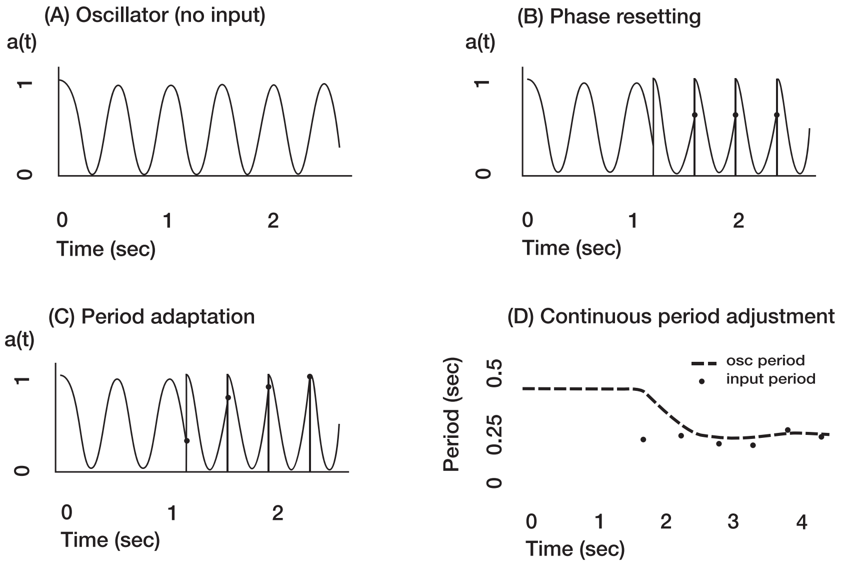


Figure 3.7: This illustration of entrainment by phase reset is taken from (Port et al., 1995, 14). (A) shows the cosinoidal oscillator with a range of $\{0, 1\}$ representing the syllable oscillator, which is not entrained yet. (B) shows the superimposition of periodic input pulses on the oscillator's activation and a phase reset to 0. (C) illustrates the step by step period adjustment to the input period, resulting in a gradual entrainment of the oscillator to the input pulse, making both periods more and more similar in course of time (D).

While coupled oscillator models are very elegant, it is still unsolved how to model the intrinsic noise present in large databases, since empirical evidence for entrainment effects have so far mostly been found (or maybe searched for!) in laboratory speech. Still, several phenomena and long-term durational properties of natural speech have been explained. However, if entrainment phenomena indeed are responsible for perceptual effects of syllable timing and stress timing, they *must* have their equivalents in natural speech as well as in laboratory speech — at least traces of them and not just across large speech samples. The reason for this is evident: Listeners do not listen to mean values of durational properties but to individual realizations which are categorized.

O'Dell and Nieminen (1999); Barbosa (2001, 2002) all see the problem that syllable timing is to a large extent determined by segmental structure. Of course, this is not incorporated in a model based on the view of a syllable sized oscillator. Both research groups make suggestions aiming to solve this problem, but is is yet unclear how this is to be achieved precisely. Even though McAuley's model aims to explain perception, the coupled oscillator approaches inherently are more production oriented - given the assumed close link between motor behaviour and time perception, such a point of view certainly makes sense. This may be the reason why these models have so far not incorporated some well-know facts about rhythm perception, such as the bias of perceiving alternations, rhythmical groupings etc. While rhythmical grouping may be explicable with respect to limitations of the stress group oscillator and a period reset, the perception bias towards rhythmical alternation does not seem to be explicable straightforwardly. In Barbosa's model, such factors are determined externally by the phonological component. However, given the fact that a tendency towards grouping and rhythmical alternation is rather universal property of human perception and production and may be subject to individual speaking style preferences (= paralinguistics), it is unclear why these phenomena should be explicable only by a language specific phonological processing.

3.3.2 Data Mining Approaches to Rhythmical Hierarchies

Coupled Oscillator models are inherently hierarchical in their classification of dependencies between oscillators which control the timing of one rhythmically relevant level of prosodic organization, e.g. the various levels of the prosodic hierarchy. One issue in rhythm research has been the description and modelling of the oscillator interactions, as we have seen in the previous section. Two more fundamental problems need to be treated in addition:

1. The estimation of the rhythmically relevant levels of prosodic organization (mora, syllable, foot, stress group, prosodic phrase?), e.g. which oscillators do we need to model?

2. Their interaction with various levels of linguistic organization, i.e. how is the interaction between the metrical structure described by the prosodic hierarchy (= metrical trees or grids, cf. Section 2.3.3) and syntax, semantics, segmental phonology etc.

O'Dell et al. (2007) make an attempt to empirically tackle the first issue while the work in Gibbon (2003a,b) is an approach to the second.

O'Dell et al. (2007) try to detect the relevant prosodic levels which determine the durations of interpause groups in Finnish conversational speech. They use a Bayesian Inference procedure built on their regression model expressed in Equation 3.9. Their best model reveals that, somewhat surprisingly, mora and stress group are the most important predictors of interpause durations — ranging before syllables and feet. Furthermore, they find that word boundaries tend to elicit a new stress group. They also show that the automatically determined stress group boundaries show a high agreement with the majority of human labellers' annotations, despite the latter showing a lot of inconsistency in their judgements. Their method may provide a very valuable tool in the detection of rhythmically relevant levels in the hierarchical organization of speech!

While O'Dell and colleagues aimed to detect the hierachical levels of prosodic organization, Gibbon (2003a,b) took another direction of research which can be regarded as being complementary to O'Dell's approach. He attempted to detect analogies between rhythmical alternations expressed as timing relations *short—long* or *long—short* and syntactic organization in English sentences. He interpretes the local timing relations as metrical relations similar to the *strong—weak* labels used in Metrical Phonology. He then calculates a kind of metrical tree based on these local timing relations, however, the trees are labelled with absolute word durations. In one version, for each local pattern of monotonically decelerating (= *short—long*) sequences, a local branching tree is generated. The node of the tree is labelled with the short duration value. This value is used recursively to build larger trees and percolates up. That way, iambic trees are generated (cf. Figure 3.9). In another version, the local trees are built on *long-short* sequences and the long value percolates up the tree. That way, trochaic trees are generated.

Afterwards, he compares the trees generated by the two different approaches with hand-annotated syntactic trees of the same data set comprising 20 sentences. The syntactic annotations were based on the judgements by six linguistically trained subjects. Then, metrical and syntactic trees were compared using a similarity index (cf. Figure 3.9). It turned out that the iambic strategy favouring decelerating structures lead to a high agreement between syntactic and rhythmical trees. Gibbon in-

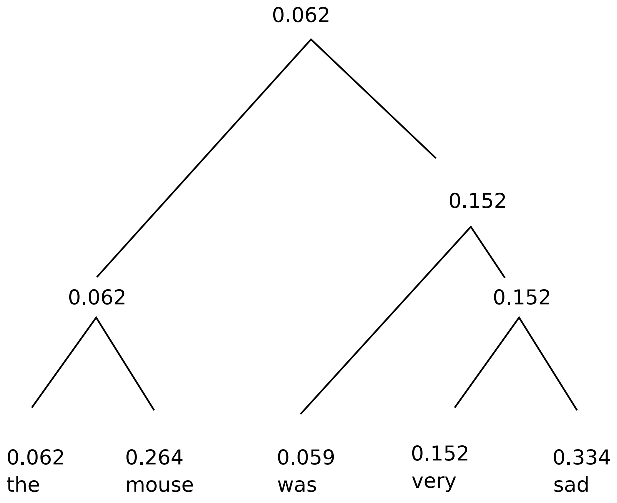


Figure 3.8: This pattern illustrates the model where word duration "metrical" trees are built recursively on *short—long* sequences with the shorter durations percolating up the tree and resembling iambic structure, which also matches the syntactic structure of the phrase. The nodes are labelled with word durations.

terpreted this as evidence for the NSR[4] on various levels of syntactic structure and a predominance am iambic rhythm in English. Indeed, the iambic pentameter has always been popular among English poets, circumstance also recognized in phonological approaches to poetic rhythm (Kiparsky (1975), Kiparsky (1977)). However, it should be kept in mind that the iambic bias is shown here solely on the level of the syntactic phrase and given a rough agreement between syntactic and prosodic phrases and the tendency of phrase final lengthening, this does not rule out trochaic patterns on other levels of English rhythmical organization. Of course, the comparative model in principle is applicable to any level of prosodic organization, but it does not provide information about the interaction between these levels. However, the method is a straightforward way to gain information about the phonetics-linguistics interface. It would be interesting to explore this method further to other rhythm-related interfaces, such as the relation between duration and moraic or foot structure.

Of course, both data mining approaches described in this section are not aimed to integrate the perception side of rhythmical grouping or the variability. O'Dell's

[4] The *Nuclear Stress Rule* (Chomsky and Halle (1968)) which stresses the last constituent in a syntactic phrase.

study also does not take into account rhythmical alternations, but to be fair, this is not what it has been developed for. Still, the relevance of the perception side of rhythm needs to be asked, because Gibbon looks at local alternations without weighing them, e.g. a sequence of $\{200ms, 202ms\}$ would be identified as *short*, *long*. The amount of acceleration or deceleration is not considered, although he explicitely intends to treat dactylic and anapaestic sequences, thus, a metrical equality of two neighboring entities ought to be considered. The integration of a perceptual variability threshold plus an option to insert ternary trees[5] might be a possible solution to this problem of this otherwise interesting approach.

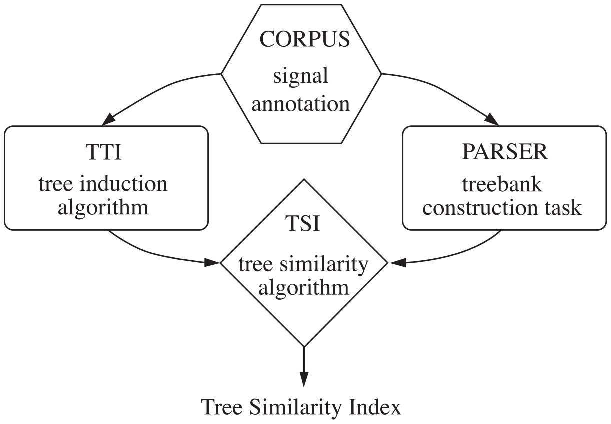


Figure 3.9: The general structure of the tree induction and comparison architecture proposed by Gibbon (2003a,b).

3.4 Conclusion

In this chapter, various attempts to quantify and measure the rhythmic organization of languages, rhythmic classes and speaking styles were presented. In summary, we can conclude that non of the proposed metrics or models of rhythmical quanitification take into account the component of metrical expectancy, which had been defined as a major influence in rhythm perception in Chapter 2 - maybe with the expception of Barbosa's coupled oscillator model who could integrate metrical expectancy with the help of his phonological component. Of course, this is not a

[5] Gibbon explicitely states that the model can generate ternary trees but leaves it open, under which circumstances.

heavy criticism since none of the other models explicitly strove to model rhythmical expectancy. Another issue not covered by any metric or model is the interrelationship between articulation rate and perceived rhythmical structure. Instead, most metrics tried to normalize for rate influences in order to be able to compare different data sets. It is yet unclear, how the role of articulation rate can be incorporated into rhythm quantifications.

The metrics of globals dispersion all seem to reflect certain properties of rhythm type, but leave open in what way they do reflect rhythmical structure, e.g. the alternation of perceived strong and weak patterns. It was shown that with the exception of dispersion metrics based on vocalic variability, the metrics were relatively unstable and can be seen as a more detailed description of phonological complexity and vocalic reduction. If these two figures are regarded as sufficient to capture rhythmic phenomena, then these measures provide valuable insight into the timing organization of various languages.

With regards to the coupled-oscillator models, one can conclude that they provide the only existing holistic approach to the modeling of rhythmic phenomena. However, they apparently are difficult to apply to real-life speech data. Many of their hypotheses have so far only been provided empirical support for in laboratory speech. The exception to this is the data-mining approach put forward by O'Dell and colleagues. They used the coupled-oscillator idea to gain insight into rhythmically relevant levels in Finnish conversational speech. Although they only model long-term properties of a language, their approach appears to be extremely helpful to identify the relevant levels of prosodic organization — especially in languages that are not fully classified in terms of rhythm.

Also, time-series analysis seems to be a fruitful approach to gain insight into rhythmical organization. However, due to the lack of interlingual studies, it seems somewhat early for drawing too many conclusions from the few results gathered with this approach. However, the (very small) tendency of rhythmical alternation has been shown to be relevant in both stress-timed English and syllable-timed French. Thus, we have at least hints that such tendencies are a universal property of (speech) rhythmic organization. Both time-series models and studies in deceleration across rhythmic units provide further support for the universal tendencies of rhythmical grouping relying on final lengthening at various levels of rhythmic organization.

Chapter 4

Multidimensional Classification of Speech Rhythm

In this chapter, a new integrative approach for rhythm research will be proposed. The apporach takes into account the insights and conclusions drawn from the previous chapters. Instead of suggesting one further rhythm metric that is trying to capture a single aspect of rhythm (e.g. temporal variability as the metrics described in section 3.1), the approach assembles a number of exploratory tools that are able to capture, vizualize and disentangle various rhythmic characteristics of a language or a specific speaking style. Such an approach is followed, as it should have become clear in the previous chapters that rhythm cannot be described as a one-dimensional property of speech, e.g. as more or less *variable* or more or less *stress* timed. Rather, rhythm should be seen as being constituted by a multitude of properties integrating into a — possibly typical — pattern. Such a pattern can be used in human language processing to deduce upcoming (rhythmic) events in perception or to plan the timing in language production. Such a pattern can be regarded as a language or speaking style specific meter. Since in Chapter 2, duration has been identified as the predominant acoustic domain of rhythm patterns, the poposed approach will start its quest in the domain of relative durations. Later, fundamental frequency will be explored as an additional acoustic parameter. The suggested collection of heuristics is divided into a data-mining approach aiming to detect rhythmic regularities and a supplementary approach aiming to describe the rhythmic patterns as long-term regularities. These can be used by a listener to make predictions concerning upcoming events in order to facilitate speech processing. The collection of heuristic tools will be exemplified referring to language data taken from two different languages, prototypically stress timed English and one prototypically syllable timed French.

4.1 Step 1: Exploring Local and Global Variability

4.1.1 Phonological Analysis

In the rhythmical analysis, it is of course useful to rely on previous phonological and phonetic analyses that may hint at hidden regularities in the respective language. A first step is the analysis of the phonological properties that predict the timing variability and rhythmic class along the stress-timed—syllable-timed continuum, e.g. phonotactic complexity and reduction (cf. Section 5). For our two prototypical example languages, this analysis leads to results represented in Table 4.1 [1]:

Typological Feature (Pro Stress Timing)	Eng	Fren
phonotactic complexity	+	-
reduction	+	-
free lexical stress	+	-
strong accentuational lengthening	+	-

Table 4.1: An overview of the typological findings correlating with stress- and syllable timing in the respective languages. Please notice that + and − only indicate tendencies. French phonotactics is more complex than that of many other known languages.

According to this very crude measure, we would now expect French to be much less variable than English which has been clearly identified as stress timed, thus, is likely to exhibit more temporal variation than French.

But with respect to rhythmical patterns, we can draw further hypotheses from our phonological analysis that go beyond mere overall variability. In Chapter 2 it was argued that a less complex phonotactics automatically leads to a higher syllable rate, thus, we expect French to have more beats per window of temporal presence

[1] Typological evidence for rhythm class that does not apply to any of the languages is not taken into account

or possibly — foot (Grouping principle 1). Also related to rate is grouping principle 5, which states that a high tempo may enhance the listener's preference to build groups with long or strong endings rather than strong beginnings. The lack of reduction may actually cause a certain amount of variability in interstress syllables, which may lead to a lack of perceived anisochrony in French sequences. Thus, combined with the grouping principles defined in Chapter 2, we are able to derive much more hypotheses related to French and English rhythm than an estimate of overall measurable variability. Still, in our next section we will also concentrate on finding evidence for the presence or lack of variability in French and English speech.

4.1.2 Visualization of Beat Timing in Time Delay Plots

The next question is how variability can be quantified and visualized straightforwardly. Many of the metrics discussed in Chapter 3 were criticized mainly based upon the fact that they only show a global tendency but are difficult to interpret. It is also unclear whether variability ought to be sought for using the local variability as in the PVI-metric or rather the global variability indicated by metrics based on dispersion. Metrics concentrating on vocalic variability have proved to give the most reliable results. However, in Section 2.2.3 it was argued that the domain of a beat in speech is likely to be the syllable. A description concentrating on vocalic intervals only disregards any syllabic consonants which may constitute a substantial number of beats in many languages such as English, German or Czech. Thus, an adequate description of beats in speech ought to search for syllables (or possibly morae, cf. Section 2.2.4) rather than vocalic intervals. Of course, the measurement of syllable duration does not provide us with an adequate correlate of the duration of a rhythmically relevant event, since the perceptual onset of a syllable is not identical to the beginning of the syllable itself. In order to estimate this beginning, the *p-center* of the syllable needs to be determined. The p-center depends on the duration of the onset and is roughly estimated with the equation proposed by Marcus (1981) (cf. Section 2.2.3). Since rhythm has to do with timing it would be practical if we could assign the beat a duration as a rough estimate of its strength or prominence. It has often been practiced to regard the interval durations between subsequent perceptual centers as equivalent to the temporal extension of beats. If this practice is transferred to speech, though, the onset of a subsequent syllable would enter the perceived strength of the previous beat. While I do not want to rule out the possibility of such an approach as being adequate, I prefer not to include the onset of a subsequent syllable in the beat duration of the previous one. At least for experienced listeners of a language, I assume for now that a syllable ending is also perceptually interpreted as the end of a beat. The coda consonants, however, ought to be included in the estimate of the beats' durations, since we know from phonological analyses (cf. Section 2.2.4) that the phonological weight, i.e. the perceptual strength of a syllable, is determined by the entire syllable rhyme. Thus, the syllable coda might play an important role in the perception of a rhythm and should not be left out in rhythmic analyses. Thus, in the subsequently presented method, "beats" or "beat durations" refer to the duration in between a syllables p-center (determined according to Marcus (1981)) and the end of the syllable's rhyme.

Rhythm metrics such as PVI and those based on dispersion (cf. Section 3.1) search for local and global variability in speech which could successfully be linked to typological rhythm classes. It would be a plus, if both types of temporal variability could be expressed simultaneously. A useful visualization tool for such an approach can be found in methods used in Nonlinear Time Series Analysis. Nonlinear Time Series Analysis have become increasingly popular for the description of deterministic, nonlinear data, especially in the description of economic and biomechanical data such as the detection of pathological vocal fold vibration (Little (2006)). The relative difficulty of obtaining significant autocorrelation at $lags > 1$ (cf. Sections 2.4.1 and 3.2.1) indicate further that linear techniques may be only partly useful in the detection of underlying regularity in speech rhythm data. For the moment being, Nonlinear Time Series Analysis will only be used in descriptive fashion. More precisely, time delay visualization, will be used in order to detect possible hidden regularities in rhythmic speech data. The idea is simple and has been used in physics in order to detect the behaviour of nonlinear, deterministic event series such as water dripping (e.g. Baier (2001)): By plotting the duration of a fundamental rhythmic event i on the x-axis vs. a subsequent fundamental rhythmic event $i + 1$ on the y-axis in a so-called time-delay cross-plot (cf. Figure 4.1), the relative timing of two subsequent rhythmic events at $lag1$ is expressed (cf. also Wagner (2006, 2007)). In order to compare durations across different speakers, speaking rates or speaking styles, all beat durations have been z-normalized[2], resulting in values where "0" indicates the mean duration and "-1" and "1" indicate the negative and positive standard deviation of the data series. Obviously, such a normalization factors out potentially important tempo-related effects of rhythm, such as the hypothesis that rate correlates negatively with perceived variability (cf. Section 2.3.5.1). Thus, such a visualization cannot be sufficient for a full exploration of speech rhythm. The following properties of a rhythmical event sequence can be identified in the proposed cross-plot:

- A tendency towards global isochrony across the data set, when the data plots cluster around the $\{0, 0\}$-co-ordinate, i.e. the point where two subsequent mean durations are indicated (cf. Figure 4.1).

- A tendency towards local isochrony as measured by the PVI, i.e. two subsequent similar beat durations is indicated in the plot whenever $beatdur_i \approx beatdur_{i+1}$, thus creating a point along the diagonal (cf. Figure 4.1).

- Typical relative durations across the data set clustering in particular areas of the two-dimensional space indicating rhythmically relevant transitions in the rhythmic pattern, e.g. they may indicate the beginnings or endings of groups, due to a local increase in time.

- A tendency for strict alternation in the rhythmic pattern. A tendency towards a regular pattern of *long-short-long-short* events is indicated by one cluster in the upper left (*short-long*) quadrant of the plot and another cluster in the lower right (*long-short*) quadrant (cf. Figure 4.2).

- Subsequent very long or subsequent very short events, indicating pronounced

[2] The z-normalized value z_i for any value x_i contained in a data series $S = x_1, x_2, x_3 \ldots$ is calculated as $z_i = \frac{x_i - \bar{S}}{\sigma(S)}$, with $\bar{S}$ expressing the mean of the data series and $\sigma(S)$ expressing the standard deviation of the data series. A problem is that the z-normalization is only useful if applied to normally distributed data. Beat durations, however, always show a significant amount of curtosis, since they certainly have a minimal duration but can become very long. Therefore, the following calculations have also been repeated based on log-transformed data showing a normal distribution, as suggested by Dellwo (2008a). The outcome of these calculations did not differ categorically from the ones presented here.

lengthening or reduction phenomena across two consecutive syllables (cf. Figure 4.2).

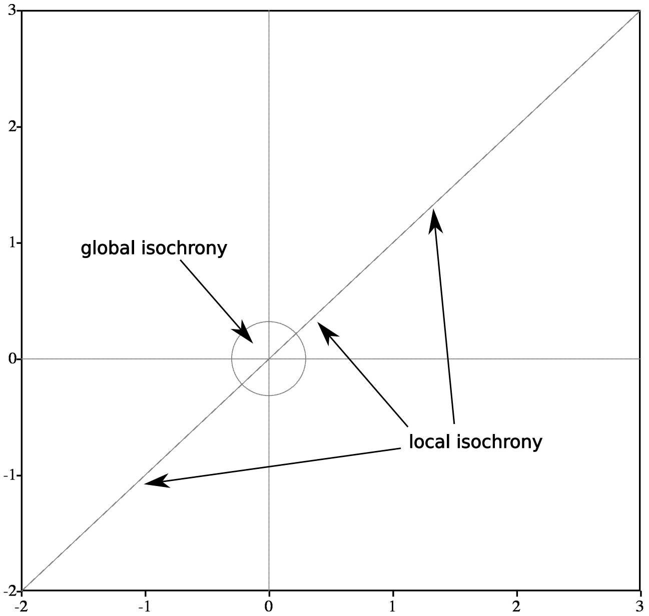


Figure 4.1: The time-delay plot representing the duration of a rhythmical event i on the x-axis and a subsequent rhythmical event $i + 1$ on the y-axis. Subsequent similar or equal durations as measured by the PVI are plotted around the diagonal. Global isochrony ought to show as a cluster concentrating around the $\{0, 0\}$-coordinate.

Relative durations also indicate the kind of grouping that is expressed by the increase or decrease in duration: The investigation of temporal patterns distinguishing iambic and trochaic grouping by Bröggelwirth (2007) and our own pilot study reported in Section 2.3.1 clearly indicated that strong relative increase in duration tends to be interpreted as final lengthening, the ending of a group, while a moderate relative increase in duration is interpreted as the beginning of a group. Thus, the plot should also indicate wether the transition between beats signals the end or the beginning of a local or global rhythmic pattern (cf. Figure 4.3).

Another possibly important timing feature that is visualized in time delay plots concerns general tendencies of deceleration that have been found to be important in earlier studies (cf. Section 3.2.2). Each local decelerating transition will appear above the diagonal, accelerating trends will appear below it.

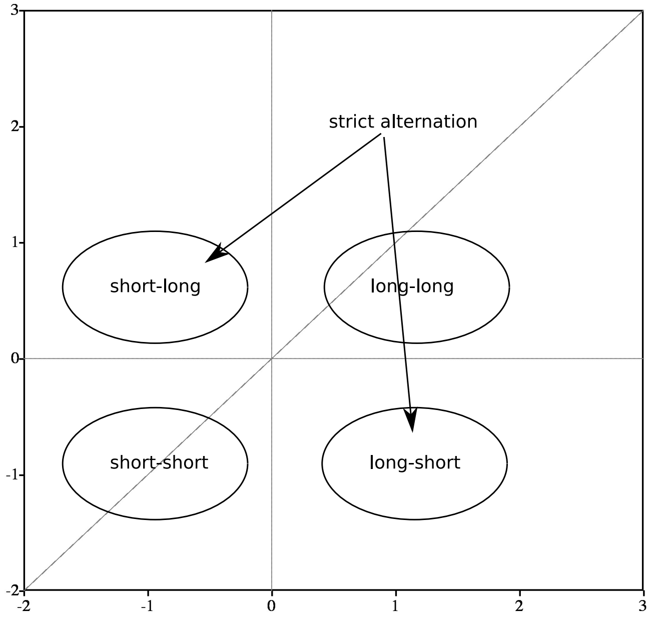


Figure 4.2: In time-delay plots, strict alternations show as two clusters, one in the lower right and one in the upper left quadrant, while sequences of short, reduced events show in the lower left, sequences of long events, show in the upper right of the diagramme.

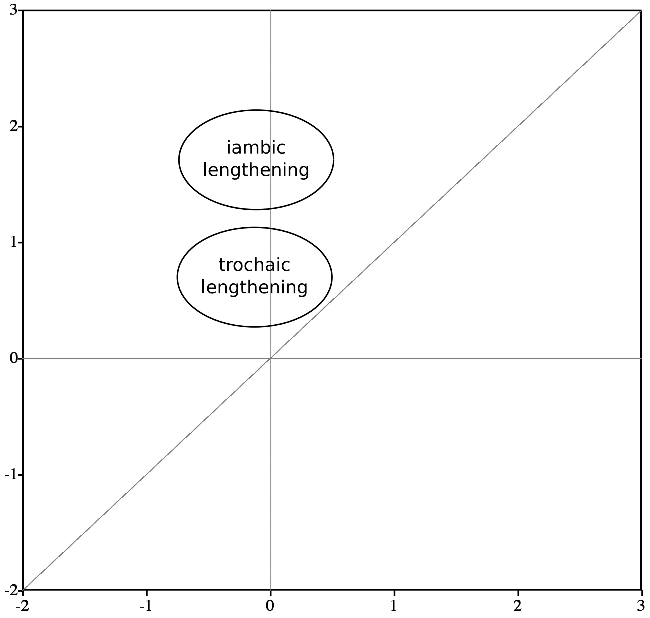


Figure 4.3: Final lengthening indicating the end of a group (= iambic lengthening) should plot in higher regions than lengthening effects indicating beginnings (= trochaic lengthening).

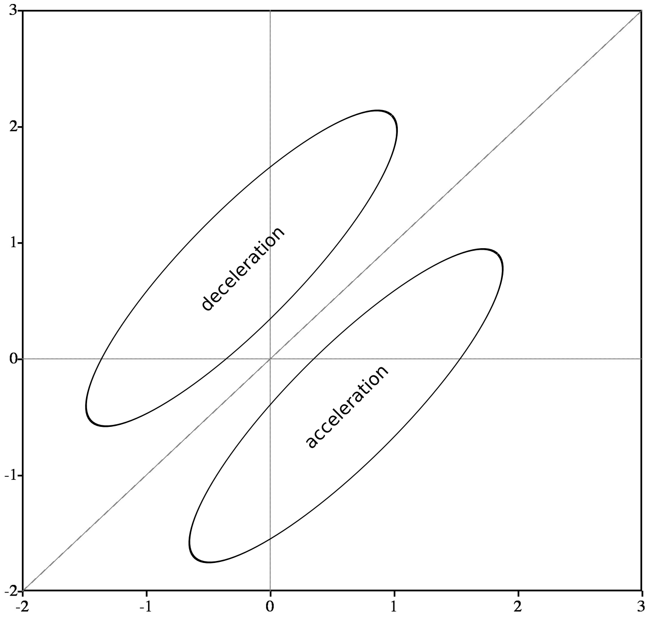


Figure 4.4: The location of deceleration and acceleration tendencies in time delay plots.

Of course, it is relatively uninteresting to get a picture of the various different temporal relations between successive beats without being able to relate these to their communicative function. In the visualization method proposed here, different colors are chosen in order to mark fuctionally different transitions between beats. Three types of functionally different transitions were chosen:

- Transitions to phrase final beats are highlighted in red. Transitions across phrase boundaries are ignored.
- Transitions to lexically stressed beats are highlighted in blue. (This includes transitions from unstressed to stressed as well as in between stressed beats.)
- Transitions to unstressed beats are highlighted in green. (This includes transitions from stressed to unstressed as well as in between unstressed beats.)

The following plots in Figure 4.5 show the differences for our two example languages English and French. The language material is taken from the BonnTempo database (Dellwo et al. (2004)). For each language, material from three speakers reading the same text passage in 5 different speech rates (approximately 450 syllables per speaker) is plotted.

The plots show clear differences between the protoypically stressed timed English and syllable timed French. For English, the plots shows a clear separation between lexically stressed and unstressed syllables that is also indicating a high degree of rhythmical alternation. Final lengthening is in most cases more pronounced than lengthening of lexically stressed syllables. Final lengthening is not always confined to the last syllable, in some cases the penultimate syllable is lengthened as well — probably due to coocurrance with lexical stress. In French, the overlap between transitions to stressed and unstressed syllables is striking — there is hardly any difference. Only final lengthening is clearly apparent in French, undermining its status as having no clearly durationally marked lexical stress with the help of a local increase induration. However, a highlighting of the end of phrases or stress groups is evident. Also, French relative durations cluster markedly around the $\{0, 0\}$-coordinate, showing a lack of durational variation in comparison with English. In total, we see that the predictions made by phonological theory could be confirmed in time delay plots. French lacks durational variation globally and hardly ever marks lexical

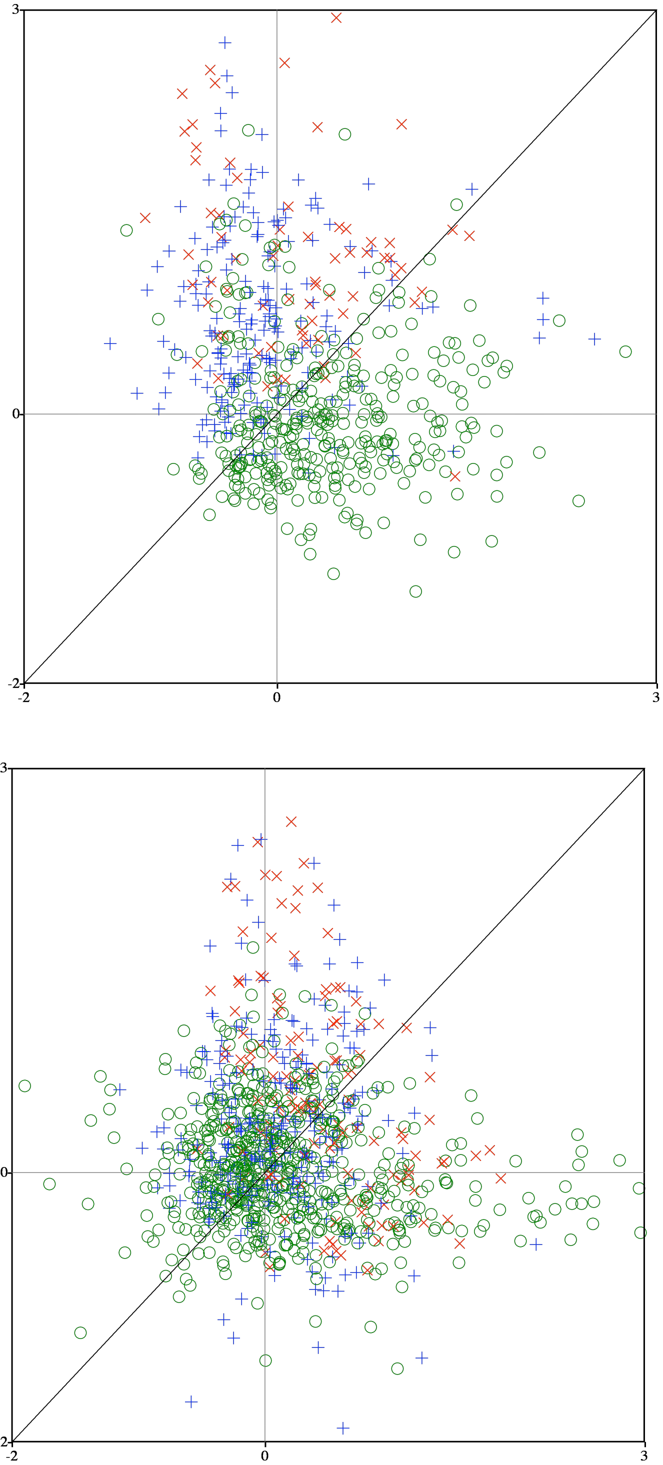


Figure 4.5: The graphs show relative beat durations in English (left) and French (right). Transitions to lexically stressed beats are marked in blue (+), transitions to unstressed ones in green (o), transitions to phrase final beats in red (x).

stress. In comparison to the plots produced by popular global rhythm measures described in Chapter 3, time-delay plots have the advantage to be directly interpretable along similar rhythm-related dimensions as have been detected in typological analyses. Of course, they have the advantage over phonological analyses that they are not limited to categorical classification. Instead, they are able to show very fine-grained timing differences on a continual scale.

