Zofia Malisz

Speech rhythm variability in Polish and English: A study of interaction between rhythmic levels

Rozprawa doktorska napisana
na Wydziale Anglistyki
Uniwersytetu Adama Mickiewicza w Poznaniu
pod kierunkiem prof. dr hab. Katarzyny Dziubalskiej-Kołaczyk

Poznań, 2013

Contents

List of Figures

List of Tables

Acknowledgements

First of all, I would like to thank my supervisor Prof. Katarzyna Dziubalska-Kołaczyk. Prof. Dziubalska-Kołaczyk not only helped and advised me on this dissertation, but also supported and followed my (slow) progress of becoming a linguist thus far, even long after I had left the School of English to seize opportunities created with her encouragement. Over the years, she has been an important inspiration to me, as an exceptional scholar, teacher, and leader.

I am indebted to Prof. Petra Wagner at Bielefeld University for her patience, trust, belief in me, for advice and very real support in the final stages. For essential discussions on ideas, directions and particulars of this dissertation and for joint work in a highly creative atmosphere.

I would also like to thank Prof. Maciej Karpiński, from whom I learnt a lot working under his supervision on several prosodic and multimodal dialogue projects, for his advice and kindness. I also thank Dr Ewa Jarmołowicz-Nowikow and Dr Konrad Juszczyk, my partners in these projects, and other colleagues from the Institute of Linguistics, Dr Katarzyna Klessa in particular. Many thanks also to Dr Marzena Żygis at Humboldt University and ZAS, Berlin for the opportunity to work together.

I have benefited a lot from my Erasmus and research grant stays in Bielefeld and Pisa, respectively. Prof. Dafydd Gibbon's clear views on the entirety of speech rhythm research, original and structured, have been a great source of knowledge. Similarly, Prof. Pier Marco Bertinetto, by kindly hosting me at his lab at Scuola Normale Superiore, enabled me to organise my thoughts and methodology at the early stages. I gained a lot from discussions with him.

Thanks to Plinio Barbosa, Michael O'Dell and colleagues, who have offered the speech rhythm community exciting options to explore in recent years, I thank them for ideas and crucial pointers.

Also many thanks go to all my colleagues at the Phonetics and Phonology Group in Bielefeld: Marcin Włodarczak for readiness, assistance and advice on R and LaTeX, and for contrarian banter. Juraj Šimko for discussions on coupled oscillator models and timing. Andreas, Barbara, Joanna at C6 for company and discussions. Hendrik Buschmeier, Spyros Kousidis and Benjamin Inden for educating team work and good spirits.

That said, all mistakes made in this dissertation are entirely mine.

Finally, many thanks go to the friends in Poznań, Bielefeld, Italy and elsewhere: Kamila, Gosia, Agnieszka, Asia, Paweł, Ania, Michał; the whole Department of Contemporary English

at the Faculty of English, AMU. Andrea and the Bono family. Per, Carlo, Veronique and Paul, Shabnam, Saeed. Saada. Asia, Florian and Jonathan.

Ralf, ich danke dir.

I was blessed with a large family. I would like to thank all of you. Szczególne podziękowania dla Mamy, Taty i siostry Marty.

Pracę dedykuję rodzinom Czarkowskich i Maliszów.

Chapter 1: Timing, duration, meter and rhythm

1.1 Introduction

The concepts of timing, duration, meter and rhythm are defined as related to the scope and purposes of this dissertation. The definitions, theoretical concepts and methodological guidelines discussed in the present chapter are used in a subsequent attempt to resolve the controversy concerning the rhythmic type of Polish, as compared to English (from accounts in literature). An experiment on spontaneous speech data in Polish is conducted using a coupled oscillator model of speech rhythm variability. The rationale behind the decision *not* to use indices of rhythmic variability based on vocalic and consonantal stretches, also known as "rhythm metrics", for this purpose, is provided as well. Subsequently, a coupled oscillator model is introduced along with terminology related to coordination dynamics, a discipline from which the model draws its formal and theoretical shape. In the final chapter, an analysis of two detailed phonetic contexts of the "voicing effect" and the "geminate effect" on preceding vowel duration is conducted. These experiments are undertaken in order to test the hypothesis of a duration balancing effect exerted by the vowel-to-vowel cycle on the constituent segments. This hypothesis underlies some assumptions of coupled oscillator models of speech rhythm which state that the vocalic cycle tends to regularise its period.

1.2 Timing and duration

The notion of "timing" in the context of speech has been used with a variety of meanings. A definition of timing is found in Kohler (2003):

> The unfolding over time of physical parameters in speech production and their

> transformation into temporal patterns in speech perception under linguistic and communicative conditions is what we refer to as timing. (Kohler 2003: 8)

Here, by reformulating and expanding the above, a similar definition and scope is proposed:

> Timing suggests the involvement of at least two activities/events unfolding in time that are temporally coordinated (organised in time and place) with each other. The study of timing describes how articulatory displacements in space are coordinated in time in speech production. The study of timing also describes how and what temporal information is structured in speech perception.

The term timing tends to be approached differently depending on the phonetician's specific goals and may be influenced by standard methodologies as used in the subfields. If the field of phonetic studies is divided into the study of speech behaviour (speech production and perception) on the one hand, and the study of the acoustic signal on the other hand, then the treatment of what counts as timing will be different. Timing has been often equated simply with duration. While inspecting a spectrogram, it is often quite easy to distinguish boundaries and patterns. A phonetician might be therefore biased to look for segments, points and boundaries that form strings of discrete intervals on a single line. Whereas the measurement of duration in the speech signal is only one of many methods of accessing the characteristics of speech timing, it often happens that a tool is equated with the real phenomenon. In general it is common to imagine time as a line extending in space or a ruler with which *distances* between events can be measured with precision (Grondin 2010). The problem with such an abstraction is that often the measured intervals are taken to be delimiting spaces between actual events. In other words, an abstraction becomes concrete and is treated as such, e.g. as real also for the speakers and hearers. However even evidence of regularities, patterns and clear boundaries in the acoustics does not always correspond to real entities. Surface timing patterns may be successfully exploited in, e.g. speech technology, traditionally more concerned with the modeling of the signal itself[1]. In speech behaviour studies however, it is essential to show that measured durations unequivocably relate to real timing phenomena, in other words, to prove they

[1]Compare also "data-driven" and "theory-driven" approaches to temporal modeling in different linguistic subdisciplines and speech technology, a distinction discussed in Gibbon (2006).

are grounded in perceptual and/or articulatory processes of interest. Thus, when speaking of timing in this work, priority will be given to identifying events and activities as they happen in real time first, and subsequently to the *traces* of these events measured as durations from the signal. Considering just the latter might be misleading.

The awareness of the non-linear relation between duration and timing is important because of the evidenced indirect relationships between what is acoustically measured and what is actually produced or perceived. For example, the discovery of the p-centre (Morton et al. 1976), that is the point of perception of a syllable different than its acoustic onset, provides one of many cases where a relation between an acoustic and a perceptual landmark is not identical. Regarding time specifically, the perception of the flow of time is essentially subjective: prospective and retrospective estimation of, e.g. waiting times depends on how many events engaging our attention and memory were experienced during that time (Grondin 2010; Wittmann 1999). Other work on explicit duration estimation, within much smaller time ranges, reveals subjective judgements of duration relations. Sasaki et al. (2002) investigated how short intervals (less than 250 msec) cause the following interval duration, in fact up to 100 msec longer, to be underestimated by subjects in perception. The effect percolates even to the overnext, third interval. The effect is known as "time shrinking" (Nakajima et al. 1992). Such effects need to be taken into account when designing studies involving explicit, *post hoc* estimation of duration. However, online, it was found that human perception is very sensitive to relationships between durations. Literature on rhythmic perception and responding (e.g. Fraisse (1963); Martin (1972)) shows that relations between event durations are preserved in human perception despite changes in absolute durations, e.g. due to tempo changes. However, as e.g. O'Dell (2003) points out, a perceptual phenomenon correspondent with the physical measurement of duration in miliseconds is rarely discussed and duration is often used indiscriminately for both (as opposed to e.g. Hz for frequency vs. Mel for pitch).[2]

[2]We find the unit *dura* used to define psychoacoustic subjective time in Fastl and Zwicker (2007).

A view taken in this thesis is that time is intrinsic to events and so it is defined by events[3], rather than the opposite (Gibson 1975). Following Gibson (1975), Jones and Boltz suggest that "events define time intervals and their inherent rhythmic patternings will affect the way in which people attend to them and judge their durations" (1989: 459). Jones and Boltz (1989) review studies, also on linguistic stimuli, suggesting that experienced duration is dependent on attentional effort or arousal associated with presented information. At the same time Jones and Boltz (1989) state that speech contains temporally highly coherent events that offer structural predictability via rhythmic patterns: the predictability inherent in spoken events provide affordances for attending to what will happen in the immediate future. In the absence of high coherence of events, e.g. no rhythmic patterns, people need to attend to and organise events locally and consequently, both future event anticipation and event duration estimation require a lot of analytic effort.

It is also evident that the way time is processed relies on different mechanisms depending on the time range, specifically, below and above one second. "The processing of smaller intervals is sensory based, or benefits from some automatic processing, whereas the processing of longer intervals requires the support of cognitive resources" (Grondin 2010: 564). People also tend to segment durations longer than 1.2 seconds into smaller intervals in order to effectively process them. The up-to-1sec range is related to most relevant prosodic intervals, from syllable to foot to intonational phrase, i.e. components of the time structure characteristic of speech and, as Jones and Boltz (1989) propose, so beneficial for quick processing or "dynamic attending".

Jones and Boltz (1989) also discuss how duration judgements depend not only on acoustic duration but also e.g. on the complexity of an event. More importantly for speech, non-temporal information such as different types of accents (based on intensity and/or pitch), is crucial for organising the perceived intervals within and between events into hierarchical time structures, e.g. into prosodic levels (Jones and Boltz 1989). This view certainly implies that what should count as, for example, a relevant rhythmic interval in speech, should ideally be judged by

[3]The founder of ecological psychology J.J. Gibson famously stated "Events are perceivable but time is not" in the title of his 1975 paper (Gibson 1975). His work initiated a perspective on time perception where the environment is the main source of time structure, as it is perceived.

native speakers from the signal. Speakers integrate acoustic features such as duration, pitch and intensity, identify a rhythmic event (beat) sequence, and hence the rhythmic interval. At the same time the intervals are split over several timescales and interact with one another to build rhythmic structures.

Moving on from acoustic duration, as related to timing events in perception, to speech production, the organisation of speech production in time has also often been interpreted to be discrete, sequential, and the timing, external to the systems. In the context of linguistic theory, traditionally, timing was implemented by means of abstract, linguistic rules that determined duration in the output. Implicitly, the time dimension was seen as having little consequence in phonology. Many phonological theories in the generative tradition do not use a temporal specification for segments at all (Keating 1990; Clements 2003, 2006). This view on timing as something extrinsic to phonology necessarily involved a "translation theory" between phonological patterns such as phonemic contrasts and their execution in time. By analogy with computation, algorithms mediated between the mental space of abstract structures and the physical execution subsystem to produce an output.

Issues with the representation of the temporal dimension of speech have often formed the battlefront on which paradigm shifts in phonetics and phonology have taken place (e.g. the collection of papers in (Perkell and Klatt 1986)). For example, in segmental phonetics, it has been known that, e.g. co-articulation is seamless and continuous. It has also been acknowledged that this characteristic property of speech production coordination is not always straightforwardly mirrored in the acoustic patterns. As Löfqvist (2010: 354) notes: "the obvious acoustic consequence [of co-articulation - ZM] is that a single temporal slice contains influences from several production units". An acoustic representation is a *flattened* one where the "several production units" generate events that can or cannot be easily told apart in the signal. Consequently, one of the central issues in segmental phonetics is the identity and representation of its most primitive element, the segment. Using this example, the difficulties and importance of accessing actual speech events are exemplified in the following position by Fowler (1980):

> It is surely more plausible to suppose that the concept of segment has material

> support. Its essential properties are manifest in the acoustic signal, although it may take a human perceptual system to detect that aggregate of properties as a significant collective. Scientists have not discovered those properties in the acoustic signal, but the reason they have not may be that they have looked for evidence of the wrong kind. They have looked for temporal discreteness when they should have looked for qualitative separateness among temporally overlapping events. And they have sought to discover abutting edges of segments perpendicular to the time axis when, perhaps, no such things are to be found. (Fowler 1980: 120)

The perils of overemphasising acoustic *horizontal* duration over *vertical* hierarchical event timing are evident. Duration variation is easily observable but should be treated as "observations on the output" (Ogden 1996; Browman and Goldstein 1992) of actual speech events that, according to some accounts, are overlapping abstract vocal tract gestures (Fowler 1980; Browman and Goldstein 1990; Saltzman and Byrd 2000; Saltzman et al. 2008).

Browman and Goldstein (1990) proposed an Articulatory Phonology where a perspective is offered of identifying phonological primitives " directly with cohesive patterns of movement within the vocal tract" (Browman and Goldstein 1990: 69), i.e. with articulatory gestures. It is the rendering of timing relationships that makes gestures become an attractive alternative to segments, i.e. traditionally static and atemporal entities. This is important, since again, events come with an inherent time dimension. Also, a lot of discussion has been since devoted to why and how temporal phonetic detail (beyond phonological length) belongs to the representation of phonological contrasts (Fowler 1980; Port and van Gelder 1995). A whole class of intrinsic timing models is defined by "the incorporation of temporality into the definition of the units themselves" (Byrd and Saltzman 2003: 156). As an alternative to this view, it can be hypothesised that motor control of speech is hard-wired and commands to individual articulators, including commands regarding timing, are sent each time a speech gesture is to be produced. But, first of all, as Port and Leary (2000) note, it is unlikely that speakers locally control the determined timing of constrictions and openings with a milisecond precision across a range of speech rates: "there is no existing model of motor control that could employ such specifications" (Port and Leary 2000: 12). Secondly, looking at higher levels of the timing hierarchy, e.g. stress-based, top-down influences in timing would have to be computed one step at a time, instead of globally.

Gestures, as defined in Articulatory Phonology on the other hand, "cohere in bundles corresponding, roughly, to traditional segmental descriptions, and (...) maintain their integrity in fluent speech" (Saltzman and Munhall 1989: 365). The coherence and integrity is expressed in inherently temporal terms of stable phase relationships within and between gestures. Such stable relationships suggest that gestures behave like functional units, forming *coordinative structures*, that in turn can be directly perceived, produced and learnt as phonological primitives. Assuming functional coupling between articulators, i.e. coordinative structures, allows to limit the degrees of freedom involved in potential movements, making the achievement of articulatory tasks easier and more efficient.

How can the existence of functional movement ensembles (coordinative structures) be evidenced and described? For example, by exploring the dynamics of speech gestures where both position and timing are involved. Mechanical perturbances to the jaw, as shown by Folkins and Abbs (1975), demonstrated how articulators pliantly "conspire" over time in order to reach a required gesture target. The study of *Intra*gestural timing describes how targets for particular gestures are attained, for example, how both lip and jaw peak velocities are coordinated in order to yield a voiceless bilabial plosive [p] (Gracco 1988; Bell-Berti and Harris 1981; Saltzman et al. 2000). Evidence for coordinative structures in the timing *between* gestures was also postulated to explain the production of serial gestures. The relative timing of opening and closing gestures, as in a sequence [ba ba], is maintained across speaking rates (Saltzman and Munhall 1989), as it is the case for movement effectuators in walking and chewing. All these motor activities are characterised by tight couplings between participating subsystems: in the case of walking, the limbs, in the case of speech articulation such as [ba ba], the consonantal and vocalic gestures.

In coordinative structures, individual articulators form functional synergies with one another, coordinate flexibly in order to perform a task. What is important within the context of this thesis, the principles governing these synergies are hypothesised to extend to higher prosodic levels (Barbosa 2006, 2007; Saltzman et al. 2008; O'Dell and Nieminen 2009; Tilsen 2009). Prosodic structure affects the spatial and temporal characteristics of individual gestures, as well as the relative coordination among different gestures. The subsystems involved,

the gestures, the mora, the syllable, the foot and the phrase operate on multiple timescales and interact. A reinterpretation of phonological representations in prosody in these terms was also proposed by Gibbon (2006):

> Phonetic events, including prosodic events, are time functions; their phonological representations are prosodic, distinctive and conditioned features. The time functions are defined over temporal domains of different characteristic durations and are associated with different 'clock' frequencies in speech production and perception (...). Levels in the discretely structured prosodic hierarchy (...) from phones to discourse units can be phonetically interpreted in terms of such domains. (Gibbon 2006: 182)

By redefining the traditional prosodic units as timescales (or time functions) they can be easily associated with characteristic frequencies (Tilsen 2009). Additionally, via relations between the subsystems (or frequencies) typical hierarchical structures are formed. Hierarchical structure is a "a time structure in which the temporal distribution of markers reveals nested time levels that are consistently related to one another at a given level by ratio" (Jones and Boltz 1989: 465). This means that it is not only the existence of levels and/or nesting that makes a structure hierarchical but a certain stability of the ratios. The ratios characteristic for relations in spoken events are also discussed in the section defining meter (Section 1.3). In metrically entrained production of speech, simple integer ratios (e.g. 2:1) turn out to be the most stable ones. Such hierarchical time structures with nesting and simple integer ratio relations (phase relationships) between levels are, as mentioned before, characteristic of "highly temporally coherent" event structures (Jones and Boltz 1989: 461). Tilsen (2009) calls the hierarchical nesting in speech *containment*, which means containment of feet within phrases, syllables within feet etc. Nesting also implies *coupling* between the levels, that is, the different frequencies of the subsystems are influencing each other's evolution in time. A sketch of a multiscale dynamical model of global timing proposed by Tilsen (2009) is presented in Figure 1.1. Similar models are implied by (Barbosa 2006; O'Dell and Nieminen 2009; Saltzman et al. 2008).