Our exploratory study can go further, however — that lack of acoustic marking of lexical boundaries by an increase in duration in French is well-known — this lack certainly stands in sharp contrast to the common notion of French being "iambic", since iambic lengthening usually would require a pronounced lengthening effect. Still, it is possible that word boundaries are still marked in French and our graph simply did not highlight the correct type of transition. Therefore, we alter the plot and instead of highlighting transitions to word final syllables, those from word final syllables to presumably unstressed syllables are plotted. The resulting graph is shown in Figure 4.6. Here, we clearly see that a lengthening effect does take place — albeit relative to the following very short beat in the upcoming foot. Thus, the intuitive impression of French both having final lengthening at word boundaries and the empirical finding that is does hardly show word final lengthening are both true. A pronounced lengthening does not take place within the foot or group but relative to the upcoming one. It is interesting that the remaining transitions are concentrated left of the diagonal. This indicates a general tendency of lengthening in French propagating throughout the foot and is limited by lexical boundaries.

4.1.3 Quantitative Interpretation of Time Delay Plots

Time delay plots are a useful tool in order to explore timing relations that are perceived as typical rhythms in speech. In this section, some suggestions concerning the quantitative interpretation of time delay plots will be made. First, we calculate the durational differences for each transition $beatdur_{i+1} - beatdur_i$. By this, we tend to get positive values for decelerating transitions and negative values for accelerating transitions. Then, a one-factorial ANOVA is calculated for each languages revealing whether there exist significant differences between the three transition types. The results confirm the visual impression that in English, the three transi-

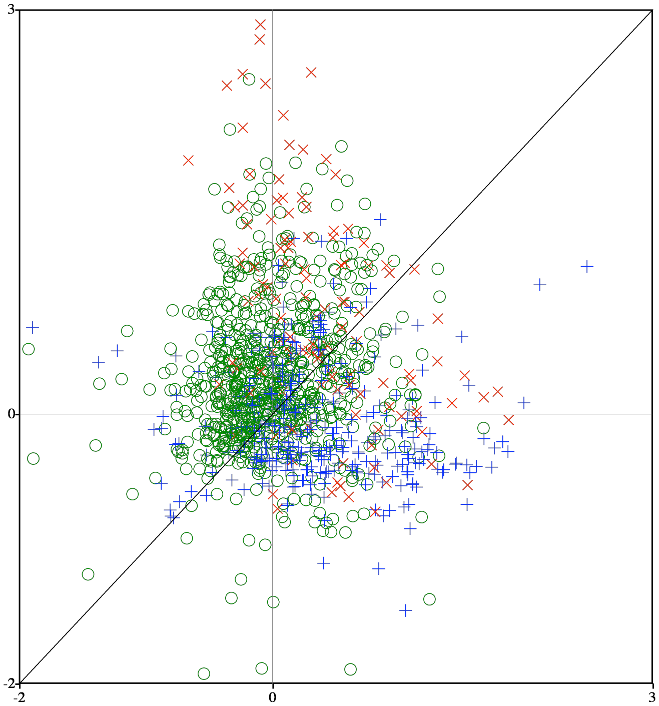


Figure 4.6: The plot shows French timing relations between transitions from word final beats to subsequent ones, marked in blue (+), transitions to other beats marked in green (o), transitions to phrase final beats marked in red (x).

tion types show highly significant differences between consecutive durations for the three transition types ($F = 223; df = 2; p < 0.0001$) while for French, there are no clear differences between the groups' variances ($F = 1.44; df = 2, p = 0.24(n.s)$) (cf. Table 4.2).

Transition Type	*Mean (z)*	*Variance*
	English	
to phrase final	1.02	0.78
to stressed	0.79	0.7
to unstressed	-0.39	1.72
	French	
to phrase final	0.4	1.5
to stressed	0.19	0.36
to unstressed	0.24	0.42

Table 4.2: The table shows the mean difference and variance between consecutive durations $beatdur_i + 1 - beatdur_i$ (z-scores) for the various transition types in our two example languages.

In order to get a better understanding concerning the nature of the differences between the various transition types per language, we calculate the mean values for each "transition type" (to stressed, to unstressed, to final) on both dimensions thus identifying the most typical transition for each language. Then we calculate the Eucledian distance between the different types of transition for each language and see whether our visual classifications are confirmed. Eucledian distance is calculated between two points in a two-dimensional space $P = (p_x, p_y)$ and $Q = (q_x, q_y)$ with the help of Equation 4.1:

$$Distance = \sqrt{(p_x - p_x)^2 + (p_y - q_y)^2} \tag{4.1}$$

It is important to keep in mind that Eucledian distance is identical to the distance we would measure with the help of a ruler. It has of course no perceptual relevance, i.e. the difference in relative timing may look quite different on a perceptual scale. Still, we can show the relevance of this metric for a typological timing distinction. The Eucledian distances between the three transition types are calculated for a number of languages contained in the BonnTempo database (Dellwo et al. (2004)), some of which have been classified as stress timed (English, German), others as syllable timed (French, Italian) while Polish timing has been notoriously hard to describe, its accentual lengthening being extremely subtle (Klessa (2006)). Results are shown in Figure 4.7, where it is also evident that there exist clear distinctions between the various languages, albeit in different dimensions. While prototypically stress timed English shows the highest distances between transitions to stressed and those to unstressed, German behaves similarly but less strongly. French shows much less difference between transitions to stressed and those to unstressed, but also a small distance between stressed, unstressed and final transitions. The difference between transitions to stressed and unstressed beats is very similar in French and Italian, both having been claimed to be syllable timed, while final lengthening in Italian has more in common with stress timed languages. Polish behaves completely different by showing hardly any effect of accentual lengthening, but a final lengthening slightly stronger than French. In total, this simple comparison delivers much direct information about language specific timing that goes beyond a binary classification of syllable and stress timing.

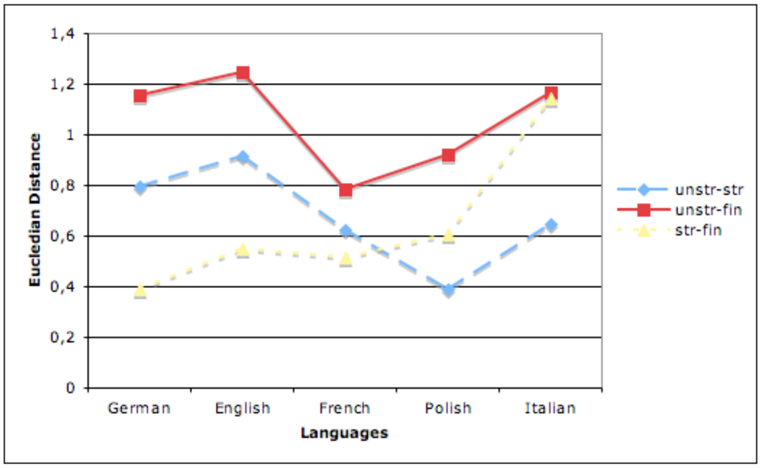


Figure 4.7: The graph shows the Eucledian Distance between the transition types *unstressed—stressed*, *unstressed—final* and *stressed—final* for our two example languages. It is evident, that there are distinctive differences for each language.

However, the Eucledian distances plotted above are mean values that do not capture the variability contained in the data or whether the distinction between the various transition types is systematic and stable enough to generalize. If a systematic relationship between relative timing and transition type exists, it could be used by a listener as a perceptual cue when processing the acoustic input. In order to test whether such a systematic relationship exists in the data, a K-Nearest Neighbor (KNN) classification[3] with $k = 5$ was performed for our two prototypical languages, English and French. The KNN-classifier was used because it also builds on Eucledian distances and thus provides a method that can be straightforwardly linked to the visual interpretation of the time-delay plots. Thus, the KNN classifier here primarily serves as an evaluation of the visual interpretation method—mainly for this reason, no model optimization e.g. by cross-validation was performed. The material used for classification was again taken from the BonnTempo database (roughly 2400 transitions per language) and it was tested whether the different transition types "to unstressed beat", "to stressed beat" and "to final beat" could be distinguished. From the material in the database, 66% was randomly chosen for training and 33% was used for testing. The hypothesis was, that such a classification would work for English which puts each transition in very different categories, while a classification would prove to be much more difficult for the French data. The results confirm this hypothesis quite clearly: While the classifier reaches 73% overall accuracy for English, it performs a lot worse for French with overall 63% accuracy. When regarding the different classes that were to be predicted, the results become even clearer. While the prediction is clearly above chance level (calculated based on the relative amounts of transitions in the data) in all three transition categories for English, and extremely successful in classifying the transitions to stressed beats, it performs a lot worse for French. Here, the classifier performs slightly better than chance but the improvement rate is clearly below the results for English, especially with respect to the transitions to beats categorized as stressed. The high accuracy for transitions to unstressed beats is the result of the large amount of such transitions in the data itself. An overview of the results is given in Figure 4.8.

Of course, our visual interpretation already indicated that we should not conclude from these results that French has not pattern at all — we can merely deduce that so far, that French does not have a timing pattern organized similar as English — with pronounced timing differences between a local unstressed and upcoming stressed or final syllable. Based on our visual interpretation in Figure 4.6, another

[3] See e.g. Dasarathy (1991) for an in-depth description of the approach

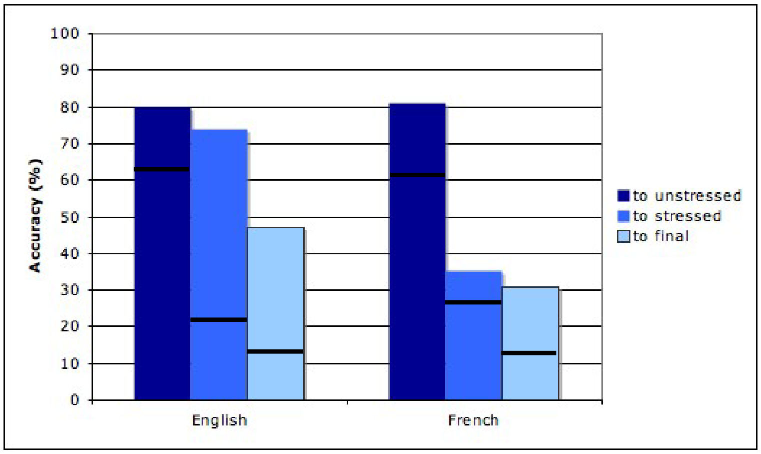


Figure 4.8: The KNN classification based on beat duration transitions was able to predict the various transition types in English well above chance level, in French the classification fails — especially for the transitions to stressed beats. The black horizontal bars indicate the chance level for each transition type based on the distribution in the original data.

KNN classification based on the same data and under the same conditions was performed to find out whether *poststress shortening* rather than stress lengthening can be detected by the classifier. Indeed, the classifier performed better in detecting poststress syllables based on their tendency to be short than in detecting stressed syllables based on their tendency to be long. Still, it only reached an accuracy of 40% (chance level: 27%) It therefore seems that finding instances of shortened syllables alone is insufficient to describe any systematic regularity in French rhythm. In consequence, another attempt to find a clear pattern in French timing is performed with respect to the distinction between deceleration and acceleration. Figure 2.19 gave the impression that French contains many sequences which are moderately decreasing and occasional sequences of long followed by short sequences. Thus, timing appears to be more smooth in French, except certain “dips” which seem to be restarts of a foot. On the contrary, the English data seems to contain more sequences showing a steep rather than moderate increase in duration. In order to confirm this impression, a simple count is performed for both languages, calculating the number plots for the four different quadrants depicted in Figure 4.3, thus filtering out predominant timing relations for both languages. The results of this count reveals that both English and French show a distribution deviating from an expected equal spreading across all four quadrants ($\chi^2; df = 3; p < 0.05$). Furthermore, they show an interesting difference in this distribution, namely a predominance for *long-long* and *long-short-* sequences in French, while English is slightly dominated by *short-long*-sequences (cf. Figure 4.9). The comparatively large number of *long-long* sequences in French may add to the impression of French having less alternations and a general tendency towards deceleration. Both languages favour *short-short-*sequences the least.

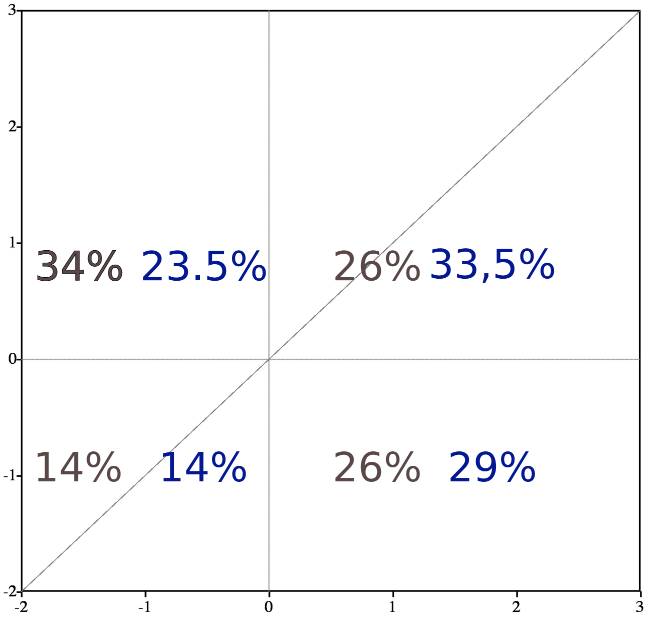


Figure 4.9: The Figure depicts the different concentrations of sequences in the different relative timing quadrants. The left (grey) number indicates the percentage of timing relations in English, the right (blue) one for French. The count indicates a slight predominance of French to favour sequences of long or long-short sequences, while English slightly favours short-long sequences.

4.2 Step 2: Exploring Long Term Timing Regularities

Figure 4.6 at least visually indicated that French timing may be organized between the *supposedly* stressed and the following beat. However, this tendency was only moderately confirmed by the KNN-classification, probably due to the large remaining overlap between beat transitions of both classes. Therefore, in this section a different strategy is proposed to look for long term timing patterns within and across foot boundaries. The foot is chosen as an important unit of rhythmical organization, since in Chapter 2 it was found to be linked to auditory time perception, namely the window of temporal present. Beats assembled into a foot, are typically perceived as belonging together thus forming a perceptual gestalt. Since the window of temporal present is apparently constrained by absolute time, ranging between 400—600 ms, we will concentrate on absolute time values rather than normalized durations as in the previous section. We are now looking at long-term characterstics of such windows, in order to detect timing regularities that listeners can learn in order to form certain expectations concerning upcoming rhythmical events. Such long-term expectations are what we defined as *meter*. Metrical hypotheses can be verified or falsified by the perceptual input, but should make good predictions in the majority of cases. It can be expected, that such hypotheses are much more often falsified in ordinary speech compared to stylized speaking styles such as poetry. However, traces of such regularities should remain. In a first approach to visualize such long term foot-internal tendencies for our two example languages, the average timing properties were plotted for feet of different sizes in order to unveil differences, similarities and typical long-term patterns. In a first step, mean durations and variation coefficients were calculated for binary, ternary and quaternary feet for our two prototypical languages. The results are depicted in Table 4.3 and Figure 4.10.

Language (Footlength)	Metric	Syll 1	Syll 2	Syll 3	Syll 4
French (2)					
	mean dur (ms)	144	208		
	var. coeff (%)	50	46		
English (2)					
	mean dur (ms)	227	169		
	var. coeff (%)	35	57		
French (3)					
	mean dur (ms)	112	159	161	
	var. coeff (%)	26	22	60	
English (3)					
	mean dur (ms)	215	122	129	
	var. coeff (%)	47	42	74	
French (4)					
	mean dur (ms)	125	162	191	181
	var. coeff (%)	44	36	44	48
English (4)					
	mean dur (ms)	169	117	155	90
	var. coeff (%)	36	29	45	44

Table 4.3: Mean durations and variation coefficients for binary, ternary and quaternary feet in English and French.

With regards to durational variation expressed in the mean-normalized variation coefficient, no clear preference for either language showing less or more is evident. The calculations show once more confirm that accentual lengthening effects in French are more subtle compared to English. However, it also shows that foot final lengthening in French is only evident for binary feet. In longer feet, the durational increase propagates throughout the foot and is not restricted to the final syllable. In English however, we can clearly see the tendency of binary alternation in longer feet, which is not present in French. Thus, it seems that the French “iambic” effect is not the result of a local lengthening of the foot final syllable, but results out of a deceleration tendency throughout the foot. Alternatively, English

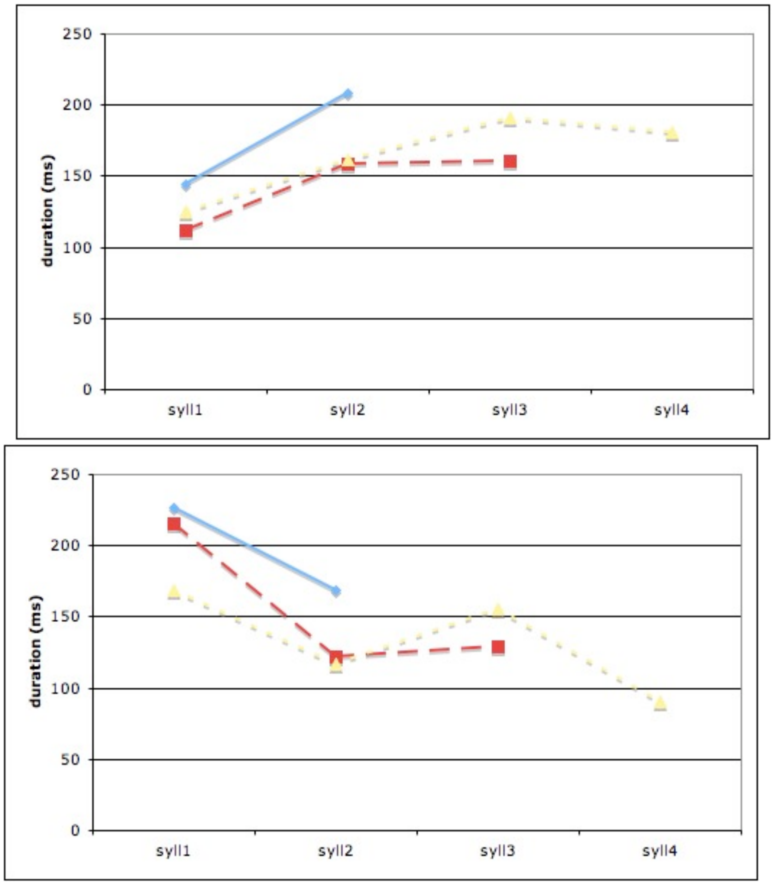


Figure 4.10: The Figures show the average durations in binary, ternary and quaternary feet for French (left) and English (right).

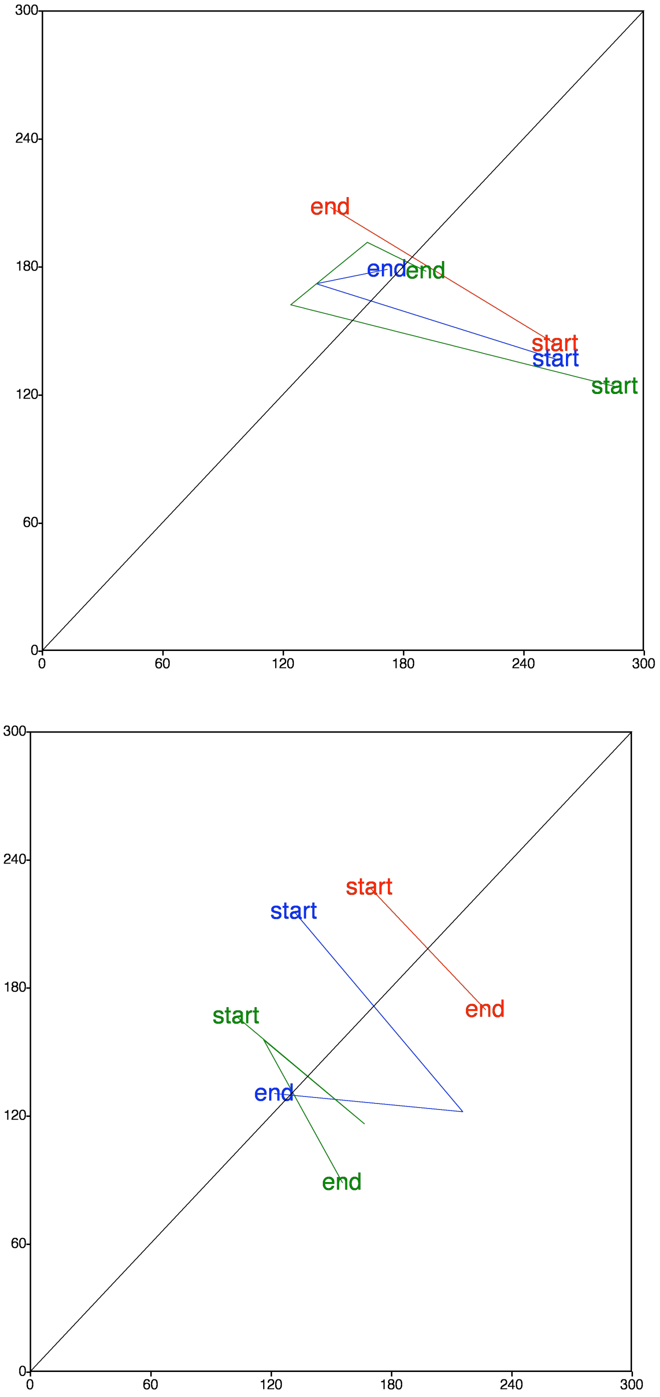


Figure 4.11: The Figures show the average relative durations in binary (red), ternary (blue) and quaternary feet (green) for French (left) and English (right). The beginnings and ends of each group are indicated with the labels *start* and *end*. French groups typically start with a transition from a long to a short syllable and then have the tendency to increase beat durations throughout the foot. English starts with a long beat followed by one or two short ones. In quaternary feet, a strict alternating pattern is evident.

is once more shown to be characterized by a strong tendency towards alternation, creating a regular trochaic or — in the case of ternary feet — dactylic pattern. It is well possible, that this foot-internal difference between French and English is responsible for many rhythm related effects. In Section 2.3.1, the perceptual effect of *time-shrinking* was introduced. Time-shrinking has the effect that in decelerating interval sequences, the consecutive intervals are perceived as more isochronous than they acoustically are. Time shrinking can propagate across several sequences but is blocked by strict alternation, as it is typically found in English. Thus, time shrinking may at least partly account for the often reported perceptual impression of French being more isochronous than English.

The phenomenon of time shrinking can be traced in Figure 4.11, where the long-term patterns are once more illustrated in the time-delay plots introduced in the previous section, only with absolute durations being shown. The lines show the shapes of relative durations followed throughout the different feet, e.g. in the English plot, the binary foot (red line) is initiated by a transition from a shorter to a longer beat and ends with a transition from the long to another short beat. Furthermore, the plots show the typical French pattern of starting with a short beat — relative to the previous foot-final one and then showing a tendency to lengthen, though subtly, throughout the foot. The foot-final beat is (on average) not longer than the penultimate one. The similarity between the patterns for French feet of different length is striking! Another rhythm related phenomenon that can be traced with the help of these long-term time-delay plots is *compensatory shortening* introduced in Section 2.4.1.2. In the case of compensatory shortening, i.e. the tendency to shorten beats as a function of the number of syllables contained in the foot, the long-term patterns would have to be distributed in different regions of the graph, with longer feet being oriented more towards the lower left quadrant. Obviously, no compensatory shortening takes place in French, since the different patterns almost overlap in the two-dimensional space. In comparison, the English long-term patterns show a very pronounced distance between stressed and unstressed beats. Also, they show the strong tendency towards alternation in the quaternary feet that is indicated by the almost perfect overlap of lines. Unstressed beats in ternary feet are almost identical in length (on average). In English, all foot internal syllables are affected by compensatory shortening, the stressed ones somewhat more than the unstressed ones. Since compensatory shortening obviously plays a role in English, a mere measurement of syllable duration as an estimate of presence of accentuation must fall too short — the timing in the vicinity of the presumably stressed syllable obviously plays an important role and needs to be taken into account when trying to find evidence for accentual lengthening. In the terminology of a coupled oscillator approach to timing, this indicates a higher coupling between feet and syllables in English compared to French. However, it is important to stress that a theory of coupling does not seem to be sufficient for a description of the foot-internal timing patterns that distinguish between French and English.

4.3 Step 3: Modeling Rhythmical Expectations

So far, we have shown a number of local and global timing differences between our two prototypical languages. In Chapter 1 the point was made that the major function of the phenomenon we call rhythm in speech is for the listener *predict* the timing and status of upcoming events and for the speaker to organize her production. When trying to model predictions, a number of statistical and machine learning tools can be used, most of which need an a priori specification of model parameters. Of course, a fully fledged model would also encompass high level linguistic information such as syntax, semantics and so forth. However, the model proposed here is more interested in the long-term regularities of timing that are learned based on input shaped by high level lingustic information but do exist independently. Naturally, our native language shapes our rhythmical expectations: The way our syntax structures the alternation of content and function words may have a strong impact on the way we expect a distribution of stressed and unstressed linguistic units. These relationships are not to be explored in this thesis — for once, because it would render a typological comparison tremendously complicated, second, because rhythmical expectations do not necessarily need the input of comprehensive language to become activated: When listening to a language that is completely incomprehensible to the listener, she may still have the impression of a characteristic rhythm—influenced by her native expectations. Thus, the rhythm may exist without the linguistic input, though it may have been shaped by it in first place.

In order to model the rhythmical expecations for our two example languages, a Bayesian Belief Network[4] was constructed based on the duration distributions collected in the Bonn Tempo database. A Bayesian Network models the conditional probabilites existing between the model variables. E.g. the probability that an event will take place in the future can modelled based on the probabilities that previous events have taken place. The conditional probabilities between the various model variables are captured by Baye's Rule, a law of probability which was first described by the mathematician and theologician Rev. Thomas Bayes (ca. 1702-1761). Bayes Rule is expressed as

$$P(b|a) = \frac{P(a|b) \times P(b)}{P(a)} \tag{4.2}$$

where $P(a)$ is the probability of a, $P(a|b)$ is the probability of a given b. We can now compute the probability that in French, the foot-final syllable is characterized by deceleration given an acceleration in the prefinal syllables in the foot ($P(final_{dec}|prefinal_{acc})$). It is built of the probability that $P(prefinal_{acc}|final_{dec})$ and the *a priori probabilities* $P(final_{dec})$ that a final syllable is decelerating and $P(prefinal_{acc})$ that the prefinal syllables are accelerating. The probabilities for the different event types were again calculated based on the material in the BonnTempo database. Thus, our model tells us what the listener may expect to come given a particular sequence of previously heard rhythmical events.

Bayesian Belief Networks are graphical models. Their illustration is built on directed acyclic graphs. Each variable, is represented by a node and the relationships between the different variables are denoted by edges indicating causal relationships between variables and the direction of the edge indicated by the arrow indicates the direction of causality. The way the *nodes*, e.g. denoting syllables, influence each other is expressed by the conditional probabilities defined between the various model *states*, e.g. whether a syllable is decelerating or accelerating.

For each language, a Bayesian Belief Model was built using three nodes for the initial, medial and final syllable of a foot. Thus, the model explains conditional probabilities for binary and ternary feet, constituting a large amount of feet across both languages. Each node can have three states, *deceleration*, *equal* and *acceleration*. These states are characterized by the conditional probabilities (=beliefs) that a syllable is audibly longer (deceleration), equally long to (equal) or shorter (acceleration)

[4] Nidermayer (2003) provides a nice introduction to the approach of Bayesian Networks

than the preceding one. The decision whether a durational deviation is audible was based on a 15%-JND threshold. Notice that this threshold is conservative, given the 10% threshold proposed by Quené (2007) for running speech. Expectations are calculated within feet, but also (indirectly) across foot boundaries, by integrating the conditional probabilities for the initial syllable relative to the preceding foot-final syllable. The foot was chosen as a unit of modeling expectations because it has been argued throughout this thesis that feet are organized within relevant timing windows.

The resulting Bayesian Networks (cf. Figures 4.12 and 4.13) convincingly show the different timing organization in both English and French resulting in different timing expectancies. For French feet, the probability is very high that it starts with a local acceleration while then decelerating towards the end. The opposite is true for English. Both languages show that the trend for deceleration is most stable around the edges of a foot while more variation is expectable in the medial syllable. The Bayesian Belief Network further undermines our previous findings that the timing organization in English is mostly based on a local deceleration on the initial foot of a group, while French is characterized by a acceleration of the beginning of a foot. In terms of grouping organization, this can be interpreted that French uses structural accents as predominant phrase internal grouping strategy — it marks endings — while English uses phenomenal accents as predominant grouping strategy — it marks beginnings. The failure to find evidence for French grouping probably has its cause in ignoring the timing relations between a group final and the postfinal syllable, since the structural accentuation is rather caused by a strong acceleration after the group final syllable than by a local deceleration of the final syllable itself. Alternatively, one could say that French marks group beginnings by a local acceleration. Experiments with the evidence for the different states show that the probablities keep stable in both Bayesian Belief Networks, thus it is likely that the probabilities cause expectations that are cognitively imprinted in the listener.

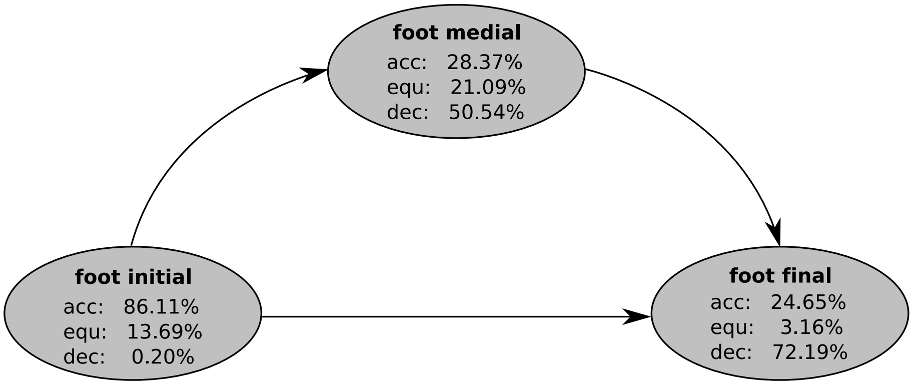


Figure 4.12: The Figure shows the graphical Bayesian Belief Network for timing expectancies in French feet. They are characterized by a strong probability to start with a syllable that is very short relative to the preceding one and have the tendency to decelerate until the end.

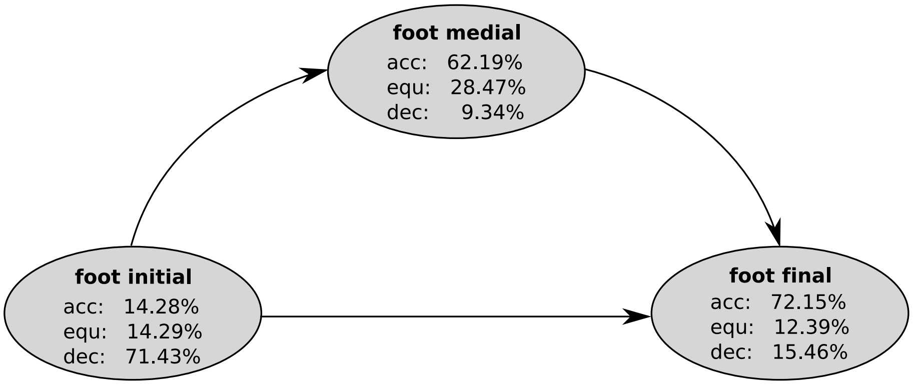


Figure 4.13: The Figure shows the graphical Bayesian Belief Network for timing expectancies in English feet. They are characterized by a strong probability to start with a syllable that is long relative to the preceding one and have the tendency to accelerate until the end of the foot.

4.4 Step 4: Investigating Rhythmic Patterns in other Acoustic Domains

So far, we have only investigated rhythmic patterns in the temporal domain. In our theoretical investigations in Chapter 2 it was concluded, that timing patterns indeed seem to constitute the fundamental grouping relationships existing between and across different beats. However, it was also concluded that timing can either lead to ambiguities with respect to grouping, e.g. the listener has to decide in the case of an increase in duration whether this marks the beginning of a new rhythmical group or indicates the end of a group. It was also suspected that languages use additional acoustic cues in order to disambiguate such patterns, e.g. by marking the beginning of a larger rhythmical group with a pitch accent and or an increase in spectral intensity (increasing prominence) but not doing so towards the end of a group. That way, a structural accent indicating an ending can be distinguished from a phenomenal one indicating the beginning of a group (cf. Section 2.3). Of course, marking the structural accent with a pitch increase but not doing so at the beginning of a group would be an alternative option of grouping — indeed, Vaissière (2002) reports exactly these two strategies existing simultaneously in French intonation, although the more likely case would be a pitch accent towards the end of a group. Certainly, there exist a multitude of possibilities how intonation, duration and further acoustic cues can interact in order to generate a rhythmic impression. One could say, that duration delivers the rhythmical frame while intonation and possibly intensity related features render this frame more complex.