In summary of the present introduction to timing and duration, first of all, the definition of timing given at the start of this chapter puts focus on hierarchies and structures where at least two or more levels, and the interactions between

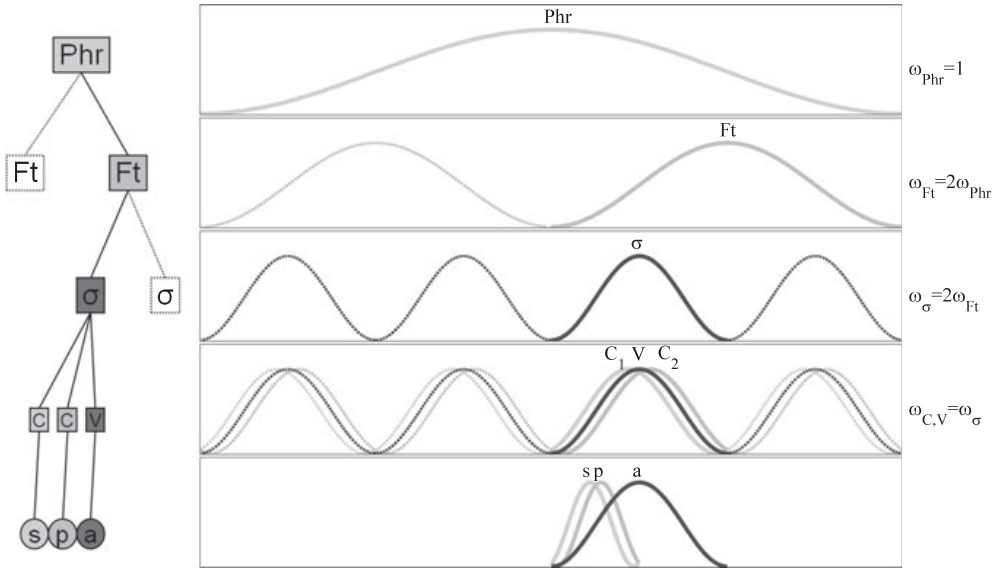


Figure 1.1: The units of the prosodic hierarchy on the left from phrase, foot, syllable to consonants, vowels and gestures. The main diagramme depicts a multitimescale model timing associated with levels of the prosodic hierarchy. The equations express the natural frequencies of the components and hierarchical nesting. Adapted from Tilsen (2009).

them, are considered, avoiding a representation of timing as duration that implies a "flat" structure. This might be argued to be just a semantic choice, however "timing as duration" suggests itself first as assuming one time scale on which speech timing operates, while it is known that it operates simultaneously on multiple time scales, in specific time ranges (or frequencies) both in perception and in production.

Secondly, indirect relationships between acoustic duration and perceived duration imply that studies of timing, be it on a gestural level or higher levels, i.e. regarding speech rhythm, should take native speaker judgement as a basis for what counts as a rhythmic interval and/or prominent event. Such an approach allows for gaining indirect access to real spoken events via convenient acoustic annotation methods that are currently available. Some consequences of analysing more or less arbitrary units to account for speech rhythm variation is touched upon in Chapter 2. As implied by Gibbon and Fernandes (2005), real events can be approximated via careful annotation of the signal. The rationale for the events and the annotation procedures is constructed according to this goal.

Thirdly, as experience from dynamical speech production models shows, the potential for functional synergies generalises from gestural detail to higher prosodic levels. The dynamical approach to speech rhythm variability is strongly connected to work on articulatory gestures and motor behaviour and will is introduced in detail in Chapter 3.

A very similar take on how surface duration relates to speech perception and production could be found recently in Turk and Shattuck-Hufnagel (2013). Their overview of speech rhythm similarly suggests to direct broadly defined speech rhythm research closer to studying what they call a "global timing profile", treated as a part of research on speech timing. To see rhythm as first and foremost, a property of speech perception and production:

> (...) "[S]peech rhythm" involves studying speech timing more generally: its control structures and processes, its perception, and the systematic relationship between phonology (both segmental and prosodic) and surface timing. This approach incorporates the wide variety of factors influencing speech timing into the ongoing search for rhythm in speech and rhythm classes for languages. It takes account of the many factors that may influence a listener's sense of the *global rhythmic profile* of a language (which we will argue might be termed the *global timing profile*, as well as a similarly wide (but not identical) variety of factors that may influence the *timing pattern of a specific utterance* in that language. (...) [E]ven though timing is not the only aspect of spoken utterance that might contribute to a listener's sense of their rhythm, we believe that an understanding of how speech timing works is a necesseary prerequisite to understanding rhythm in all its possible meanings [all emphases theirs - ZM]. (Turk and Shattuck-Hufnagel 2013: 94)

1.3 Meter

In the present thesis meter will be defined as understood in Port (2003) and Cummins and Port (1998): superficially, as patterns of integer-ratio timings such as 2:1, 3:1 etc. In Metrical Phonology for example, such metres would be represented symbolically as a metrical grid of binary, *swsw*, or ternary, *swwsww* patterns. Cummins and Port (1998) demonstrated experimentally how salient events in speech are biased towards these integer-ratios by attracting perceptual attention to the pattern and by influencing the motor system. Port (2003); Cummins and Port (1998) also postulated neurocognitive oscillators (see Section 3.1.2 for

a formal definition of an oscillator) that generate pulses to which vocalic onsets of stressed syllables are attracted. Moreover, as Port (2003) and Cummins and Port (1998) showed, speakers are able to control the extent to which the metrical patterns constrain their timing.

Evidence for *metrical attractors* (see Section 3.1.2 for a formal definition of an attractor) was given by Cummins and Port (1998) by means of the "harmonic timing effect". Subjects were presented with a phrase with two stressed syllables such as "big for a duck" as a stimulus. They were asked to repeat the phrase and align the first stressed syllable with a regularly generated tone A and the second stressed syllable with a tone B. The tone B was varied in a randomly uniform fashion between 0.2 and 0.8 target phase angles of the A-A cycle. The variable of interest was the phase of the second stressed syllable "duck" relative to the repetition cycle, as located by the subjects, compared to the target phase, which was varied as above. The results showed that native speakers of English tended to "lock" into the $1/3$, $1/2$ and $2/3$ fractions of the repetition cycle for early, mid and late target phase angles respectively. This means, as Port (2003) explains, that given one periodicity, i.e. the phrase repetition cycle, other periodicities emerge, at harmonic fractions of the lower frequency cycle. The emergent structure supports the notion that a phrase cycle is coupled to the foot cycle (cf.Jones and Boltz (1989); Tilsen (2009) above). The simple harmonic phases at which the coupling is stable, constitute attractors for the system. These phases are in fact stable also across speech rates (Cummins and Port 1998).

The concept of metrical hierarchies as the abstract "skeletal" rhythmical structure in language, an organisational structure, was the focus of much research in the Metrical Phonology framework (Prince 1983; Hayes 1984; Selkirk 1984). An empirical validation of the concept was achieved by Cummins and Port (1998) with the speech cycling tasks described above. It was also independently extended in studies within the dynamical approach to speech rhythm (Barbosa 2006; Port and Leary 2000; Port et al. 1999). The progress made with the dynamical approach accounts of metrical structure is based on the grounding of alternation and relative prominence in a real time process in speech. Thereby it merges the formal advantages of the metrical grid (serial order, sequence of alternating "beats") and the metrical tree (structure, nesting of metrical levels) as well as adds the contin-

uous, non-discrete dimension operationalised by means of relative phase, in one model.

Speech cycling tasks reveal how salient events are biased to metrical attractors that influence the motor system and guide attention. But as Tajima and Port (2003) point out, languages obviously differ with respect to the units which count as salient (prominent). And so, cross-linguistic differences in performing the speech cycling task are to be expected.

Tajima et al. (1999); Tajima and Port (2003) report on a speech cycling task with English, Japanese and Arabic speakers. They use a simplified version of the speech cycling task where the speakers repeat a phrase aligning the first stressed syllable of the phrase with consecutive beats of a metronome, at different speech rates. Each phrase exploited a different stress pattern, matched between languages. The metronome beats and the vowel onsets of the aligned stressed syllables were extracted. Relative phase was measured in two ways: one by taking the repetition cycle as reference ("external phase"), as in Cummins and Port (1998), and second, by taking the interstress interval between the first and last stressed syllable in the repetition as reference ("internal phrase"). The analysis of the external phase showed that Japanese subjects placed the phrase *final* syllable stably at the simple harmonic phase of 0.5, while Arabic and English speakers tended to produce the final *stressed* syllable there. Internal phase comparisons showed that Arabic speakers approximated the 0.5 point with higher variation than English speakers. Since in case of the internal phase measurement, the 0.5 phase is halfway between the first and third stressed syllables, it is a good approximator of isochrony between these beats. The authors conclude that English under these conditions turned out to be more "stress-timed" than Arabic, while Japanese represented a different metrical strategy altogether. The authors concluded that the intuitions regarding traditional rhythmic types received new support in the form of a constrained production task.

How do speakers of Polish behave when faced with such a task? Malisz (2005) presented preliminary results on the simplified speech cycling task as in Tajima et al. (1999), with Polish speakers. Sentences exhibiting different metrical patterns were used as stimuli, for example: *ssws* "daj Basi dom" ("give Basia a house"), *swsws* "Dosia zgubi go" ("Dosia will lose him") etc. First of all, it was

found that Polish speakers had no difficulties performing the task, i.e. they were also susceptible to the "pull" of metrical attractors in their production. Results for the external phase measurement were presented for two speakers. For the two example stimuli above, two modes, at 0.5 and 0.75 were apparent for both speakers. The $2:1$ (0.5 phase angle) metre was more often produced with the *ssws* pattern and the $3:1$ (0.75) with the *swsws* one. As a function of metronome rate, the tendency was to place the stressed syllable approximately at the half of the cycle in slow trials and two thirds as tempo increased, with both stimuli. The results suggest that Polish speakers behave similarly to English and Arabic speakers. Lack of internal phase measurements precluded Malisz (2005) from finding a more precise location of Polish rhythm relative to the two languages studied before in this paradigm.

Tajima and Port (2003) in fact proposed that the relative temporal stability of one syllable versus another in different languages, as revealed by the task, may be used as a diagnostic for rhythmic linguistic types. The types in this case would be determined by what counts as the prominent unit retaining the most stability in coupling to the metre, in this experimental paradigm.

1.4 Speech rhythm

A definition of speech rhythm that is commonly accepted in the research community is not available (Rouas et al. 2005). The notion of rhythm in general, even though intuitively graspable, may refer to a range of phenomena, e.g. isochronous sequences of monotone beeps, complex rhythms that arise from regularly occurring pitch changes (notes) that create an impression of structure, loud and quieter sounds that seem to appear and disappear in alternation. And in fact, the words "rhythm" or "rhythmic" were often used in such disparate contexts in literature. However, already from such descriptions crucial notions can be picked out, notions that constrain the idea of rhythm: regularity, alternation and the co-existence of different types of events that build a rhythmical structure. One attempt recently to define rhythm formally in the context of speech has come from Gibbon and Gut (2001):

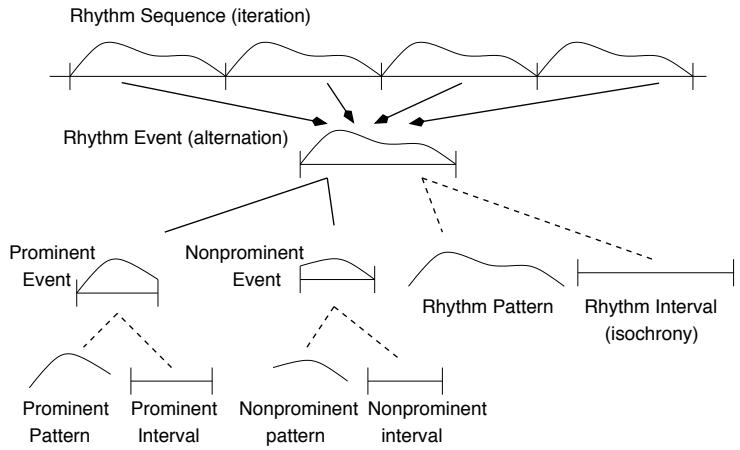


Figure 1.2: A schematic depiction of units and events that contribute to the definition of speech rhythm as proposed by Gibbon (2006). Adapted from Gibbon (2006)

> Rhythm is the recurrence of a perceivable temporal patterning of strongly marked (focal) values and weakly marked (non-focal) values of some parameter as constituents of a tendentially constant temporal domain (environment).

Gibbon and Gut (2001) distinguish between the internal temporal pattern of prominence distinctions and the external rhythmic environment here. The focal-nonfocal internal pattern is instantiated by the latter factor, the external rhythmic environment, i.e. rhythmic units, such as the syllable, the foot, etc.

The notion was further constrained in the following way in Gibbon and Gut (2001):

> Rhythm is the directional periodic iteration of a possibly hierarchical temporal pattern with constant duration and alternating strongly marked (focal, foreground) and weakly marked (non-focal, background) values of some observable parameter.

Figure 1.2 presents the schematic depiction of each of the concepts and units used in the definition above. This diagramme generalises the rhythm event from the rhythm sequence, subsequently decomposes the sequence into events, patterns and intervals with attention to overlaps, such as those between events and intervals. Regarding overlaps between events and intervals, we find a similar differentiation in Guaitella (1999), who states that notions of a "marked element" and "marking element" are often confused. In other words, according to Guaitella

(1999) prominent elements, are often likened to elements located at the boundary of some group.

The general approach to rhythm that will be pursued further, was first defined by Cummins and Port (1998). In their work, speech rhythm is treated as a case of macro-timing coordination on multiple levels, as suggested in Section 1.2. As Cummins and Port (1998) propose:

> Rhythm is viewed here as the hierarchical organization of temporally coordinated prosodic units. We show that certain salient events (beats) are constrained to occur at particular phases of an established period, and we develop an argument that the establishment of this period serves a coordinative function. This is a radical departure from the conventional treatment of rhythm in phonetics, which has been concerned primarily with the search for simple isochrony. (...) Our claim is that rhythm in speech is functionally conditioned. It emerges under just those speaking conditions in which a tight temporal coordination is required between events spanning more than one syllable. Linking disparate motor components together into a single temporal structure, or rhythm, greatly simplifies the problem of coordination among the many parts. (Cummins and Port 1998: 145)

As will be further elucidated in the course of this thesis, the view taken by Cummins and Port (1998) is derived from a) phonetic research on the duration of stress groups and/or feet and syllable compression, b) from Metrical Phonology and from c) modern developments bridging methods used in, e.g. motor coordination dynamics and speech production. With a) and b) providing comprehensive linguistic motivation and constraints on the functionalities of the resulting system, c) situates the approach in a broader context of coordination dynamics that extends to many other domains of cognitive science and provides formal tools that are generalisable to many other domains of human behaviour.

Since the specific notion of speech rhythm employed in this dissertation will be defined in stages throughout the thesis, the remainder of this section is confined to a sketch of some general approaches to speech rhythm. Detailed reviews of recent literature on this topic are located in Chapters 2 and 3.

Early studies of speech rhythm concentrated on absolute duration patterns, especially on simple ones, i.e. on potential isochrony of a prosodic unit that is dominant in a language of a given speech rhythm "type". Pike (1945) proposed a taxonomy where well-studied languages such as Spanish and English served as

examples of distinct types: the former syllable-timed, i.e. with syllable recurrence, and stress-timed, i.e. with stress recurrence. This idea has been taken further by Abercrombie (1991) who claimed French and Spanish rhythm arises from isochronous syllable sequences and English rhythm from isochronous inter-stress intervals. The general assumption was that rhythm in speech is characterised by a sequence of strictly regularly appearing events, similar to dripping water or machine gun shots (Couper-Kuhlen 1993). These events would be marking elements, delimiting the boundary of a group, an interval (Guaitella 1999). This claim of a rhythmical taxonomy in which languages were classified into rhythm types based on stress or syllable isochrony was not attested by subsequent experimental research, e.g. (Roach 1982; Lehiste 1970) (but see also Jassem et al. (1984) in Section 3.1.1).

A more prominence based approach was taken by phoneticians and phonologists who pointed out the significance of salient elements rather than succession in the construction of rhythm. It was believed that regular recurrence is not necessary, if there is even a single alternation in prominence. The representation of prosodic units related to each other in terms of salience i.e., where some must be more prominent than others, and their organisation, became the basis for Metrical Phonology. In physical terms, loudness (amplitude), pitch height, fluctuations of fundamental frequency were found by "non-temporal" phoneticians especially worth investigating as relevant for stress and prominence patterns.

Generally, the notions of alternation and succession, interval timing and prominence-based timing were competing for some time in the speech rhythm debate. Allen (1975) insisted on considering the two factors as equally important in describing speech rhythm and thus touched on a challenging issue of relating both to each other in a conjunctive account of speech rhythm. The challenge has survived until now: "Rhythm coding requires an enriched temporal, sequential representation as well as a hierarchical structure based on discrete units" (Keller and Keller 2002: 9).