Obviously, such various types of grouping may be exploited language or speaking style dependently. To illustrate this point, Figures 4.14 and 4.15 show the acoustic realization of two poetic lines taken from German poetry, one iamb and one trochee, respectively. Both lines were read by a professional actor. It is striking that in the trochaic versions, fundamental frequency and durational increase indicating foot beginnings are often disparate, while in the iambic version they are mostly parallel. While this is only anecdotal evidence, it may be a hint to an important distinction causing the impression of two different meters. A real disentangling the contributions of various acoustic cues to rhythm is not carried out in this thesis — however, two proposal for further investigations in this area are made in the following:

- It is important to take into account the contributions of various acoustic parameters to mark both beginnings and ends of rhythmical groups. Maybe, the end marking is enhanced by a pitch event or the beginning or both.
- It is probable that different strategies of boundary markings are employed at different prosodic levels, e.g. a language may decide to mark beginnings rather than ends at foot level, but ends rather than beginnings at phrase level. Of course, both is possible — the realization strategies may of course differ.
- It may be useful to look for positive correlations between various acoustic parameters at sensitive spots in order to perform such an analysis on languages about which little is known, e.g. beginnings or ends of content words, syntactic phrases etc.

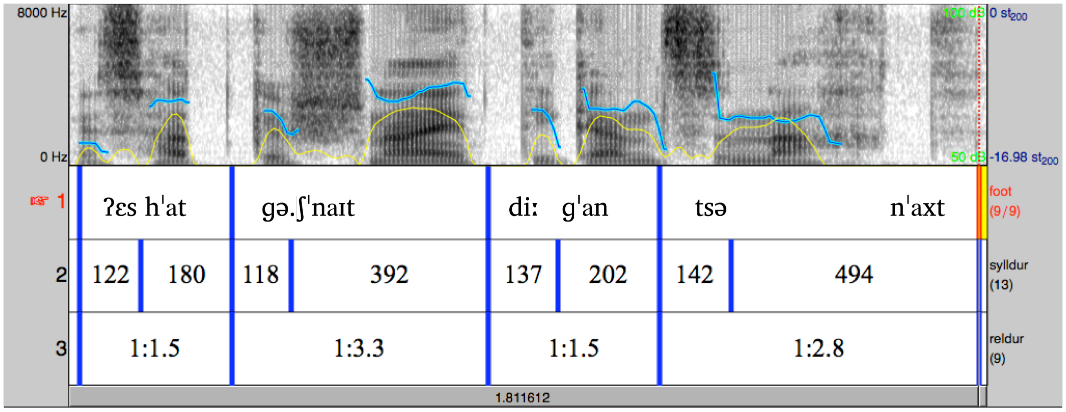


Figure 4.14: The Figure depicts a professional author's production of a German poem with an iambic meter (one line). The iambic foot final stress is produced with a pronounced increase in duration relative to the preceding unstressed syllables. Increase in duration is accompanied by an increase in intensity and fundamental frequency.

In the following subsection we will explore how our two example languages employ fundamental frequency as a means of rhythmical grouping.

4.4.1 Fundamental Frequency and Grouping Strategies

Our previous analyses detected duration cues indicating rhythmical grouping in our two example languages. We will now explore whether the groups thus identified

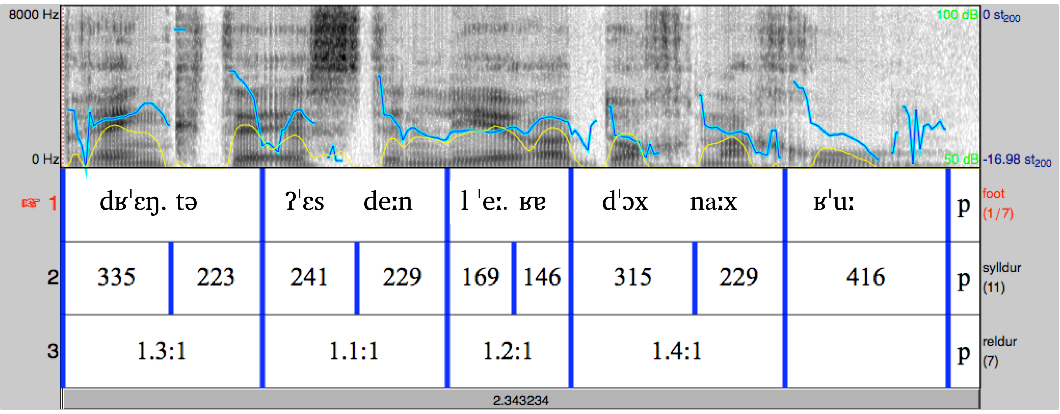


Figure 4.15: The Figure depicts a professional author's production of a German poem with a trochaic meter (one line). The trochaic foot initial stress is produced with an increase in duration, albeit less strong compared to the iambic lengthening. The prominence intensifying acoustic parameters intensity and fundamental frequency play an unclear role.

based on relative durations across important boundaries receive further acoustic enhancement via fundamental frequency cues. Thus, our two-dimensional picture will be added a third dimension, in which the fundamental frequency changes will be captured. The dynamic pitch change between the two successive beats is calculated as the semitone difference between the average fundamental frequency of the different beats (cf. Equation 4.3). By regarding the relative difference between two subsequent beats, increases and decreases of fundamental frequency are treated alike, i.e. low "valley" accent are taken into account as well. It is possible that late accents with reach their peak after the accented syllable are certainly not captured adequately by this simple approach.

$$diff(st) = 12 \times \log_2 \frac{\bar{f0}_i}{\bar{f0}_{i+1}} \tag{4.3}$$

In a first step, the relative fundamental frequencies are calculated for both languages across the same transitions that had lead to the most stable patterns in the previous duration ananlyses, i.e. for English, transitions to foot initial, stressed beats, transitions to phrase final beats and transitions to all other, unstressed beats were compared. For French, transitions from groupfinal (stressed) to groupinitial (unstressed) beats, transitions to phrase final and transitions to all other beats were compared. In a first step, mean differences were compared across the different groups (cf. Table 4.4). For English, an ANOVA ($df = 2; F = 49,0; p < 0.0001$) including a post-hoc Tukey test revealed that — as expected — all three transition types indeed show a highly significant different f0-dynamics. Transitions to phrase final beats show the strongest frequency dynamics, followed by transitions to stressed beats. Transitions to unstressed beats show the least difference in fundamental frequency change. For French however, the ANOVA only reveals a clear difference between the phrase final and the other transitions. This could be taken as evidence that the only rhythmically relevant transition type which is intonationally marked is located at phrase boundaries. However, just as with respect to duration, another strategy is attempted to unveil the rhythmical pattern in French lying below the level of the prosodic phrase. So far, the f0-differences were assumed to be organized along the same dimensions as durations. Another possibility is, however, that they are organized independently: While duration slowly increases until the end of a foot and then making a reset across the foot boundary, fundamental frequency might be mark the end of a foot rather than the transition to a new one. Therefore, for French — analogue to the English examples — f0-differences between the (prefinal) unstressed and (foot final) stressed syllables were computed as well. Now, a direct comparison of the two transition types indeed reveals a significant (two-tailed t-test; $p < 0.05$) difference between both transition types: Transitions to French foot final beats show a stronger fundamental frequency dynamics than those to unstressed beats. Summing up, French rhythmic patterns are more complicated than English ones: While in English, duration and fundamental frequency changes can be called to behave synchronously, showing the strongest dynamics at the transition to stressed or phrase final beats, French is more complex: Here, duration increases subtly until reaching the stressed, foot final syllable which is also marked by an increase in f0. The strongest dynamical change, however, takes place between the foot final and the foot initial beat. Thus, in French, fundamental frequency and duration dynamics can be said to be asynchronous.

	English	
Transition Type	*Mean (st)*	*Standard Dev. (st)*
to phrase final	2.4***	1.9
to stressed	1.5***	1.2
to unstressed	1.3***	1.3
	French	
to phrase final	2.2***	1.9
from stressed	1.5	1.8
from unstressed	1.6	1.6
to stressed	1.7*	1.6
to unstressed	1.5*	1.6

Table 4.4: The table shows the differences in mean fundamental frequency differences (in semitones) for the various transition types in our two example languages. Within languages, significant differences are starred (*= $p < 0.05$; ***= $p < 0.001$))

The tendencies are now illustrated by adding a third dimension of f0-differences to the time delay plots introduced earlier. For the different transition types, the plots will show a typical clustering — however, while the English plots show a unique pattern both in the dimension of duration and f0 for the three transition types identified as rhythmically important, the French plots only show a trend in the direction of a fundamental frequency difference when plotting transitions to stressed beats.

4.4.2 Fundamental Frequency and Rhythmical Expectancy

Another important aspect of rhythmical function that may be investigated in a domain other than timing relates to the aspect of rhythmical expectancy. In Section 2.4.1.2 we reported about production studies by Port and collaborators who found that rhythmically stressed beats in a prosodic phrase are preferrably aligned at anchor points dividing the phrase into simple integer intervals. Such anchor points were argued to be language specific. Of course, such production strategies are important for perception or rhythmical expectancy as well because they provide the listener with important information about possible future occurrances of important (= stressed) information. In the production studies by Port and colleagues, speech was produced under laboratory conditions using highly stylized monosyllabic stressed words and a task where speakers had to align words to the high tones of a metronome beat. However, if their reported tendencies are indeed tendencies of language production and perception, they should remain stable or visible at least to a certain extent in less constrained but still fairly controlled language production data such as read speech. It is furthermore unclear, whether such alignment in speech is also governed by intonational phenomena. In order to get a glimpse whether such tendencies really exist in read speech, the production data in the BonnTempo database was examined for our two example languages.

The timing of stressed beats normalized relative to the duration of the embedding intonational phrase was plotted — a stressed syllable (p-center beginning) aligned with the beginning of an intonation phrase receives the value 0, a syllable aligned with the end of it receives the value 1. In reality, these cases do not occur because phrase intial syllables following a pause and syllable final stresses are ignored due to the complex interactions with preinitial pauses, final lengthening effects, boundary tones etc. Also, a p-center is never identical with the beginning of the syllable itself, so there must be a time span between the phrase beginning and p-center placement. Next the absolute difference between the mean fundamental frequency in the stressed syllable relative to the preceding (for English) or following one (for French) was calculated in octaves. By calculating the difference rather than measuring wether the stressed beat was produced higher, low tones are allowed to produce the impression of an increase in prominence as well as high tones. The choice for the leftward and rightward comparison was based on the relative duration patterns showing a high between prestressed and stressed in English, but the strongest contrast between stressed and poststressed syllable in French.[5] The f0 differences are plotted on the y-axis for each stressed syllable beat that is aligned somewhere within the embedding phrase. The resulting plots are depicted in Figure 4.18. The plots indeed unveil differences in alignment between both languages.

[5] In the previous section it was argued that foot finally, French is characterized by an increase in fundamental frequency dynamics. However, it is assumed that the alignment preferences identified by Port and collaborators are more adequately detected under recurrance to duration properties, which is for French the transition from foot final to subsequent foot initial beat.

While French shows stong concentrations around 0.2, 0.5 and 0.6 of the embedding phrase, English speakers produce an early stress prior to 0.2 (probably due to the trochaic tendency of English) and show concentrations aroung 0.5 and 0.75. However, the English productions are concentrated less in the center of the embedding phrase. While such plots are certainly interesting, one should be careful to avoid an overinterpretation — given the strong influence of syntax and text on the timing of feet, it is important to look at more comparable material for a cross-language analysis than that contained in the BonnTempo database, which consists of translations and is not built on a premise of maximising phonological similarity. However, the method could serve well to study intralanguage stylistic variation.

4.5 Step 5: Investigating the Perceptual Reality of Rhythmic Regularities

After having detected certain language specific timing patterns and modelling pertinent timing expectancies, it will be explored to what extent the typical patterns create language specific rhythmical auditory impressions. In order to find out more about this relationship, two perception experiments were carried out. It was examined whether listeners were able to identify languages based on their typical timing patterns.

4.5.1 Stimulus Preparation

In order to have delexicalized but still rhythmically meaningful stimuli, binary, ternary and quaternary feet consisting of sequences of the syllable *ba* were recorded by a female native speaker of German. The different feet were then manipulated to match the typical foot durations of both French and English resulting in the identified patterns of decelerating or accelerating feet. After this manipulation process, there existed three typical feet for each language, a binary (2), a ternary (3) and a quaternary (4) one. In addition, a binary final foot was created based on median durations for both languages. In order to eliminate any influence of f0, the stimuli were given a flat f0-contour of 155Hz and an identical average intensity. All manipulations of duration and fundamental frequency were carried out with the help of the PSOLA algorithm (Charpentier and Moulines (1989)) implemented in the software package Praat (Boersma and Weenink (2008)). After manipulation, the average feet were concatenated into sequences of three feet each. The first two feet in each stimulus were either binary (baba), ternary (bababa) or quaternary (babababa) (3x3 stimuli for each language), the last foot was always the binary final foot. This resulted in 18 different stimuli altogether, consisting of different sequences of *ba* (see Table 4.5 for an overview).

Foot Pattern	Stimulus
222	baba—baba—baba
232	baba—bababa—baba
322	bababa—baba—baba
332	bababa—bababa—baba
342	bababa—babababa—baba
432	babababa—bababa—baba
442	babababa—babababa—baba
422	babababa—baba—baba
242	baba—babababa—baba

Table 4.5: The different stimuli generated for both languages in the perception task.

4.5.2 Experiment 1: Language Identification

In a first pilot experiment, the 18 stimuli were presented to 14 undergraduate students in an identification task. 11 students were native speakers of German, 3 were highly proficient in German but spoke a different native language (English, Russian, Romanian). All students claimed to have a good or very good knowledge of English, while 8 had no or only very little knowledge of French. The stimuli were presented in randomized order and presented twice. After each stimulus, the subjects had to decide whether it was rather French or English. The results show no ability to identify the intended language just based on average timing relationships. Rather, a certain bias to judge the stimuli as French was obvious for several subjects. When only taking into consideration those subjects that claimed to have a substantial proficiency in both English and French, there exists a marginally significant tendency to identify the stimuli better than chance. Overall, the subjects found the task very hard and reached only an identification accurracy around chance level. Obviously, the usage of delexicalized, unnatural stimuli was too difficult — also taking into account that previous investigations in the perceptual reality of timing were always restricted to discrimination rather than identification tasks. Besides, the classroom situation that was used as experimental setting must be judged as suboptimal. The somewhat better performance of the subjects with a proficiency in French indicates that it may be important to have a very good intuition of a language's prosody — however, given the poor overall results, no clear conclusions can be drawn from this study.

4.5.3 Experiment 2: Language Identification through Discrimination

In a follow-up experiment using the stimuli set of the previous experiment was used in a language identification task that built on a abiliy to discriminate between English and French. This design was chosen because it would enable subjects to build their judgement partly on a probably better proficiency in either language. Notice that this task is still more challenging than the one usually employed in rhythm discrimination tasks, e.g. the experiments by Ramus et al. (1999); Ramus and Mehler (1999); Ramus et al. (2003). Here, subjects usually had to decide whether two delexicalized stimuli belong to the same language or not. However, no language identification was demanded— it is of course evident that testing newborns or animals as in the studies by Nazzi et al. (1998); Rincoff et al. (2005); Toro et al. (2003) renders such an identification task impossible.

In the second experiment, subjects were asked to listen to pairs of stimuli, with each stimulus pair consisting of one French and one English stimulus. Each pair was identical with respect to the order and number of syllables per foot, e.g. the French *baba—bababa–baba* (2-3-2) was compared to the corresponding English stimulus. The subjects' task was now to decide which of the two was the French one. The idea behind this experiment is that even if listeners have difficulties deciding whether a given stimulus is French or English, they may be able to say that a stimulus is "more French or English" than a comparable one.

In a pretest, the stimulus pairs were presented to three subjects all of whom reported to be able to perceive a difference between the two stimuli per pair. However, they still reported difficulties of assign these perceived differences to a characteristic French or English prosody. It would have been an option to insert a learning phase before assessing the perception data in order to accustom subjects to a particular French or English timing. However, this would have lead to the learning of a pattern that would have been labelled "French" or "English" without testing whether the patterns have anything to do with the languages themselves. As an alternative, more natural method to boost identification, it was decided to enhance the language specific rhythmic impression through a few intonational characteristics based on our analysis results for fundamental frequency dynamics. In the intended French stimuli, each foot final syllable as added a pitch accent. The pitch accent contour was modelled by a linear interpolation between a low point starting at the syllable's onset, a point at the f0-topline halfway through the syllable and a point at the f0-baseline at the end of the syllable. The increase in pitch was 3 semitones for both languages. No other manipulation, e.g. a declination or downstep was performed. The final foot was equipped with a pitch accent on the initial syllable for both languages. This was done in order to prevent listeners from a classification strategy purely based on final falling vs. final rising or prefinally or finally stressed stimuli. As before, the manipulation was carried out with the help of the PSOLA algorithm implemented in Praat. For the listening test, subjects with and without skills in French were chosen. It was however a prerequisite to report a certain intuition of "what French sounds like". All subjects reported a good or even excellent knowledge of English. Prior to the experiment the subjects had to listen to original versions of both English and French sentences and the corresponding delexicalized *bababa*-stimuli. The delexicalization procedure chosen for the training set differed from the one proposed in Ramus and Mehler (1999) which has become quite popular in rhythm research (e.g. White et al. (2007); Dellwo (2008b)). Ramus exchanges every consonantal interval with the consonant [s] and every vocalic interval with an [a]. While this spares him from possibly tricky syllable segmentation the result may not properly reflect the beat structure of the corresponding utterance: In languages allowing both syllables without codas and syllables without onsets, e.g. *CV, VC, V...* there may be sequences such as *CV-VC*. In such cases, two beats would be represented by the sequence [saːs]. Another problem may occur given a language that allows for syllabic consonants which would be represented as [s] instead of a vowel. In order to get a more adequate representation of the original rhythmic pattern, the delexicalization strategy employed here uses the consonant [b] for the time span from syllable onset to the beginning of the phonologically determined perceptual center. The remaining part of the syllable is represented as the vowel [a]. In case of bisyllabic consonants, these are cut in half in order to estimate both the perceptual center and the syllable rhyme duration. If applicable, syllable boundaries are determined by the Maximum Onset Principle.

12 native speakers of German participated in the perception experiment, about the half of them reported to have some knowledge of French, albeit on average their proficiency in French was rather low. All subjects reported to have average to very good knowledge of English and reported no hearing difficulties. The results show that listeners are able to identify the language specific pattern significantly better than chance level ($\chi^2; df = 1; p < 0.02$) with an accuracy of approximately 60%. While this result does not seem to be overwhelming, given the difficulty of the task, the unnatural stimulus material and the fact most of the subjects were not proficient in French, the ability to identify a language based on a very abstract and rudimentary representation of its prosodic characteristics shows that our approach was able to capture at least some of the important aspects of French and English rhythm. A closer look at the individual results show that one subject who is not proficient if French but who used to live in Belgium for a while and is a musician, performed almost perfectly (30 correct answers out of 36 stimuli), while one subject without high proficiency in either English or French quite consistently misclassified intended English as French. While the former subject certainly was able to perceive the typical pattern, the latter subject was not able to match the prosodies to the languages, but he was able to perceive a systematic difference as well. As mentioned before, it is very likely that a mere discrimination task would have brought about much better results but would not have shown that the model indeed captures the characteristics of the languages under examination.

An interesting finding was that subjects often felt the need to describe their impressionistic hypotheses used to identify stimuli as either French or English. Some typical reactions are collected in the following list:

- The French were the swingy ones
- I classficied the staccato, abrupt ones as English
- English was more monotonous and rigid
- French was more monotonous and more regular
- English had stressed beginnings

From the reactions we see, that the labels listeners may give to a particular impression may not always be conclusive, e.g. despite the fact that subjects identified the majority of the stimuli correctly, some of them perceived intended French as monotonous, while others experienced intended English as monotonous. This finding should be kept in mind when coming across oversimplistic perceptual labels trying to capture a particular speech rhythm as *isochronous*, *machine gun*, *swingy* etc. Without a well-defined semantics of the phenomena referred to by such labels, they may denote a variety of prosodic, even contradicting acoustic patterns.

4.6 Summary: A Multidimensional Rhythm Identification Based on Grouping and Variation

In this section we proposed a multidimensional rhythm analysis in order to compare the rhythmic characteristics of two prototypical languages, English and French that have traditionally been characterized as being rhythmically distinct in most respects. In course of this comparison, several features of these languages were identified that go beyond a mere analysis of global variation, as has often been proposed by recently popular rhythm metrics. Instead, the proposed approach aims to identify global and local variation, relative timing at important boundaries (foot, phrase), deceleration and acceleration trends, compensatory shortening, tempo, alternation and placement of pitch accents within the phrase. The analysis identified a multitude of differences that characterize the rhythmic structure of our two example languages beyond a mere binary classification into stress timed and syllable timed. As a key method, time-delay plots were presented in order to visualize relative timing properties of speech. The patterns that were identified can serve as a basis of language classification and probably provide good anchor points for timing expectancies that enhance language planning, production and perception processes. A listening experiment further showed that subjects are able to identify languages based on the abstract timing patterns that were the result of our analyis. Furthermore, it was shown with the help of a Bayesian Belief Network that the timing expectations remain quite stable even given unexpected input, i.e. a syllable is shorter or longer than predicted by the probability analysis. Thus, top down-expectancies regarding rhythm can be quite stable after a certain amount of training. Also, with respect to expectancies, a first analysis of pitch accent placement within the phrase gave insight to possible anchor points that can guide a listener's attention to particular spots in an utterance where important information is likely to be produced by the speaker. Thus, beyond presenting a set of analytical methods in order to investigate rhythmic characteristics of languages, it was further shown how such regularities might enhance the cognitive processing of speech.

An overview of the identified characteristics is presented in Table 4.6.

Characteristic	**English**	**French**
Overall variability	more	less
Global isochrony	less	more
Local isochrony	none	none
Grouping strategy	acceleration	deceleration
Group boundary marking	initial lengthening	postfinal shortening
Final lengthening at phrase boundaries	more	less
Compensatory shortening	yes	no
Alternation	yes	no
Tempo	slower	faster
Pitch accents on stressed syllables	yes	yes
Accent distribution	diffuse	central
Pitch accent anchors	0.1, 0.5, 0.75	0.3, 0.5, 0.6
Perceptual phenomena		
Time Shrinking	no	yes
Tempo rel. Variability (grouping)	more	less

Table 4.6: Overview of rhythmic characteristics for English and French.

Definitely, the analysis proposed in far from complete and certainly not without flaws. An important aspect that has not been taken into account is the perceived local speech rate that can be explained as a function of linguistic information and absolute tempo. Given the influence of tempo on rhythm perception, this aspect ought to be taken into account more closely in follow-up studies. However, given the difficulty to compare perceived local speech rate across different languages such an approach needs a very different kind of data. Another issue that has only been briefly tackled is the influence of prominence lending acoustic features such as fundamental frequency and spectral intensity. Especially the alignment of pitch accents which have been completely ignored in this study ought to be looked at since they may provide valuable cues to the listener concerning the ends and beginnings of rhythmic groups. The somewhat bold hypothesis that the thus identified patterns play an important role in the cognitive processing of speech, of course need further empirical investigation as well. Having said this much, it can be stressed that even based on duration patterns and a rudimentary intonation analysis, many valuable insights into the charactistics of speech and language rhythm have been assembled in this chapter. The presented characterization goes beyond previous metrics and is able to draw a more complete and complex pattern of speech rhythm. In the upcoming chapter, the approach will be applied to various areas of rhythm related research.

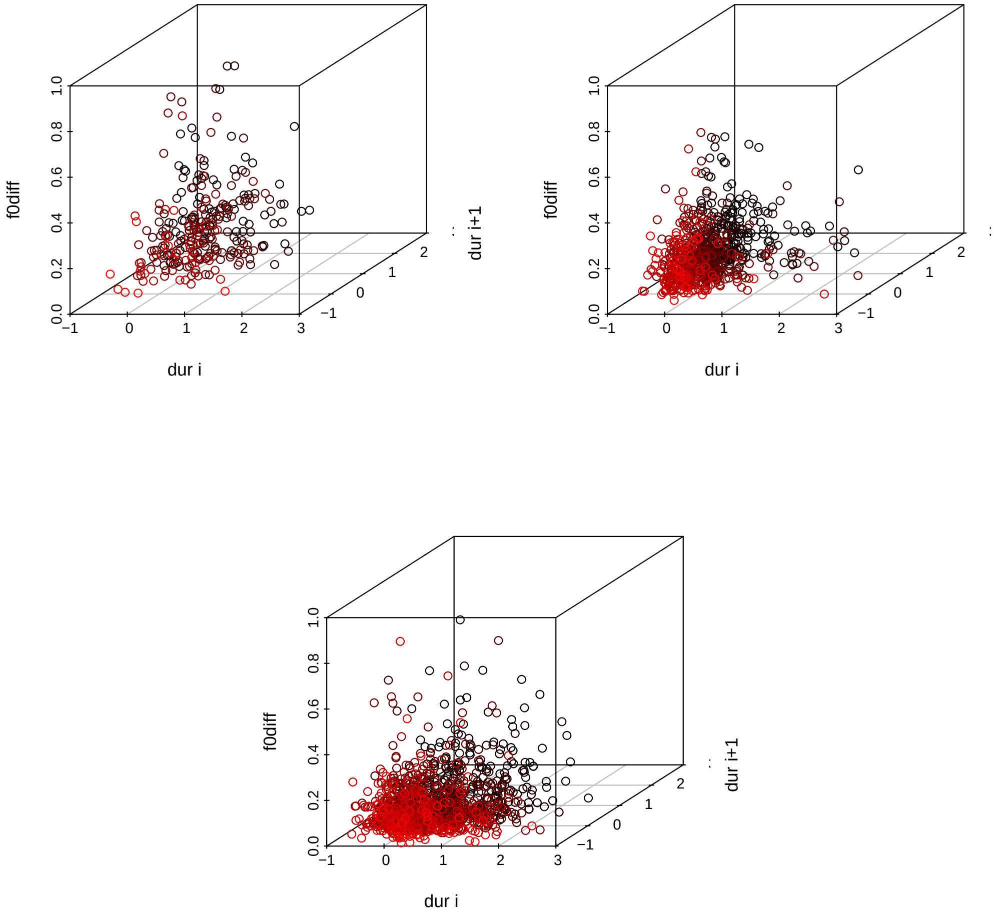


Figure 4.16: Time delay plots which have been added information of fundamental frequency difference for the various rhythmically relevant transitions in English. The x-axis describes the duration (z-score) of a beat i, the y-axis the duration of a subsequent beat $i + 1$, the z-axis shows the f0-difference between the two subsequent beats i and $i + 1$ in octaves. In addition to the durational variation, the graphs show differences in f0-dynamics for the different transition types. The upper left graph shows transitions to phrase final beats (strong difference in f0), the upper right graph shows transitions to stressed beats (less f0-dynamics than in the phrase final condition), the bottom graph shows transitions to unstressed beats (less f0-dynamics than in the other two conditions).

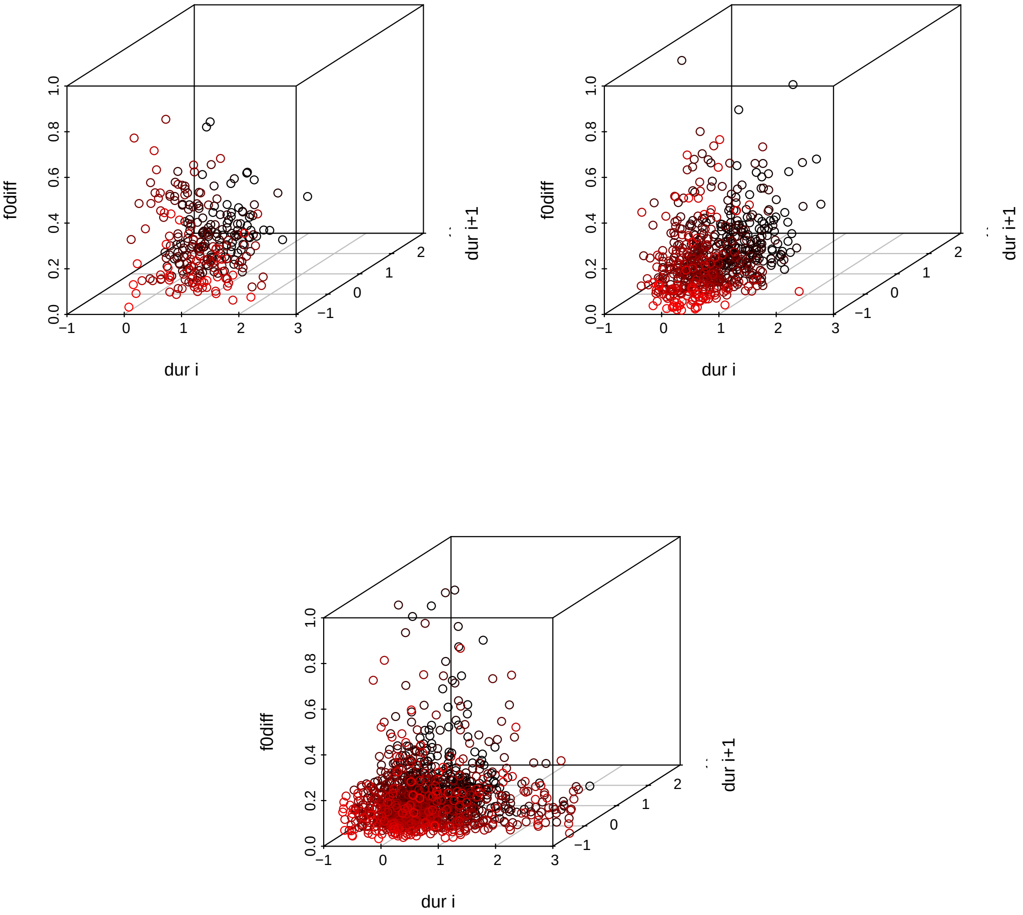


Figure 4.17: Time delay plots which have been added information of fundamental frequency difference for the various rhythmically relevant transitions in French. The x-axis describes the duration (z-score) of a beat i, the y-axis the duration of a subsequent beat $i + 1$, the z-axis shows the f0-difference between the two subsequent beats i and $i + 1$ in octaves. The graphs show differences in f0-dynamics for the three transition types, but only the transitions to phrase final beats also show a unique timing pattern. The upper left graph shows transitions to phrase final beats (strong difference in f0), the upper right graph shows transitions to stressed beats (less f0-dynamics than in the phrase final condition), the bottom graph shows transitions to unstressed beats (less f0-dynamics than in the other two conditions).

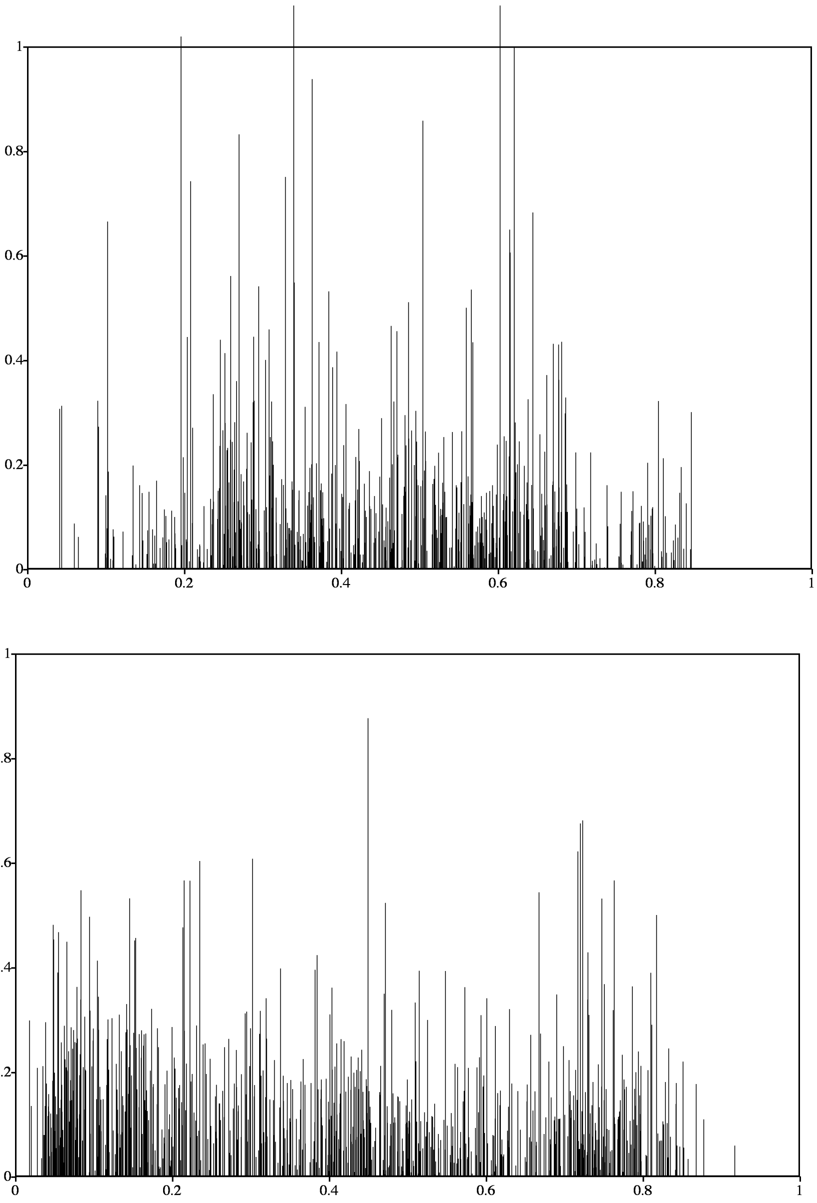


Figure 4.18: The Figures show differences in alignment of stressed syllables relative to an embedding intonation phrase. The x-axis shows the relative time within the intonation phrase, the duration of each phrase is normalized to 1. Each line depicts the timing of a stressed beat. The y-axis shows the difference of the stressed beat to the unstressed previous (English) or following (French) syllable in octaves.