Guaitella (1999) considers the distinction of the *metric* approach and the *rhythmic* approach to speech rhythm. The metric approach concentrates on assimilation and standardisation of intervals, focusing on regularity and could be associated with the philosophy behind isochrony studies. The rhythmic approach

highlights dissimilation, focuses on the contrasts between events and could be associated with the Metrical Phonology tradition. As Guaitella (1999) observes, Fraisse (1963) and Gestalt theorists would argue that the "human perceiver is torn between making elements similar and reinforcing the difference between already marked elements" (Guaitella 1999: 510); this particular statement would put Allen (1975), calling for a conjunctive account of prominence and duration, among the advocates of rhythmic and metric synergy, along with, e.g. Fraisse (1963).

More recently, we find a classification of existing speech rhythm theories in Gibbon and Fernandes (2005). Proposals to measure rhythm representation in the acoustic signal are subsumed under Physical Rhythm Theories (henceforth PRT):

> The PRT standpoint [is] that there are indeed physical cues to rhythm (by no means a necessary assumption): 1. The signal provides cues for synchronising with the constrained activities which produced it. 2. Cues to rhythmical organisation can be detected by distributional analysis of physical measurements. 3. But: careful subjective annotation approximates to a criterion for emergent phenomena. (Gibbon and Fernandes 2005: 3289)

In this way, a set of constraints within the Physical Rhythm Theory for speech rhythm representation is provided. It will prove to be very useful in the discussion of the so called "rhythm metrics" in Chapter 2 and in the evaluation of their predictive power. This set of constraints is also strongly related to the tenet according to which models and theories used in this thesis try to operate: what is measured is not in itself evidence for or against speech rhythm, patterns and events identified on the surface, need to be related to real production or perception phenomena (cf. Section1.2 and Turk and Shattuck-Hufnagel (2013)). Gibbon and Fernandes (2005) suspect that acoustic patterns corresponding to rhythmic events can eventually be found and that subjective annotation of the signal may adequately represent the events.

At the same time Gibbon and Fernandes (2005) posit Emergent Rhythm Theories (ERT) which, at the present moment, encompass dynamical rhythm models that concentrate on rhythmic speech behaviour (discussed in Chapter 3) as well as Metrical Phonology (Hayes 1984). The definition by Cummins and

Port (1998) given above is subsumed under Emergent Rhythm Theories. The ERT states that:

> Rhythm is an emergent perceptual construct based on the coordination of many different temporal activities due to the interaction of a variety of different physiological and cognitive systems.

In other words, Gibbon and Fernandes (2005) define rhythm here as an emergent phenomenon that arises from a number of factors: phonetic, phonological and discourse related. They also note that the factors have been selectively treated by researchers according to what suits their particular goal and analysed accordingly from the many possible multiple points of view by phonologists, phoneticians and speech processing engineers independently. Such a situation makes comparisons between resulting models difficult. Interestingly, it has been recently argued that even from the point of view of applications in speech technology the approach to prosodic phenomena that involves analysis and modelling of the acoustic signal does not produce the desired results, e.g. naturally sounding speech synthesisers. Xiu (2008), by putting forward an interesting hypothesis, pointed out that prosodic models based on the signal fail to capture naturalness. His hypothesis states that most of what is studied within the realm of prosody, including speech rhythm, is an epiphenomenon emerging from obligatory articulation constraints and functional information coding. Instead, he posits more attempts in speech research at modelling "the articulatory encoding of communicative functions"(Xiu 2008: 24). Such an approach actually indicates that also work in speech technology should push towards advances in what Gibbon and Fernandes (2005) call Emergent Rhythm Theory. With the difference that Xiu (2008) actually suggests to drop the notion of rhythm altogether (as an epiphenomenon). Here, we acknowledge his point of view as one that helps draw attention to the pitfalls that the search for rhythm representation in the signal has posed so far, and at the same helps encourage the study of rhythmical speech behaviour.

However, one more logical possibility exists, namely the opposite to Li Xiu's claims, where an inherent tendency towards rhythmicity (e.g. as a coordinative device) is there in the first place (Dziubalska-Kołaczyk 2002), and will only be perturbed or modified by "articulatory constraints and information coding".

Similarly eurythmy principles, as defined by Hayes (1984) in Metrical Phonology and dynamical models, the latter discussed at length in Chapter 3, provide evidence, from within and outside language, for top-down rhythmic "pressures" on speech. Such a view of rhythm as hierarchy with interactions of different nature is quite characteristic of existing phonological and dynamical models in general, as Gibbon (2006) notes.

1.4.1 Speech rhythm profiles of Polish and English [4]

Subsequent sections serve to provide some standard background on the rhythmic profiles of both languages, as a starting point to further discussion of rhythmic variability in Polish and English. Additionally, since several issues concerning the correlates of lexical and phrasal stress in Polish are apparent, a few facts and new findings about the acoustic correlates of lexical stress and prominence in Polish will be reviewed in the forthcoming sections as well.

Descriptions of Polish rhythmic strategies found in the literature are often impressionistic or inconclusive. Within the space of canonical rhythm types, Polish has been placed between stress- (Rubach and Booij 1985) and syllable-timing (Hayes and Puppel 1985) and consequently, is often described as "mixed" (Nespor 1990). English enjoys the status of a prototypical stress-timed language with a phonological structure (Dauer 1983) that conventionally corresponds to descriptions of this type: it reduces vowel duration in unstressed syllables systematically as well as possesses a complex syllable structure with complex onsets. It also lengthens stressed syllables and hence creates an acoustically clear pattern of prominences that delimit inter-stress intervals. For Polish however, some phonological characteristics suggested to correlate with distinct rhythm types (Dauer 1983) point to a mixed type: large consonant clusters and no vowel reduction. Regarding the former, however, frequencies of complex syllable types in Polish were found to be rather low: relatively simple syllables (CV, CCV, CVC, CCVC) predominated in a large corpus analysed by Klessa (2006). Regarding the latter, phonetic vowel reduction ranging from centralisation to deletion was observed by (Rubach 1974) and, especially for unstressed high vowels, by Sawicka (1995).

[4] A part of this Subsection appeared in Malisz, Żygis and Pompino-Marschall (2013).

Nowak (2006b) demonstrated the applicability of the target undershoot model Lindblom (1963) to the variability of Polish vowels, especially in the context of consonants with a palatal component ("soft" consonants). In general however, Polish seems to exhibit a "limited durational variation of [...] vowels vis-à-vis many other languages", as reiterated by Nowak (2006b: 378) and suggested previously by, e.g. Jassem (1962) and Lindblom (1963). The lack of a phonological vowel length contrast certainly contributes to this characteristic. English vowels exhibit notably more variability in duration due to a length contrast, vowel reduction and vowel duration systematically contributing to syllable prominence (cf. Kim and Cole 2005).

Global and local linear measures of segmental variability, i.e. "rhythm metrics", discussed in Chapter 2, have been used to classify languages according to the traditional rhythm taxonomy, with very limited success. Nonetheless, in a study by Ramus et al. (1999) a short text read by four Polish speakers exhibited high standard deviation of consonantal intervals (ΔC) and a low proportion of vocalic intervals (%V). Also, a very low value of the vocalic variability index (ΔV) was obtained. The combination of the above segment-based values placed Polish out of the parameter space delimited by the canonical stress-timed and syllable-timed types and motivated Ramus et al. (2003) to suggest devising a rhythmic "category of its own" for the language. English consistently clustered in the expected locations for a stress-timed language in the same studies (Ramus et al. 1999, 2003), however it is important to take all caveats regarding speech tempo, inter-speaker variability and text influences on these scores discussed in depth in Chapter 2.

Some recent studies indicate that the syllable might be the domain to look at when characterising the rhythm of Polish. Gibbon et al. (2007) study of a Polish corpus found nPVI values for syllable duration to be lower than what had been typically found for Polish segmental intervals. The result suggested a greater regularity of syllabic intervals relative to segmental ones, despite large consonant clusters admitted in the phonology. Gibbon et al. (2007) proposed that the tendency towards syllable isochrony could be accounted for by a) a Zipf effect: "large clusters are rare", or b) compensatory effects operating within the syllable domain, or c) the lack of vocalic quantity contrasts and the general "inflexibility"

of Polish vowels, both contributing to consonant-vowel ratios within syllables that are closer to unity. A tendency for uniform durations of syllables in corpus studies was also found using different methods by Wagner (2007). The above results highlight the contradictory patterns postulated in previous qualitative, phonological studies of rhythm in Polish, and it is difficult at this stage to determine its exact nature.

1.4.2 Acoustic correlates of lexical stress in Polish[5]

As shown by Domahs et al. (2012), the neurophysiological expectation of lexical stress on the penultimate syllable manifests itself in negative event-related potentials in the EEG and constitutes a very robust characteristic of rhythmical processing in speakers of Polish. Fixed, quantity insensitive stress on the penult is one of the most characteristic features of Polish prosody. However, accounts differ as far as the acoustic correlates of lexical stress in the language are concerned. Traditionally, Polish lexical stress has been described as "dynamic", that is, acoustically primarily correlated with overall intensity. It was assumed that stressed syllables are articulated with greater vocal effort and perceived as louder. It is the relative differences in loudness that define Polish stressed and unstressed syllables rather than pitch movements or duration.

Consequently, early observational studies such as Dłuska (1950) claimed that a slight rise in loudness is the primary correlate of Polish stress. Jassem (1962), however, has shown on the basis of acoustic measurements that it is in fact pitch movement that is ranked as the most salient correlate of lexical stress in Polish. His study involved spontaneous and read material, including isolated words and sentences. Dogil (1999) collected recordings of three speakers who replied to questions designed to elicit broad, narrow and no focus on a target word in a sentence. His results showed that in the position of no focus, primary stress in the target word is characterised by the highest f_0 with a sharp pitch slope. The results appear to confirm Jassem's (1962) findings. Under broad focus however, as Dogil (1999) proposes, "a position for the association with the nuclear pitch-accent morpheme of a sentence" is only "pointed to" by lexical stress. This means

[5]A version of the following subsections appeared in Malisz and Wagner (2012).

that lexical stress in Polish is best represented by a model where it is context dependent, "potential" and strongly interacts with the intonational structure of a sentence, such as the one suggested by Abercrombie (1991).

Notably, in all the above studies, duration has no or only weak influence on stressed vowels, contrary to, e.g. most Germanic languages. Jassem (1962) estimated the duration ratio of stressed to unstressed vowels at 1.17 Nowak (2006b) in a large corpus study on vowel reduction in Polish found a similar relationship of 1.22. Klessa (2006) analysis of a corpus of spontaneous and read speech built for speech synthesis purposes also quotes values that amount to a ratio of approx. 1.2. However, when vowels in prepausal syllables were excluded, the ratio in Klessa's work equals 1.1, while for English, this value equals two for monophthongs (Crystal and House 1988).

Secondary stress has received some attention and is impressionistically agreed to exist. In words longer than three syllables, secondary stress falls on the first syllable. Acoustically, Dogil (1999) showed that relatively longer duration and a fully articulated vowel characterise syllables receiving secondary stress. However, a perceptual study by Steffen-Batogowa (2000) has found no systematic evidence of secondary stress. The acoustic status of secondary stress in Polish has also has been questioned recently Newlin-Łukowicz (2012). A common process occurs, as described by Dogil (1999), where under narrow focus, primary stress shifts from the canonical penult onto the first syllable, i.e. "in Polish a single word, when under focus, switches the prominence values of primary and secondary stress" (Dogil 1999: 286).

Crosswhite (2003) showed that an acoustic measure linked to spectral tilt (the difference between the perceived loudness in phons and sound intensity level in dB), was significantly affected by stress in Polish, Macedonian and Bulgarian. The author is aware of no other studies related to the effect of the slope of the spectrum on stress in Polish.

1.4.3 Acoustic correlates of prominence in Polish

Malisz and Wagner (2012), on the basis of four dialogues from the same Polish corpus analysed in the present dissertation (see Subsections in 3.3.1 for details on

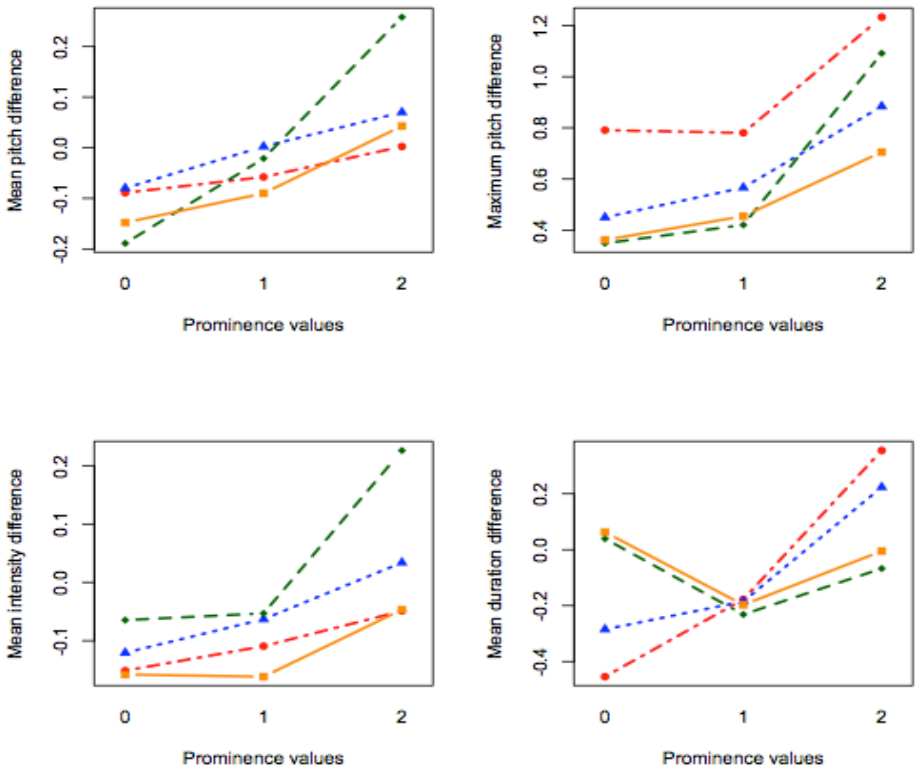


Figure 1.3: Panels clockwise from top left a) pitch difference, b) maximum pitch difference, c) mean intensity difference and d) mean duration difference values by subject for three prominence values 0: no prominence, 1: weak prominence and 2: strong prominence. Adapted from Malisz and Wagner (2012).

the corpus), studied three levels of perceptual prominence (no prominence, weak prominence, and strong prominence) and their relation to a number of acoustic features. Overall, non-prominent syllables were distinguished from all prominent ones by maximum pitch and mean intensity difference. Between weakly and strongly prominent syllables, duration was also a significant predictor. The results are presented in Figure 1.3.

The study suggests that acoustic correlates of prominence in Polish manifest themselves largely in phrase accentuation structure, as suggested by Dogil (1999), not in the lexical stress domain. Overall intensity, duration, and pitch movement are good correlates of phrase accent. Lexical stress is weakly expressed acoustically, especially in the duration dimension, a clear difference from English. Similar results can also be found in Newlin-Łukowicz (2012).

Chapter 2: Rhythm metrics[1]

A new interest in rhythm taxonomies arose when metrics for quantifying typological prosodic distinctions between languages emerged a decade ago (Ramus et al. 1999; Low et al. 2000; Grabe and Low 2002). The particular typology that constituted the basis for implementation was first formulated by Dasher and Bolinger (1982) and later elaborated upon by Dauer (1983) and Bertinetto (1988). It was based on the view that language phonotactics and syllable structure determine rhythm classes. In general, the criteria of classification in the taxonomy are defined by phonological language specific properties, such as syllable structure and vowel reduction, taken cumulatively to locate languages on a scale from syllable- to stress-timed. This way, segment durations as well as phonotactic segment configurations become the proposed determinants of rhythmic language type. The metrics proponents stressed that their main aim was to provide a phonetic account for a largely impressionistic but also possibly phonologically motivated grouping (possibly a continuum) of syllable- to stress-timing. The account would also ideally be mathematically explicit. In the most fundamental way, the metrics can be treated as "formulas that seek to quantify consonantal and vocalic variability and use this quantification to classify languages rhythmically" (Arvaniti 2009: 47). However, as they are mainly used to test and evaluate the descriptive power of the standard classification and, at the same time, the effectiveness of measurement is evaluated taking the same classification for granted, there is a serious danger of circular reasoning. Most importantly, the Dauerian "rhythm class hypothesis" has not been itself independently tested. It is also often implied, if only by the systematic use of the term "*rhythm* metrics", that the measures can be treated high dimensionally as models of speech rhythm and evaluated as such (for early criti-

[1] A preliminary version of this chapter appeared in Malisz (2006).

cisms see Gibbon and Gut (2001) or Cummins (2002)). By adopting this stance, fundamental questions need to be investigated: what exactly do the "rhythm" metrics measure? Is it speech rhythm? To what extent can the measures support and be part of a general model of rhythm? Are they sufficient to account for rhythmical phenomena in production and perception on their own? Already at their inception, the original analyses by Ramus et al. (1999) and Grabe and Low (2002) revealed inconsistencies in the results stemming from empirical problems. Given all of these issues, this chapter will first discuss the methodological caveats and the consequences of the metrics approach for the rhythmic type hypothesis. Also, and more importantly, some of the metrics will be evaluated from a theoretical point of view, in the context of adequate rhythm modelling.

Indirectly, this chapter serves to motivate the decision *not* to use rhythm metrics as indicators of speech rhythm type of Polish on theoretical and methodological grounds. The criticisms will cue in the exposition of what Gibbon and Fernandes (2005) call Emergent Rhythm Models and the implementation of one of the models with Polish spontaneous speech data in Experiment 1.

2.1 Methodological problems

Seeking a phonetic validation for the phonologically based standard rhythm continuum hypothesis, studies such as Ramus et al. (1999) and Grabe and Low (2002) used consonantal and vocalic stretches as the base units for the calculation of classificatory statistics. This was done in order to adopt direct phonetic correlates, contrary to language-dependent phonological units such as the foot or mora. Consequently, an approach was proposed by Ramus et al. (1999) where purported rhythm types were implemented by vowel and consonant interval variability. Indices used in Ramus et al. (1999) were:

a) %V, the percentage of vocalic intervals in a sample, and

b) two indices of variability: the standard deviation of vocalic intervals, ΔV, and consonantal intervals, ΔC.

All indices were calculated globally over predefined stretches of annotated spoken material. In the original study, a corpus of five sentences per eight languages

per four speakers each was analysed. The main achievement accredited to Ramus et al. (1999) was to provide quantitative equivalents of *some* of Dauer's phonological parameters of rhythm types. The dispersion of studied languages in the plane defined by the ΔC dimension and %V dimension is presented in Figure 2.1. The particular combination of parameters corresponds to the traditional accounts of rhythm classes.

Dauer's criteria for classification, in fact, did not include only purely temporally defined features (duration, quantity) but also pitch (intonation, tone), quality based parameters of vowel and consonant reduction as well as the function of stress placement: fixed or free (Dauer 1983). Consequently, out of several phonological dimensions suggested by Dauer to have an influence on rhythm type classification, the qualitative distinction between vowels and consonants were preserved, and only the duration-related surface effects were quantified. As will be noted again later, even the readily observable parameter of stress placement could not be captured by the metrics proposed by Ramus et al. (1999). With global measures, any specific alternation based effects are lost. The connection between high variability of vocalic intervals (ΔV) and free stress in e.g. stress-timed languages needs to be made independently on the basis of previously known facts.

The Pairwise Variability Index (PVI), first formulated in Low et al. (2000), differs from the indices proposed by Ramus et al. (1999) by describing (binary) sequential rather than global duration variability. The first formula below is used with raw durations, where *m* is the number of analysed units and *d* is the duration of a given unit *k*. The raw PVI formula is mostly used for consonantal durations that are known not to vary considerably with speech rate. The reason to use a non-normalising formula with consonantal intervals was also to preserve syllable structure information such as long consonantal clusters etc. that might add to language-specific duration profiles.

$$rPVI = \left[\sum_{k=1}^{m-1} |d_k - d_{k+1}|/(m-1)\right] \tag{2.1}$$

The PVI compares durations of consecutive intervals locally, that is, the differences between sequential pairs of units are calculated and averaged. This averaged distance measure can be used with different adjacent units in fact, not

only vowels and consonants but also syllables, CV groups or other acoustically defined units (see Section 2.1.6). The value of the index is lower when consecutive intervals are more equal in duration. The more the index value approaches 0, the closer to isochrony the intervals are. The second formula, presented below, is used to normalise intervals that vary with speech rate, e.g. usually vocalic intervals.

$$nPVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right] \tag{2.2}$$

Table 2.1 contains a list of recent modifications to the two most popular techniques described above as well as original proposals of new methods of measuring rhythm. Some of them will be discussed below in more detail. As can be seen from the table, metrics based on the PVI have become most popular.

Both PVI based metrics and the measures proposed by Ramus et al. (1999) have been currently used to distinguish between stages in first language acquisition, bilingual and monolingual (Bunta and Ingram 2007; Vihman et al. 2006), second language acquisition (Ordin et al. 2011; White and Mattys 2007a), different dialects and accents (Ferragne and Pellegrino 2004, 2007; Meireles et al. 2010; White and Mattys 2007b,a), different musical styles (Patel and Daniele 2003; Patel 2010), forensic phonetics and speaker recognition (Mary and Yegnanarayana 2008; Dellwo and Koreman 2008; Dellwo et al. 2012), and in language pathology (Liss et al. 2009). The currently noted usefulness of metrics in e.g. speaker recognition exposes at the same time the weaknesses of the techniques to robustly distinguish between language types on the basis of temporal intervals, a distinction the measures initially set out to quantify (but see results on automatic language recognition in Rouas et al. (2005)). Careful control of rate, dialect, accent, gender etc. is needed in order to reliably distinguish between languages using metrics. Methodological issues concerning "rhythm" metrics (or "rhythm measures", (Barry et al. 2003)) largely overlap with theoretical considerations and can be summarised as follows:

a) speech rate

b) choice of materials

c) elicitation style

Table 2.1: A list of some popular metrics, chronologically from the top

Name	Statistic	Source
No name given	Percentage deviation of foot duration from tone unit duration, divided by the number of feet per tone unit	Roach (1982)
RIM	Rhythm Irregularity Measure: Sum of absolute log value of ratios between one unit's duration and the next	Scott, Isard and de Boysson-Bardies (1986)
VI	Variability Index (of syllables)	Low (1994); Deterding (1994, 2001)
%V	Percentage of vocalic intervals	Ramus, Nespor and Mehler (1999)
ΔV	Standard deviation of vocalic intervals	
ΔC	Standard deviation of consonantal intervals	
CrPVI	Raw pairwise variability index (PVI) of consonantal intervals, see Equation 2.1	Low, Grabe and Nolan (2000); Grabe and Low (2002)
VnPVI	Normalised PVI of vocalic intervals see Equation 2.2	
CnPVI	Normalised consonantal PVI	
PVI-CV	PVI of consonant and vowel groups (pseudosyllables)	Barry, Andreeva, Russo, Dimitrova and Kostadinova (2003)
med_CrPVI	Median CrPVI	Ferragne and Pellegrino (2004)
med_VnPVI	Median VnPVI	
YARD	PVI of normalised syllable durations	Wagner and Dellwo (2004)
nCVPVI	Normalised PVI of consonant and vowel groups (pseudosyllables), syllables and feet	Asu and Nolan (2005)
VarcoΔC	ΔV/mean vocalic duration	Dellwo (2006)
VarcoΔV	ΔC/mean consonantal duration	
CCI	Modified PVI where each interval is divided by the number of segments comprising it	Bertinetto and Bertini (2007/2008)
Vdur/Cdur	Ratio of vowel duration to consonant duration	Russo and Barry (2008)

d) interspeaker variability

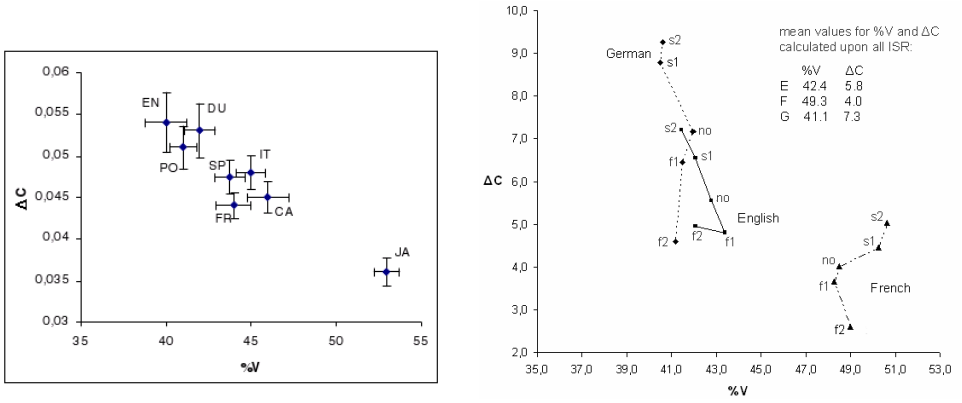
e) segmentation strategy

f) choice of rhythmic correlates

2.1.1 Speech rate

Normalisation for speech rate was pointed out several times as an important factor influencing the results and robustness of metrics. The problem was originally raised by Ramus (2002) in his response to Grabe and Low (2002) who obtained a different dispersion of languages in the hypothetical rhythmic space using similar criteria of measurement. Ramus claimed that the largely disparate results of Grabe and Low (2002) came from the lack of normalisation for rate in Grabe and Low's corpus. In fact, the nPVI proposed by Grabe and Low (2002) itself includes normalisation of the used intervals. Ramus however, despite being aware of the built-in normalisation, calculated PVI scores using data from Ramus et al. (1999), where rate was controlled by averaging sentence duration (3 sec) and the number of syllables per sentence (15 to 19). The outcome of the new calculation indicated that, when the same normalised corpus was used, both methods produced very similar language clusterings. In his discussion of speech rate effects, Ramus (2002) proposed using perceptual judgements of rate by speakers within a language in order to establish a "speech rate norm" that would enable cross-linguistics comparisons (cf. Gibbon 2003).

Dellwo and Wagner (2003) conducted a study that could be seen as going towards satisfying Ramus's desiderata. The study was designed to investigate the effect of speech rate on %V and ΔC scores. The intended speech rate that subjects produced when following instructions regarding tempo ("read the text normally, slowly, even slowlier, fast, even faster") was compared with the number of syllables per second. All subjects showed a change in the measured syllable per second rate as they introduced the intended rate change. The speakers, however, also showed that the rate effect is different depending on language and speaker specific characteristics. The authors investigated speakers of English, French and German in the study. It turned out, as the authors note, that French seemed "to



(a) (b)

Figure 2.1: a) Values of ΔC (in msec) and %V for eight languages in Ramus et al. (1999): Catalan (CA), Dutch (DU), English (EN), French (FR), Italian (IT), Japanese (JA), Polish (PO) and Spanish (SP). Adapted from Ramus et al. (1999); b) values of ΔC (in csec) and %V at 5 intended speech rates in Dellwo and Wagner (2003): normal (no), slow (s1), very slow (s2), fast (f1) and very fast (f2) for German, English and French. Adapted from Dellwo and Wagner (2003).

provide the greatest freedom" (Dellwo and Wagner 2003: 473) in terms of how many syllables per second the native speakers were able to produce. English and German clustered close to each other by allowing "less freedom" compared to French in terms of syllable rate along the indended tempos. The differences could be explained by the languages' individual syllable structures. This result suggests that a simple syllable rate metric is potentially able to distinguish between purported rhythmic groups, looking at the emerging clustering along this parameter in Figure 2.2. In this light, Ramus's method of ensuring that all studied sentences, in all studied languages, be of the same rate of syllables per second, as it turned out, miss one potential typological variable: overall average rate with which a given language is spoken. This possibility was hinted at by Ramus (2002), but Dellwo and Wagner (2003) provided more evidence to support the hypothesis that inherent rate differences can also underlie the perceived rhythmic variability.

However, crucially for the discussion of the original metrics, the results of the analysis by Dellwo and Wagner (2003) using the %V and ΔC indices revealed that %V is not greatly affected by intended rate change: deceleration attempts showed a slight decrease in %V for English and German and a slight increase

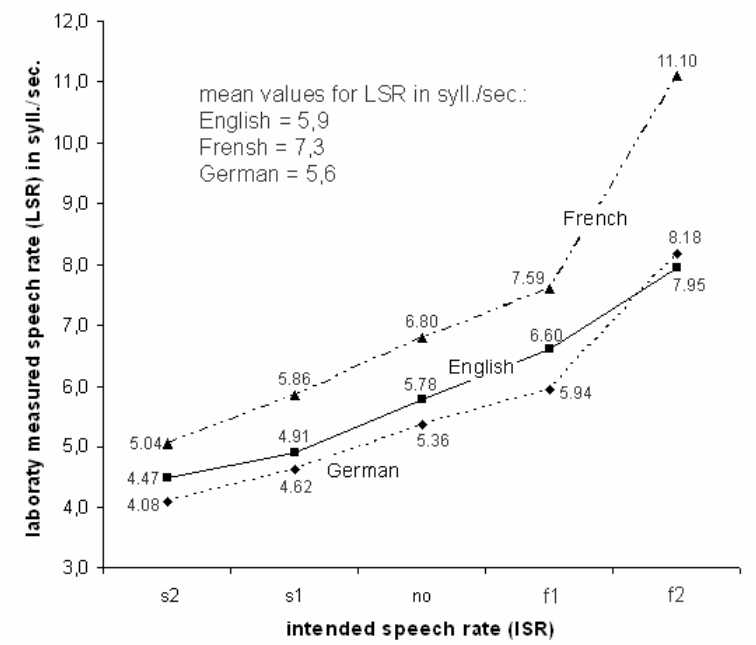


Figure 2.2: Adapted from Dellwo and Wagner (2003). Intended speech rate vs. laboratory speech rate in syll./sec. Speakers of French, English and German.

for French. Russo and Barry (2008) and Keane (2006) also found that %V was the most tempo-resistant parameter in their data. They also showed that it is the most language-distinguishing measure. %V expresses the relative frequency of consonant and vowel intervals and so the above results might reflect more careful consonant articulation in slower rates in English and German and vowel insertion in French.

Interestingly, in Dellwo and Wagner (2003), the range of differences across tempos in terms of percentage points was greater than the differences between languages in Ramus et al. (1999). For example, the difference range between tempos in French was 2.3%, while Ramus et al. (1999) report a 1.5% difference in %V between Dutch and Spanish. As can be seen in Figure 2.1, ΔC values calculated in the same study cluster English and French in a way corresponding to the results of Ramus et al. (1999). However, contrary to %V values, ΔC varies considerably across intended tempos and the value given to French by Ramus et al. (1999) is in fact the value for slow French speech. These results suggest that type placement might depend on tempo as well as imply that the switching between

different timing types might be demonstrable both within and across languages, depending on speech rate.

Indeed in Barry et al. (2003) the authors state that "tempo-differentiated analyses clearly show a tendency for languages to converge with increasing tempo towards what has been defined as a 'syllable-timed' position" (Barry et al. 2003: 2696). Building on these claims, Russo and Barry (2008) found that vocalic variability measures (ΔV and vocalic nPVI) show higher values at slower rates and lower values at higher rates. These results led Russo and Barry (2008) to conclude that "rhythm values are a function of articulation rate" (Russo and Barry 2008: 422). Speakers not only approximate to syllable-timing in fast speech but also vary the length of their vowels more in slow speech, a durational phenomenon that has been considered characteristic of stress-timed languages. It is possible, therefore, that the grouping effect of stress in slow speech as well as the regularising isosyllabic effect in fast speech are universal features present in all languages that have stress (see also section 3.1). An interesting question remains about the relationship between phonological structure in a given language and the threshold points along the speech rate parameter at which the language starts moving its timing towards one type or another. It seems that observing segmental variability in the speech rate dimension might reveal important differences, e.g. as to the rate of change in timing strategies in different languages. The results could be more interesting than finding "the norm", the one typical timing type for each language.

In the meantime, not only corpus design but also normalisation procedures in rhythm measures have attempted to factor out the inherent component of speech rate from duration data. If rhythm metrics are meant to isolate the language-specific rhythmic factors, rate has to be controlled, if not accounted for, in the used model. Speech rate appears to be treated here as a factor that merely introduces noise into a typological ideal that an optimal corpus should reflect in order to be useful for rhythmic analysis. However, the evidence quoted in this section can also be looked at from a different point of view: speech rate is a dynamic parameter in the produced timing of utterances that interacts with the temporal structure inherently in a constrained manner. More importantly, it is possible to incorporate it into a model of speech rhythm variability, as will be shown in Section 3.1. Such a decision requires a different look at the rhythm type hypothesis,

one that allows for having both syllable- and stress-timing within the rhythmic repertoire of a language. Studies within the rhythm metrics paradigm can also provide support for this alternative: if timing types are a function of speech rate (as shown by Russo and Barry (2008)) that can be varied consistently by speakers of all languages (as shown by Dellwo and Wagner (2003)), then it follows that degrees of both timing types can be observed in each of these languages.

2.1.2 Corpus materials

The choice of the text used in constructing the corpora on which metrics analyses are performed can influence the results. The problem was already hinted at by Gibbon (2003, 2006), and the effect was experimentally examined by Arvaniti (2009). Gibbon (2006) noted in his discussion of the PVI that the measure cannot capture all possible alternation patterns, both between and within languages. Despite giving the possibility of locally comparing two adjacent intervals in a binary fashion, the PVI cannot describe structures such as unary and ternary (anapaestic and dactylic) rhythms. As an example for a unary timing alternation, Gibbon (2006) gives the sentence: "This one big fat bear swam fast near Jane's boat". This sentence could also serve as an example of an approximately syllable-timed structure[2] that is possible in English. On the other hand, "Jonathan Appleby wandered around with a tune on his lips and saw Jennifer Middleton playing a xylophone down on the market-place" features a more complex ternary structure with a *sww* syllable pattern and could at the same time describe a typically stress-timed structure. Neither of these non-binary patterns can be described by the PVI algorithm. Gibbon's examples at the same time hint at the possibility that the text used to construct corpora serving as a basis for the calculation of metrics should not be arbitrarily chosen and could be manipulated to demonstrate its effect on the scores.

Arvaniti (2009) in fact devised three sets of texts in English and Spanish that exploited the potential of producing "strongly syllable-timed" and "strongly stress-timed" utterances in each of these languages. Example sentences are provided in Table 2.2.