Chapter 5

Applications

In this chapter, various applications for the rhythm analysis method introduced in the previous chapter are presented, illustrated and discussed. The chapter starts with an example of rhythmic classification in a typological framework going beyond previous analytical methods and being able to provide a more in-depth rhythmic characterization of languages. Then, language varieties are discussed, more explicitely, the influence of L1 on L2 rhythmical timing will be discussed as well as a comparison of two language varieties, namely European and Cameroon French, is presented. In another section, the application of the methodology on the investigation of highly stylized speech is illustrated. First, a speaking style known for its rhythmicity and entrainment effect, sermon style, is analyzed. Second, German poetic speech is analyzed. It will become evident that the regularities of stylized speech are instantly unveiled with the help of time-delay cross-plots. In a last section, possible applications of the rhythmical analysis method in the field of speech technology will be discussed. Since it has already been shown in the previous chapter, that a visual interpretation of the time delay plots lead to similar results as in-depth statistical tests and automatic classification, the presented analyses will be mainly descriptive in this chapter.

5.1 Typological Classifications

In this section, the multidimensional classficiation method will be employed in order to investigate the rhythmic pattern of a variety of languages. That way, a fuller picture will be drawn that goes beyond an impoverished binary classification along the stress-timed—syllable-timed continuum. In the previous chapter 4 some relative timing characteristics for German, Italian and Polish were already given in addition to the languages examined in-depth, namely English and French. When adding these three languages to our characterization matrix, differences and similarities will be revealed not captured by previous classifications. In line with English, German is usually classified as stress timed. It marks lexical accents with increases in duration and pitch accents, reduces unstressed syllables and shows compensatory shortening. Italian, classified as syllable timed, shares some, but not all features with French: It has little reduction, but shows strong accentuation in both duration increase and pitch accentuation. Polish, the last language under discussion, is probably the most difficult case: It has been notoriously difficult to be classified along the stress-timed—syllable-timed dichotomy. It has often been claimed to be "rhythmically mixed", since it does reveal some increase in duration on stressed syllables and some compensatory shortening effects, but less than in English. Its lack of reduction and a fixed lexical stress on the penultima also indicate a tendency towards syllable timing. Phonotactic complexity and the resulting variability may enhance the stress timing effect. Thus, in addition to our prototypical languages, we are now looking at different data. Our method ought to reveal the characteristics, similarities and differences between the various languages along further dimensions, hopefully receiving a fuller picture of the language specific rhythms.

5.1.1 Variability Analysis

Our time-delay plots (5.1) clearly reveal differences in relative timing for our three additional languages. While German looks similar to English, there is still significantly more overlap between the regions of transitions to unstressed and stressed syllables. Thus, the distinction between stressed and unstressed is marked less in German — German shows a less consistent acceleration tendency after stressed syllables. Italian, however, much unlike French, marks stressed beats strongly by duration. However, unlike English and (to a lesser extent German) due to the lack of reduction the transitions to unstressed syllables do not cluster below the diagonal, thus not indicating a strict trend towards acceleration in transitions to unstressed syllables. They do cluster around the $0, 0$-co-ordinate, indicating a global trend towards isochrony, but only for the transitions to unstressed syllables. To put it in a nutshell, Italian beats are more or less "isochronous" unless they are strongly accented. Another finding for Italian is the extreme deceleration trend at phrase boundaries. These are captured in a region entirely of their own and are marked by pronounced lengthening of both the final and the penultima syllable. Although this is partly explicable with the trend towards stressing the penultima in most Italian words, this trend is still much more pronounced than in languages showing penultima stress very often or always as well, i.e. German and Polish (also see Wagner (2007)). If the marking of phrase finality and stress are both indicators of stress- vs. syllable timing, our classification now puts English and German on the end of stress-timing, while French and Polish are on the other end. With respect to phrase finality marking, Polish looks more stress timed, while French looks more stress timed than Polish with respect to stress marking. Unlike French, plotting the transitions from stressed to poststressed beats in Polish does not reveal any "hidden" pattern (Figure not shown)[1]. The most striking "syllable timed" feature of Italian obviously is its lack of reduction. These tendencies had already been mentioned in the previous chapter in the calculation of Eucledian Distances (cf. Figure 4.7 on page 158).

[1] It is possible, though, that Polish shows significant effects in fundamental frequency in postaccented syllables (Grazyna Demenko, personal communication). This has not yet been examined in this thesis.

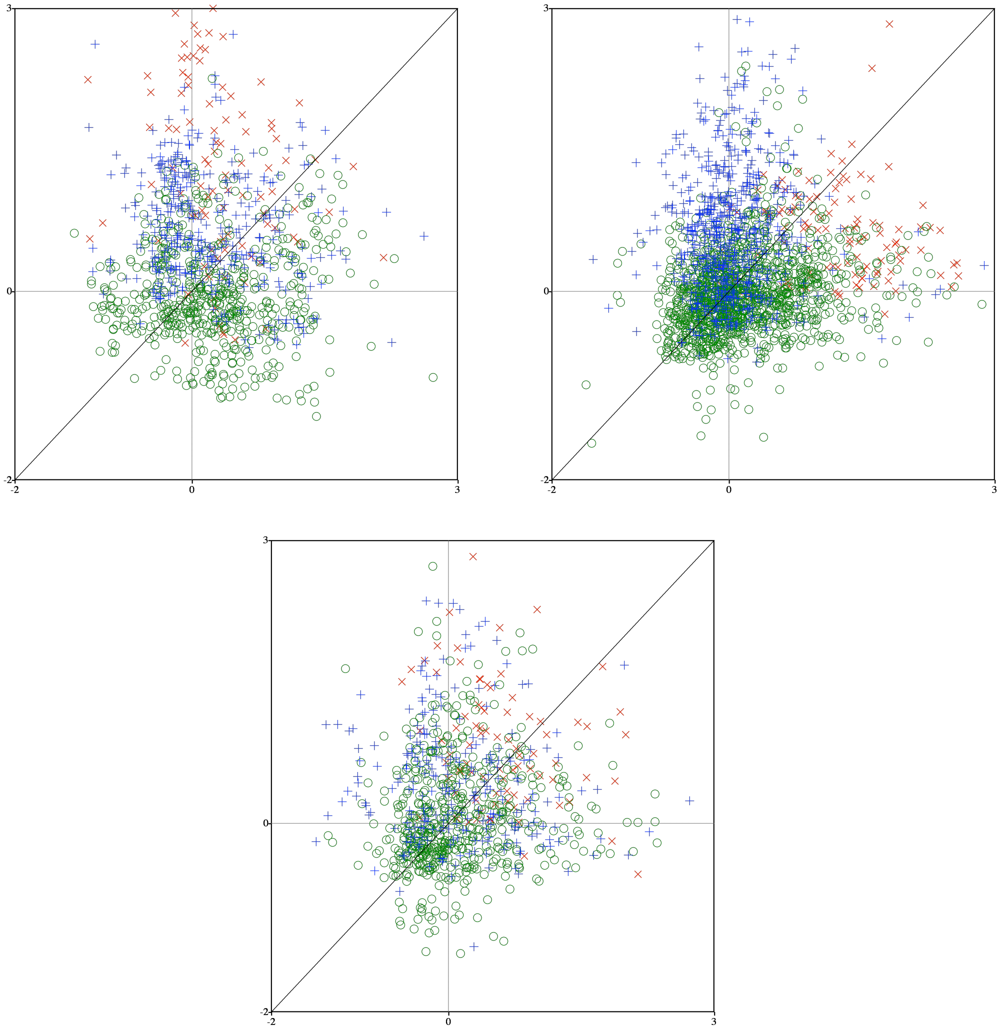


Figure 5.1: Time-delay plots of stress-timed German (upper left), syllable timed Italian (upper right), and hitherto unclassified Polish (lower). As before, transitions to lexically stressed beats are plotted in blue (+), transitions to phrase final beats are plotted in red (x), transitions to unstressed syllables are plotted in green (o).

5.1.2 Long Term Patterns

With regards to long-term temporal organization, our approach regarded whether a language shows a tendency of deceleration or acceleration within structures. Also, the level of compensatory shortening and overall tempo is taken into account in order to receive a fuller picture of the rhythmical tendencies. Average timing in feet binary, ternary and quaternary feet is compared in time-delay plots for German, Italian and Polish (cf. Figure 5.2). These plots reveal once more that a dichotomy of stress and syllable-timing is insufficient. While the German material does show a number of similarities with the English patterns, there are also a number of differences. First, the distance between stressed and unstressed syllables in German seems to be far less pronounced in binary feet. Also, the tendency of compensatory shortening is less regular, with stressed beats in ternary feet being almost as long as those in binary feet. Compensatory shortening is more evident in unstressed beats, maybe indicating a tendency of stronger reduction as a function of the number of syllables in a foot. Maybe the most striking difference between German and English, however, is the German tendency of increased duration on foot final beats but a lack of alternation. Thus, ternary feet show an overall pattern of *long—short—long* instead of a dactylic one. In quaternary feet, no alternation is evident. Instead, a *long—short—short—long* pattern is revealed, again showing a tendency of final lengthening at foot level. For Italian, the picture reveals an overall higher tempo compared to stress timed languages like English and German. This feature seems to be typical for most languages classified as syllable timed and explains much of the lack of perceived variation, as discussed in Chapter 2. The often-found claim of Italian having no compensatory shortening at all, however, is falsified. Beats in ternary and quaternary feet are much shorter than those in binary feet. However, there hardly exists a difference between ternary and quaternary feet. Maybe, this lack of further compensatory compression is a result of the overall high tempo not allowing for further reduction. Interestingly, Italian, like German, shows a tendency towards foot final lengthening in quaternary feet but no tendency towards strict alternation. For Polish, it can be once more confirmed that there exists only a subtle tendency towards a systematic lengthening of stressed syllables. A similarity between Polish and Italian seems to be that compensatory shortening is not further intensified if feet become longer, although all beats are affected by compensatory shortening comparing binary and other feet. In Polish, a possible explanation for this phenomenon may be the comparatively high phonotactic complexity and lack of reduction tendencies in Polish preventing a further reduction. Thus, different reasons may lead to similar phenomena, since it was hypothesized that in Italian, the high overall tempo plus lacking reduction prevents an unlimited compensatory shortening. Also, a pronounced tendency for foot final lengthening can be traced in the Polish patterns, while strict alternation seems to be restricted to English feet so far. Another interesting point is the fact that the long Polish feet are preceded by a long previous footfinal beat. Since only phrase internal beat patterns are observed, this cannot be the consequence of a previous phrase final lengthening phenomenon. Thus, it can be concluded that foot final lengthening can even be stronger than lexical stress induced lengthening if the stressed beat is affected by compensatory shortening. Polish also shows a tendency of deceleration throughout the foot, which enables the phenomenon of time shrinking and may lead to the impression of isochrony and a lack of variation. Summing up, while German shares many characteristics with English, it shares a slight tendency of foot final lengthening with French and shows a lack of strict alternation. Italian shares a high tempo, a lack of reduction and a slight tendency of foot final lengthening with French, but the strong accentuation of foot initial syllables with English and German. Polish shares a strong tendency of deceleration, a lack of reduction and a lack of accentual lengthening with French, but a moderate tempo and a lot of temporal variation with English. For neither language it can be deduced that they can be characterized along a stress-timing—syllable-timing continuum, if English and French are regarded as protoypical cases of either class. The overall patterns for the three different languages are also further illustrated in Figures 5.2 and 5.3.

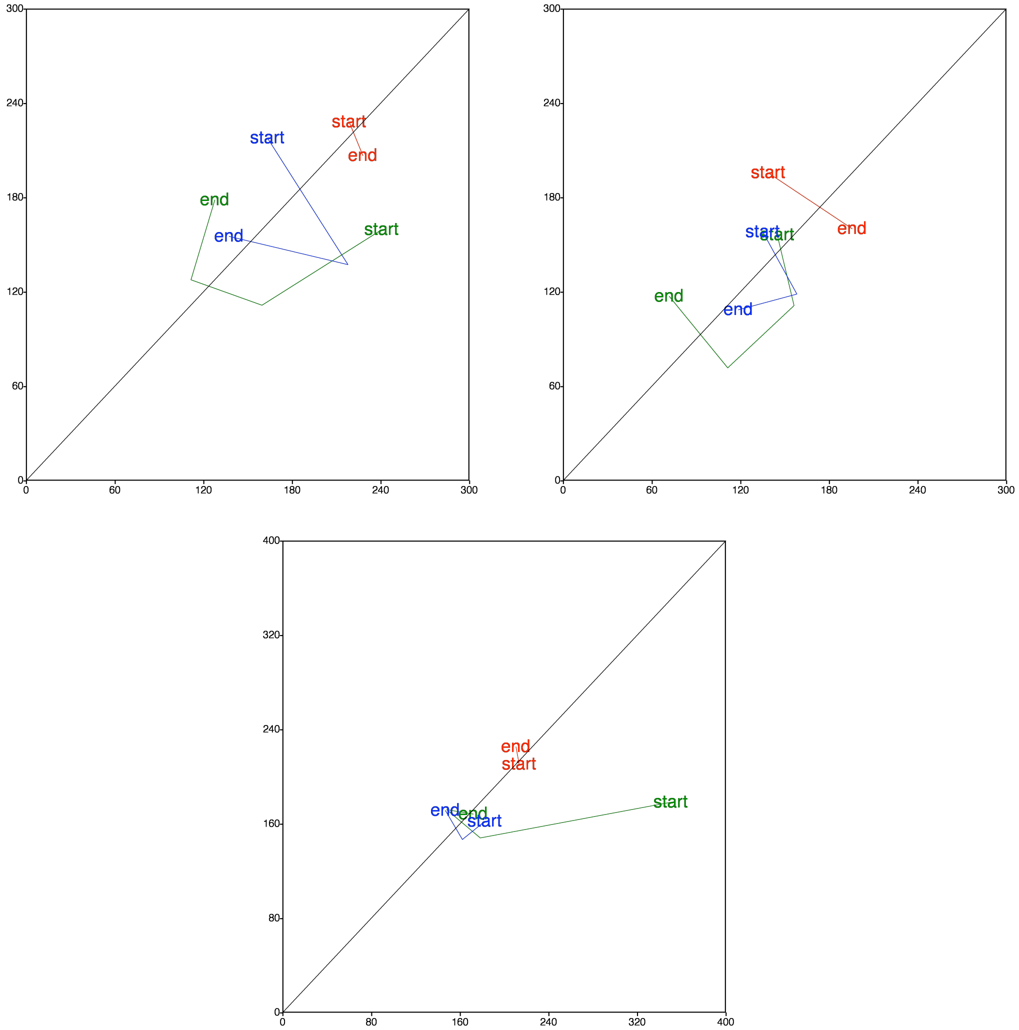


Figure 5.2: Long-term ime-delay plots of stress-timed German (upper left), syllable timed Italian (upper right), and hitherto unclassified Polish (lower). The plots show the long-term timing characteristics of binary (red), ternary (blue) and quaternary (green) feet. Notice that the Polish plot has a different time axis!

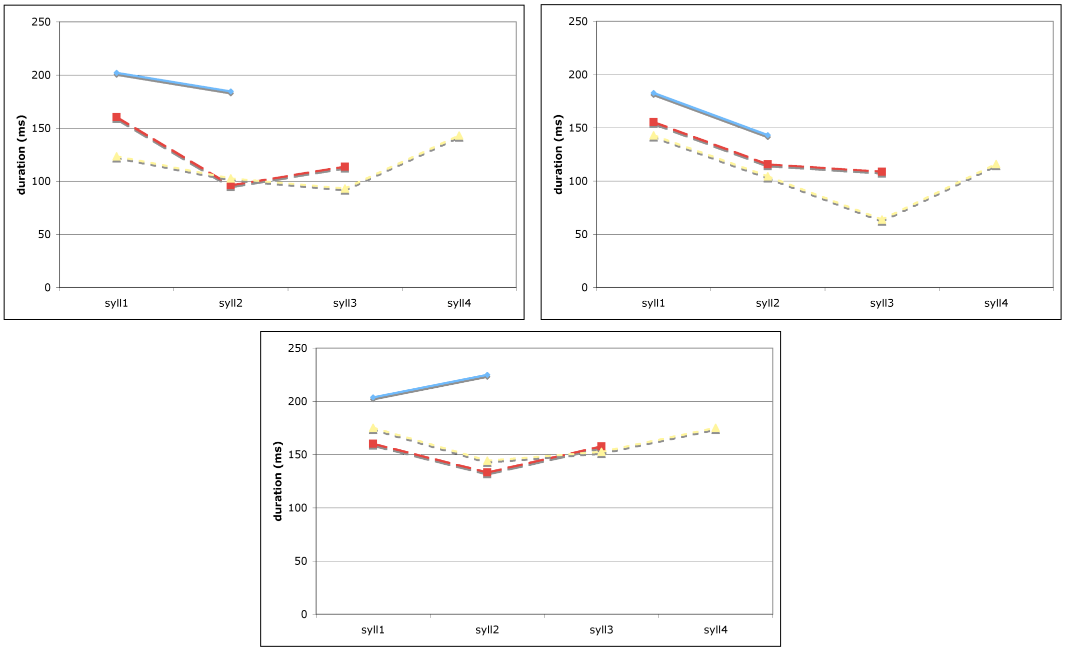


Figure 5.3: Long-term timing patterns across binary, ternary and quaternary feet for German (upper left), Italian (upper right) and Polish (lower).

5.1.3 Intonational Characteristics

In order to get a fuller picture of the rhythmical characteristics of our example languages, the intonational characteristics of Polish, Italian and German will be addressed in addition to the previous durational analysis. First, three-dimensional time-delay plots (introduced in Section 4.4.1) showing the f0-difference between successive beats will be employed to get a fuller picture of the intonational dynamics around rhythmically relevant borders. Both the plots and a one-way ANOVA examining the f0-differences reveals that there exist high or highly significant differences for f0-dynamics for the different transition types in all three languages[2]. Transitions to final beats show the most, transitions to stressed syllables less and transitions to unstressed beats the least f0-movement. The resulting plots are shown in Figures 5.1.3, 5.1.3 and 5.1.3.

[2] Results of the one-way ANOVA to investigate differences for the three groups are for German: $F = 22,2; df = 2; p < 0.0001$, for Italian: $F = 6.15; df = 2; p < 0.01$, for Polish: $F = 6.18; df = 2; p < 0.01$

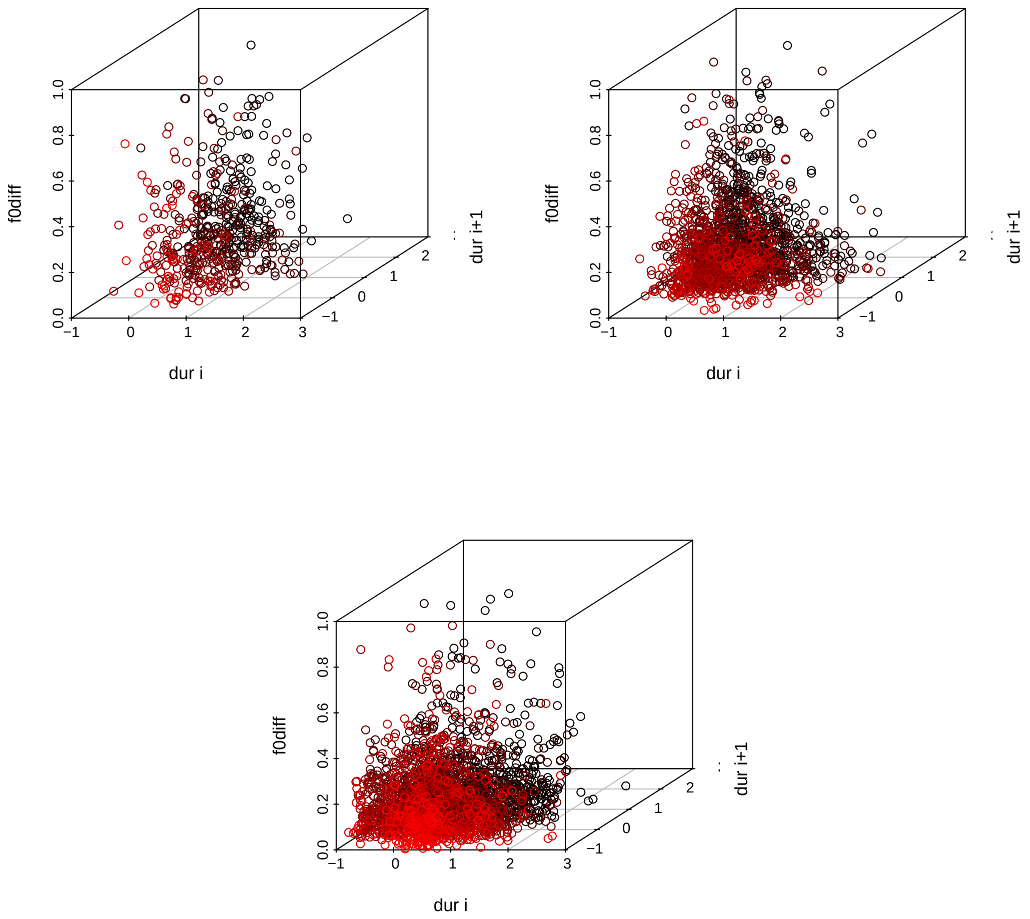


Figure 5.4: Time delay plots which have been added information of fundamental frequency difference for the various rhythmically relevant transitions in German. The x-axis describes the duration (z-score) of a beat i, the y-axis the duration of a subsequent beat $i + 1$, the z-axis shows the f0-difference between the two subsequent beats i and $i + 1$ in octaves. The graphs show differences in f0-dynamics for the three transition types. The upper left graph shows transitions to phrase final beats (strong difference in f0), the upper right graph shows transitions to stressed beats (less f0-dynamics than in the phrase final condition), the bottom graph shows transitions to unstressed beats (less f0-dynamics than in the other two conditions).

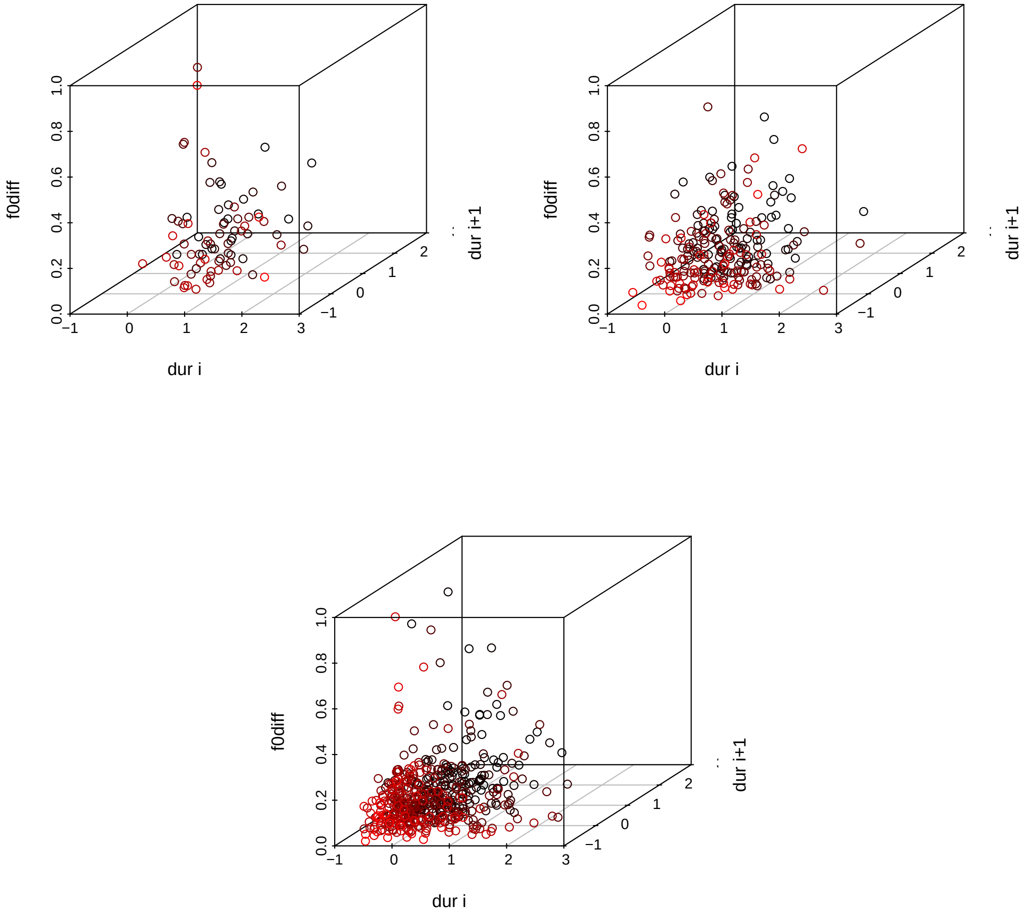


Figure 5.5: Time delay plots which have been added information of fundamental frequency difference for the various rhythmically relevant transitions in Polish. The x-axis describes the duration (z-score) of a beat i, the y-axis the duration of a subsequent beat $i + 1$, the z-axis shows the f0-difference between the two subsequent beats i and $i + 1$ in octaves. The graphs show differences in f0-dynamics for the three transition types. The upper left graph shows transitions to phrase final beats (difference in f0 not stronger than in the lexical stressed condition), the upper right graph shows transitions to stressed beats (similar f0-dynamics as in the phrase final condition), the bottom graph shows transitions to unstressed beats (less f0-dynamics than in the other two conditions).

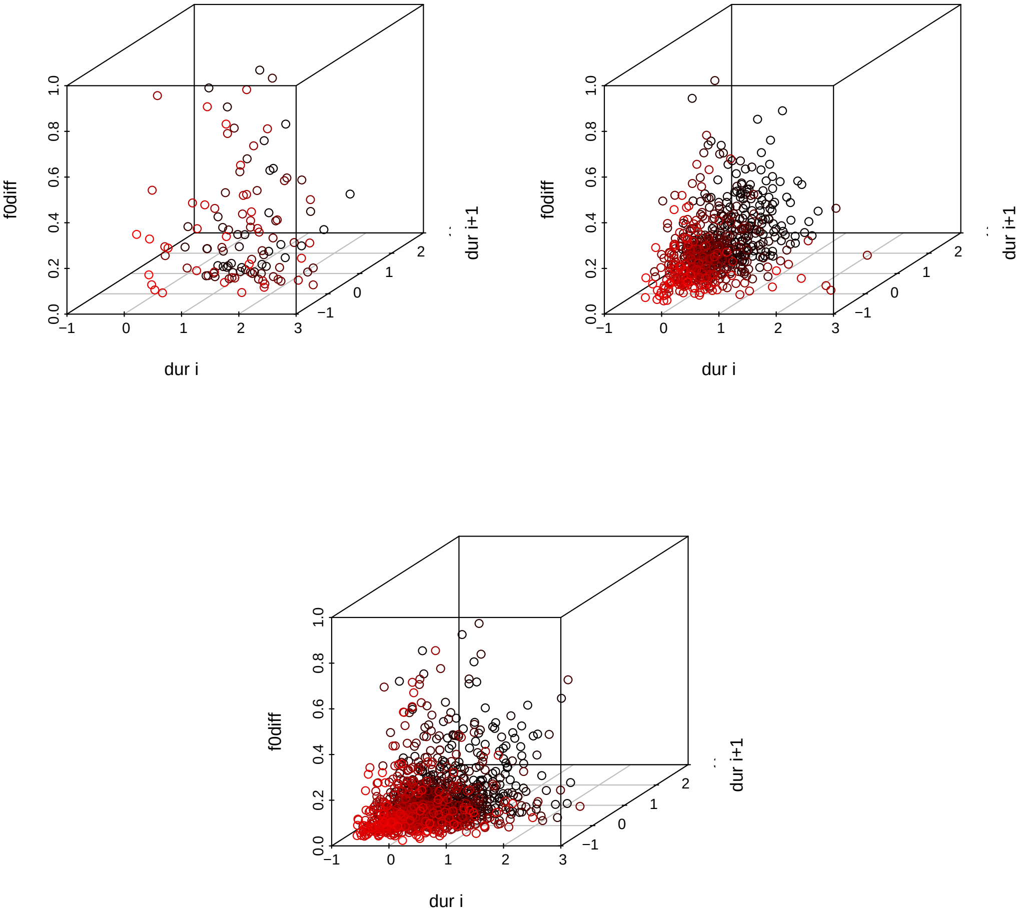


Figure 5.6: Time delay plots which have been added information of fundamental frequency difference for the various rhythmically relevant transitions in Italian. The x-axis describes the duration (z-score) of a beat i, the y-axis the duration of a subsequent beat $i + 1$, the z-axis shows the f0-difference between the two subsequent beats i and $i + 1$ in octaves. The graphs show differences in f0-dynamics for the three transition types which are stronger than in the other languages. The upper left graph shows transitions to phrase final beats (strong difference in f0), the upper right graph shows transitions to stressed beats (less f0-dynamics than in the phrase final condition but still pronounced), the bottom graph shows transitions to unstressed beats (less f0-dynamics than in the other two conditions).

In a last analysis step, the preferred feet distributions across intonational phrases together with pitch dynamics are compared for the three different languages. The results are shown in Figure 5.1.3. Due to the different amount of data available for the different languages (three speakers for Polish, four for Italian, but 15 for German), the German plots show higher concentrations. Again, language specific preferences can be traced. While for all languages, there exists a tendency of placing the first foot when about 1 tenth of the intonational phrase has elapsed (0.1), German and Italian show high concentrations around 0.3 and 0.6, which is similar to the results of Port and colleagues for their American English data. However, Italian lacks a further concentration around 0.5, which can only be found in the German data. In addition, both Italian and German show a high concentration of foot placement around 0.75. The Polish data is difficult to interpret due to lack of material. However, it also shows a high density around 0.1, 0.3, 0.5 and 0.7 and shows similarities to German foot placement preferences.

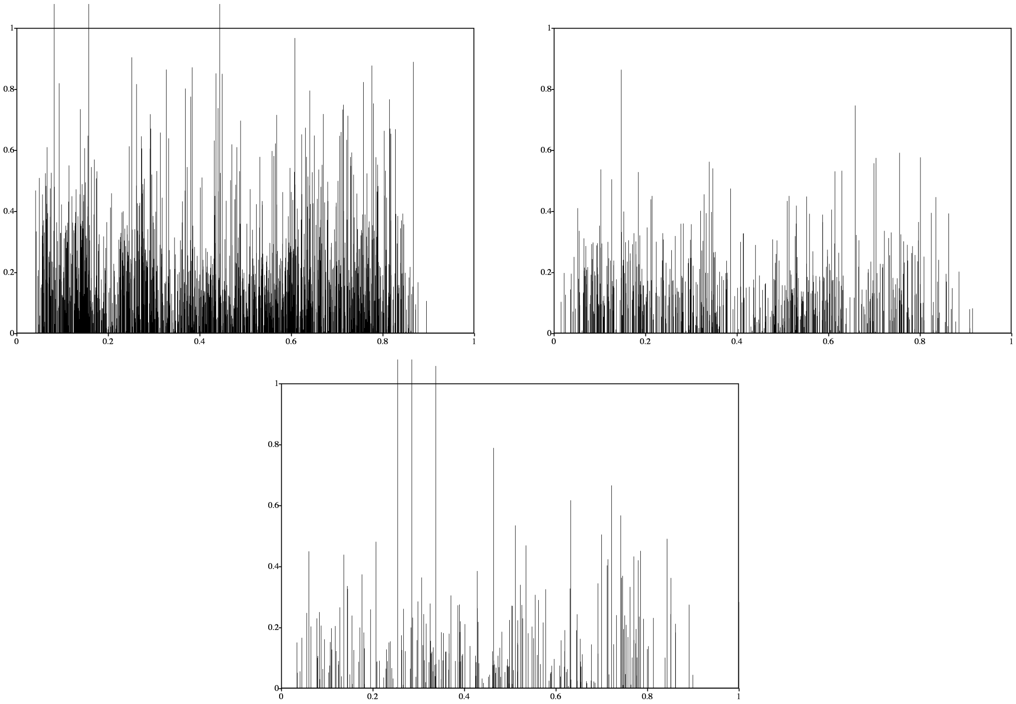


Figure 5.7: The plots show the distributions of foot beginnings relative to an embedding intonational phrase on the x-axis and the pertinent f0-difference relative to the mean f0 in the preceding beat in octaves is plotted on the y-axis.