[2]It is very likely that speakers would introduce a higher level stress grouping into this largely isochronous syllable sequence. The implausibility of "syllable-timing" as a notion is also discussed by Arvaniti (2009) and in Section 3.1.1.

Table 2.2: Examples of "syllable-timed", "stress-timed" and uncontrolled English sentences used by Arvaniti (2009).

Timing type	Example sentences
"stress-timed"	*The production we increased by three fifths in the last quarter of 2007.*
"syllable-timed"	*Lara saw Bobby when she was on the way to the photocopy room.*
uncontrolled	*I called Gatsby's house a few minutes later, but the line was busy.*

Arvaniti hypothesised that since metric scores measure durational variability they should reflect the highly stress- or syllable-timed structure of the used materials. Most importantly though, if the metrics indeed describe crosslinguistic variability, the uncontrolled materials should pattern with the scores of the "stress-timed English corpus", and the uncontrolled Spanish material should yield scores similar to the "syllable-timed Spanish corpus". Unfortunately, as Arvaniti phrases it, "most disturbingly perhaps, metric scores can be affected by the choice of materials, so that, independently of a language's accepted rhythmic type, more stress-timed materials can yield scores that are closer to those of stress-timed languages" and vice versa. This effect is apparent in Figure 2.3 (note that ΔC is plotted here on the X axis and %V on the Y axis), where Spanish stress-timed material clearly clusters with uncontrolled English material and vice versa.

Most recently, Prieto et al. (2012) and Wiget et al. (2010) investigated the effect of specifically designed sentences on the scores. Prieto et al. (2012) provide results based on materials where syllable type (CV, CVC and mixed) served as the variable. Speakers of three languages, English, Spanish and Catalan, were asked to read the materials. The results showed that syllable type had a strong effect on the interval proportion and deviation measures on the one hand, but on the other, the PVI was robust against materials containing open or closed syllables while still showing a distinction between Spanish, Catalan and English. Wiget et al. (2010) again demonstrated the sensitivity of the PVI to metrical structures inherent in the sentences they used: "the ordering of strong and weak syllables is (...) critical for nPVI-V" and "could be seen as a strength" (Wiget et al. 2010:

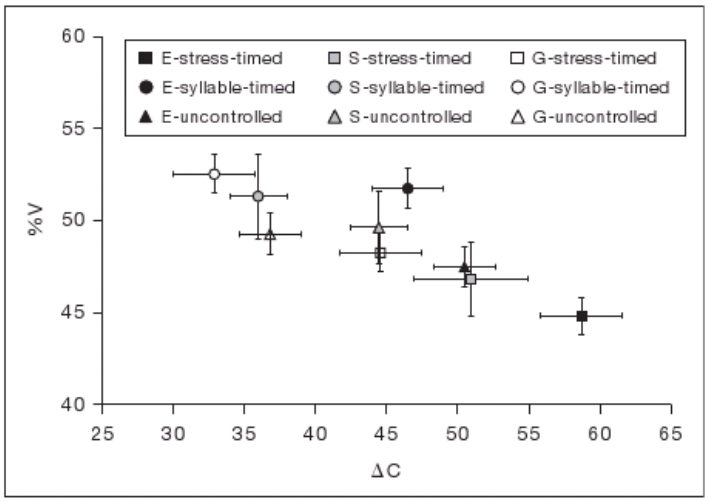


Figure 2.3: Results for two rhythm metrics indices: ΔC and %V for stress-timed, syllable-timed and uncontrolled text materials in three languages, English (E), Spanish (S) and German (G) as found by Arvaniti (2009) (note that ΔC is plotted here on the X axis and %V on the Y axis). Adapted from Arvaniti (2009).

1565) of the measure, as they suggest. The problem persists in that the sensitivity, as defined by the formula, describes binary metrical structures only. Wiget et al. (2010) at the same time suggest that corpus materials used to calculate metrics should be sufficiently large and exhibit a "language-typical range of metrical structure". Again, the measures discussed here were originally designed to *reveal* the "language-typical" rhythmic features in a mathematically explicit way. However, as it seems, it is not known what is measured unless independent qualitative reasearch defines the language-typical structures first.

Gibbon's and Arvaniti's remarks on text material support the observations by Cummins (2002) who discussed the general validity of the stress- and syllable-timed taxonomy. He pointed out that both types of timing can be produced, sometimes with stylistic intent, by speakers of e.g. English. Certain types of discourse, such as preaching and political speeches make ample use of both types of rhythm, Cummins (2002) noted. Barbosa (2000) draws attention to the fact that Pike had already observed both syllable- and stress-timing is utilised in English in certain speech styles and in singing (Pike (1945: 71), in Barbosa (2000)). Therefore metrics, to make strong typological claims about rhythm type, can only point out ten-

dencies within languages towards certain types of timing, given that caution with the design of corpus collection is taken, as succinctly demonstrated by Arvaniti (2009) and other studies discussed in this section.

2.1.3 Elicitation style

As indicated above, style of delivery may change the timing relationships in the production of speech. Keane (2006) reports on a study of formal and colloquial Tamil where three metrics techniques were used: the Ramusian parameters, the PVI formulas and Deterding's VI (see Table 2.1) where syllabic constituents are used to calculate the index. The two stylistic varieties of Tamil were found to differ significantly with regard to interval metrics. The results constitute strong counterevidence for the use of metrics as indicators of language type. However, the two styles of enunciation in Tamil differ greatly to the extent of being considered two different "languages", as Keane (2006) notes. Nonetheless, the results also point out the interaction of elicitation style and speech rate. Consonantal measures of variability (the consonantal raw PVI and ΔC values) distinguished between the two styles best in Keane (2006). Given that formal Tamil tends to be slower than the colloquial variety, Keane's results are in line with Dellwo and Wagner (2003), where ΔC values pointed to the difference between slowly spoken French and other tempos. Whether style and speech rate could be potentially conflated as factors influencing duration remains to be seen.

Arvaniti (2009) used three elicitation styles: read sentences, read running speech and spontaneous speech in her study. She hypothesised that while in read speech the typological differences should be maximised, in the spontaneous style, due to universal reduction processes, languages should cluster together more in terms of timing patterns, thus blurring the typological distinctions. Indeed, in spontaneous speech, despite the appearance of language clusters in the expected locations (closer to one another in the postulated typological space), most pairwise comparisons between the languages used failed to reach statistical significance. This result led Arvaniti (2009) to conclude that "score differences among languages disappear in spontaneous speech" (Arvaniti 2009: 54). One has to wonder whether a similar effect of conflation of timing strategies across languages would

have been observed due to variation induced by speech rate (as discussed in Section 2.1.1), if very slow speech had been used by Arvaniti.

2.1.4 Interspeaker variability

Arvaniti (2009) also noticed that, typologically considered, score differences, as revealed by different metrics used in her study (PVI, Ramusian indices, Varcos), are not statistically significant. The result was explained by high interspeaker variability in the data where individual speaker scores did not cluster in distinct language groups. Moreover, across metrics, the individual results show different patterns, pushing the given languages once towards the "expected" language group and once in the other direction. Also in Keane's study (Keane 2006) the differences between speakers, as quantified by the metrics, often exceeded those separating different languages. Keane suggests at the same time that the "heavy speaker dependence of the (...) measures may be largely attributable to differences in speech rate, both between speakers and within the speech of individuals" (Keane 2006: 325). Given the results by Dellwo and Wagner (2003), where the rate factor, compared with the language factor, obscured the linguistically determined difference, Keane's suggestion is certainly valid. The problem of the speaker effect on language scores posed by Arvaniti (2009) could then be explained by rate differences as well. Given the speaker factor, the fact that the original studies by Grabe and Low (2002) used single speakers for each studied language, weakens the general validity of their results.

2.1.5 Segmentation strategy

As most interval measures and PVI measures rely on durations extracted from an annotated corpus, the issue of an appropriate and consistent segmentation of the speech signal cannot be overlooked. There have been attempts to evaluate the usefulness and accuracy of manual segmentation conducted for the purpose of measuring rhythm. Some studies have empirically demonstrated inconsistencies between annotators in this context (Wiget et al. 2010), however the magnitude of the effect did not exceed the material (cf. Section 2.1.2) and speaker effect (cf. Section 2.1.4). Both Grabe and Low (2002) and Ramus (2002) discussed this

issue in the context of compatibility of their respective classification attempts. The intervocalic raw PVI calculated in Grabe and Low (2002) clustered Japanese, English and German together. Grabe and Low (2002) noted that the result might be due to the inclusion of devoiced vowels in the measured intervocalic intervals; devoiced vowels constituted 16% of all vowels in their Japanese material.

All the methods of measuring duration in the context of rhythm discussed in this chapter rest on a comparable and robust method of delimiting consonantal and vocalic intervals. Of course this point is strongly related to a theoretical question whether consonant and vowel stretches *are* the appropriate units of segmentation in case of "measuring" speech rhythm. This point is elaborated on in Section 2.1.6 below. From a methodological point of view however, for the methods themselves, standards of segmentation need to be established so that the rather subjective process of interval annotation can be depended on empirically. As mentioned, e.g. in Section 2.1.4, large corpora seem to be necessary to reliably measure durations with a typological project in mind. However, even with a standard protocol at hand, annotating large amounts of data is very laborious. Some studies discussed the problem proposing automatic annotation methods instead (Galves et al. 2002; Loukina et al. 2009; Wiget et al. 2010).

Wiget et al. (2010) compared the performance of automated phone alignment (where an orthographic transcription of the data is provided and boundaries are identified on the basis of the transcription using a speech recognition algorithm) with segmentation prepared by humans. The metrics score results based on automated segmentation produced similar results to the ones based on manual segmentation. The authors concluded that automatic segmentation based on statistical rather than acoustic properties of the signal, given enough training data, will replace the time-consuming and often inefficient manual segmentation for this purpose. Loukina et al. (2009) used a purely acoustic segmentation procedure where an algorithm computed time series of loudness and aperiodicity from the signal. After smoothing and normalisation the resulting segmentation corresponded to vowel and sonorant intervals, obstruent intervals and pauses. A classification algorithm was used to compare combinations of several metrics parameters, most of them included in Table 2.1. The assumption of the independent validity of the rhythm class distinctions was again necessary to evaluate the performance of the

machine learning algorithms. The best classifying performance was achieved by measures based on normalised vocalic intervals where speech rate was also added as an additional dimension: 47% of the data was correctly classified for median vocalic PVI, 48% for vocalic duration to consonantal duration ratio, 48% for %V, 48% for normalised vocalic PVI.

Loukina et al. (2009) discussed the advantages of performing an automatic segmentation based on the acoustic properties of the signal without applying phonological rules (that often play a role in the judgements taken by manual annotators): it mimics the perception and classification of unknown languages and the perception of speech by pre-lexical infants. A similar approach was taken by Galves et al. (2002) in their reinterpretation of vocalic and consonantal stretches in Ramus et al. (1999) as sequences of sonorous and obstruent intervals (see below). The approach in Galves et al. (2002) was motivated by the original premises in the Ramus et al. (1999) study, where infant language discrimination patterns were the primary goal to be modeled. Whereas all the above factors interfering with manual segmentation can be controlled for in rigorous data collection and annotation, the potentially compromising factors discussed below are more of a fundamental nature.

2.1.6 The choice of rhythmic correlates

From the criticism of the phonology based rhythmic taxonomy a conclusion was drawn that "more effective measurements to account for how rhythm is extracted by the perceptual system" (Ramus et al. 1999: 269) were needed. Therefore several questions can be raised: a) how far are consonant and vowel stretches the relevant units for the *perception* of rhythm? Also, it is important to ask if b) the metrics approach suggests that consonantal and vocalic intervals are the abstract prosodic constituents that take part in rhythmical *linguistic* processing in general. Similarly, c) is the segmental domain of consonants and vowels the right one for a model of rhythm *production*?

Regarding a), it may be the case indeed with prelinguistic infant speech processing where the argument is that abstract prosodic constituents have not yet formed. Ramus et al. (1999) motivated their choice of correlates on the basis

of studies on language recognition by children. Psycholinguistic research showed that infants pay more attention to vowels and are able to count them independently of syllable structure. It was found that prosodic cues alone can be used by infants to discriminate between rhythm classes (Nazzi et al. 1998). Since one cannot expect infants to use abstract phonological categories such as the foot or syllable in foreign language discrimination without prior linguistic knowledge, Ramus et al. (1999) concluded, a phonetic correlate was needed to account for children's perception of rhythm. This decision was also motivated by basic prosodic facts: vowels carry accent, most of the durational load and voicing.

It could be argued however, even in the case of children, that the choice of consonants and vowels rather than an acoustic category with similar properties, e.g. sonority vs. obstruency, is not uncontroversial. Spence and Freeman (1996) report on an adult identification study of intrauterine recordings where only 33.5% of English CVC and VCV utterances could be identified. They also review other work which confirms that the attenuation of high frequencies *in utero* removes most of phonemic information, leaving frequency contour, voice quality and amplitude variation information available to the fetus (Querleu et al. 1988). It is highly unlikely that fetuses start learning phonemic distinctions in the womb, such as differences between nasals, liquids and vowels. Therefore, it is possible that temporal distinctions that are in fact identified by newborns are based on sonority vs. obstruency and, possibly, between voiced and unvoiced intervals.

We find a sonority based metrics approach in Galves et al. (2002), motivated in similar terms. The segmentation strategy used in Ramus et al. (1999) was changed here by directly and automatically extracting intervals from the signal as a function of sonority, thus circumventing also, e.g. hand-labelling controversies (cf. Section 2.1.5). Galves et al. (2002) acoustically measured sonority contrasting it with obstruency, providing a more coarse-grained rather than detailed phonetic distinctions, as this way e.g. nasals, liquids and vowels are collapsed into one category. Galves et al. (2002) obtained a positive linear correlation between the rough sonority vs. obstruency measures and %V and ΔC measures as found in Ramus et al. (1999). This way, they were able to represent infant phonotactic processing abilities at around 6-9 months of age, when babies only start to learn the details in spectral properties of consonants. Since the Ramus et al. (1999)

study was primarily designed to illustrate infant language discrimination abilities based on prosodic cues, it seems the approach taken by Galves et al. (2002) is an improvement as far as this goal is concerned, if only in terms of greater rigour in the choice of acoustic correlates.

Contrary to Galves et al. (2002), Steiner (2003) was able to show that including consonants en bloc into the analysis, misses some consonant type dependent effects on the durational output statistics. Steiner analysed German, French, English and Italian corpus data using the following data preparation strategy: the signal was tagged according to sonority scale criteria to arrive at six types of intervals (instead of two): vowels, approximants, laterals, nasals, fricatives and stops, including separate labels for syllabic nasals and laterals. Next, the Ramusian parameters were extracted and compared using different combinations of interval types. The results showed that the combination of parameter types that maximally separates the standard syllable- from stress-timed languages is not %V (proportion of vocalic intervals) and ΔC (standard deviation of consonantal intervals) (Pearson's $r = 0.822$) but %l and %n (Pearson's $r = 0.902$), the percentage of laterals and nasals in the given corpora. The rhythm continuum hypothesis rests on a qualitative analysis of linguistic properties of adult language phonology and therefore a more finely grained analysis, such as the one in Steiner, seems more appropriate to represent typological distinctions of this nature. Specifically, Steiner's results suggest that "the functional load of consonant classes is not homogenous" (2003:6) so that the phonotactic profile pertaining to rhythmical classification needs to be more detailed than the one found in Ramus et al. (1999). This is certainly probable if, again, the premise is accepted that the only way to evaluate a rhythm metric is by how well it is able to classify the data into (untested) rhythm classes.

Incorporating a phonological approach to the choice of units Ramus et al. (1999), in a similar vain, introduce a caveat in a footnote to their paper: "We are aware, of course, that the consonant/vowel distinction may vary across languages, and that a universal consonant/vowel segmentation may not be without problems" (Ramus et al. 1999: 271). The potential choice of finer phonemic distinctions raises problems. The only differentiation in terms of characteristics other than duration, i.e. spectral, is the general consonant/vowel distinction. The poten-

tial contribution of intensity or segment quality and their relation to duration and prominence has been explicitly omitted, also in perceptual discrimination experiments. For example, for the purpose of perceptually evaluating the parameters in Ramus et al. (2003) the data was resynthesised using the "flat sasasa" technique, which eliminates intonation, prominence alternations resulting from segmental quality and distribution of energy in the signal[3]. Such a strategy was necessary, the authors claimed, for extracting solely the durational effects which were equated with "rhythmic cues" in their work. However, the remaining problem of the significance of other correlates was not missed by the authors who commented that:

> Languages differ in the way they use duration and intensity to signal phonological properties such as stress or quantity. It can therefore not be excluded that a similar quantitative, cross-linguistic study of intensity variations might provide yet another dimension for the study of rhythm classes. (Ramus et al. 2003: 341)

Widening the inventory of units of analysis could be beneficial if, in the context of rhythm perception, input units could include acoustic correlates, other than durational, into the analysis, e.g. pitch, loudness and segmental quality. If one of the aims of phonetic implementation of rhythm classes is to find out how rhythms are being extracted perceptually, one needs to include the missing correlates to construct a more complete measure. Theoretical implications would also have to be taken into consideration, e.g. regarding the kind of typological distinctions the measure would imply when correlated with durational indices. In fact, languages differ in the extent they make use of different correlates of generally considered prominence. However, introducing levels of rhythmic magnitude that prominence distinctions bring in, would quickly expose the inability to handle hierarchical relations such as "weak prominence" and "strong prominence" that arise, e.g. from minor phrasal accents and major phrasal accents respectively.

Keane (2006) uses a PVI based measure of loudness to determine a difference between Tamil and English. The loudness values were extracted by averaging over 10ms steps within 50ms windows in a given interval's spectral power density profile. Since loudness as a prominence correlate in Tamil is known to vary

[3] A method based on speech resynthesis proposed in Ramus et al. (2003) used to delexicalise utterances: consonants are replaced by the sibilant fricative /s/ and vowels with /a/ and the pitch is flattened. Sound samples can be found at: http://www.lscp.net/persons/ramus/resynth/ecoute.htm

only with segmental quality within a word, Keane (2006) expected that English, where stress is realised also with a loudness variability, will show higher values. Indeed, the normalised loudness coefficient differentiated between the two languages, adding an intensity based dimension to the difference between strong and weak syllables.

A study that directly addresses the concern of the "necessary but not sufficient" aspect of duration in accounting for rhythm is Lee and Todd (2004). They propose an *auditory prominence hypothesis* which distinguishes languages on the basis of the "greater variability in the auditory prominence of their phonetic events (in particular vowels)" (Lee and Todd 2004: 227). They concentrated, like Ramus, on building a model that does not rely on complex phonological information, but is able to segment the speech signal into primitive phonetic events, however, including the prominence level of these events. Exploiting the "time-intensity patterning" in the signal, they simulated the human auditory nerve responses by using a bank of low-pass filters with a range of time-constants used to smoothe the responses. Peaks in the output of such a simulation were then mapped on what the authors call "the rhythmogram", where each event found on the rhythmogram has a prominence value assigned to it. The prominence value is determined by three parameters: intensity, duration and frequency. Their model offers a good representation of a perceived rhythmic structure. At the same time it offers a fuller and more sophisticated modelling of auditory processing of prominence, incorporating temporal integration as well as non-durational parameters.

Cumming (2011) acknowledges that there are many other correlates that belong to the percept of speech rhythm, apart from duration: tonal properties, amplitude, spectral balance and spectral properties, such as formant structure. Cumming (2011) combines the weighted cues of f_0 and duration and uses the PVI to reflect the perceived pattern of prominences, or perceptual rhythm, in Swiss German and two varieties of French. The language specific cue weights were derived empirically by Cumming (2011) from a perceptual experiment. Participants judged several intonationally and temporally manipulated stimuli according to perceived level of "deviance" from what they find natural. These perceptions turned out to be different for each of the languages studied. By undertaking this effort, Cumming (2011) acknowledged the need for a fuller range of perceptual

correlates in the rhythm metrics paradigm and importantly, implemented them by relating the correlates to real differences in the linguistic perception of rhythm.

Finally, since the original dichotomy was based on the isochrony of prosodic units, i.e. syllables and interstress intervals (the foot approximately), one could ask why not quantify the units that are in fact participating in the prosodic hierarchy. The PVI in particular can be used with any base intervals that are seen fit to be compared in a pairwise fashion. The answer probably lies in the original motivations for Grabe and Low (2002) and Ramus et al. (1999), where a quantification based on phonetic equivalents of the phonological rhythm typology by Dauer (1983) were adopted and these referred mostly to vowels.

Asu and Nolan (2006), by taking this perspective, came to interesting conclusions about duration typologies on the example of Estonian. Estonian possesses very characteristic duration features: short, long and extralong phonological vowel contrast in particular, geminates. Quite opposite to, for example, Rouas et al. (2005) where the aim was to find a universal unit enabling automatic annotation and classification of multilingual data, in Asu and Nolan (2006) and Asu and Nolan (2005) the approach is to first analyse and characterise duration patterns in a specific language by testing out the candidate prosodic units in the PVI paradigm. The PVI analysis is to reveal the optimal units that embody both duration regularities and take part in the rhythmical structure in that language. It was concluded by Asu and Nolan (2006) and Asu and Nolan (2005) that it is the syllable and the foot that best characterises Estonian duration patterns. PVIs based on these units exhibited a combination of low interspeaker variability for syllabic and interstress intervals (Std.Dev. of 1.4 and 0.5 respectively, compared to 3.5 in case of the normalised vocalic PVI) with PVI values lower than the ones based on vocalic intervals. These facts pointed out that there is a higher regularity in intersyllabic and interstress interval patterns, compatible with phonological evidence and observations about Estonian.

2.2 Summary

First of all, rhythm metrics, as speech rhythm models and approaches to language classification suffer from the lack of independent motivation. Rhythmic category

is assumed *a priori* and the methods are adjusted to fit the hypothesised model (itself not testable yet). The evaluation of linear metrics, often proceeds according to the objective of how neatly the standard rhythmic categories are separated out from the data. The whole approach rests on the assumption that rhythm classes constitute strongly attested objective observations. The question of how well *both* the rhythm class hypothesis, and the metrics implementing it, match some putative criteria necessary for an accurate modelling of rhythm in general is not addressed.

In fairness, some authors that have worked with rhythm metrics (Wiget et al. 2010), have made an attempt at a definition of the type of temporal structure the rhythm metrics describe:

> These rhythm metrics are intended to capture the stable differences between and within languages in degree of temporal stress contrast (...). Given this assumption, such metrics are best at gauging 'contrastive' rhythm, i.e., the balance of strong and weak elements in speech, rather than 'dynamic' speech timing, i.e., the temporal arrangement of groups of sounds according to a higher level structure. (Wiget et al. 2010: 1559)

The above distinction is a step towards constraining the scope of metrics. The authors also provide a list of good practices and assumptions that need to guide the appropriate use of the measures. However, as argued here, even if the above definition of speech rhythm is adopted, "contrastive rhythm" includes strong-weak contrasts, i.e. binary ones, maximally, when nPVI is used. So Gibbon (2003: 292): "The binary model is too strict: durations in real speech do not follow a simple long–short–long–short pattern; nor are they strictly linearly organised".

Moreover, even if a rhythmically based classification of languages is possible, it is clear that it would need to be multidimensional. As indicated by the limitations of the canonical rhythm types space that is capable to contain only prototypical languages, clearly, nonprototypical languages at least point to the inclusion of other dimensions. Loudness, pitch, spectral features, i.e. correlates of prominence, as well as metrical patterns (unary, ternary rhythms) and speech rate need to be all considered as factors supplementing duration in the construction of a rhythmic typology. Speech rate inherently interacts with temporal structure. Observing variability in the speech rate dimension might reveal important differences, e.g. how languages change their timing. In fairness, it has been evident

from the limited success of rhythm metrics that the information contained in the vocalic portions of the signal holds at least part of the answer to the question whether language discriminating rhythm *can* be measured at all. Parameters based on the *vowel* (percentage, variability) have been the most robust and indicative of cross-language variability. The vowel and the vowel cycle is certainly a significant event in rhythmical processing.

Additionally, most rhythm metrics capture only the global, linear aspects of prosodic structure. The lack of structure precludes them from being adequate models of rhythm[4]. Empirical problems exist in the choice of phonological units and corresponding acoustic correlates, as well as in the parameters and in the eventual statistic measures performed on the segmented data.

Also, clear definitions of what duration, rhythm and speech timing are necessary, along with statements about which of these phenomena the quantitative methods measure, as found, e.g. in Cumming (2011). Work on detection of rhythm types in the signal has so far been misleading in that regard in that constraints of perception and production on speech rhythm have not been fully accounted for. Minimally, such attempts, potentially useful in technological applications, should be kept separate from studies of rhythm that focus on speech behaviour. Meanwhile, a purely signal driven answer to the question of speech rhythm cannot be given without formulating hypotheses about, e.g. perceptual event structure.

[4]A proposal found in Bertinetto and Bertini (2007/2008) is an attempt at a synthesis of advancements in Articulatory Phonology, coupled oscillators modeling of rhythm and other studies on the topic, including rhythm metrics. The authors depart from their "rhythm metrics" inspired model, the Control/Compensation Index (CCI), however, their ultimate rhythm model proposal is inherently hierarchical, since it postulates that speech rhythm variability, among and within languages, is specified by two structural levels, the phonotactic level and the sentential level. The two levels in themselves are characterised by a coupling between a) consonantal and vocalic gestures and b) syllable peak (p-centre) and accentual cycles.

Chapter 3: Dynamical models of speech rhythm

3.1 Coupled oscillator models of speech rhythm

Several authors have considered the timing relationships *between* rhythmic levels, i.e. syllables and feet, drawing attention to the relative coordination in a hierarchical perspective. In speech we deal with relatively few prosodic subunits (segments, syllables, feet, phrases) that form a single, complex system of interactions on multiple timescales, as was briefly introduced in Chapter 1. The coordination between the subunits is probably based on rhythmic principles. Periodicity on the level of speech articulation is also assumed. This facilitates the coordination of speech production in time on the one hand, on the other, rhythmic constituents impose structure to serve a similar goal, i.e. coordinating inherently periodic syllabic units with stress-feet within which they are nested. Stress-feet, in turn, serve grouping functions in perception and planning functions in production (Tilsen 2011). Both these functions need to be taken into account and inform the study of speech rhythm, so far observable, but not strictly detectable in the acoustic signal. These functions lie at the core of the models of speech rhythm variability considered in this chapter (Barbosa 2006, 2007; O'Dell and Nieminen 1999).

First, it will be argued that the way in which the syllable and the variability of foot duration relate to each other is an important descriptor of rhythmic timing strategies and duration patterns found in Polish, English and other languages. At the very least, the linear measures discussed in Chapter 2 could benefit from relating relevant rhythmic interval durations to each other, as shown in e.g. Asu and Nolan (2006) (see further in Section 3.1.1). Subsequently, it will be shown that hierarchical coupling (interaction) between prosodic levels, and hence rhythmic variability in speech, can be modelled by coupled oscillators.

The chapter is structured as follows: a review of phonetic work that alludes to the coordination between subsystems will be presented first, while explicit accounts of rhythmic variability based on coupled oscillator models will ensue. Basic principles behind dynamic modeling of timing will be introduced as well. The chapter concludes with an experiment on Polish data using a coupled oscillators model of rhythm variability with speech rate differentiation. Hypotheses concerning the characterisation of inter-level rhythmic timing in Polish spontaneous speech will also be presented. The goal is to clarify the status of rhythmic timing strategies characteristic for Polish. The nature of these strategies has been subject to controversy when standard methods of quantifying rhythmic variability were used, as was summarised in Section 1.4.1.

3.1.1 Phonetic accounts of hierarchical timing

The reader may recall that rhythm metrics have been criticised, e.g. by Gibbon (2006) and Cummins (2002), for the lack of hierarchical structure (Section 2.1), among other things. "Flat" models are not compatible with what is observed in speech regarding the temporal structure of both speech perception and production (cf. Section 1.2). The models omit important characteristics such as hierarchical nesting of components on multiple timescales that contribute to the functional benefits of rhythmical structures (Jones and Boltz 1989). Rhythm metrics instead imply a limited rhythmic structuring on the level of vowel and consonant alternations only, seen as peaks and troughs in the signal (Gibbon 2006)[1]. Clearly, such high frequency alternation is not the only one that exists in languages that have stress. In both English and Polish at least one immediately superior level of lower frequency alternation induced by stressed syllables exists. These two higher and lower frequency cycles are usually conceptualised by syllables (or syllable-sized units) and interstress intervals respectively.

In a study by Asu and Nolan (2006), PVI (Grabe and Low 2002) indices were used with both syllable and foot intervals in Estonian and compared to segment-based index values. Syllabic and foot-based nPVI scores turned out to be

[1] Or as Cummins (2002) puts it: "Where is the bom-di-bom-bom in %V?... The discrete basis for the suggested taxonomy can be argued to be grounded in segmental inventories and syllabic phonotactics." (Cummins 2002: 2)

more robust than segmental ones. While vocalic intervals showed high variability between speakers, foot and syllable retained consistency across speakers. In conclusion, Asu and Nolan (2006) pointed out that there might be different degrees to which the isosyllabic and isoaccentual forces are exerted in a given language.

Asu and Nolan (2006) discussed their results by essentially reinstating the significance of language specific syllable dynamics within a foot. The way the two levels relate to each other and influence each other is an important descriptor of rhythmic variability between languages and within languages; it is the interplay between the unit higher in the prosodic tree, such as the foot, and a lower unit, such as the syllable that negotiates particular patterns. Asu and Nolan (2006) wonder whether the to-be-compressed units, nested within a superior unit, receive equally distributed compression "duty" or whether there is some language specific principle upon which the duration distribution rests:

> The results presented above suggest that languages need not, as in the traditional dichotomy, either (like English) squash their unstressed syllables to achieve approximate foot-isochrony, or (like French) keep their syllables fairly even and not bother about foot timing. They could also equalise their feet to some degree, but share the 'squashing' more democratically in polysyllabic feet. Estonian, with its strong stress but near absence of vowel quality reduction in unstressed syllables, and despite its three-way quantity contrast which sporadically curtails syllable-equality, may be at base such a language. (Asu and Nolan 2006: 251)

Patterns and relations like the ones considered above would be hard to represent using just a single dimension such as e.g. vocalic interval variability, without a reference to a higher unit. The position Asu and Nolan (2006) take from the perspective of "rhythm metrics" indirectly refers to approaches to e.g. polysyllabic shortening and rhythmic variability that have been undertaken before, e.g. Eriksson (1991). In such approaches, the interaction of interstress interval duration and the size of the syllabic material in that interval are analysed by means of simple and multiple regression.

O'Dell (2003) discusses the well known observation that "smaller units such as segments or syllables tend to become shorter in duration as more of them are incorporated into a higher level timing unit"(O'Dell 2003: 105). Such a relation between interstress intervals and constituent syllables is known as "polysyl-

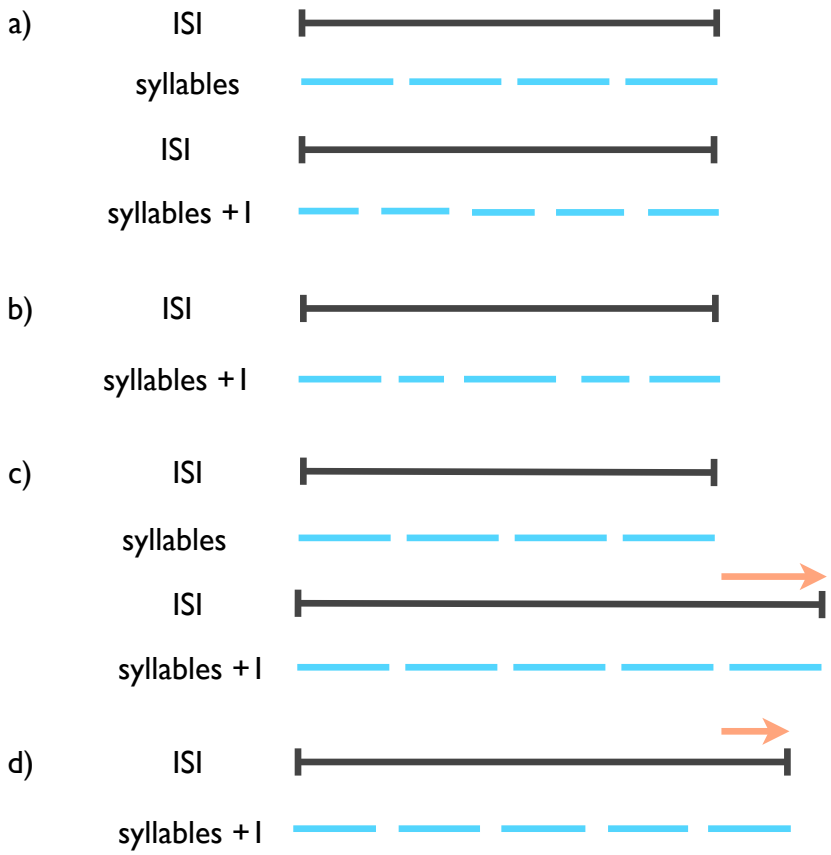


Figure 3.1: Schematic depiction of possible relations between the duration of an inter-stress interval (ISI) and the number of syllables in that interval.

labic shortening", "stress-timed shortening" (Beckman and Edwards 1990) or, as O'Dell (2003) proposes, "rhythmic gradation".

Bouzon and Hirst (2004) discuss different approaches to isochrony in the context of higher timing levels interacting with lower levels. While the strong isochrony hypothesis expects relevant rhythmic units (feet or syllables) to maintain equal duration in a strict fashion, the weak hypothesis requires only a tendency for the units to always be of equal duration. The consequences for the subunits are therefore also different. The compression effect observed in the subunits is less dramatic within the requirements of the weak hypothesis.

Table 3.1: Simple linear models of interstress interval duration as a function of the number of syllables for five languages. Adapted from Eriksson (1991). Note that r denotes the correlation coefficient of the models.