Our three languages can be characterized with respect to their rhythmic characteristics in our multidimensional analysis procedure which is presented in Table 5.1.

The characteristics of English and French are also shown for reasons of a better comparison to more prototypical cases of language rhythm. The overview furthermore clarifies that a classification along the stress-timed—syllable-timed dimension falls too short. Even along the previously popular metrics such as variability and compensatory shortening show more than a simple binary classification for the various languages. Italian, e.g. shows a little bit of compensatory shortening, but definitely not as little as French does, while Polish behaves similar to French with respect to a deceleration strategy across the foot. However, Polish is much slower than French and shows a lot more timing variation, probably due to its rather complex but variable phonotactics. Thus, the claim that Polish behaves like syllable-timing French in some respects and like stress-timing English in other respects, is certainly true. However, the same can be said for Italian which is usually called syllable-timed. Even German, although being more similar to English in many respects, is certainly not rhythmically identical to it, i.e. it does show less strict alternation than English. It order to get a more complete picture of a language's rhythm, several dimensions need to be taken into account and clarified. Probably the most important conclusion from this chapter is to see that languages which have been difficult to classify in the traditional rhythmic dichotomy can still be assigned a stable rhythmic pattern. Polish is certainly rhythmic, but it is rhythmic in a way that is different from both French and English.

Characteristic	**English**	**German**	**Polish**	**Italian**	**French**
Overall var.	+	+	+	+-	-
Global isochr. (tendency)	-	-	-	+-	+
Local isochr.	-	-	-	+	-
Grouping via	accel.	accel.	decel.	accel.	decel.
Group boundary	initial length.	initial length.	neither	initial length.	postfinal short.
Final lengthening (phrase)	+	+	+-	++	+-
Final lengthening (foot)	-	+	+	+-	+
Comp. shortening	+	+	+-	+-	-
Alternation	+	-	-	-	-
Tempo	slow	slow	very slow	very fast	fast
Pitch marking	yes	yes	yes	yes	yes
Pitch distribution	diffuse	diffuse	diffuse	diffuse	compact
Stress anchors	0.1, 0.5, 0.75	0.1, 0,3, 0.6, 0.7	0.1,0.3,0.5,0.7	0.1,0.3,0.6	0.3, 0.5, 0.6
Perceptual phenomena					
Time Shrinking	no	no	yes	no	yes
Tempo rel. Variability	more	more	more	less	less

Table 5.1: Overview of rhythmic characteristics of English, German, Polish, Italian and French. The table is organized in such a way that the prototypical stress timed languages are on the left while the prototypical syllable timed ones are on the right. Polish has been put in the middle since it has often been labelled as being rhythmically mixed in between both classes.

5.2 Investigating Varieties

In this chapter, the multidimensional analysis will be carried out on two cases of non-native speech rhythm in order to investigate the level of L1 influence on L2 rhythm patterns. In the first scenario, French L2 productions by German L1 speakers are examined. In the second scenario, material from one speaker is examined whose native language is a West African tone language, but who grew up using French on a daily basis, at work, at school and university. The speaker considers herself bilingual and her French pronounciation audibly differs from European French.

5.2.1 The influence of L1 German on L2 French

Taking into consideration our typological analysis in Section 5.1, we can derive various hypotheses concerning a possible influence of an L1 German on L2 French. Of course, we should not expect a kind of L2 rhythm that has no traces of the target L2 whatsoever. Such a pattern is highly unlikely given the close relationship between the segmental level and the resulting timing properties of a language, i.e. much of the typical rhythm comes "for free" if a learner manages to produce the L2 more or less correctly on the segmental level. Still, some influences of the L1 rhythm may still show in a target language. Consequently, such traces ought to be identifyable with the help of our multidimensional analysis. Based on our typological analysis, we can derive a variety of hypotheses concerning possible features of an L2 French spoken by native speakers of Germans which are contained in the following list:

- L2 French spoken by Germans should be more variable and show less tendencies towards global isochrony

- It should have less typical deceleration throughout the foot

- It will show a strong foot final lengthening but no postfinal shortening, because the foot final syllable is produced as a typical German accent

- There may be untypical compensatory shortening

- Tempo will be significantly slower, however, this may be explicable with general production difficulties rather than a transplantation of a native rhythm to an L2

- Stress anchors may shift to the "German" values

In a first analysis step, a time-delay plot of L2 French is interpreted based on the data of eight speakers reading the French material of the Bonn Tempo database. The time delay plots instantly show a tremendous difference between the native French data and the L2 data. The Germans definitely introduced an accentual lengthening typical for their native language. In Figure 5.8, the native French and German plots are compared to the L2 French ones. As predicted by our hypotheses, the L2 French data shows more temporal variability and the foot final beats are interpreted as accents by the native Germans and are produced in a style similar to native German, with a pronounced lengthening.

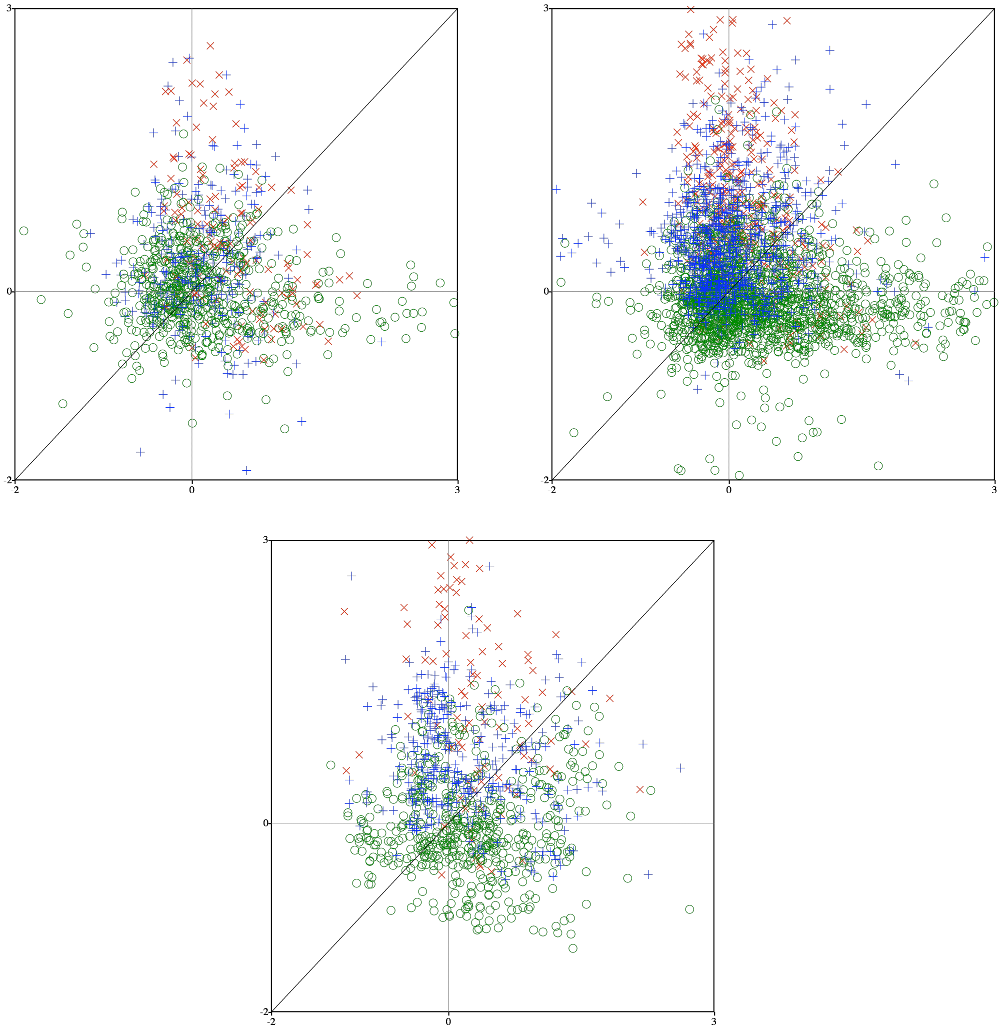


Figure 5.8: The time delay plots show native French timing relations (top left) in comparison with native German (bottom) and L2 French produced by native speakers of German (top right). Transitions to foot final beats are shown in blue (+), transitions to phrase final beats are red (x), all other transitions are green (o).

Compensatory shortening and deceleration trends are better examined in long-term timing plots building on non-normalized durational data. In our long-term time-delay plots (Figure 5.9), it is evident that the L2 French productions are much slower than the native ones. As stated above, this is expected and not necessarily interpreted as the impact of the L1 on an L2. However, this may have an impact on the perception of L2 French showing more variability as it introduces more groups or feet on the level of perception. Indeed, a study by Wagner and Dellwo (2005) found that fast L2 French is usually perceived as less accented than slow L2 French by native speakers. However, tempo is not the only rhythmic factor revealed. The plot once more shows a strong tendency to lengthen the foot final syllable which is quite typical in German longer feet as well. However, the plot also reveals that the overall timing pattern with its deceleration is mimicked quite well by the Germans — even if the timing relations, especially foot final lengthening are clearly exaggerated. Compensatory shortening effects cannot be traced in the L2 plots either.

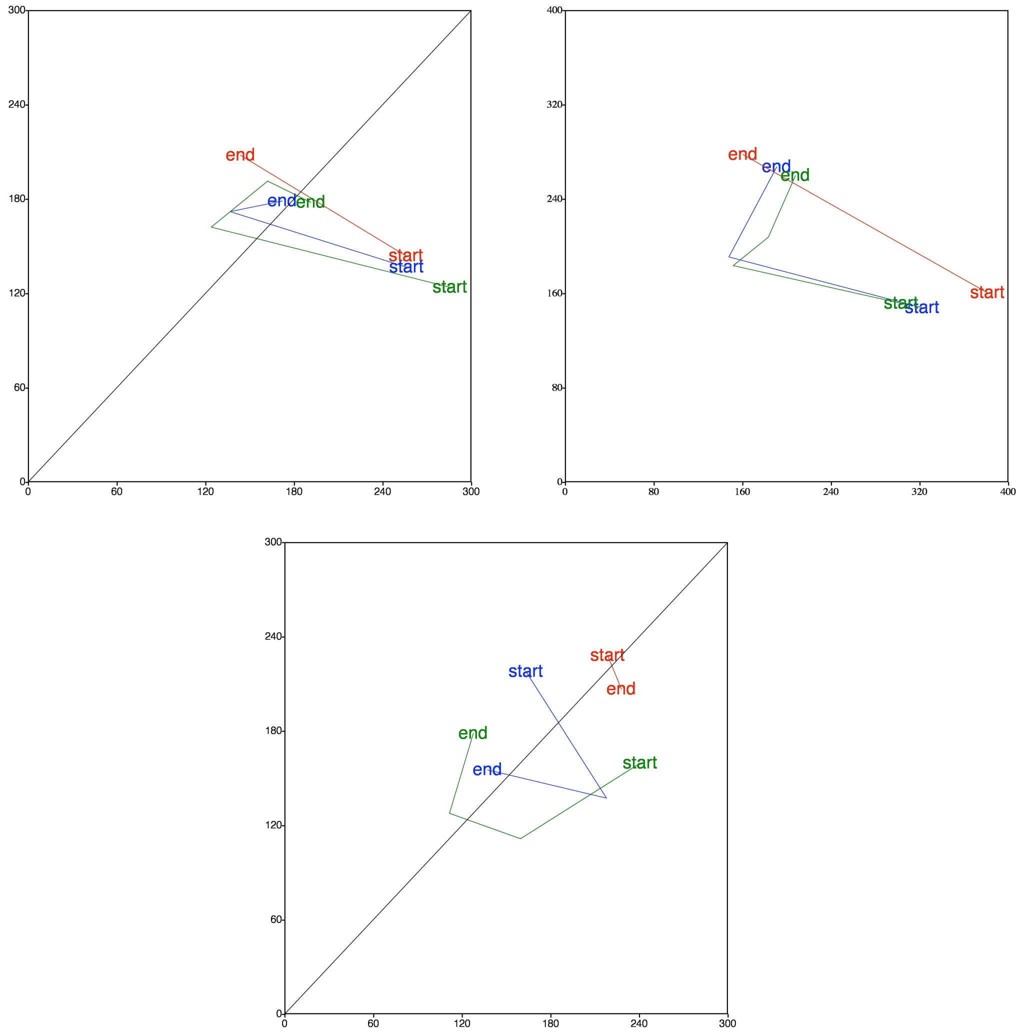


Figure 5.9: The time delay plots show long term timing patterns for binary, ternary and quaternary feet in native French (top left) in comparison with native German (bottom) and L2 French produced by native speakers of German (top right). Notice the different scale of the L2 French plot!

In Figure 5.10 the median beat durations for the most frequent foot types are compared. Here again, striking differences between native and non native French timing patterns are revealed. While the French deceleration trend throuought the foot is characterized by only a moderate increase in duration which becomes almost nonexistent towards the end, the German data clearly exaggerates the deceleration trend of French. The German L2 French shows a trend towards a linear increase in duration across the foot. The L2 data may be the result of the perceptual impression of French deceleration throughout the entire foot. This trend is mimicked by the non-natives. However, the final increase in duration may be the combined result by assigning the foot final beat the status of an accented syllable. This trend may be further intensified by a general trend of foot final deceleration in German or an interpretation of the foot final syllable as iambic. This last interpretations seems even more likely given the circumstance that the lengthening effect in the L2 is even stronger than accentual lengthening in the German data. Since the German accents tend to be trochaic, but iambic lengthening in German is stronger than trochaic lengthening, the exaggerated lengthening effect may be the result of a "rhythmic misinterpretation" by the German listeners. Summing up, it can be concluded that Germans are able to master some of the characteristics of French rhythm like the general trend of deceleration and the lack of compensatory shortening. However, they do transplant some characteristics of German onto the L2 by exaggerating foot final lengthening and treating it similar to an accented (German) syllable.

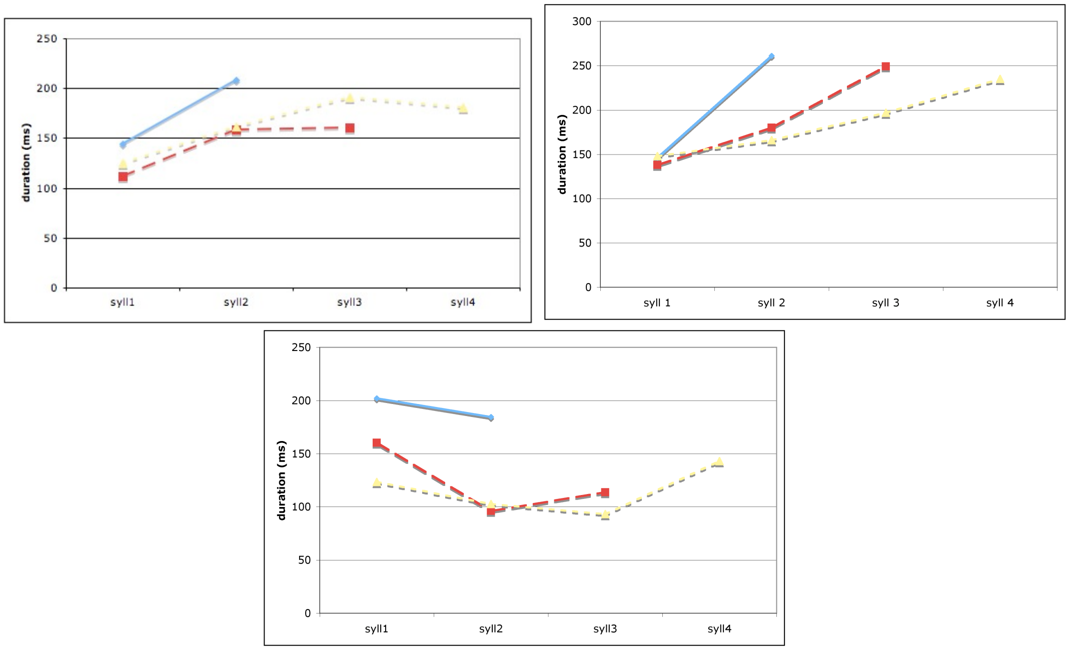


Figure 5.10: The plots show median duration patterns for binary, ternary and quaternary feet in native French (top left) in comparison with native German (bottom) and L2 French produced by native speakers of German (top right).

In a last analysis step, the timing of stressed beats within the embedding intonation phrase is examined. It is possible, that this timing is influenced by native German preferred "anchors" of foot timing. However, when looking at the plot in Figure 5.11, these show some similarity with the native French patterns with their particular concentration in the center of the embedding phrase. A closer look reveals a couple of differences between the native French and non-native French utterances. While in the non-native utterances, there is a concentration around 0.2 and 0.4, while the native early anchor lies around 0.3. It is possible, that this timing misplacement is indeed influenced by the native German preference of placing stressed syllables around 0.15 and 0.4. While the non-native patterns do show slight concentrations around 0.5 and 0.6, similar to the native French ones, they show a further concentration around 0.7 which cannot traced at all in the native French utterances, but shows up in the native German ones. Summing up, the anchoring of stressed beats within the embedding intonation phrase seems to be partly influenced by rhythmical preferences of the native language as well.

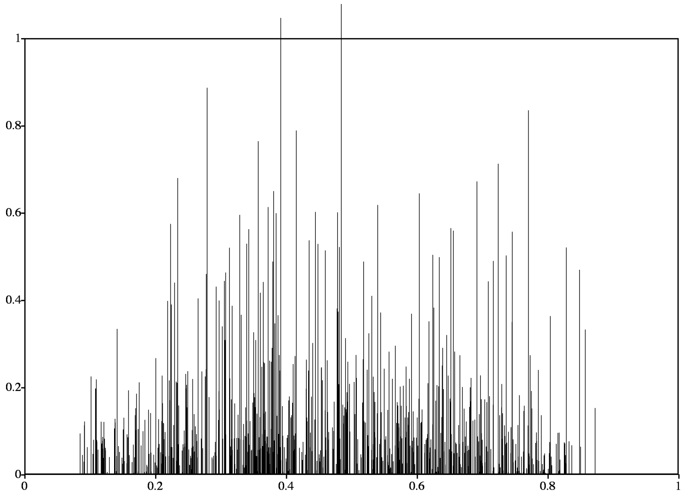


Figure 5.11: The figures show the alignment of stressed beats within an embedding intonational phrase for L2 French utterances produced by native speakers of German. The x-axis shows the relative timing within the intonation phrase, the y-axis shows the difference of the stressed beat to a previous one in octaves.

Summing up, the non-native French uttered by the native speakers of German that are contained in our database has the characteristics described in Table 5.2:

Characteristic	German	L2 French (L1 German)	French
Overall var.	+	+	-
Global isochr. (tendency)	-	-	+
Grouping via	accel.	decel.	decel.
Group boundary	initial length.	final length.	postfinal short.
Comp. shortening	+	-	-
Tempo	slow	slow	fast
Accent distribution	diffuse	centered	centered
Stress anchors	0.1, 0.3, 0.6, 0.7	0.2,0.4,0.5,0.6,0.7	0.3, 0.5, 0.6

Table 5.2: Comparison of the rhythmic characterstics of native French, non native French uttered by speakers of native German and native German. Some influences of the L1 on the L2 can be clearly traced, other possible influences do not show. Only those characteristics where L1 German and L2 French differ are shown.

5.2.2 Comparison of Two Varieties of French

With respect to a comparison of European French and Cameroon French, few clear hypotheses can be derived because unfortunately, the (second) native language of our database speaker is not known. It is however known that her background is a West African Tone language. We do therefore expect a very different behaviour of the fundamental frequency alignment. However, a detailed analysis of the tonal features will be far beyond the scope of this thesis and we will not look into these aspects in detail. With regards to timing, the Cameroon French productions look very much unlike European French — the plots show a lengthening of the foot final beats (cf. Figure 5.12) and a tendency of durational variation stronger than in European French. Furthermore, the long-term patterns reveal a deceleration strategy propagating throughout the foot, similar to European French, however, the long-term pattern reveals that foot-final lengthening becomes less strong with an increasing number of syllables per foot (cf. Figure 5.13). No compensatory shortening is evident. Also, the f0-analysis shows strong tonal dynamics at phrase boundaries, but less strong tonal deviation at foot boundaries (cf. Figure 5.14). The overall tempo is remarkably shorter than in European French. Thus, our analysis instantly identified several parameters of difference between the two varieties of French but also many similarities. The most striking differences seem to be the overall slower tempo, the more pronounced foot final lengthening and the differently organized fundamental frequency dynamics. Given only one speaker in our analysis, of course, these results should rather be interpreted as a first step towards a fuller picture.

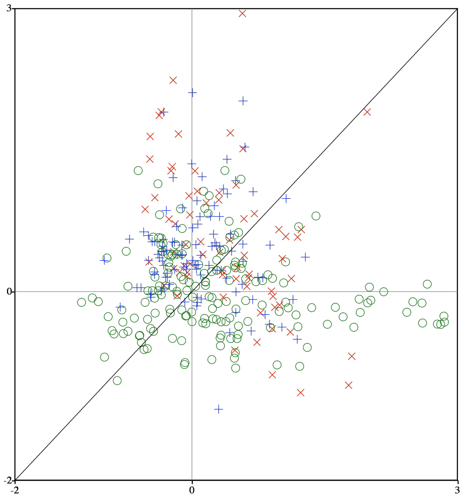


Figure 5.12: The time delay plots show Cameroon French timing relations produced by one speaker. Transitions to foot final beats are shown in blue (+), transitions to phrase final beats are red (x), all other transitions are green (o).

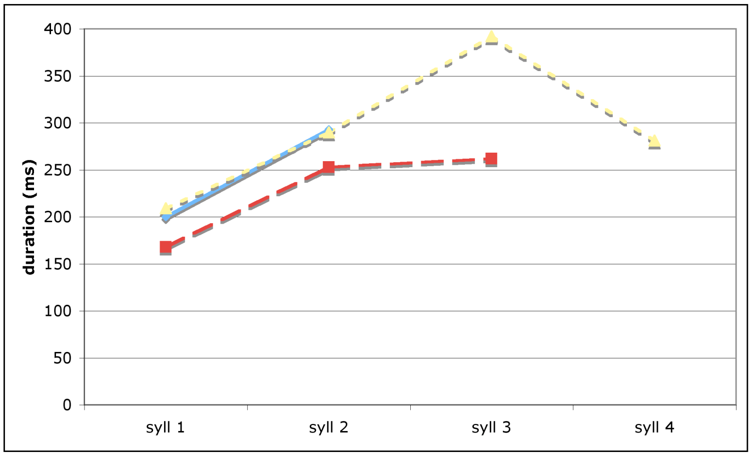


Figure 5.13: The plots show median duration patterns for binary, ternary and quaternary feet in Cameroon French.

5.3 Speaking Styles

5.3.1 Speech Rate

It is well known that a change in speech rate is not only achieved by a mere compression of the articulation gestures. An increase in speech rate may lead to complicated articulatory processes, e.g. a stronger overlap of articulation gestures leading to segmental elision on the phonetic surface (e.g. Kröger (1996)). Other strategies to accelerate speech is the omission of pauses, thereby enabling an increase in rate without accelerating the articulatory movements or to accept the circumstance of not reaching an articulatory target. The typical tradeoff between articulatory precision and articulatory effort is modelled in the H&H-theory by Lindblom (1990). On the level of prosody, an increase in rate is usually attributed a minimization of pauses and pause durations and a flat fundamental frequency contour (Kohler (1986); Rietveld and Gussenhoven (1987)). A decrease in rate tends to be accompanied by more and longer pauses and a higher f0-excursion (e.g. see the overview in Trouvain (2004)). With regards to rhythm, the effects are less examined. In Dellwo and Wagner (2003); Barry and Russo (2003), various rhythm metrics such as PVI and the Ramus Metric were identified as being strongly influenced by speech rate. While this can be seen as a weakness of the metrics, it could also mean that the various rates show a rate dependent rhythm, just as intonation, pause structure and articulation are influenced by rate. The multidimensional rhythm analysis is less endangered to show tempo related effects which are mere artifacts of the chosen metric since it only relies on methods based on mean durations etc. where it is desirable, i.e. when absolute tempo related characteristics are investigated. In the remainder of this section, the interaction between tempo and rhythmical realization is examined for two languages, namely German and French. The analysis is only performed on those languages of which there are more than 5 speakers in the database and because they are very dissimilar in rhythmic structure. The Bonn Tempo database is built on data on different tempos which are intended by the speakers themselves, e.g. each speaker made his or her own choice with reagards to a *normal*, *high* or *slow* tempo. The highest tempo recorded for each speaker in the database is the subjective maximal speech producable by the speakers which is still comprehensive. It has been described in Dellwo and Wagner (2003) that almost all speakers managed to produce versions of different tempos that fulfilled the desired properties, i.e. the intended maximal speech was faster than any other versions. In the subsequent analysis of fast vs. slow speaking styles, the two intended slow (moderate and slowest) are compared against the two fast conditions (faster than normal and fastest) across the different speakers for each language.

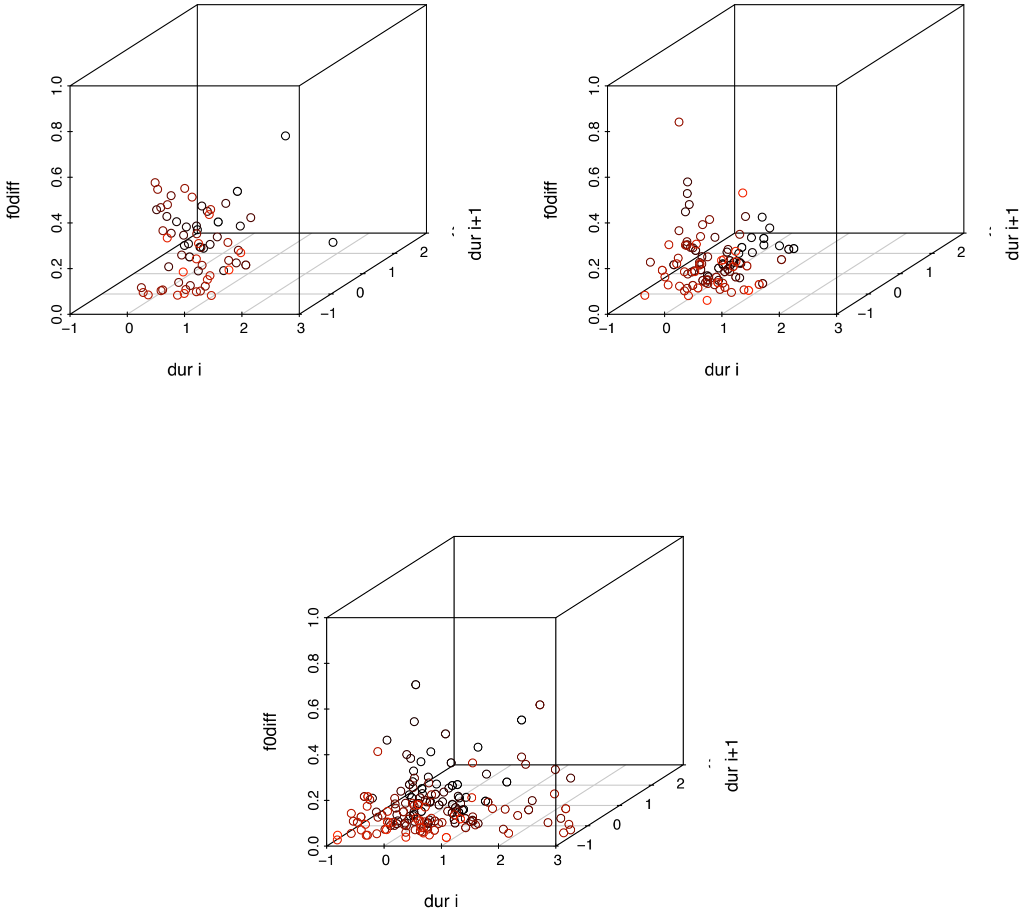


Figure 5.14: The figures show the alignment of stressed beats within an embedding intonational phrase for French utterances produced by a native speaker from Cameroon for phrase final transitions (top left), foot final transitions (top right) and all other transitions (bottom). The x-axis shows the relative timing within the intonation phrase, the y-axis shows the difference of the stressed beat to a previous one in octaves.

5.3.1.1 Fast vs. Slow German

With regards to relative transitions at foot and phrase boundaries, fast German does not differ much from slow German (Figure 5.15). However, the long-term analyses (cf. Figures 5.16 and 5.17) reveal differences. While of course the average beat durations are shorter in the fast versions, it is the slow versions revealing deviations from the typical foot patterns starting with a long beat followed by short beats and only a slight foot final lengthening effect. The fast versions show these properties as well, only reducing the foot final lengthening. Thus, the slow tempo does not have the effect on speakers to stay to a rigid pattern but apparantly causes a certain *freedom of timing*, especially showing in a stronger tendency towards foot final lengthening. This effect may go hand in hand with the higher frequency of intonation phrases in slow speech. The foot final lengthening effect may indicate potential phrase boundaries which are not yet realized as such, but are on their way of becoming so. The principle of compensatory shortening is obeyed only in the slow versions. With respect to intonation, the fast version keep the difference between transitions to stressed and unstressed beats — transitions to foot initial beats show a higher excursion. However, the slow versions do not show a significant difference in f0-excursion (cf. Figure 5.18). Here again, the lack of time compression obviously lead to a higher freedom of intonational marking, automatically leading to a higher amount of variability and possibly idiosyncratic expression.
Summing up, is interesting that to German speakers the relative timing properties are not sacrificed for a higher speaking rate when asked to maximize it. Part of the reason for this finding may lie in the Weight-to-Stress Principle, causing complex syllables to be stressed preferrably in German. In the slow versions, obviously the foot final lengthening effect is exaggerated, while the fast versions disobey compensatory shortening and mimimize foot final lengthening. These findings indicate that the relative foot initial lengthening is a stable and probably important property of German speech rhythm.

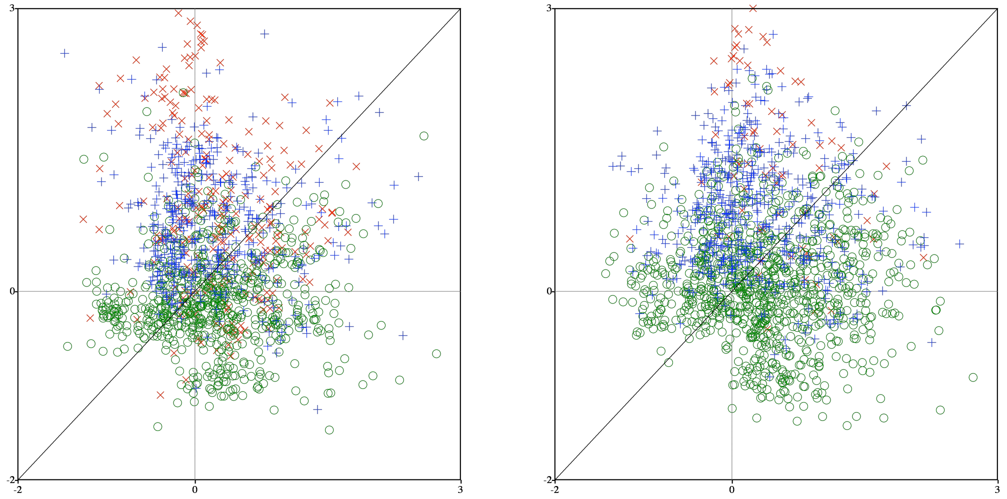


Figure 5.15: The time delay plots show slow (left) and fast (right) German timing relations. Transitions to foot final beats are shown in blue (+), transitions to phrase final beats are red (x), all other transitions are green (o).

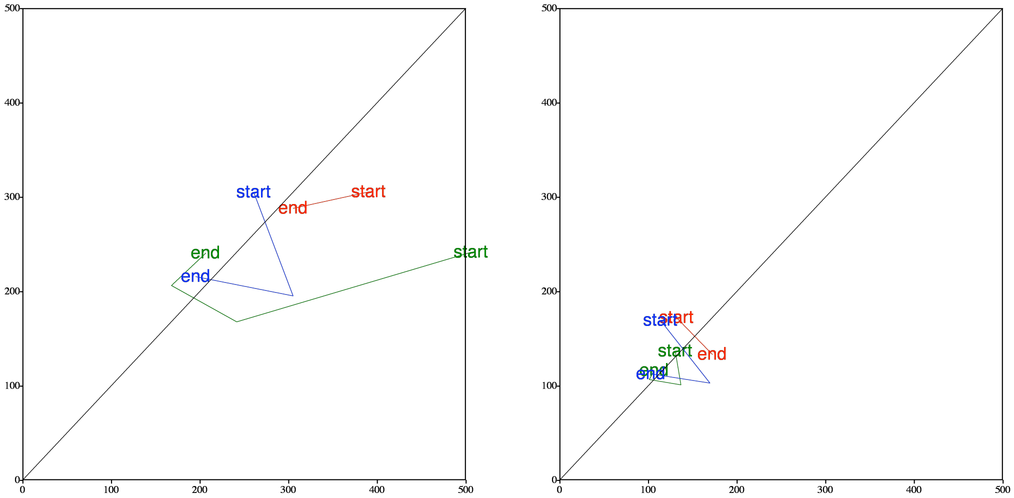


Figure 5.16: The time delay plots show long term timing delay patterns for binary, ternary and quaternary feet in slow (left) in comparison with fast German (right).