Language	Regression equation	Corr. coeff.
English	$I = 201 + 102n$	$r = .996$
Thai	$I = 220 + 97n$	$r = 0.973$
Spanish	$I = 76 + 119n$	$r = 0.997$
Greek	$I = 107 + 104n$	$r = 1.0$
Italian	$I = 110 + 105n$	$r = 1.0$

Figure 3.1 shows schematically how the duration of a higher unit, e.g. an interstress interval (henceforth ISI), may be influenced by adding a subconstituent, e.g. a syllable. The depiction is simplified with regard to the first, stressed syllable which, at least in English, is assumed to be longer than the others. Figure 3.1 shows absolute interstress interval isochrony in a), with uniform compression of syllables as a syllable is added. Picture b) shows absolute isochrony on the interstress interval level, as a subunit is added, as well as one of many logical possibilities of a non-uniform compression of the subunits. Figure c) illustrates a schematic case of syllabic interval isochrony or, in other words, a proportional increase of the superior interval duration with increasing number of subconstituents. Subplot d) shows a positive correlation between the units, similarly to c), but some compression of the syllables as a syllable is added occurs, i.e. there is a negative correlation between the duration of a subunit and the number of subunits.

Which of the schematic patterns in Figure 3.1 however, has been evidenced in speech? Let us concentrate on the complexity effects on ISI duration first. Bouzon and Hirst (2004) analyse several levels in British English and study the complexity (number of subconsituents) in a higher unit such as: syllables in a foot, phones in a syllable, feet in an intonational unit, etc. They test the strong isochrony hypothesis, i.e. that complexity should not affect the duration of the higher unit. For all levels studied, they find a positive correlation between the number of constituents and the duration of the unit, meaning that strict isochrony clearly does not occur. However, as expected, they find a negative correlation, i.e. some compression of the subunits on any level of structure. Such relations correspond to pattern d) in Figure 3.1.

Eriksson (1991) notes that the original strong isochrony hypothesis means

that interstress interval duration is independent of the number of syllables. As he and others showed (Beckman 1992; Bouzon and Hirst 2004), this is clearly not the case in all languages traditionally classified as stress- or syllable-timed that were studied by these authors. Beckman (1992) states that: "(...) in every language for which we have such data, the intercept of the regression line fitted to such a plot yields a non-zero intercept. That is, there is always at least one durational effect that consistently occurs exactly once somewhere within the stress group or prosodic phrase (...)" (Beckman 1992: 459). As the above implies, we will most likely find relations such as d) in natural language (again, disregarding for a moment the lack of information on how exactly duration is spread over constituent syllables). In fact, strict inter-stress interval (or foot) isochrony, as in a), has never been confirmed, nor has a strictly proportional syllable-timed model, as in c). In fact d) describes a pattern midway between a) and c) and appears to reflect reality.

As Eriksson (1991) shows, there are interesting conclusions to be drawn from the d) pattern. It is apparent that the linear increase in stress group duration as a function of the number of syllables does not behave in exactly the same way in all the analysed languages. Table 3.1 presents Eriksson's regression analysis of mean stress group duration predicted by the number of syllables comprising the group (from one to four) in five languages: English, Thai, Spanish, Greek and Italian. The slope coefficient expresses the effect of adding a syllable on stress group duration. It is noticeable that the slope coefficient is approximately the same for all languages, i.e. the rate of duration increase is the same in both putative rhythmic type language groups (English and Thai vs. Spanish, Greek and Italian). As Eriksson explains, there is also an "initial value" to which a largely stable slope coefficient value is added. That initial constant, i.e. for the intercept, is different in the two hypothesised rhythm groups. It is approximately 100 msec for syllable-timed languages and approximately 200 msec for stress-timed languages. Given these generalisations, Eriksson (1991) proposes a model for rhythmic variability of the form:

$$I = k + 100 * N \tag{3.1}$$

where k is a constant in which 200 msec characterises stress-timed languages

and 100 msec syllable-timed languages. O'Dell and Nieminen (2009) note that Eriksson's analysis suggests that there is an "underlying unity in the rhythms of different languages" (O'Dell and Nieminen 2009: 179) running along the lines of the traditional rhythm type dichotomy. But it does so without implying that the principles are to be found in simple, one level isochrony.[2]

Next, the effects on the subunits need to be briefly considered. The linear increase of interstress interval duration as a function of syllable number, as Eriksson explains, does not necessarily mean that there is no compression of syllables happening within the interval. In fact he demonstrates that formally speaking, both stressed and unstressed syllables can be compressed and still satisfy the assumption that the increase in interval duration will be linear. One possible logical example of how the duration duty might be distributed is shown in pattern b). The above analyses assume an symmetrical pattern of compression among the components within the unit as the number of components increases (Saltzman et al. 2008). Consequently, the specific patterns of compression and expansion *within* the unit, as in the difference between pattern a) vs. b) in Figure 3.1, cannot be decribed using this method. It can be discussed, if in a language such as English, sensitive to duration as a marker of prominence, a case as in Figure 3.1 b) would already induce a restructuring of inter-stress intervals and their number by an "insertion" of prominence onto the expanded syllables. Indeed, what was actually found in English, is that as syllables are added to a foot, it is the stressed syllable that shortens and the unstressed syllables remain stable (Kim and Cole 2005). To recall the quote by Asu and Nolan (2006) above, the duration distribution duty (the "squashing") in English, seems to rest actually on the stressed syllable rather than on the unstressed ones. None of the subunits expands and so no duration based prominence pressures are exerted up until the next stressed syllable beginning the next ISI.

The author is not aware of similar studies of compression effects within polysyllabic feet for Polish. Some effects of a superior unit on the subconstituents in Polish are subject of investigation in Chapter 4, namely, the effect of the vowel-

[2]As it seems, if one insisted, based on the mean values of a few speakers that were used in Eriksson (1991), this measure seems to offer itself as a "rhythm metric" of global duration variability with a hierarchical component, and so it could be added to the set of available formulas in Table 2.1, with similar caveats to its validity applied.

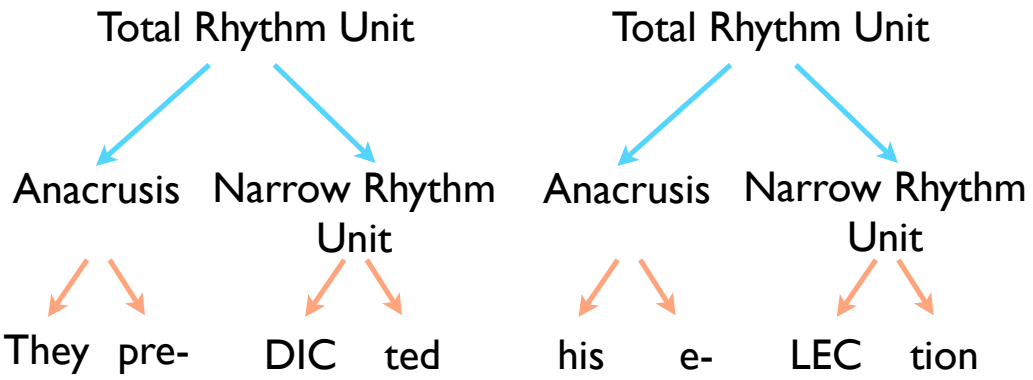


Figure 3.2: The rhythmic units by Jassem et al. (1984). Adapted from Bouzon and Hirst (2004)

to-vowel cycle (the phonetic syllable) on the constituent segments. In this case, the vowel-to-vowel cycle is hypothesised to exert a duration balancing effect on component segments of different durations, as suggested by Barbosa (2006, 2007), as part of his coupled oscillator model of speech rhythm.

Units other than the syllable or interstress interval were considered as a basis of a rhythmic interaction in English. It appears that, at least for English, the Narrow Rhythm Unit (NRU) as proposed first by Jassem et al. (1984) is likely to demonstrate a language-specific greater "mass" in relation to the subunits.

Figure 3.2 shows how the rhythmic units posited in Jassem et al. (1984) are constructed. The size of the Narrow Rhythm Unit (henceforth NRU) depends on the number of syllables in the unit, and the NRU's left boundary is always a stressed syllable. The NRU, however, is different from the Abercrombian foot (Abercrombie 1991): "The foot is effectively a cognitive unit of planning or of perception, whereas the rhythmic unit [the NRU - ZM] is one which is physical and measurable" (Tatham and Morton 2002: 393). An anacrusis (henceforth ANA) is defined as: "a syllable or sequences of syllables (...) characterised by being as short as possible. (...) the ANA always precedes the NRU and belongs to that NRU" (Jassem et al. 1984: 60). The Anacrusis, the iambic element of (usually) a sequence of unstressed syllables, plus the following NRU, form a Total Rhythm Unit.

The syllable relations within an NRU, as Jassem et al. (1984) define, are

essentially as the pattern in d) in Figure 3.1: a two-syllable NRU is longer than a monosyllabic one, but it is distinctly less than twice the monosyllable length, all durations relative to a given tempo (Jassem et al. 1984: 206). Jassem et al. (1984) also postulate that the constituent syllables in the NRU are of approximately equal length and in fact find isochrony within the unit. Given these findings also the patterning of syllable durations within the NRU would tend to conform with pattern d) in Figure 3.1. As illustrated above, the Narrow Rhythm Unit is of roughly fixed duration contrary to the Anacruses, which are of variable length, and proportional to the number of segments within them.

In Jassem et al. (1984), similarly to Eriksson (1991), models were compared by regressing the foot, the ANA and the NRU duration respectively on the number of phones in the given unit. Syllable duration was excluded as unviable due to problems with syllable parsing in the study. Jassem et al. (1984) divided phones into specific classes and used mean phone durations. Simple regression analyses detected minimal phone isochrony in Anacruses and strong isochrony in Narrow Rhythm Units. Following the results, the authors stated that the special statistical status of the two rhythm units in English, the Anacrusis and the NRU, should be recognised and the Anacrusis be excluded from estimation of rhythm unit durations. As well as that, it was the near equal length of the NRUs, in their analysis, that gave the impression of isochrony in English. Jassem et al. (1984) is so far a unique method that successfully found a measure of isochrony in the acoustic signal in English and provided a model of its rhythmic structure.

However, the isochrony is not strict, as Bouzon and Hirst (2004) showed for British English by regressing the count of subunits on their duration, as they appeared inside various rhythmic superunits, among others, the foot and the NRU. They confirmed that the strongest negative correlation between the number of phones and their duration exists in the case of phones belonging to an NRU. Units within the Abercrombian foot exhibit patterns that correspond to a "midway" between the NRU and ANA, since the foot includes both units into its duration.

Jassem et al. (1984) set out to provide and evaluate an adequate measurement procedure that could account for perceptions of rhythmic variability in languages and an assumed isochrony effect in English. Rhythm as a notion in Jassem et al. (1984) was not considered to be synonymous with isochrony, Jassem et al.

(1984) were looking for phonetic evidence that isochrony exists as one of many possible effects of rhythmicality, which they were unable to explain as such.

In summary, the studies reviewed in this section acknowledge that there have to be degrees of both formal syllable- and stress-timing strategies in, at least, all languages that have lexical stress. They also suggest that in order to account for the variability of the strategies between and within languages, some relation between the stress group (or a unit equivalent to it) and the syllable (or a unit equvalent to it) needs to be expressed. It is also evident that a cycle of an even lower frequency than the repeating stress group has an effect on the system. The quantification of rhythmic variability can be achieved with methods and models that take the structure building function of stress and the cyclical function of the syllable into account. The above supports the notion that rhythm is produced on the prominence level but has to be reconciliated with the phasing of segments in syllables and structural constraints stemming from phonology and phonotactics. This view is taken in the experiment conducted in the present chapter. Aspects of rhythmic constituency effects on segments are tested in Chapter 4.

3.1.2 Coordination dynamics

As has been suggested in Section 3.1, relatively few prosodic subunits in speech (segments, syllables, feet, phrases) form a single, complex system of interactions on multiple timescales. A few examples from phonetic studies on the interaction between interstress intervals and syllables as descriptors of rhythmic variability between languages were presented thereafter. In this subsection, the coupled oscillator models of rhythmic variability to be employed in experiments on data in Polish will be introduced. As a prerequisite, this goal requires the presentation and discussion of concepts and methods that apply to the model.

First, it will be argued that the existence of meter (rhythmic structure) and the employment of rhythmic strategies such as syllable and stress-timing are a manifestation of *coordination* between prosodic units.

Rhythmic behaviour lies in the nature of coordination processes in biological systems. Biological systems, from bodies as a whole, including brains, to speech articulators are able to move through their existence purposefully despite

being very complex. A whole interdisciplinary field of study called synergetics (Greek: "working together") (Haken 1982) explores how coordination patterns dynamically unfold in time and allow for functions and structures in complex organisms to arise.

Coordination patterns or synergies in biological systems arise spontaneously as a result of self-organisation when a system is motivated by an intention to reach a particular goal, or effectively continue a functional process. Synergies are therefore functionally ordered and may characterise all possible natural behaviours where many component subsystems operate at different time scales. This essentially means that the study of coordination dynamics strives to model not only movement and other low-dimensional phenomena but also cognitive functions. Kelso's classic "Dynamical Patterns" (Kelso 1995) provides an accessible introduction to how synergetics explain motor coordination as well as cognition. A comprehensive overview of the study on coordination, its history and some extensions into linguistics can be found in Turvey (1990), a recent debate on complex systems approach to cognitive science can be found in Gray (2012). It is in fact hypothesised that cognition is the behaviour of a dynamical system (Elman 1995; Gray 2012) [3].

Dynamical systems[4] are "those state-determined systems whose behavior is governed by differential equations. Dynamical systems in this strict sense always have variables that are evolving continuously and simultaneously and which at any point in time are mutually determining each other's evolution" (Port and van Gelder 1995: 5). The solutions to differential equations provide the behaviour of the system for all time, given the initial conditions, that is, starting values assigned to the variables (components of the system). Contrary to Newtonian mechanics, the theory of nonlinear dynamical systems is not looking for values of position or

[3]The "dynamical programme" in cognitive science assumes that the classical rationalist mind-brain divide is unnecessary. The consequences for linguistics are manifold. Malisz (2004) reviews the implications for functional and formal linguistic theories that stem from the programme: dichotomies such as mind-brain, planning-execution, grammar-motor or phonetics and phonology are abandoned in the framework. If only because a transformation of the set of natural numbers (discrete) to a set of real numbers (continuous) for any type of one-to-one correspondence between its elements is impossible, as Barbosa (2006: 7) notes.

[4]The mathematical theory was set by Henri Poincaré (1854-1912) and the name Dynamical Systems Theory normally refers to his fundamental work on differential equations.

velocity in a given point in time but for global features over longer periods. In other words, it does not seek formulae for each solution but it studies the collection of all solutions for all time. The advantage of dynamical systems approaches to complexity is that by looking for parameters that describe the whole system, simple behaviour can be observed, simpler than when looking for regularities on any participating level (Kelso 1995).

The geometrical description of a system's behaviour (a *topology*) evolving in time will be used here as a qualitative characterisation of the theory. *Trajectories* show how variables of the system will evolve over time. "A trajectory plots a particular succession of states through the state space and is commonly equated with the behaviour of the system" (Eliasmith 1995: 22). *State space* or *phase space* represents all possible states of the system and their evolution in time. *Attractor* is a definition of an area of stability (a *limit set*) in the phase space to which trajectories tend to go and where they are likely to stay. If the set of governing equations (for example, a physical law, such as gravitation) and a state on the trajectory is known, the theory may predict the behaviour of the system. In other words, we need to specify the rule for change (the *dynamic*) and the state of the system in order to predict the behaviour.

It is characteristic of coordination dynamics that instabilities of the system's motion play an important role in establishing the parameters that govern the behaviour. In synergetics, it is at the point of instability that the dynamics of self-organisation in biological systems can be observed. A *bifurcation* is a qualitative change in the dynamics of a system when a certain parameter value is reached. We may also say that when a qualitative change in the type of the attractor occurs, the system has undergone bifurcation (Norton 1995; Gray 2012: 57). For the explanation of self-organisation and coordination of movement, it is important to identify the *collective variable* of the system. There are two types of parameters involved in the system's evolution: the *control parameter* is an external parameter. For example, the temperature regulated by a chemist in a laboratory influences the behaviour and motion of a liquid and serves as an external parameter. Instabilities are created by control parameters which "move the system through its collective states" (Kelso 1995: 45). The collective variable, on the other hand, expresses the interaction between components. It can be identified when found near a bi-

furcation where instability causes reorganisation, a switch to a different pattern. "Relevant degrees of freedom [variables], those characterizing emerging patterns in complex systems, are called collective variables in synergetics" (Kelso 1995: 16).

A much cited cited example of a biological system whose behaviour is described by coordination dynamics is the Haken-Kelso-Bunz model (Haken et al. 1985). Haken et al. (1985) examined the wagging of index fingers at different rates. They found that at slow speeds there are two finger oscillation patterns (*bistability*) that reflect preferred movement coordination: the in-phase pattern and the anti-phase pattern. The fingers are in phase when they both move in the same direction, they are out of phase when, e.g. the index finger of the right hand moves rightwards and the index finger of the left hand moves leftwards. With finger wagging, after a critical movement rate is reached the anti-phase pattern looses its stability and there is a sudden phase shift into the in-phase movement. This shift is a bifurcation under the influence of the movement rate parameter. However once the most stable, in-phase pattern is reached, slowing down the speed does not cause a switch back to the anti-phase pattern: bistability is merely possible again when the rate is decreased below the critical point (*hysteresis*). The collective variable in this example is the relative phase between two oscillatory systems, the fingers. The dynamics of the collective variable (or, as it is also called, the *order parameter*) is an equation describing the coordinated motion of the system. The equation may have simple (*fixed point, limit cycle*) or complicated solutions (*chaos*).