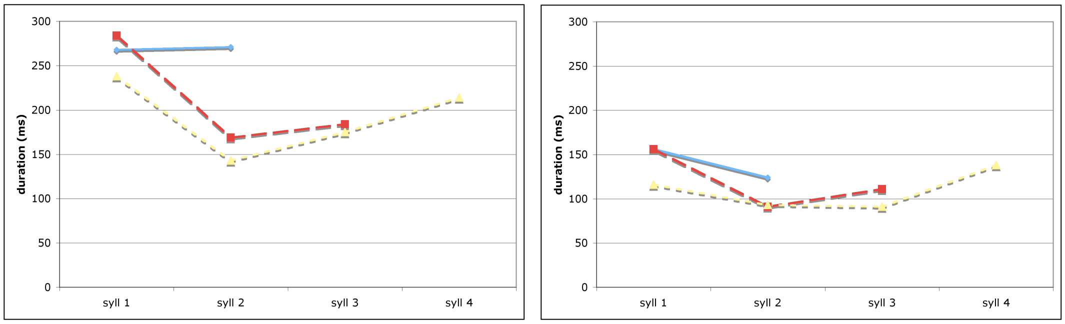


Figure 5.17: The time delay plots show absolute long term timing patterns for binary, ternary and quaternary feet in slow (left) in comparison with fast German (right).

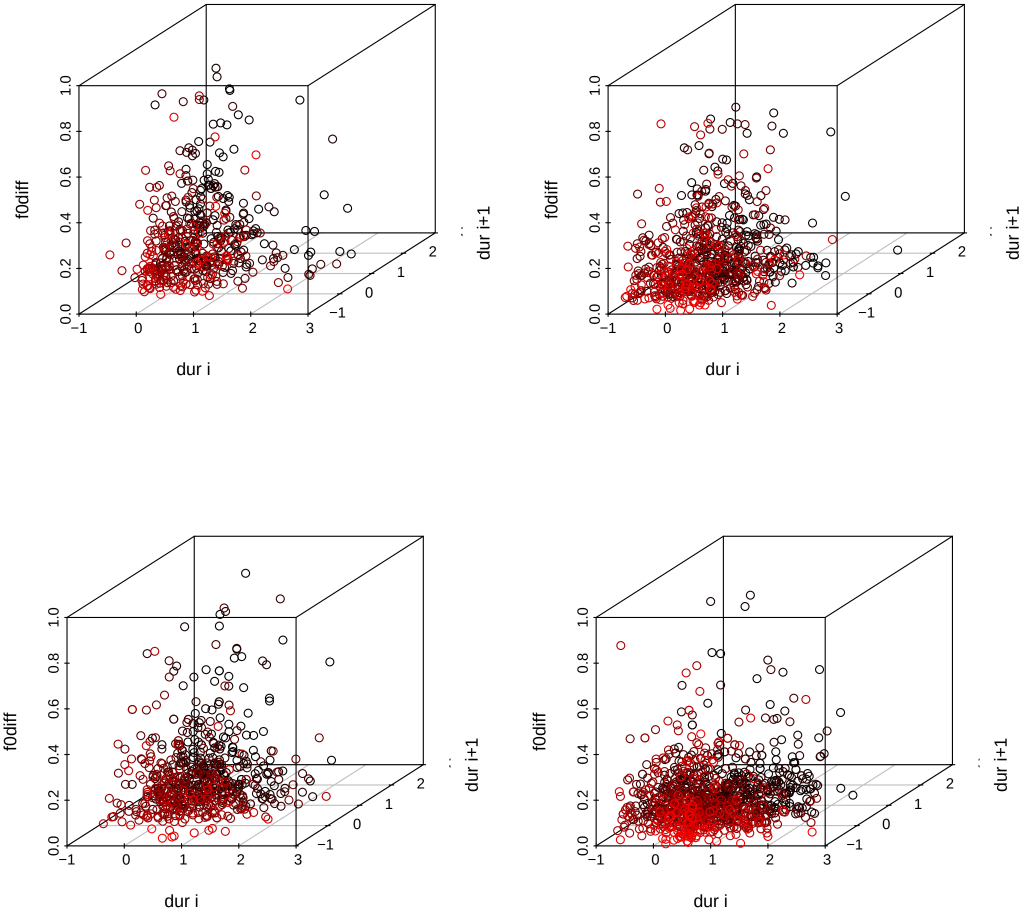


Figure 5.18: The time delay plots show slow (top) and fast (bottom) timing relations and f0-difference compared to the previous syllable in German speech. Transitions to foot initial beats are shown on the left, transitions to nonfinal, unstressed beats are shown on the right hand side. Durations (z-scores) of an event i are shown on the x-axis, durations of an event $i + 1$ on the y-axis, and f0-differences to the previous beat are shown on the z-axis (in octaves).

5.3.1.2 Fast vs. Slow French

Similar to German, relative timing at foot boundaries is not affected much when comparing fast and slow French (cf. Figure 5.21). It is possible that given that these timing features of French are rather subtle already, they are not affected much by an increase in rate. Striking, however, is the lack of phrase final lengthening in fast

French. It is here that speakers save much time when increasing their rate, since the characteristic phrase final lengthening is literally nonexistent any longer. The overall structure of fast French is in general more variable when relative durations are regarded. This may be caused by the fact that at high rates, speakers are basically bound by the segmental structure, any freedom of shaping the beat in a particular way is minimalized. On the level of absolute timing, the beat structures becomes very similar at high rates, hardly any rhythmical differences can be made out across feet of different length. However, the slow foot structure seems to introduce a slight variation as a function of foot length that might be interpreted as a subtle tendency towards final compensatory shortening (cf. Figure 5.22). The timing across the feet with their characteristic decelerating pattern becoming less strong towards the end of the foot is present in both tempos (cf. Figure 5.20). Interestingly, the slow versions do not show a significant difference in f0-excursion (cf. Figure 5.18). Here, as in German, the lack of time pressure obviously lead to a higher freedom of intonational marking, automatically leading to a higher amount of variability and possibly idiosyncratic expression. Summing up, fast French shows stable relative timing at foot boundaries and within feet. With respect to intonation, slow speech allows for more variability, while fast speech shows more rigidity.

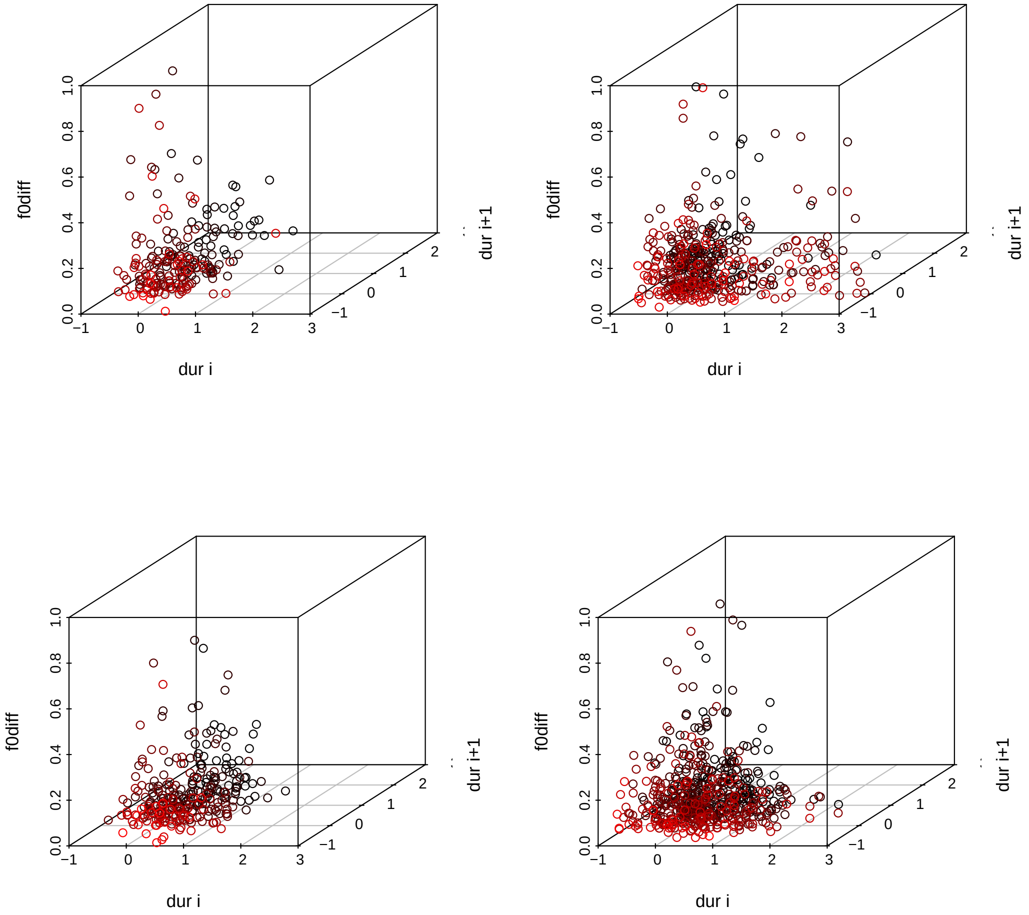


Figure 5.19: The time delay plots show slow (top) and fast (bottom) timing relations and f0-difference compared to the previous syllable in French speech. Transitions to foot initial beats are shown on the left, transitions to nonfinal, unstressed beats are shown on the right hand side. Durations (z-scores) of an event i are shown on the x-axis, durations of an event $i+1$ on the y-axis, and f0-differences to the previous beat are shown on the z-axis (in octaves).

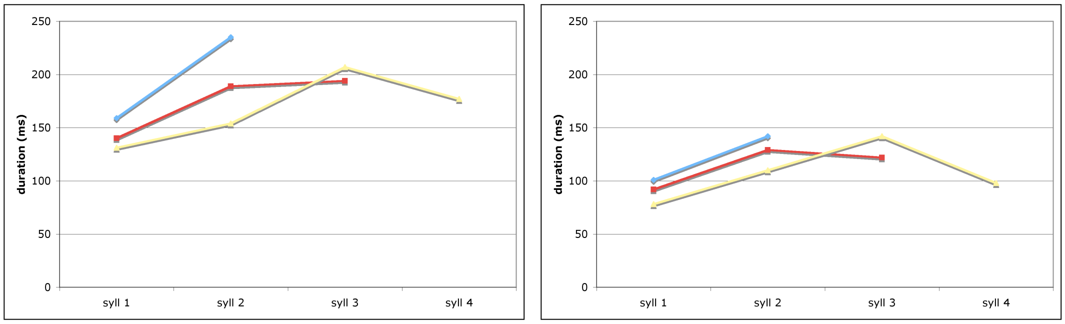


Figure 5.20: The time delay plots show absolute long term timing patterns for binary, ternary and quaternary feet in slow (left) in comparison with fast French (right).

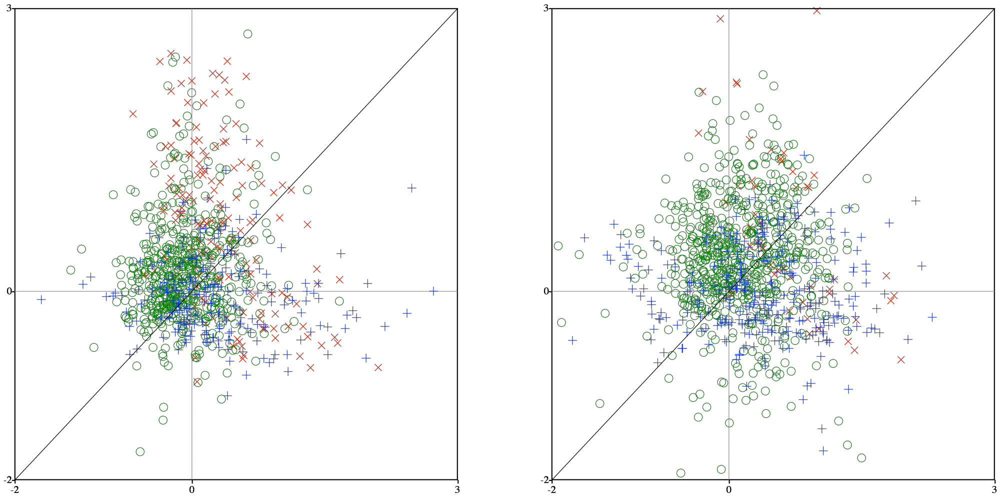


Figure 5.21: The time delay plots show slow (left) and fast (right) French timing relations. Transitions to foot initial beats are shown in blue (+), transitions to phrase final beats are red (x), all other transitions are green (o).

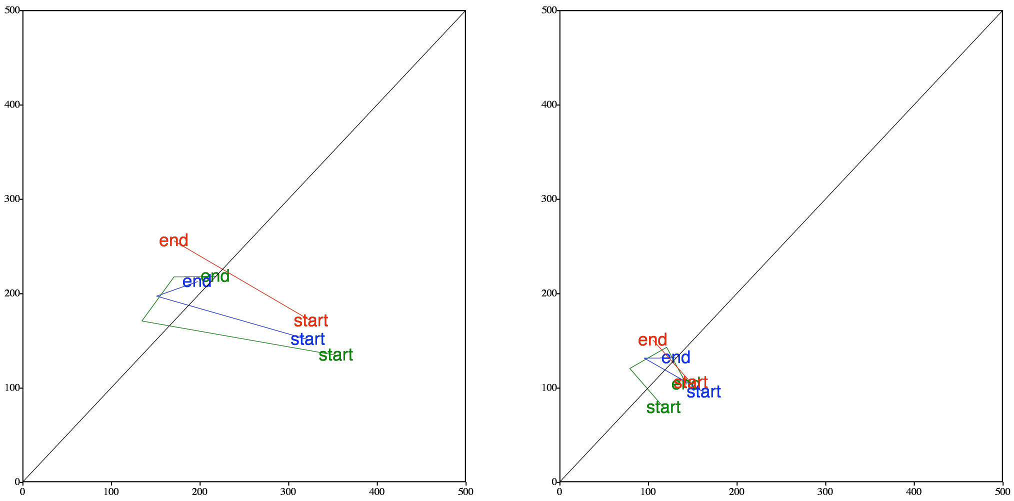


Figure 5.22: The time delay plots show long term timing delay patterns for binary, ternary and quaternary feet in slow (left) in comparison with fast French (right).

5.3.2 Rhythmically Stylized Speech: Sermons and Poetry

5.3.2.1 The Rhythm of "Entraining" Speech

Sermons which are produced in the style of speech that is particular to the evangelist movement are often called to be "entraining" — it should be no surprise if these qualities also show in the rhythmic structure of these sermons. In order to examine the style of "entraining" sermons, 41 prosodic phrases all produced by the same speaker were randomly chosen from a large online sermon database (Köthe (2008)). The speaker is a German preacher presiding an evangelist church in Stuttgart (Germany). He speaks a standard variety of German showing a very slight regiolectal color. In blogs and other online sources, his sermons are described as highly emotional and entraining. We are now testing the hypothesis that his effect may at least be slightly correlated to a unique rhythm used by the preacher. A time-delay plot (cf. Figure 5.23) reveals a striking difference to read German which was examined above. There exists a strong tendency towards isochrony and little overall variability between beats. Also, there exists a strong tendency of local isochrony, i.e. the isochrony of consecutive beats, which is visible by the strong orientation of plots along the diagonal. Of course, there are execeptions and some stressed beats are remarkably longer than others, however, the relative pattern between both types of transitions is very different from what we have seen previously, for both fast and slow German. Still, the overlap between the green and blue regions are still less compared to the French data. This reveals, that the speaker obeys the timing constraints of German, but does so in a very unique manner. Since the auditory impression reveals that the speaker overemphasizes few, probably important words, he employs a speaking style much more regular than in read speech. This impression is undermined even further when looking at the Eucledian distances between the different transition types. Here, the productions in sermon style reveal distances between transitions to stressed and unstressed and phrase final beats which are much closer to the less variable French than to highly variable German (cf. Figure 5.24). Summing up, a short analysis shows that the speaking style employed by the evangelist preacher differs highly from the standard German rhythmical pattern by being much more regular and isochronous. Few beats, which are probably co-occuring with important words, are lengthened strongly. Phrase final beats, however, are not lengthened systematically, either. Other language specific timing relations, like long-term foot patterns, are obeyed and closely resemble German read speech (Figure not shown). It is very likely that the strong uniformity of the speaking style adds to the impression of entrainment often experienced by the listeners. The analysis of this particular speaking style provides further evidence for the initial claim that rhythm — although partly determined by external timing constraints of the language spoken — can still be shaped rather independently by the speaker.

5.3.2.2 Poetic Speech

Poetry being the rhythmical speaking style *per se* should reveal at least some of its perceptual regularity with the help of our visualization method. In order to investigate poetic speech, the database constructed within the APROPOS-project (Bröggelwirth (2005)) is used. The APROPOS database contains read German poetic speech of both actors and nonprofessional speakers. The database contains various meters such as trochees, dactyls, iambs and doggerels and 2-3 poems per meter. In the present study, only iambs and trochees (binary rhythms) were examined with the help of the database. Since there are significant differences between the annotations in the APROPOS and the BonnTempo database, slight deviations from the

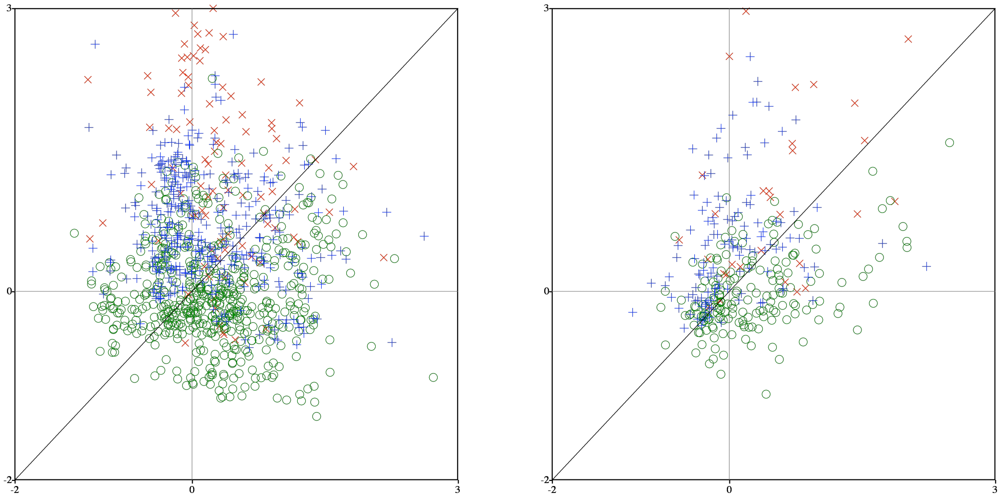


Figure 5.23: The time delay plots show timing relations in read German (left) and "sermon style" German. Transitions to foot initial beats are shown in blue (+), transitions to phrase final beats are red (x), all other transitions are green (o).

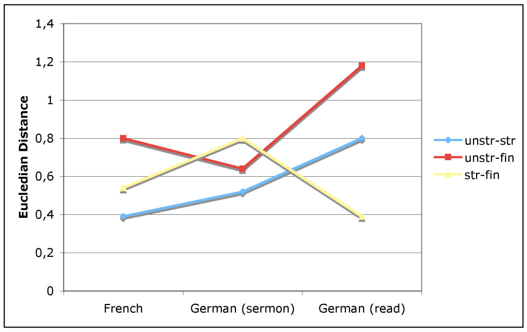


Figure 5.24: The figure depicts Eucledian distances between the different transition types for read French, "sermon style" German and read German.

originally proposed analysis process had to be made. Due to the lack of segmental annotations, syllable durations instead of beat durations were used in the analysis. Foot boundaries were determined based on the prominence annotations according to Fant and Kruckenberg (1989) contained in the database. This prominence scale consists of 32 levels and stressed beats are usually expected to receive a prominence value > 20. Thus, any syllable which had been assigned such a high prominence value was interpreted as the beginning of a new foot. This procedure is certainly advantageous in comparison with the text based foot detection method that was used in the BonnTempo database, since it is performance based. Phrase boundaries were detected based on pauses. This only causes a minor problem, since poetic speech is characterized by pause insertion in between phrases. Since poetic speech may be characterized by strong speaker idiosyncrasies and given the large amount of material, it was decided to analyze various speakers individually. The iambic and trochaic productions of two professional and two non-professional speakers were examined with the help of time-delay analyses (Figures 5.25 and 5.26). Concerning the structure of poetic speech, the following hypotheses are formulated:

1. The patterns show less variation, especially the transitions to unstressed syllables should have a tendency towards isochrony.

2. Iambic (final) lengthening should be stronger than trochaic (initial) lengthening of stressed syllables, i.e. there should be a considerable overlap between the iambic foot lengthening and phrase final lengthening.

3. There should be a better separation between different transition types compared to read speech data.

4. It can be expected that the actors show a more regular pattern than the non-professional speakers.

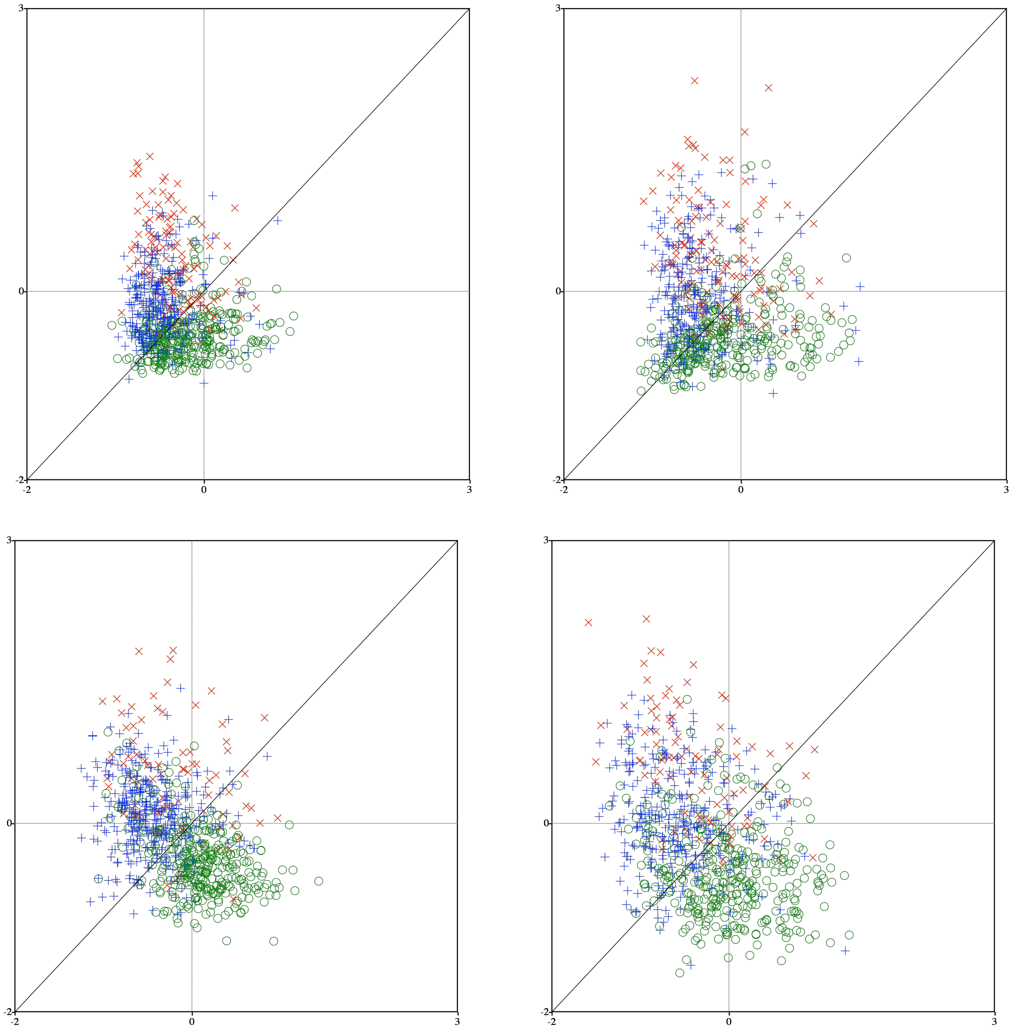


Figure 5.25: The time delay plots show timing relations in German poetic iambs. The top figures are produced by actors, the bottom figures show productions from nonprofessional speakers. Transitions to foot initial beats are shown in blue (+), transitions to phrase final beats are red (x), all other transitions are green (o).

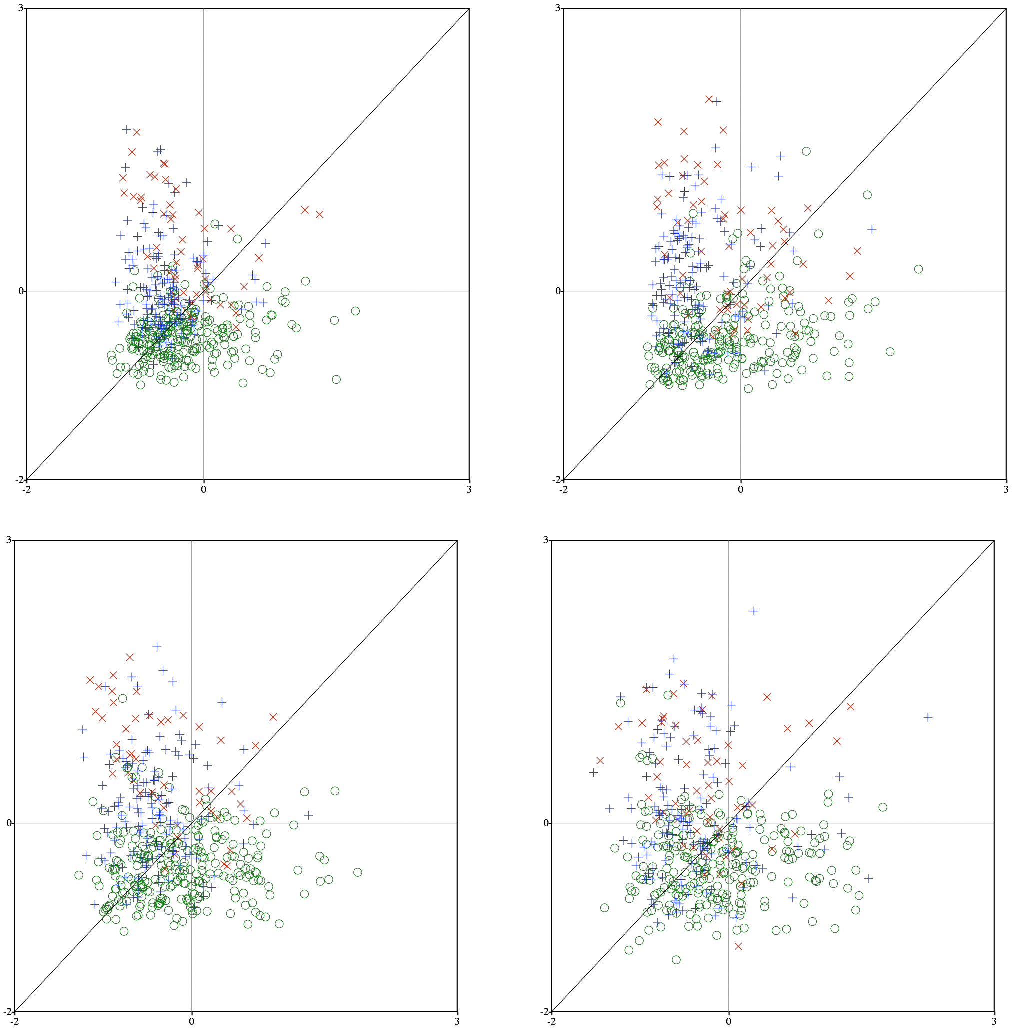


Figure 5.26: The time delay plots show timing relations in German poetic dactyls. The top figures are produced by actors, the bottom figures show productions from nonprofessional speakers. Transitions to foot initial beats are shown in blue (+), transitions to phrase final beats are red (x), all other transitions are green (o).

The plots reveal the following things: The actors and nonactors show characteristic differences. An overall impression is that the actors produce more condensed plots and show less variability. When comparing dactyls and iambs, the iambs seem to be more regular than dactyls, producing a better split between all transition types. Reasons for this phenomenon are not easy to find and would be mostly based on speculation. Another interesting finding is that especially the iambic pattern of nonactors are concentrated in the lower right and upper left quadrant, thus, they describe a very regular alternating pattern. It is very likely that this indicates a certain tendency towards the “droning” that is characteristic for a reading style of poetry without paying enough attention to the semantics of the poem — a “function follows form” approach to reading. Contrary to this, the professional actors show a concentration in the lower left quadrant, which shows sequences of rather short syllables. At times, they deviate from this pattern and produce pronounced lengthening effects. Still, their iambs show few overlaps. These findings can be interpreted in such a way that they are able to keep up the important relative timing pattern at all times, but produce sequences of very short syllables, independent of their metrical status. They deviate from this very regular, short and staccato style in certain places, which are probably semantically relevant for the poem’s content. It is very likely, that such syllables are also performed with a pronounced pitch accent (not investigated here). Hypothesis 2 seems to be falsified by the productions of both the actors and non-actors. Iambs do not show the strong final lengthening effect that was expected. Instead, it obviously is the relative timing making the difference between dactylic and iambic lengthening. The iambs of unstressed events tend to be shorter, thus creating a stronger contrast relative to the stressed beats. Apart from this finding, the major difference between iambs and dactyls lies in the overall higher variability in dactyls. Apart from this, few timing differences between the two meters can be detected. To conclude, most of our hypotheses received support: The actors produce patterns of less overall variability but less strict alternation. Poetic speech shows a lot less overlap between the various transition types “to stressed”, “to unstressed” and “to final” beats.

5.4 Speech Technology

5.4.1 Possible Applications in Speech Synthesis

Speech synthesis systems produce artificial speech out of an abstract representation, e.g. a sequence of phones supplemented with some prosodic information, e.g. the location of stresses and prosodic phrase boundaries. Often, but not always, this abstract representation has been generated out of an orthographic representation. All current approaches have in common that they need to predict the phonetic properties of the target utterance as closely as possible, e.g. its prosodic and segmental features need to be defined. Based on this description, the utterance can either be acoustically generated in parametric synthesis or built on prerecorded speech units. In the former approach, of course, rhythmic features are entirely built from scratch, in the latter, nowadays more popular approach, the realization of the rhythmic-prosodic characteristics depends on the method. Either, the units are entirely chosen based on their segmental appropriateness and the prosodic characteristics, i.e. duration and fundamental frequency, are modified in a separate step. This is usually the case if the database of prerecorded units is very small (typically diphones) and contains only one instance of a segmental unit fitting into the segmental context. In this case, of course, the prosodic properties are entirely generated based on the model prediction. However, current state-of-the-art models of speech synthesis employ the technique of *unit selection*. This means that a sequence of synthetic speech is built entirely out of prerecorded speech units and not prosodically modified. While the prerecorded units may have the size of words or even short phrases in a closed domain such as weather forecast, the units will usually be smaller, typically consisting of half-phones, in open domains such as reading machines for the blind, where the segmental chains to be generated can be very complex and may include proper names and foreign words. In the unit selection process, the algorithm choses a chain of units matching the model predictions best. This selection process is based on a cost function weighing all unit deviations from the calculated target utterance and the transitions between units. Since an optimal quality of the segmental structure is crucial to obtain a high degree of intelligibility, it is often treated as more important by the cost function. As a consequence, the prosodic characteristics of the synthetic speech often deviate significantly from the model prediction. Of course, the unit selection architecture would still allow for modification of the prosody — similar to diphone based synthesis. However, such an approach is rarely followed due to possible signal distortions introduced by the signal manipulation process itself. The question answered here is, whether the analysis tools described in the previous chapters may still come into play in enhanced speech synthesis models. First of all, though, it has to be pointed out that current models for duration prediction are still far from perfect — it seems that given the current predominance of unit selection procedures, work on duration prediction has become less popular after a rather active phase in the 1990s, in which rule based and statistical methods were integrated (e.g. van Santen (1994)) and CART-based[3] methods were optimized for duration prediction purposes. Lately, also Bayesian Networks have been used for the purpose of duration prediction (Goubanova and King (2008)). The lack of research activity may be due to the circumstance that in unit selection, a rough estimate of segmental or syllable duration is sufficient for an adequate quality given an optimal database. Our analyses do not allow for a straightforward model that can be directly integrated into existing duration prediction models. However, the theoretical insights concerning the different rhythmical strategies may come into play in multilanguage or multidialect systems using the same voice data (e.g. Traber et al. (1999)) or in applications where various speaking styles, e.g. poetic speech vs. “normal” read speech are to be modelled by integrating the characteristic timing relations into the cost function and preferring those transitions between successive units matching the correct relative timings across important prosodic boundaries.

5.4.2 Possible Applications in Speech Recognition

Applications in the large field of speech recognition are probably more straightforward than those in the field of synthesis. As we have seen, the relative timing between neighboring rhythmic events is characteristic for different languages. Given the availability of robust methods to automatically detect syllable boundaries or perceptual centers in speech, it is possible to describe these characteristics without a detailed recognition of the segmental structure of a given input. Thus, the characteristics may be very useful in order to enhance preprocessing modules in speech recognition engines which identify the spoken language or language variety before starting the subsequent segmental recognition process. Recognition can be notibly enhanced if the spoken variety is known to the recognizer in advance. However, many such approaches mainly concentrate on intonational features, e.g. Dizdarevic et al. (2004). One approach building on the basic insights presented here has already been implemented by Timoshenko and Höge (2007) while other pseudo-rhythmical approaches of language identification are built on the static relationship between

[3] CART is an abbreviation for *classification and regression tree* and is described in Breiman et al. (1984).

consonantal and vocalic intervals (Rouas et al. (2003); Rouas (2005)). It is furthermore possible to use the identified characteristics in automatic speaker recognition applications or interactive CALL applications, where the learner's performance is automatically analyzed and feedback is provided to the learner. Relative timing has also been identified as an important cue in automatic emotion classification (Schuller and Rigoll (2006)). It is possible that the classification method described in this thesis may be useful to enhance systems of automatic emotion recognition as well.

Chapter 6

Summary and Outlook

This thesis covered the following main points:

- A motivation was given why the subject of rhythm is an interesting one in the field of phonetics, linguistics and cognitive science in general.
- It summarized various factors influencing rhythm perception and production. Issues like temporal processing, beat perception, temporal windows of auditory perception, tempo perception, psychological aspects of rhythmical grouping, phonological aspects related to rhythm were presented and discussed.
- Aspects of perceptual rhythmic grouping were related to many findings of phonetic and phonological analyses thus leading to a better understanding of many perceptual phenomena in rhythm.
- An overview about the various aspects that need to be taken into account in rhythm research such as low level cognitive processing, rehearsal, language specific top-down expectancies is given.
- Several metrics aiming to classify linguistic or phonetic rhythm are examined with regards to their usefulness and their limitations. The contention was made that a full description of a language's rhythm cannot be performed concentrating on just one aspect of rhythmic phenomena such as overall variability, compensatory shortening etc. Instead, the multidimensionality of rhythm must be taken into account in order to get a full account of rhythmic phenomena.

- A set of heuristics and tools for the characterization of rhythm was presented, mostly based on time-delay plots that allow an easy and straighforward data-mining approach to detect rhythm related phenomena. The heuristics were used to establish a number of characteristics for two rhythmically prototypical languages, namely English and French. This characterization goes beyond a mere binary classification into stress timed vs. syllable timed — a strategy often employed in previous analyses. The thus detected language specific patterns were used to build abstract prosodic patterns which could be identified by listeners as either English or French above chance level.

- The multidimensional characterization has shown to offer an approach that is able to perform a typological analysis. Instead of characterizing hitherto rhythmically unclassifyable languages as "rhythmically mixed" along a stress-timed—syllable-timed continuum, the method performs a multidimensional description of rhythmical phenomena. The method is useful as well to perform speaking style analyses.

- The flexibility, simplicity and universality of the proposed method may help it becoming a useful tool in the L2-classroom and in the performance enhancement of speech synthesis and recognition systems.

Of course, due to restrictions in time and space, not all the important points were covered. One of the key issues seems to be the way that tempo perception can be integrated into rhythm models. The way that timing was treated in the proposed method definitely does not provide an adequate representation of perceived duration. However, even if for a given language an adequate model of time perception were successfully integrated, this would not be applicable for cross-language comparisons, since it is far from clear how tempo and rhythm is perceived in foreign languages. The trade-off between rhythmical top-down expectations and the way they are influencing our rhythm perception is another issue hardly tackled so far. However, we are able to prove that rhythmic patterns are relatively stable within a given language and it is very likely that our stabilized expectations influence our perception to a substantial extent. The many interfaces between rhythm and speaking style, idiolect and speaker attitudes or emotions go far beyond the scope of this book — however, I hope to have shown that rhythm is, even if complicated, possible to describe, especially if it is regarded in terms of a pattern rather than a general impression.

Probably the most important two points I wanted to make with this book are the following ones

- Timing and rhythm are important aspects of prosody that may exist independently of segmental structure, just like a language's intonation may exist independently of intrinsic pitch and tonal contrast.
- Rhythmic patterns are stable — they provide us with easily accessable information cues concerning the timing of upcoming events, the identification of the language spoken, the speaking style or speaker. Due to their relative independence of segmental input, they should remain stable even in situations of a noisy channel. Thus, they probably aid communication in a way that has so far underestimated. The usefulness of rhythmic features in applications such as dialogue modeling, speech technology and cross-lingual conversation is evident. Therefore, it is about time to explore these issues in a more systematic fashion.

List of Tables

List of Figures

Bibliography

Abercrombie, D. (1967). *Elements of general phonetics*. Aldine Publishing Corporation, Chicago.

Allen, G. (1972). The location of rhythmic stress beats in english: An experimental study i & ii. *Language and Speech*, 15:72–100.

Allen, G. and Hawkins, S. (1979). Trochaic rhythm in childen's speech. In Hollien, H. and Hollien, P., editors, *Current Issues in the Phonetic Sciences*, pages 927–933. John Benjamins.

Allen, J. (1975). Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3:75–86.

Altmann, H., Batliner, A., and Oppenrieder, W. (1989). *Zur Intonation von Modus und Fokus im Deutschen*. Niemeyer, Tübingen.

Andreeva, B., Barry, W., and Steiner, I. (2007a). The phonetic exponency of phrasal accentuation in french and german. In *Proceedings of INTERSPEECH 2007*, pages 1010–1013, Antwerp, Belgium.

Andreeva, B., Barry, W., and Steiner, I. (2007b). Producing phrasal prominence in german. In *Proceedings of the XVIth International Congress of the Phonetic Sciences*, pages 1209–1212.

anonymous (2007). English language learning and teaching. http://en.wikipedia.org/wiki/English_language_learning_and_teaching
newblock Pronunciation.

Arnold, D. (2008). Zum Einfluss der Erwartungshaltung auf die Eahrnehmung deutscher Eilbenprominenz - eine experimentalphonetische Studie. Master's thesis, Universität Bonn.

Arnold, D. and Wagner, P. (2008). The influence of top-down expectations on the perception of syllable prominence. In *Proceedings of the ISCA Workshop on Experimental Linguistics (ExLing 2008)*, Athens, Greece.

Aschersleben, G. (2000). Zeitliche Steuerung einfacher motorischer Handlungen. In *Rhythmus. Ein interdisziplinäres Handbuch*, pages 137–158. Bern: Huber.

Asu, E. and Nolan, F. (2006). Estonian and english rhythm: a two-dimensional quantification based on syllables and feet. In *Proceedings of Speech Prosody*. OS1-5_0229.

Asu, E. L. and Nolan, F. (2005). Estonian rhythm and the pairwise variability index. In *Proceedings FONETIK 2005, The XVIIIth Swedish Phonetics Conference*, pages 29–33. Department of Linguistics, Göteburg University.

Auer, P. (1993). Is a rhythm-based typology possible? A study of the role of prosody in phonological typology. Technical Report 21, KontRI Working Paper, Universität Konstanz.

Auer, P. and Uhmann, S. (1988). Silben- und akzentzählende Sprachen. *Zeitschrift für Sprachwissenschaft*, 7:214–259.

Baddeley, A. (2000). The episodic buffer: a new component of working memory. *Trends in Cognitive Science*, 4:417–423.

Baddeley, A. and Hitch, G. (1974). Working memory. In Bower, G., editor, *The psychology of learning and motivation: Advances in research and theory*, volume 8, pages 47–89. New York: Academic Press.

Baier, G. (2001). *Rhythmus - Tanz in Körper und Gehirn*. Reinbek, Rowohlt.

Barbosa, P. (2000). Syllable-timing in Brazilian Portuguese: Uma crítica a roy major. *D.E.L.T.A.*, 16(2):369–402.

Barbosa, P. (2002). Explaining cross-linguistic rhythmic variability via a coupled-oscillator mkodel of rhythm production. In *Proceedings of Speech Prosody*, pages 163–166.

Barbosa, P. and Madureira, S. (1999). Toward a hierarchical model of rhythm production: Evidence from phrase stress duration. In *Proceedings of the XIVth International Congress of the Phonetic Sciences*, volume 1, pages 297–300, San Fracisco, USA.

Barbosa, P. A. (2001). Generating duration from a cognitively plausible model of rhythm production. In *Proceedings of Eurospeech*, volume 2, pages 967–970, Aalborg, Denmark.

Barry, W., Andreeva, B., Russo, M., Dimitrova, S., and Kostadinova, T. (2003). Do rhythm measures tell us anything about language type? In *Proceedings of the XVth International Conference of the Phonetic Sciences*, pages 2693–2696, Barcelona, Spain.

Barry, W. and Russo, M. (2003). Measuring rhythm. Is it separable from speech rate? In Mettouchi, A. and Ferré, G., editors, *Actes de Interfaces Prosodiques*, pages 15–20. Nantes: Université Nantes.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2007). The impact of f0 extraction errors on the classification of prominence and emotion. In *Proceedings of the XVIth International Congress of the Phonetic Sciences*, pages 2001–2004, Saarbrücken, Germany.

Baumann, S. (2006). *The Intonation of Givenness - Evidence from German*, volume 508 of *Linguistische Arbeiten*. Tübingen: Niemeyer.

Beckman, M. and Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic consistency. In *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech*, pages 152–178. Cambridge University Press.

Benguerel, A.-P. and D'Arcy, J. (1986). Time warping and the perception of rhythm in speech. *Journal of Phonetics*, 14:231–246.

Bertinetto, P. (1989). Reflections of the dichotomy ”stress” vs. ”syllable-timing”. *Revue de Phonétique Appliquée*, 91-93:99–130.

Bertrán, A. P. (1999). Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design*, 2:103–130.

Bilmes, J. (1993). Timing is of the essence: Perceptual and computational techniques for representing, learning, and reproducing expressive timing in percussive rhythm. Master's thesis, Massachussetts Institute of Technology.

Boersma, P. and Weenink, D. (2008). Praat: Doing phonetics by computer (version 5.0.13). http://www.praat.org.

Bolton, T. (1894). Rhythm. *American Journal of Psychology*, 6:145–238.

Boltz, M. (1989). Rhythm and “good endings”: Effects of temporal structure on tonality judgments. *Perception and Psychophysics*, 46(1):9–17.

Bond, Z. and Fokes, J. (1985). Non-native patterns of English syllable timing. *Journal of Phonetics*, 11:407–420.

Bond, Z., Markus, D., and Stockmal, V. (2003). Prosodic and rhythmic patterns produced by native and non-native speakers of a quantity language. In *Proceedings of the XVth International Congress of the Phonetic Sciences*, pages 527–530.

Bouzon, C. and Hirst, D. (2004). Isochrony and prosodic structure in british english. In *Proceedings of Speech Prosody*, Nara, Japan.

Bregman, A. S. (1990). *Auditory Scene Analysis: The perceptual organization of sound*. Cambridge, MA: MIT Press.

Breiman, C., Friedman, L., Ohlson, J., and Stone, R. (1984). *Classification and Regression Trees*. Pacific Grove: Wadsworth International Group.

Breuer, S., Fancuzik, K., and Demenko, G. (2006). Analysis of polish segmental duration with cart. In *Proceedings of Speech Prosody*, pages PS2–05_0264, Dresden, Germany.

British_Council and BBC (2002). Teaching English. http://www.teachingenglish.org.uk.

Broadbent, D. and Ladefoged, P. (1958). Auditory perception of temporal order. *Journal of the Acoustical Society of America*, 31:1539.

Bröggelwirth, J. (2007). *Ein rhythmisch-prosodisches Modell lyrischen Sprechstils*. Doctoral Dissertation, Universität Bonn.

Brown, G. (1983). Prosodic structure and the given/new distinction. In Cutler, A. and Ladd, D., editors, *Prosody: Models and Measurements*, volume 14 of *Language and Communication*, pages 67–77. Berlin, Springer.

Bruhn, H. (2000a). Kognitive aspekte der Entwicklung von Rhythmus. In Müller, K. and Aschersleben, G., editors, *Rhythmus. Ein interdisziplinäres Handbuch*, page 227–244. Bern: Huber.

Bruhn, H. (2000b). Zur definition von Rhythmus. In und G. Aschersleben, K. M., editor, *Rhythmus - ein interdisziplinäres Handbuch*, pages 41–56. Bern: Huber.

Bröggelwirth, J. (2005). A rhythmic-prosodic model of poetic speech. In 2397-2400, editor, *Proceedings of INTERSPEECH*, Lisbon, Portugal.

Buder, E. and Eriksson, A. (1999). Time-series analysis of conversational prosody for the identification of rhythmic units. In *Proceedings of the XIVth International Congress of the Phonetic Sciences*, San Francisco.

Buxton, H. (1983). Temporal predictability in the perception of English speech. In Cutler, A. and Ladd, D., editors, *Prosody: Models and Measurements*, volume 14 of *Language and Communication*. Berlin, Springer.

Campbell, N. (1992). Syllable-based segmental duration. In Bailly, G., Benoit, C., and Sawallis, T., editors, *Talking Machines. Theories, Models, and Designs*, pages 211–224. Elsevier Publishers.

Campbell, N. (1995). Loudness, spectral tilt and perceived prominence. In *Proceedings of the XIIIth International Congress of the Phonetic Sciences*, Stockholm, Sweden.

Campbell, N. (1999). A study of japanese speech timing from the syllable perspective. *Journal of the Acoustical Society of Japan*, 3:29–39.

Cauldwell, R. (2002). The functional irrhythmicality of spontaneous speech: A discourse view of speech rhythms. *Applied Language Studies*, 2(1):1–24.

Charpentier, F. and Moulines, E. (1989). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. In *Proceedings of EUROPEECH*, pages 13–19, Paris:ENTS.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Cambridge, MIT Press.

Clarke, E. (1999). Rhythm and timing in music. In Deutsch, D., editor, *The Psychology of Music*, pages 473–500. New York: Academic Press, second edition.

Classé, A. (1939). *The Rhythm of English Prose*. Basil Blackwell, Oxford.

Claßen, K., Dogil, G., Jessen, M., Marasek, K., and Wokurek, W. (1998). Stimmqualität als Korrelat der Wortbetonung im deutschen. *Linguistische Berichte*, 174:202–245.

Clements, G. (1990). The role of the sonority cycle in core syllable. In Beckman, M. E. and Kingston, J., editors, *Papers in Laboratory Phonology I*, pages 283—333. Cambridge: CUP.

Cohn, A. (2003). Phonological structure and phonetic duration: The role of the mora. *Working Papers of the Cornell Phonetics Laboratory*, 15:69–100.

Cooper, A., Whalen, D., and Fowler, C. (1986). P-centers are unaffected by phonetic categorization. *Perception and psychophysics*, 39:187–196.

Cooper, G. and Meyer, L. (1960). *The Rhythmic Structure of Music*. Chicago: University Press.

Couper-Kuhlen, E. (1986). *An introduction to English prosody*. Tübingen, Niemeyer.

Crystal, D. (1969). *Prosodic Systems and Intonation in English*. London, Cambridge University Press.

Crystal, T. and House, A. (1988). Segmental durations in connected speech signals: Current results. *Journal of the Acoustical Society of America*, 83:1553–1573.

Cummins, F. (2002). Speech rhythm and rhythmic taxonomy. In *Proceedings of Speech Prosody 2002*, Aix en Provence, France.

Cummins, F. (2005). Interval timing in spoken lists of words. *Music Perception*, 22(3):497–508.

Cummins, F. and Port, R. (1998). Rhythmic constraints of stress timing in english. *Journal of Phonetics*, 26:145–171.

Cutler, A. (1980). Syllable omission errors and isochrony. In *Temporal Variables in Speech. Studies in Honor of Frieda Goldmann-Eisler*, pages 183–190. Mouton: The Hague, Paris, New York.

Cutler, A. (1994). The perception of rhythm in language. *Cognition*, 50:79–81.

Dalla-Bella, S. and Peretz, I. (2005). Differentiation of classical music requires little learning but rhythm. *Cognition*, 96:B65–B78.

Dancovičova, J. and Dellwo, V. (2007). Czech speech rhythm and the rhythm class hypothesis. In *Proceedings of the XVIth International Conference of the Phonetic Sciences*, pages 1241–1244, Saarbrücken, Germany.

Darwin, C. and Donovan, A. (1980a). Perceptual studies of speech rhythm: Isochrony and intonation. In *Spoken Language Generation and Understanding*. Dordrecht: D. Riedel Publishing Company.

Darwin, C. J. (1997). Auditory grouping. *Trends in Cognitive Sciences*, 1(9):327–333.

Darwin, C. J. and Donovan, A. (1980b). Perceptual studies of speech rhythm: isochrony and intonation. In Simon, J., editor, *Proceedings of NATO ASI on Spoken Language Generation and Understanding*.

Dasarathy, B. V., editor (1991). *Nearest Neighbor (NN) Norms: Patterns Classification Techniques*.

Dauer, R. (1983). Stress timing and syllable timing reanalyzed. *Journal of Phonetics*, 11:51–62.

Dauer, R. (1987). Phonetic and phonological components of language rhythm. In *Proceedings of the XIth International Congress of the Phonetic Sciences*, volume 5, pages 447–450. Tallinn: Academy of Sciences.

Davies, M. and Plumbley, M. (2005). Beat tracking with a two state model. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages iii/241–iii/244.

de Manrique, A. and Signorini, A. (1983). Segmental duration and rhythm in spanish. *Journal of Phonetics*, 11:117–128.

de Saussure, F. (1916). *Cours de linguistique générale*. Payot, Paris.

Delattre, R. (1966). A comparison of syllable length conditioning among languages. *International Review of Applied Linguistics*, 4:183–198.

Dellwo, V. (2003). Rhythm and speech rate: A variation coefficient for c. In *Proceedings of the 38th Linguistic Colloquium*, Budapest, Hungary.

Dellwo, V. (2008a). *The influence of speech rate on speech rhythm*. PhD thesis, Universität Bonn.

Dellwo, V. (2008b). Influences of language typical speech rate on the perception of speech rhythm. In *Abstract Book of the Workshop on Empirical Approaches to Speech Rhythm*, UCL, London. CHC.

Dellwo, V., Aschenberner, B., Wagner, P., Dankovicova, J., and Steiner, I. (2004). BonnTempo-corpus and BonnTempo-tools: A database for the study of speech rhythm and rate. In *Proceedings INTERSPEECH 2004*, pages 777–780, Jeju Island, Korea.

Dellwo, V., Fourcin, A., and Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. In *Proceedings of the XVIth International Conference of the Phonetic Sciences*, pages 1129–1132, Saarbrücken, Germany.

Dellwo, V. and Wagner, P. (2003). Relationships between speech rate and rhythm. In *Proceedings of the 15th International Conference of the Phonetic Sciences*, pages 471–474, Barcelona, Spain.

Demenko, G. (2003). Phrase time structure modeling for speech synthesis purposes. In *Proceedings of the XVth International Congress of the Phonetic Sciences*, Barcelona.

Dennet, D. (1991). Real patterns. *Journal of Philosophy*, LXXXVIII:27–51.

Desain, P. (1992). A (de)composable theory of rhythm perception. *Music Perception*, 9(4):439–54.

Deterding, D. (2001). The measurement of rhythm: a comparison of singapore and british english. *Journal of Phonetics*, 29:217–230.

di Cristo, A. (1998). Intonation in french. In Hirst, D. and di Cristo, A., editors, *Intonation Systems: a survey of twenty languages*, pages 195–218. Cambridge University Press, Cambridge, England.

Dietze, G. (1885). Untersuchungen über den Anfang des Bewußtseins bei regelmäßig aufeinanderfolgenden Schalleindrücken. *Philosophische Studien*, 2:362–394.

Dixon, S. (2001). An empirical comparison of tempo trackers. In *8th Brazilian Symposium on Computer Music*, Fortaleza, Brazil.

Dizdarevic, V., Hagmüller, M., Kubin, G., Pernkopf, E., and Baum, M. (2004). Prosody-based recognition of german varieties. In *Proceedings of ICASSP*, volume 1, pages 929–932.

Dogil, G. (1995). Phonetic correlates of word stress. In der Hulst, V., editor, *Word Prosodic Systems in the Languages of Europe*. De Gruyter, Berlin.

Donovan, A. and Darwin, C. (1979). The perceived rhythm of speech. In *Proceedings of the IXth International Congress of the Phonetic Sciences*, volume 2, pages 268–274, Copenhagen.

Dupoux, E., Pallier, C., Galles, N. S., and Mehler, J. (1997). A destressing 'deafness' in french? *Journal of Memory and Language*, 36:406–421.

Dupoux, E., Peperkamp, S., and Galles, N. S. (2001). A robust method to study stress „deafness“. *Journal of the Acoustical Society of America*, 3:1606–1618.

Ebbinghaus, H. (1885/1964). *Memory: A contribution to experimental psychology*. New York: Dover.

Eckert, H. and Barry, W. (2002). *The phonetics and phonology of English pronunciation*. Wissenschaftlicher Verlag Trier.

Eriksson, A. (1991). *Aspects of Swedish speech rhythm*. PhD thesis, University of Göteborg, Sweden.

Eriksson, A., Grabe, E., and Traunmüller, H. (2002). Perception of syllable prominence by listeners with and without competence in the tested language. In *Proceedings of Speech Prosody, Aix en Provence, France*.

Eriksson, A., Thunberg, G., and Traunmüller, H. (2001). Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing. In *Proceedings of EUROSPEECH*, pages 399–402, Aalborg, Denmark.

Fant, G. and Kruckenberg, A. (1989). Preliminaries to the study of swedish prose reading and reading style. *STL-QPSR*, 2/1989:1–83.

Fant, G. and Kruckenberg, A. (1996). On the quantal nature of speech timing. In *Proceedings of ICSPL '96*, volume 4, pages 2044–2047.

Fay, W. H. (1966). *Temporal Sequence in the Perception of Speech*. The Hague: Mouton & Co.

Ferragne, E. and Pellegrino, F. (2004). Rhythm in read british english: Interdialect variability. In *Proceedings of Interspeech*, pages 1573–1576, Jeju Island, Korea.

Font, C. and Mestre, R. (1991). Compensatory shortening in spanish spontaneous speech. In *Proceedings of ESCA*, volume 16, pages 1–5.

Fourakis, M. and Monahan, C. B. (1988). Effects of metrical foot structure on syllable timing. *Language and Speech*, 31:283–306.

Fowler, C. (1983). Converging sources of evidence on spoken and perceived rhythms of speech: cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General*, 112:386–412.

Fowler, C. A. (1977). *Timing Control in Speech Production*. Bloomington, Indiana: Indiana University Linguistics Club.

Fraisse, P. (1963). *The psychology of time*. NY: Harper and Row.

Fraisse, P. (1982). Rhythm and tempo. In Deutsch, D., editor, *The Psychology of Music*, pages 149–80. Academic Press, New York.

Friberg, A. and Sundberg, J. (1995). Time discrimination in a monotonic isochronous sequence. *Journal of the Acoustical Society of America*, 98(5):2524–31.

Frota, S., Vigario, M., and Martins, F. (2002). Discrimination and rhythm classes: Evidence from portuguese. In *Proceedings of Speech Prosody*, Aix en Provence, France.

Fry, D. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27:765–768.

Fry, D. (1958). Experiments in the perception of stress. *Language and Speech*, 1:126–152.

Gasser, M. and Eck, D. (1996). Representing rhythmic patterns in a network oscillators. In *Proceedings of the Fourth International Conference on Music Perception and Cognition*, pages 361–6, Montreal, Quebec. Faculty of Music, McGill University.

Geiser, E., Schmidt, C., Jancke, L., and Meyer, M. (2006). Neural correlates of rhythm processing in speech perception. In *Proceedings of Speech Prosody*, Dresden, Germany.

Gibbon, D. (2003a). Computational modelling of rhythm as alternation, iteration and hierarchy. In *Proceedings of the International Congress of the Phonetic Sciences*, volume 3, pages 2489–2492, Barcelona, Spain.

Gibbon, D. (2003b). Corpus-based syntax-prosody tree matching. In *Proceedings of EUROSPEECH*, pages 761–764, Geneva, Switzerland.

Gibbon, D. and Gut, U. (2001). Measuring speech rhythm. In *Proceedings of EUROSPEECH*, Aalborg, Denmark.

Gibbon, D. and Williams, B. (2007). Timing pattens in welsh. In *Proceedings of the XVIth International Conference of the Phonetic Sciences*, pages 1249–1251, Saarbrücken.

Gilbert, A. C. and Boucher, V. J. (2006). Syntax and syllable count as predictors of french tonal groups: Drawing links to memory for prosody. In *Proceedings of Speech Prosody*, pages PS3–06, Dresen, Germany.

Gilbert, A. C. and Boucher, V. J. (2007). What do listeners attend to in hearing prosodic structures? investigating the human speech-parser using short-term recall. In *Proceedings of Interspeech*, pages 430–433, Antwerp, Belgium.

Glover, H., Kalinowski, J., Rastatter, M., and Stuart, A. (1996). Effect of instruction to sing on stuttering frequency at normal and fast rates. *Perceptual and motor skills*, 83(2):511–522.

Goedemans, R. (1998). *Weightless Segments*. PhD thesis, Rijksuniversiteit te Leiden, The Netherlands.

Goldmann-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. London: Academic Press.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., Rosen, S., and Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*, 99(16):10911–10916.

Goto, M. and Muraoka, Y. (1995). Real-time rhythm tracking for drumless audio signals. In *Proceedings of IJAI-95 Workshop on Computational Auditory Scene Analysis*, pages 68–75.

Goubanova, O. and King, S. (2008). Bayesian networks for phone duration prediction. *Speech Communication*, 50(4):301–311.

Grabe, E. and Low, E. L. (2000). Acoustic correlates of rhythm class. In Gussenhoven, C., editor, *Laboratory Phonology 7*, pages 515–546. Berlin: Mouton de Gruyter.

Grabe, E., Post, B., and Watson, I. (2000). Acoustic correlates of rhythm in english and french four-year-olds. *Oxford University Working Papers in Linguistics, Philology & Phonetics*, 5:7–17.

Grice, M. and Baumann, S. (2002). Deutsche intonation und GToBI. *Linguistische Berichte*, 191:267–298.

Grice, M., Ladd, D., and Arvanati, A. (2000). On the place of phrase accents in intonational phonology. *Phonology*, 17:143–185.

Gussenhoven, C. and Rietveld, A. (1992). Intonation contours, prosodic strength and preboundary lengthening. *Journal of Phonetics*, 20:283–303.

Gut, U., Urua, E., Adouaku, S., and Gibbon, D. (2001). Rhythm in West-African tone languages: a study of Ibibio, Anyi and Ega. In *Typology of African Prosodic Systems Workshop*, Bielefeld, Germany.

Haken, H., Kelso, J., and Bunz, H. (1985). A theoretical model of phase transitions in human hand movements. *Biological Cybernetics*, 51:347–356.

Handel (1989). *Listening: An introduction to the perception of auditory events*. Cambridge, UK: Cambridge University Press.

Handel, S. (1992). The differentiation of rhythmic structure. *Perception & Psychophysics*, 52:497–507.

Handel, S. (1993). The effect of tempo and tone duration on rhythm discrimination. *Perception and Psychophysics*, 54(3):370–382.

Hayes, B. (1994). *Metrical stress theory: priciples and case studies*. University of Chicago Press: Chicago.

Heldner, M. (2001). Spectral emphasis as an additional source of information in accent detection. In *Proceedings of the ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding*, pages 57–60, Red Bank, NJ.

Hellström, A. and Rammsayer, T. (2004). Effects of time-order, interstimulus interval, and feedback in duration discrimination of noise bursts in the 50- and 1000-ms ranges. *Acta psychologica*, 116:1–20.

Henke, S. (1993). *Formen der Satzakzentuierung und ihre Beitrag zur Satzbedeutung in deutschen Aussagesätzen*. Wissenschaftlicher Verlag Trier.

Heuft, B. (1999). *Eine prominenzbasierte Methode zur Prosodieanalyse und -synthese*, volume Sprache, Sprechen, Computer. Peter Lang, Frankfurt a. M. (also doctoral dissertation, Universität Bonn, Germany).

Hirschfeld, U. and Stock, E. (2004). Aussprache. In Pabst-Weinschenk, M., editor, *Grundlagen der Sprechwissenschaft und Sprecherziehung*, pages 31–48. Ernst Reinhard Verlag, Stuttgart.

Hirsh, I. (1959). Auditory perception of temporal order. *Journal of the Acoustical Society of America*, 31:759–767.

Hirst, D. and Bouzon, C. (2005). The effect of stress and boundaries on segmental durations in a corpus of authentic speech (british english). In *INTERSPEECH-2005*, pages 29–32.

Hoequist, C. (1983). Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica*, 40:203–237.

Howell, P. (1984). An acoustic determinant of perceived and produced anisochrony. In den Broecke, M. V. and Cohen, A., editors, *Proceedings of the Tenth International Congress of the Phonetic Sciences*, pages 429–433. Dordrecht: Foris.

Howell, P. (1988). Predicton of p-center location from the distribution of energy in the ampitude envelope: Part i & ii. *Perception and Psychophysics*, 43:90–93 & 99.

Howell, P. (2004). Cerebellar activity and stuttering: Comments on max and yudman (2003). *Journal of Speech, Language and Hearing Research*, 47(1):101–111.

Hyman, L. (1985). *A Theory of Phonological Weight*. Dordrecht: Foris.

Höhle, B., Bijeljac-Babic, R., Nazzi, T., Herold, B., and Weissenborn, J. (2007). The development of language specific prosodic preferences during the first half year of life and its relation to later lexical development: Evidence from german and french. In *Proceedings of the 16th International Conference of the Phonetic Sciences*, pages 1529–1532, Saarbrücken, Germany.

Ikoma, M. (1993). Stress-timed rhythm in german speech: a study in acoustic phonetics. *Sophia Linguistica*, 33:197–216.

Isačenko, A. and Schädlich, V. (1966). Untersuchungen über die deutsche satzintonation. *Studia Grammatica*, 7:7–64.

Ivry, R. and Hazeltine, R. (1995). Perception and production of temporal intervals across a range of durations: Evidence for a common timing mechanism. *Human Perception*, 21:3–18.

James, L. (1940). *Speech signals in telephony*. University College London.

James, W. (1890). *The principles of psychology*. New York: Henry Holt.

Janker, P. (1996). Evidence for the p-center syllable-nucleus-onset correspondence hpothesis. *ZAS Papers in Linguistics*, 7:94–124.

Janker, P. and Pompino-Marschall, B. (1991). Is the p-center influenced by tone? In *Proceedings of the XIIth International Congress of the Phonetic Sciences*, volume 3, pages 290–293.

Janker, P. M. (1993). *Sprechrhythmus, Silbe, Ereignis ? Eine experimentalphonetische Untersuchung zu den psychoakustisch relevanten Parametern zur rhythmischen Gliederung sprechsprachlicher Äußerungen.* PhD thesis, Universität München. zugleich: Forschungsbericht des Instituts für Phonetik und Sprachliche Kommunikation der Universität München, 33.

Jassem, W. (1952). *Intonation in Conversational English*. PhD thesis, Polish Academy of Science, Warsaw.

Jassem, W., Hill, D., and Witten, D. (1984). Isochrony in english speech: its statistical validity and linguistic relevance. In Gibbon, D. and Richter, H., editors, *Intonation, Accent and Rhythm. Studies in Discourse Phonology*, pages 203–225. Walter de Gruyter, Berlin, New York.

Jensen, C. and Tøndering, J. (2005). Choosing a scale for measuring perceived prominence. In *Proceedings of Interspeech*, pages 2385–2388, Lisbon, Portugal.

Jessen, M., Marasek, K., and Claßen, K. (1995). Acoustic correlates of word stress and the tense/lax opposition in the vowel system of german. In *Proceedings of the XIIIth International Congress of the Phonetic Sciences*, volume 4, pages 428–432, Stockholm, Sweden.

Jones, D. (1918, 1976a). *An outline of English phonetics*. Cambridge, Cambridge University Press, 9 edition.

Jones, M. R. (1976b). Time, our lost dimension: Toward a new theory of perception, attention, and memory. *Psychological Review*, 83:323–355.

Jongsma, M. L., Desain, P., and Honing, H. (2004). Rhythmic context influences the auditory evoked potentials of musicans and nonmusicians. *Biological Psychology*, 66:129–152.

Kager, R. (1999). *Optimality Theory*. Cambridge University Press.

Kaiki, N., Takeda, N., and Sagisaka, Y. (1992). Linguistic properties in the control of segmental duration for speech synthesis. In Bailly, G., Benoit, C., and Sawallis, T., editors, *Talking Machines: Theories, Models, and Designs*, pages 255–264, Elsevier: Amsterdam.

Kalveram, K. T. (2000). Stottern: Eine rhythmusstörung in einer hierarchisierten handlungssteuerung? In K. Müller, G. A., editor, *Rhythmus - ein interdisziplinäres Handbuch*, pages 191–217. Bern: Huber.

Kato, H., Tsuzaki, M., and Sagisaka, Y. (1997). Acceptability for temporal modification of consecutive segments in isolated words. *Journal of the Acoustical Society of America*, 101:2311–2322.

Katz, D. (1948). *Gestaltpsychologie*. Basel: Schwabe.

Kaufmann, S. (2006). Wie erklären psycholinguistische Produktionsmodelle das Stottern? Master's thesis, Universität Bonn.

Kehrein, R. (2002). *Prosodie und Emotionen*. Tübingen: Niemeyer.

Keller, E. and Zellner, B. (1995). A statistical timing model for french. In *Proceedings of the XIIIth International Congress of the Phonetic Sciences*, volume 3, pages 302–305, Stockholm.

Keller, E. and Zellner, B. (1996). A timing model for fast french. In *York Papers in Linguistics*, volume 17, pages 53–75. University of York.

Keller, E., Zellner, B., Werner, S., and Blanchoud, N. (1993). The prediction of prosodic timing: Rules for final syllable lengthening in french. In *ESCA Workshop on Prosody*, pages 212–215, Lund, Sweden.

Keller, E., Zellner-Keller, B., and Local, J. (2000). A serial prediction component for speech timing. In Sendlmeier, W., editor, *Speech and Signals - Aspects of Speech Synthesis and Automatic Speech Recognition*, volume 69 of *Forum Phoneticum*, pages 41–49. Frankfurt: Hektor.

Kien, J. and Kemp, A. (1994). Is speech temporally segmented? comparison with temporal segmentation in behaviour. *Brain and Language*, 46:662–682.

Kiparsky, P. (1975). Stress, syntax and meter. *Language*, 51:576–616.

Kiparsky, P. (1977). The rhythmic structure of english verse. *Linguistic Inquiry*, 8:189–247.

Klapuri, A., Eronen, A., and Astola, J. (2006). Analysis of the meter of musical signals. *IEEE Transactions on Speech and Audio Processing*, 14(1):342–355.

Klatt, D. (1979). Synthesis by rule of segmental durations in english sencences. In Lindblom, B. and Öhmann, S., editors, *Frontiers of Speech Communication Research*, pages 287–299. London: Academic Press.

Klessa, K. (2006). *Analiza iloczasu głoskowego na potrzeby syntezy mowy polskiej. (Polish Segmental Duration Analysis for the Purpose of Speech Synthesis)*. PhD thesis, Uniwersytet im. Adama Mickiewicza, Poznań.

Kochanski, G., Grabe, E., and Coleman, J. (2005). Loudness predicts prominence: fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118:1038–1054.

Koffka, K. (1909). Experimentelle Untersuchungen zur Lehre vom Rhythmus. *Zeitschrift für Psychologie*, 52:1–109.

Koffka, K. (1935). *Principles of Gestalt Psychology*. New York: Harcourt.

Kohler, K. (1986). Parameters of speech rate perception in german words and sentences: duration, pitch movement and pitch level. *AIPUK, Universität Kiel*, 22:137–177.

Kohler, K. (1991). Terminal intonation patterns in single-accent utterances of german: Phonetics, phonology and semantics. *AIPUK, Universität Kiel*, 25:115–185.

Kohler, K. (2006). What is emphasis and how is it coded? In *Proceedings of Speech Prosody 2006*, Dresden, Germany.

Kohler, K. J. (1982). Rhythmus im Deutschen. *Arbeitsberichte, Institut für Phonetik der Universität Kiel (AIPUK)*, 19:89–105.

Kohler, K. J. (1983). Stress-timing ans speech rate in German. A production model. *Arbeitsberichte, Institut für Phonetik der Universität Kiel*, 20:7–53.

Kohler, K. J. (2003). Neglected categories in the modelling of prosody: pitch timing and non-pitch accents. In *Proceedings of the 15th International Conference of the Phonetic Sciences*, pages 2925–2928, Barcelona, Spain.

Kohler, K. J. (2005). Form and function of non-pitch accents. AIPUK 35a, Universität Kiel.

Kotby, M., Moussa, A., El-Sady, S., and Nabieh, A. (2003). A comparative study between certain behavioral methods in treatment of stuttering. In Zohny, A. and Ruben, R., editors, *Oto-Rhino-Laryngology*, volume 1240 of *International Congress Series*, pages 1243–1249. Elsevier.

Krampe, R. T., Engbert, R., Kliegl, R., and Fuchs, J. (2000). Koordination und synchronisation der Hände beim rhythmischen Timing. In Müller, K. and Aschersleben, G., editors, *Rhythmus. Ein interdisziplinäres Handbuch*, pages 163–183. Bern: Huber.

Kruckenberg, A. and Fant, G. (1993). Iambic versus trochaic patterns in poetry reading. In *Nordic Prosody VI*, pages 123–135, Stockholm.

Krull, D., Traunmüller, H., and van Dommelen, W. (2003). Local speaking rate and perceived quantity. *PHONUM*, 9:41–44.

Kröger, B. J. (1996). Zur phonetischen Realisierung von Sprechtempoänderungen unter Einbeziehung von artikulatorischer Eeorganisation: Artikulatorische und perzeptive Untersuchungen. In Gibbon, D., editor, *KONVENS*, pages 171–185. de Gruyter.

Köhlmann, M. (1984). Bestimmung der Silbenstruktur von fließender Sprache mit Hilfe der Rhythmuswahrnehmung. *Acustica*, 56:120–125.

Köthe, S. (2008). Sermon Online. Die Predigt-Datenbank. Online Ressource. http://www.sermon-online.de

Ladd, D. (1996). *Intonational Phonology*, volume 79 of *Cambridge Studies in Linguistics*. Cambridge University Press.

Large, E. W. and Kolen, J. F. (1994). Resonance and the perception of musical meter. 6(2+3):177–208.

Lea, J. (1980). The association between rhythmic ability and language ability. In Jones, F., editor, *Language disabilities*, pages 217—230. Lancaster, MA: MIT Press.

Lee, C. S. and McAngus-Todd, N. P. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition*, 93:225–254.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Mass, and London: MIT Press.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5:253–263.

Lehiste, I. (1990). Phonetic investigation of metrical structure in orally produced poetry. *Journal of Phonetics*, 18:123–133.

Lenneberg, E. H. (1967). *Biological Foundations of Language*. New York: John Wiley & Sons, Inc.

Lerdahl, F. and Jackendoff, R. (1983). *Generative theory of tonal music*. Cambridge, MA: MIT Press, 2nd, revised edition edition.

Levelt, W. (1994). Do speakers have access to a mental syllabary? *Cognition*, 50:239–269.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8:249–336.

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Marchal, A., editor, *Speech Production and Speech Modeling*. Dordrecht: Kluwer Academic Publishers.

Lintfert, B. and Schneider, K. (2005). Acoustic correlates of contrastive stress in German children. In *Proceedings of Interspeech 2005*, pages 1177–1180, Lisbon, Portugal.

Little, M. A. (2006). *Biomechanically Informed Speech Signal Processing*. PhD thesis, University of Oxford.

Lleó, C., Rakow, M., and Kehoe, M. (2007). Acquiring rhythmically different languages in a bilingual context. In *Proceedings of the XVIth International Congress of the Phonetic Sciences*, pages 1545–1549, Saarbrücken, Germany.

Low, E. L., Grabe, E., and Nolan, F. (1999). A contrastive study of prosody and lexical stress placement in Singapore English and British English. *Language and Speech*, 42(1):39–56.

Low, E. L., Grabe, E., and Nolan, F. (2000). Quantitative characteristics of speech rhythm: Syllable timing in Singapore English. *Language and Speech*, 43(4):377–401.

Marbe, K. (1904). *Über den Rhythmus der Prosa*. Ricker'sche Verlagsbuchhandlung: Leipzig.

Marcus, S. (1981). Acoustic determinants of perceptual center (p-center) location. *Perception and Psychophysics*, 30:247–256.

Martin, J. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behaviour. *Psychological Review*, 79(6):487–509.

Martin, J. (1979). Rhythmic and segmental perception are not independent. *Journal of the Acoustical Society of America*, 5:1286–1297.

Martin, P. (2002). Regional variations of sentence intonation in french: The continuation contour in parisian french. In *Proceedings of Speech Prosody*, pages 483–486, Nara, Japan.

Max, L. and Yudman, E. M. (2003). Accuracy and variability of isochronous rhythmic timing across motor systems in stuttering versus nonstuttering individuals. *Journal of Speech, Language and Hearing Research*, 46:146–163.

McAngus Todd, N. P. (1994). The auditory "primal sketch": A multiscale model of rhythmic grouping. *Journal of New Music Research*, 23(1):25–70.

McAngus Todd, N. P. and Brown, G. J. (1996). Visualization of rhythm and meter. *Artificial Intelligence Review*, 10:253–73.

McAuley, J. (1993). Learning to perceive and produce rhythmic patterns in an artificial neural network. Technical report, Department of Computer Science, Indiana University.

McAuley, J. D. (1995). *Perception of Time as Phase: Towards and Adaptive Oscillator Model of Rhythmic Processing*. PhD thesis, Indiana University.

Mengel, A. (2000). *Deutscher Wortakzent: Symbole, Signale*. Books on Demand.

Meyer, E. (1898). *Die neueren Sprachen*, volume 6, chapter Die Silbe, pages 479–493.

Meyer, E. (1903). *Englische Lautdauer: Eine experimentalphonetische Untersuchung*. Harrasowitz: Uppsala, Leipzig.

Mixdorff, H. and Widera, C. (2001). Perceived prominence in terms of a linguistically motivated quantitative intonation model. In *Proceedings of EUROSPEECH 2001*, Aalborg, Denmark.

Mooshammer, C. and Harrington, J. (2005). Linguistic prominence and loudness: a systematic comparison between lexical word stress, sentence accent and vocal effort. In *Workshop Between Stress and Tone (BeST)*, Leiden, The Netherlands. Abstract of Talk.

Morton, J., Martin, S., and Frankish, C. (1976). Perceptual centers (p-centers). *Psychological Review*, 83:405–408.

Möbius, B. (1993). *Ein quantitatives Modell der deutschen Intonation. Analyse und Synthese von Grundfrequenzverläufen*, volume 305 of *Linguistische Arbeiten*. Tübingen, Niemeyer.

Möbius, B. and van Santen, J. P. (1996). Modeling segmental duration in german text-to-speech synthesis. In *Proceedings of the International Conference on Spoken Language Processing*, volume 4, pages 2395–2398, Philadelphia, PA.

Nazzi, T., Bertoncini, J., and Mehler, J. (1998). Language discrimination by newborns: towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3):756–766.

Neisser, U. (1974). *Kognitive Psychologie*. Stuttgart: Klett.

Nespor, M. (1990). On the rhythm parameter in phonology. In Roca, I., editor, *Logical Issues in Language Acquisition*, pages 157–175. Dordrecht, Foris.

Nespor, M. and Vogel, I. (1986). *Prosodic Phonology*. Dordrecht, Foris Publications.

Nidermayer, D. (2003). An introduction to Bayesian networks and their contemporary applications.http://www.niedermayer.ca/papers/bayesian/bayes.html.

Noel, P. (2006). Integrating quantitative meter in non-quantitative metrical systems: The rise and fall of the german hexameter. *Metrica*, 1:1–12.

Nooteboom, S. (1998). The prosody of speech: Melody and rhythm. In Hardcastle, W. J. and Laver, J., editors, *Handbook of Phonetic Sciences*, Blackwell Handbooks in Linguistics, pages 641–673. Blackwell, Oxford.

Näätänen, R. (1992). *Attention and Brain Function*. Hillsdale, NJ: Erlbaum.

Nöth, E., Batliner, A., Kuhn, T., and Stallwitz, G. (1991). Intensity as a predictor of focal accent. In *Proceedings of the XIIth International Congress of the Phonetic Sciences*, pages 403–406, Aix en Provence, France.

O'Dell, M., Lennes, M., Werner, S., and Nieminen, T. (2007). Looking for rhythms in conversational speech. In *Proceedings of the XVIth International Congress of the Phonetic Sciences*, pages 1201–1204. ID 1382.

O'Dell, M. and Nieminen, T. (2001). Speech rhythms as cyclical activity. In Ojala, S. and Tuomainen, J., editors, *Papers from the 21st Meeting of Finnish Phoneticians*, pages 159–168, Turku.

O'Dell, M. L. and Nieminen, T. (1999). Coupled oscillator model of speech rhythm. In *Proceedings of the XIVth International Congress of the Phonetic Sciences*, volume 2, pages 1075–1078, San Francisco, USA.

Oerter, R. and Bruhn, H. (1998). Musizieren. In Bruhn, H. and Rösing, H., editors, *Musikwissenschaften. Ein Grundkurs*, page 330–348. Reinbek: Rowohlt.

Opitz, M. (1624). *Buch von der Deutschen Poeterey*. David Müller, Breslau.

Ota, M., Ladd, D., and Tsuchiya, M. (2003). Effects of foot structure on mora duration in japanese? In *Proceedings of the 15th International Conference of the Phonetic Sciences*, pages 459–563, Barcelona, Spain.

Paradis, C. and Deshaies, D. (1990). Rules of stress assignment in Quebec French. *Language Variation and Change*, 2:135–154.

Patel, A., Iverson, J., and Rosenberg, J. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America*, 119(5):3034–3074.

Patel, A., Löfqvist, A., and Naito, W. (1999). The acoustics and kinematics of regularly-timed speech: A database and method for the study of the p-center problem. In *Proceedings of the 14th International Congress of the Phonetic Sciences*, volume 1, pages 405–408.

Patel, A. D. and Daniele, J. R. (2003a). An empirical comparison of rhythm in language and music. *Cognition*, 87(1):B35–B45.

Patel, A. D. and Daniele, J. R. (2003b). Stress-timed vs. syllable-timed music? a comment on huron and ollen (2003). *Music Perception*, 21(2):273–276.

Peper, C., Beek, P., and van Wieringen, P. (1995). Frequency induced transitions in bimanual tapping. *Biological Cybernetics*, 73:301–309.

Peperkamp, S. and Dupoux, E. (2002). A typological study of stress d́eafness.́ In Gussenhoven, C. and Warner, N., editors, *Papers in Laboratory Phonology*, volume VII, pages 203–240. Berlin, Mouton de Gruyter.

Peperkamp, S., Dupoux, E., and Galles, N. S. (1999). Perception of stress by french, spanish and bilingual subjects. In *Proceedings of EUROSPEECH*, volume 6, pages 2683–2686.

Peters, B. (2006). *Form und Funktion prosodischer Grenzen im Gespräch*. PhD thesis, Universität Kiel.

Pfitzinger, H. (1999). Local speech rate detection in German speech. In *Proceedings of the XIVth International Congress of the Phonetic Sciences*, volume 2, pages 893–896, San Francisco, USA.

Pfitzinger, H. R. (2001). Phonetische analyse der sprechgeschwindigkeit. *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der Universität München*, 38:117–264. (also doctoral thesis at the Universität München).

Pierrehumbert, J. and Hirschberg, J. (1990). The meaning of intonation contours for the interpretation of discourse. In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in Communication*, pages 271–311. Cambridge, Mass.: MIT Press.

Pike, K. (1945). *The Intonation of American English*. University of Michigan Press, Ann Arbor.

Plank, T. (2005). *Auditive Unterscheidung von zeitlichen Lautheitsprofilen*. PhD thesis, Universität Regensburg.

Pointon, G. (1980). Is spanish really syllable-timed. *Journal of Phonetics*, 8:293–394.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the p-center phenomenon. *Journal of Phonetics*, 17:175–192.

Pompino-Marschall, B. (1990). *Die Silbenprosodie: Ein elementarer Aspekt der Wahrnehmung von Sprachrhythmus und Sprechtempo*. Tübingen: Max Niemeyer Verlag.

Pompino-Marschall, B. (1995). *Einführung in die Phonetik*. de Gruyter Studienbuch. Berlin, New York, Walter de Gruyter.

Port, R. (2003). Meter and speech. *Journal of Phonetics*, 31:599–611.

Port, R., Cummins, F., and Gasser, M. (1995). A dynamic approach to rhythm in language: Toward a temporal phonology. Technical report, Indiana University Cognitive Science Program, Bloomington, IN.

Port, R. F. (1990). Representation and recognition of temporal patterns. 2(1-2):151–76.

Portele, T. and Heuft, B. (1997). Towards a prominence-based speech synthesis system. *Speech Communication*, 21:61–72.

Povel, D.-J. and Essens, P. (1985). Perception of temporal patterns. *Music Perception*, 2(4):411—440.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction in generative grammar. Technical Report 2, Rutgers University Center for Cognitive Sciences.

Ptok, M. (2006). Stottern - Pathogenese und Therapie. *Deutsches Ärzteblatt*, 103(18):A 1216–1221.

Pérez, P. E. (1997). *Consonant Duration and Stress Effects on the P-Centers of English Disyllables*. PhD thesis, University of Arizona.

Pöppel, E. (1990). Unmusikalische Grenzüberschreitungen? *Universitätsforum*, 5:105–124.

Pöppel, E. (1994). Temporal mechanisms in perception. *International Review of Neurobiology*, 37:185–202.

Pöppel, E. (1997). A hierarchical model of temporal perception. *Trends in Cognitive Sciences*, 1(2):56–61.

Quené, H. (2007). On the just noticable difference for tempo in speech. *Journal of Phonetics*, 14(2):149–158.

Quené, H. and Port, R. (2005). Effects of timing regularity and metrical expectancy on spoken word perception. *Phonetica*, 61(1):1–13.

Rammsmayer, T. (2000). Zeitwahrnehmung und Rhythmuswahrnehmung. In und G. Aschersleben, K. M., editor, *Rhythmus - ein interdisziplinäres Handbuch*, pages 83–106. Bern: Huber.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart rhythmic and intonational cues. *Annual Review of Language Acquisition*, 2:85–115.

Ramus, F., Dupoux, E., and Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *Proceedings of the 15 ICPhS, Barcelona*.

Ramus, F. and Mehler, J. (1999). Language identification with suprasegmental cues. *Journal of the Acoustical Society of America*, 105:512–521.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292.

Rao, S., Mayer, A., and Harrington, D. (2001). The evoluation of brain activation during temporal processing. *Nature Neuroscience*, 4:317–323.

Rapp, K. (1971). A study of syllabic timing. *Speech Transmission Laboratory, Royal Instit. Technology Quarterly Status and Progress Report*, 1:14–19.

Rasch, R. (1988). Timing and synchronisation in ensemble performance. In Sloboda, J. A., editor, *Generative Processes in Music: The Psychology of Performance, Improvisation and Composition*, pages 70–90. Clarendon Press, Oxford.

Reyelt, M., Grice, M., Benzmüller, R., and Mayer, J. (1996). Prosodische Etikettierung des Deutschen mit ToBI. In Gibbon, D., editor, *Natural Language and Speech Technology. Results of the 3rd KONVENS conference*, pages 144–155, Berlin, Mouton de Gruyter.

Rietveld, A. and Gussenhoven, C. (1987). Perceived speech rate and intonation. *Language and Speech*, 15:273–285.

Rincoff, R., Hauser, M., Tsao, F., Spaepen, G., Ramus, F., and Mehler, J. (2005). The role of speech rhythm in language discrimination: further tests with a non-human primate. *Developmental Science*, 8(1):26–35.

Riper, C. V. (1986). *The treatment of stuttering*. Englewood Cliffs, Prentice Hall.

Roach, P. (1982). On the distinction between 'stress timed' and 'syllable timed' languages. In Crystal, D., editor, *Linguistic controversies, Essays in linguistic theory and practice*, pages 73–79. London: Edward Arnold.

Rooth, M. (1996). Focus. In Lappin, S., editor, *The Handbook of Contemporary Semantic Theory*, pages 271–297. London:Blackwell.

Rosen, S. and Howen, P. (1987). Is there a natural sensitivity at 20ms in relative tone-onset time continua? In *The psychophysics of speech perception*, pages 199–209. Boston: Martinus Nijhoff.

Rouas, J. (2005). Modeling long and short-term prosody for language identification. In *Proceedings of INTERSPEECH*, pages 2257–2260, Lisbon, Portugal.

Rouas, J., Farinas, J., and Pellegrino, F. (2003). Automatic modelling of rhythm and intonation for language identification. In *Proceedings of the 15th International Congress of the Phonetic Sciences*, pages 567–570, Barcelona, Spain.

Sagisaka, Y. (1999). Nihongo onin no jukanchou seigyo to chikaku (Japanese sound duration control and perception). *Gengo*, 28:51–56.

Saito, S. and Ishio, A. (1998). Rhythmic information in working memory: effects of concurrent articulation on reproduction of rhythms. *Japanese Psychological Research*, 40(1):10–18.

Sasaki, T., Suetomi, D., Nakajima, Y., and ten Hoopen, G. (2002). Time-shrinking, its propagation, and gestalt principles. *Perception and Psychophysics*, 64(6):919–931.

Schmitz, H.-C. and Wagner, P. (2006). Experiments on accentuation and focus projection. IKP-Arbeitsbericht NF 21, Universität Bonn.

Schneider, K. and Möbius, B. (2007). Word stress correlates in spontaneous child-directed speech in German. In *Proceedings of INTERSPEECH*, pages 1394–1397, Antwerp, Belgium.

Schreuder, M. (2006). *Prosodic Processes in Language and Music*. PhD thesis, Groningen, The Netherlands.

Schuller, B. and Rigoll, G. (2006). Timing levels in segment based speech emotion recognition. In *Proceedings of INTERSPEECH - ICSLP*, pages 1695–Wed2BuP.8.

Schulze, H.-H. (1989). The perception of temporal deviations in isochronic patterns. 45(4):291–6.

Scott, D., Isard, S., and de Boysson-Bardies, B. (1985). On the measurement of rhythmic irregularity. *Journal of Phonetics*, 13:155–162.

Scott, S. (1993). *P-centers in Speech: An Acoustic Analsis*. PhD thesis, University College London.

Selkirk, E. (1982). The syllable. In van der Hulst, H. and Smith, N., editors, *The Structure of Phonological Representations (Part II)*, pages 337—383. Dordrecht: Foris.

Selkirk, E. (1984). *Phonology and Syntax. The Relation between Sound and Structure*. Cambridge University Press: Cambridge, Mass.

Selkirk, E. O. (1995). Sentence prosody: Intonation, stress and phrasing. In Goldsmith, J., editor, *Handbook of Phonological Theory*. Oxford.

Shen, Y. and Peterson, G. (1962). Isochronism in English. *Studies in Linguistics, Occasional Papers*, 9:1–36. University of Buffalo.

Silipo, R. and Greenberg, S. (1999). Automatic transcription of prosodic stress for spontaneous english discourse. In *Proceedings of ICPhS-99*, San Francisco, CA, USA.

Silipo, R. and Greenberg, S. (2000). Prosodic stress revisited: Reassessing the role of fundamental frequency. In *Proceedings of the NISP Speech Transcription Workshop*, College Park, MD.

Sluijter, A. and van Heuven, V. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100:2471–2485.

Spitznagel, A. (2000). Geschichte der psychologischen Rhythmusforschung. In Katharina Müller, G. A., editor, *Rhythmus - Ein interdisziplinäres Handbuch*. Bern, Göttingen, Toronto, Seattle: Verlag Hans Huber.

Stachowiak, H. (1973). *Allgemeine Modelltheorie*. Wien: Springer.

Steiner, I. (2003a). On the analysis of rhythm through acoustic parameters. Master's thesis, Universität Bonn.

Steiner, I. (2003b). A refined acoustic analysis of speech rhythm. In *Proceedings of LingColl*. Budapest, Hungary.

Stockmal, V., Markus, D., and Bond, D. (2005). Measures of native and non-native rhythm in a quantity language. *Language and Speech*, 48:55–63.

Streefkerk, B. (2002). *Prominence. Acoustic and lexical/syntactic correlates*. PhD thesis, Universiteit van Amsterdam, The Netherlands.

Swanson, L. A. and Leonard, L. B. (1991). Vowel duration in mothers' speech to young children. *Journal of Speech and Hearing Research*, 35:617–625.

Tajima, K. (1998). *Speech Rhythm in English and Japanese. Experiments in Speech Cycling*. PhD thesis, Indiana University, Bloomington, Indiana, USA.

Tajima, K., Zawaydeh, B., and Kitahara, M. (1999). A comparative study of speech rhythm in Arabic, English and Japanese. In *Proceedings of the XIVth International Congress of the Phonetic Sciences*, San Francisco, USA.

Tamburini, F. (2006). Reliable prominence identification in english spontaneous speech. In *Proceedings of Speech Prosody 2006*, pages PS1–9–19, Dresden, Germany.

Tamburini, F. and Wagner, P. (2007). On automatic prominence detection for german. In *Proceedings of Interspeech 2007*, pages 1809–1812, Antwerp, Belgium.

Timoshenko, E. and Höge, H. (2007). Using speech rhythm for acoustic language identification. In *Proceedings of the International Conference in Speech Signal Processing*, pages 182–185, Antwerp, Belgium.

Toro, J., Trobalon, J., and Sebastian-Galles, N. (2003). The use of prosodic cues in language discrimination tasks by rats. *Animal Cognition*, 6(2):131–136.

Traber, C., Huber, K., Nedir, K., Pfister, B., Keller, E., and Zellner, B. (1999). From multilingual to polyglott speech synthesis. In *Proceedings of EUROSPEECH*, pages 835–838.

Trouvain, J. (2004). *Tempo Variation in speech production: Implications for speech synthesis*, volume 8 of *PHONUS*. Saarbrücken: Institute of Phonetics, Saarland University. also doctoral dissertation.

Tuller, B. and Fowler, C. (1980). Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, 27:277–283.

Turk, A., Juczyk, P., and Gerken, L. (1995). Do English-learning infants use syllable weight to determine stress? *Language and Speech*, 38(2):143—158.

Uldall, E. (1971). Isochronous stresses in R.P. In Hammerich, L., editor, *Form and Substance*, 205-210. Copenhagen: Akademisk Forlag.

Underhill, A. (1994). *Sound Foundations*. London, Heineman.

Vaissière, J. (2002). *Problems and Methods in Experimental Linguistics*, chapter Cross-linguistic prosodic transcription: French vs. English.

van der Hulst, H. (1984). *Syllable Structure and Stress in Dutch*. Dordrecht: Foris.

van Dommelen, W. (2006). Quantification of speech rhythm in Norwegian as a second language. *Working Papers of the Department of Linguistics and Phonetics, Lund University*, 52:33–36.

van Donselaar, W., Koster, M., and Cutler, A. (2005). Exploring the role of lexical stress in lexical recognition. *Quarterly Journal of Experimental Psychology*, 58A:251–273.

van Heuven, V. and Menert, L. (1996). Why stress position bias? 100(4):2439–2451.

van Santen, J. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer, Speech and Language*, 8:95–128.

Vayra, M., Avesani, C., and Fowler, C. (1983). Patterns of temporal compression in spoken Italian. In *Abstracts of the Tenth International Congress on the Phonetic Sciences*, pages 541–546.

Vennemann, T. (1995). Der Zusammenbruch der Quantität im Spätmittelalter und sein Einfluss auf die Metrik. *Amsterdamer Beiträge zur Älteren Germanistik*, 42:185–223.

Volín, J. and Skarnitzl, R. (2007). Temporal downtrends in Czech read speech. In *Proceedings of Interspeech*, pages 442–446, Antwerp, Belgium.

von Essen, O. (1956). *Grundzüge der hochdeutschen Satzintonation*. Ratingen, Düsseldorf: A. Henn-Verlag.

von Holst, E. (1937). Vom wesen der Ordnung im Zentralnervensystem. *Naturwissenschaften*, 25:625–631;641–647.

Vorberg, D. and Wing, D. (1994). Modelle für Variabilität und Abhängigkeit bei der zeitlichen Steuerung. In Heuer, H. and Keele, W., editors, *Psychomotorik*, volume 3 of *Enzyklopädie der Psychologie, Kognition*, pages 223–320. Göttingen: Hogrefe.

Wachsmuth, I. (1999). Communicative rhythm in gesture and speech. *Lecture Notes in Computer Science*, 1739:277ff.

Wachtel, S. (2000). *Schreiben fürs Hören: Trainingstexte, Regeln und Methoden*. Konstanz: UVK Medien.

Wagenknecht, C. (1999). *Deutsche Metrik. Eine historische Einführung*. München, Beck, 4 edition.

Wagner, P. (1999). The synthesis of German contrastive focus. In *Proceedings of the XIVth International Congress of the Phonetic Sciences*, San Francisco, USA.

Wagner, P. (2000). Rhythmic alternations in german read speech. In *Proceedings of Prosody*, pages 237–245, Poznan.

Wagner, P. (2002). *Vorhersage und Wahrnehmung deutscher Betonungsmuster*. PhD thesis, Universität Bonn.

Wagner, P. (2005). Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates. In *Proceedings of Interspeech*, pages 2381–2384, Lisbon, Portugal.

Wagner, P. (2006). Vizualization of speech rhythm. In *Proceedings of the Workshop in Speech Signal Annotation, Processing and Synthesis*, Poznan, Poland.

Wagner, P. (2007). Visualizing levels of rhythmic organization. In *Proceedings of ICPhS 2007*, Saarbrücken.

Wagner, P., Breuer, S., and Stöber, K. (2000). Automatische Prominenzetikettierung einer Datenbank für die korpusbasierte Sprachsynthese. In *Fortschritte der Akustik, DAGA*, Oldenburg, Germany.

Wagner, P. and Dellwo, V. (2004). Introducing yard (yet another rhythm determination) and re-introducing isochrony to rhythm research. In *Proceedings of Speech Prosody*, pages 227–230.

Wagner, P. and Dellwo, V. (2005). Objective measures for the influence of L1 on L2 speech rhythm. Unpulished Ms.

Wagner, P. and Paulson, M. (2006). Stress patterns of complex German cardinal numbers. In *Proceedings of Speech Prosody*.

Warner, N. and Arai, T. (2001a). Japanese mora-timing: a review. *Phonetica*, 58(1/2):1–25.

Warner, N. and Arai, T. (2001b). The role of the mora in the timing of spontaneous japanese speech. *Journal of the Acoustical Society of America*, 109:1144–1156.

Warren, R. (1999). *Auditory Perception: A new analysis and synthesis*. Cambridge: University Press.

Weardon, J. H. (2004). Applying the scalar timing model to human time psychology: Progress and challenges. In Helfrich, H., editor, *Time and Mind II: Information processing perspectives*, pages 21–39. Göttingen: Hogrefe & Huber.

Weinert, S. (2000). Sprach- und Gedächtnisprobleme dysphasisch-sprachgestörter Kinder: Sind rhythmisch-prosodische Defizite die Ursache? In und G. Aschersleben, K. M., editor, *Rhythmus - ein interdisziplinäres Handbuch*, pages 255–283. Bern: Huber.

Wenk, B. and Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, 10(2):193–216.

Wenk, B. J. (1982). Speech patterns in time: The French (rhythmic) connection. *International review of the aesthetics and sociology of music*, 13(2):191.

White, L. and Mattys, S. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35(4):501–522.

White, L., Mattys, S., Series, L., and Gage, S. (2007). Rhythm metrics predict rhythm discrimination. In *Proceedings of the XVIth International Conference of the Phonetic Sciences*, pages 1009–1012, Saarbrücken, Germany.

Whitworth, N. (2002). Speech rhythm production in three German-English bilingual families. *Leeds Working Papers in Linguistics and Phonetics*, 9:175–205.

Wightman, C., Shattuck-Hufnagel, S., and Ostendorf, M. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, 91:1707–1717.

Wing, A. and Kristofferson, A. (1973). Response delays and the timing of discrete motor responses. *Perception and Psychophysics*, 14:5–12.

Wittmann, M. and Pöppel, E. (1999/2000). Temporal mechanisms of the brain as fundamentals of communication - with special reference to music perception and performance. *Musicae Scientiae*, pages 13–28.

Woodrow, H. (1951). Time perception. In Stevens, S. S., editor, *Handbook of Experimental Psychology*, chapter 32, pages 1224–36. Wiley and Sons, New York.

Wundt, W. (1911). *Grundzüge der psychologischen Physiologie*, volume 3. Leipzig: Wilhelm Engelmann.

Yabe, H., Tervaniemi, M., Sinkkonen, J., Huotilainen, M., Ilmoniemi, R., and Näätänen, R. (1998). Temporal window of integration of auditory information in the human brain. *Psychophysiology*, 35:615–619.

Zec, D. (1995). Sonority constraints on syllable structure. *Phonology*, 12:85–129.

Zwicker, E. (1982). *Psychoakustik*. Heidelberg: Springer.

Zwicker, E. and Fastl, H. (1999). *Psychoacoustics: Facts and models*. Berlin: Springer.