Extrapolating from this very simple phenomenon of interlimb coupling, intergestural coordination of speech has been modelled. One example of phase transitions as observed in speech is found in the experiment by Tuller and Kelso (1991). Subjects were asked to repeat the syllables /pi/ and /ip/ at varying rates. Glotal adductions and abductions were observed by transillumination and examined in relation to lip aperture in time. Stetson (1951) observed that at fast rates a repeated syllable such as "eeb" is heard as a sequence of "bee"'s. Tuller and Kelso (1991) replicated this result with naive listeners and articulatory measurements. Tuller and Kelso (1991) showed that indeed a phase shift in gesture production occurs that corresponds to listener's identifications of the syllable shape. The ar-

ticulatory phase was determined by relating the glottal opening peak within the lip opening cycle. CV syllables started off at 40 degrees after peak closure while VC syllables at 20 degrees. VC production shifted articulatorily after a short phase of zero degree timing to a stable 40 degree timing under speech rate manipulation.

The coordinated motion of the system may yield simple solutions. Those solutions, i.e. states to which the system tends to go and settle in, are called attractors. There are different types of attractors. Only the ones currently used to model speech behaviour will be introduced in this thesis. The system involved in the production of consonantal gestures is postulated to follow *point attractor* dynamics. Often a given system will settle over time in a stable state or cycle of states (*orbit*). The point attractor describes stable dynamics: the system, regardless of the initial conditions (the starting point), ends up in the same stable state. Also when perturbed, it eventually returns to the stable state.

An example from speech articulator motion for the existence of attractors are e.g. perturbation studies of consonant articulation. Kelso et al. (1972) showed that when the jaw was unexpectedly tugged downwards during a production of /baez/ the tongue still produced a gesture reaching the target. Bilabial closure was also attained in the production of /baeb/ both in the perturbed and unperturbed conditions. A quick reorganisation of the system to reach a particular goal, despite different initial conditions, is characteristic of the existence of *coordinative structures* introduced in Section 1.2.

The *limit cycle attractor* (a circle on the x-y plane) is especially interesting from the point of view of rhythmic behaviour. Rhythmic behaviour can be modelled in terms of oscillations. An oscillator behaves periodically which means that each value of a periodic function must repeat every n time units. Instead of representing an oscillator's behaviour as a time series, it is possible to illustrate it as a phase portrait, which combines position and velocity to show the phase space of the system. Figure 3.3 shows an idealised example of a time series translated to a phase portrait. Oscillatory behaviour strictly follows limit cycle attractor dynamics, in the idealised case, the dynamics of the so called *harmonic oscillator*, where no friction restricts its motion, as in the Figure 3.3.

The evolution of a system in time can be observed in phase portraits in the phase space. The basic difference between the linear time series rendition of

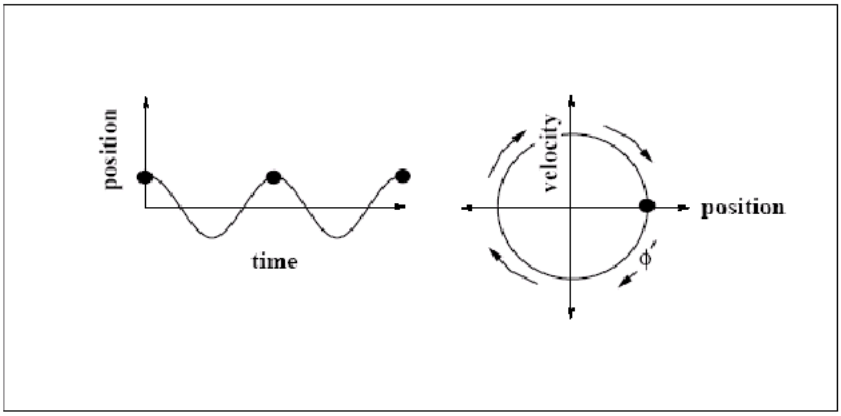


Figure 3.3: Time series of an oscillator's motion (left) and a phase-portrait (right) which combines position and velocity to show all possible states (the phase space). Each phase specifies a fraction of the oscillator's cycle. After McAuley (1995: 49)

time and the phase portrait is that the absolute duration in the former model is exchanged for relative duration. Relative duration is the comparison of one duration with another. Time dimension is intrinsic to the units and systems by application of relative timing and phase relationships. There is no external clock mechanism. Figure 3.4 introduces the concept of the phase portrait as an alternative to the usual Newtonian time series rendition of events in time on the basis of a linguistic example. A simple sequence of CV syllables, mentioned earlier (Section 1.2), /baba/, represents a sequential oscillatory behaviour in speech articulation based mainly on mandibular oscillation.

Another concept extremely relevant for the modeling of rhythmic behaviour is entrainment. *Entrainment* happens when two oscillating systems having different periods assume the same periods, although different phase locking is possible. A simple example of entrainment would be the rhythmic behaviour in music and dance where people entrain their movements to the beat. In this case, the beat in the music provides the periodic signal dancers get "coupled" to. Their bodies start performing in sync with the beat, *period synchrony* is achieved (i.e. the intervals between one step and the other is synchronised with the intervals between beats) and often also *phase synchrony*, i.e. one beat: one step, although other stable phase relationships, such as $1 : 2$ or other, are also possible. *Self-entrainment*, where "the oscillators in question are part of the same physical system" is of

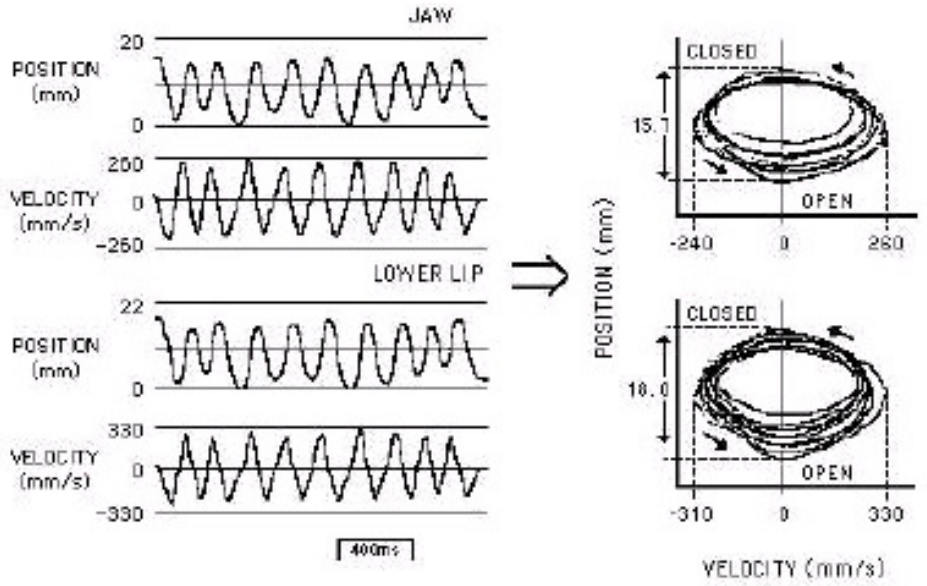


Figure 3.4: On the left: displacement of the jaw and the lower lip in milimeters in time plus jaw and lip velocities in mm/s. On the right: phase portraits of jaw and lip position and velocities. After Kelso (1995: 49).

greater importance for the modelling of speech. And so, for example, a jogger's breath after some time will lock with a steady step. Port et al. (1999) give these and many other examples of self-entrainment in everyday activities as well as from perception and speech. Self-entrainment is a stable mode of behaviour, that is, organisms are likely to exhibit this type of behaviour as the preferrable one; locking oscillating systems with each other is an "attractive" and natural thing to do.

Port et al. (1999) also stress that there exists evidence for self-entrainment between cognitive systems, not only between motoric systems such as the limbs. The example of dancing can then be seen rather as self-entrainment between the motor and auditory system where the latter informs the former. In fact, it is not the beat of music per se that entrains the movement, but the perception of it. This way, entrainment can be seen not only as a result of physical forces influencing the dynamics of two oscillatory systems but in fact, perceived periodic information (be it visual, tactile, auditory) can be used for coupling. The finger wagging experiment underlying the Haken-Kelso-Bunz model described above, as well as other

studies on human interlimb coordination also between subjects, provide direct, not only anecdotal, evidence for self-entrainment. As far as speech is concerned, the dynamics can run even deeper. The speech cycling tasks (Cummins and Port 1998) introduced in Section 1.3 on meter provide another example of entrainment in speech. In this case, salient events in speech are entrained to metrical attractors, that as Cummins and Port (1998) propose, may correspond to pulses provided by neurocognitive oscillators. The "magnet-like" properties of entrainment (Saltzman et al. 2008) are characteristic of nonlinear ensembles of coupled oscillators.

3.1.3 Cyclic events and structuring events in speech

Rhythmic movements are very common in behaviour, especially human behaviour, and highly rhythmic, almost periodic, speaking can be easily observed in poetry, chant and singing. Periodic behaviour constitutes in fact a very simple form of coordination, as cyclic events are a natural form of control (Port 2013). Turvey (1990) argues that because coordinated activities unfold in time, they should be analysable as a sum of periodic contributions. All in all, he concludes that "facts combine to make periodic movement the basis of any theory constructed to account for patterns of coordination" (Turvey 1990: 941).

Caution should be exercised however when talking about periodicity in the context of speech rhythm research. What is implied in the present section does not involve isochrony, i.e. surface periodicity. As Turk and Shattuck-Hufnagel (2013) recently suggest:

> (...) [H]ow persuasive is the evidence for this claim that typical communicative speech involves periodicity? On our view, given the extensive evidence that normal conversational speech is not periodic on the surface, i.e., that no constituent recurrs at regular temporal intervals, it is still very much an open question whether or not speech is (1) controlled using periodic control structures, and/or (2) perceived as periodic. Although the latter claim is widely assumed, it has been subjected to remarkably little empirical testing, and so must be taken as a hypothesis rather established observation. (Turk and Shattuck-Hufnagel 2013: 93)

What is postulated in the following models is a periodic control structure mostly implemented by the vocalic cycle.

Cyclic events can be nested within each other into hierarchical structures: a well known property of prosodic structures (Port 2013; Jones and Boltz 1989). Speech needs to be coordinated on multiple levels with many degrees of freedom. Coordinative structures emerge as control systems in circumstances where there is an interaction among the system components. The coupling of different kinds of oscillation at simple harmonic ratios (1/3, 1/2, 2/3) is a ubiquitous and intrinsic property of animal control systems (Port et al. 1999) and is found in speech. Such coordinative structures, defined as "self-organised, softly assembled (i.e. temporary) sets of components that behave as a single functional unit" (Shockley et al. 2009: 313), have an inherent tendency to behave periodically (Kelso 1995; Turvey 1990). In a set of prosodic components that behave as a single functional unit, therefore, at least *tendencies* towards periodic behaviour, temporary and self-organising, should be observed. Strong periodic tendencies are observed in speech production that is highly constrained, i.e. especially coordinated, as Cummins and Port (1998) have shown experimentally in speech cycling tasks.

In the quotation above, Turk and Shattuck-Hufnagel (2013) rightly observe above that spontaneous communicative speech rarely involves straightforwardly observable periodicities (be it on the surface or in motor structures). Looking at less constrained and regular styles of speaking, we see that: "subsystems can be postulated which could exhibit simple oscillatory behaviour in isolation but may exhibit more complex behaviour when allowed to influence each other normally" (O'Dell 2003: 103). In other words, we might observe very simple oscillatory behaviour under certain conditions on any single level, by repeating a simple CV syllable for example, or when we constrain one level strictly to the other, as e.g. in chanting. However, when these cycles interact freely, they may exhibit complex behaviour such as in rhythmic and quasi-rhythmic patterns of spontaneous speech.

Also, on the lower level scale, meaning has to be communicated by means of sound contrast and sound combinations. In other words, phonotactic structure of the given language influences the behaviour on the subsyllabic level to a large extent. Hence, speech oscillations inherently posited to have regular periods (*eigenfrequencies*) are subject to paradigmatic perturbations (Barbosa 2006). Barbosa (2006, 2007) subscribe a tendency for periodicity to the vowel-to-vowel cycle. This hypothesis will be revisited in Chapter 4. Some proposals and mod-

els based on coupled oscillators that integrate intra- and inter-level timing exist (Bertinetto and Bertini 2007/2008; Nam et al. 2010; O'Dell and Nieminen 2009). However a comprehensive discussion is beyond the scope of this dissertation[5].

Moreover, the cycles are also constrained by their "environment" just like moving bodies: by intention, context and function, that is by higher level linguistic functions. Communicative context influences, e.g. information structure and the patterning of e.g. pitch accents[6]. Given all these and other complications it has been difficult to find simple periodic behaviour manifested in spontaneous speech acoustics, i.e. strict isochrony, especially on single levels of coordination, such as in syllable or foot duration. Barbosa (2006) points out how searching for absolute periodicity (isochrony) and/or lack of inclusion of hierarchy[7] into rhythm models could be the reason for the lack of results (cf. Fowler (1980)).

Regarding structure building, even in a purely isochronous train of syllables, as suggested for syllable-timing, structure is anyway perceived. Such a sequence of syllables is usually heard as "tic-toc, tic-toc" (Arvaniti 2009). This way Arvaniti (2009) points out the inherent "implausibility of syllable timing". As Arvaniti (2009) stresses, if syllable-timing is based on isochronous production of syllables, then speakers would essentially be "striving to produce an acoustic effect that their listeners will discard" (Arvaniti 2009: 60). Mechanisms of relative prominence and grouping (perceptual principles) need to apply and indeed do apply in both cases of putative syllable- or stress-timing. Essentially, some structure needs to be imposed on regularity, if only because of Gestalt principles, or, as argued here, structure emerges as at least two systems interact and are func-

[5]Bertinetto and Bertini (2007/2008) note that the "content" of the segmental interacts with the syllable peak oscillator. While the phonotactic level I in their model addresses properties that directly relate to segment relations and traditional syllable structure, the syllable peak oscillator, also called the syllabic oscillator, should not be confused with level I as it is primarily a "nucleus-(peak)-carrier" that serves the organisation of segments on the most basic rhythmic level. It is meant as the most basic cyclical event in speech with the highest frequency, as also suggested above in in this section. Bertinetto and Bertini (2007/2008) suggest that the vowel-to-vowel cycle is physical, i.e. directly derivable from physiological and articulatory mechanisms. The accentual oscillator is more abstract, meaning it arises from phonological (stress and accent characteristics) correlated with pragmatic constraints.

[6]Cf. Byrd and Saltzman (2003) and Saltzman et al. (2008) for boundary adjacent phenomena in the dynamical framework.

[7]"Identificar ritmo com oscilação sem estrutura": identifying rhythm with oscillation, without structure (Barbosa 2006: 56).

tionally ordered. To represent the structuring aspect, annotation for the purposes of rhythmic analysis has to reflect real time patterning of relative prominences, as they are perceived by native speakers. Such procedure is implemented in the subsequent experiment (see Subsection 3.3.1.3 for details of Rhythmic Prominence Interval annotation).

Given all these and other semantic and pragmatic "regulatory" caveats, a complex timing system is observed in speech that slips away from simple formalisms such as absolute isochrony on any single level. However, as will be demonstrated next, coupled oscillator models employ *tendencies* towards periodicity in the signal as manifestations of *underlying* cyclical control structures in a hierarchical account of speech rhythm.

3.1.4 A coupled oscillator model of rhythm variability

O'Dell and Nieminen (1999, 2009) developed a coupled oscillator model which is able to account for the rhythmic variability posited for those languages that were traditionally described using the syllable-timed, stress-timed scale (Dauer 1983; Nespor 1990). The model relies in parts on the analysis performed by Eriksson (1991). Eriksson (1991) demonstrated a linear relationship between the duration of the stress group (the interstress interval) and the number of syllables contained in the stress group by using the following function:

$$I = a + bn \tag{3.2}$$

where *n* is the number of syllables contained in a stress group, and *a* is a constant term. The details of Eriksson's analysis were presented earlier in Section 3.1.1.

In essence, the results of the linear regression analysis conducted by Eriksson show that languages differ only in the constant term a. In an illustration of Eriksson's results by Barbosa (2000), language types can be distinguished depending on the stress group duration pattern as a function of the number of syllables determined by the equation. Figure 3.5 presents the models of with functions for the two traditional typological poles: syllable and stress timing. Consequently, Barbosa (2006, 2000, 2007) argues that there is a degree of typologically distinct isochrony possible in the production of speech but based on the interaction be-

tween at least two levels: the syllabic and the accentual (stress). The types of rhythmic strategies found, be it close to syllable timing or stress timing, are not qualitatively distinct but can be demonstrated on a continuous scale using the parameter of "weight", i.e. coupling between the two levels.

As discussed in Section 3.1.3 we do find cyclic events in speech prosody such as the syllable and foot. Cyclic events can be described by oscillators: "Any process that tends to repeat itself regularly can generally be described as an oscillator"(O'Dell and Nieminen 1999: 180). Coupled oscillator models of speech rhythm (Barbosa 2006, 2007; O'Dell and Nieminen 1999, 2009) utilise at least two universal oscillators: the syllabic oscillator and the phrase stress oscillator that operate at distinct timescales (cf. Cummins and Port 1998; Tilsen 2009). The "task" of the syllabic oscillator is to keep pace with the vowel onset sequence. Vowel onset sequences specify "phonetic syllables" (and p-centres). The phrase stress oscillator specifies both prosodic phrasing and prominence.

The interaction between the tendency to preserve the periodicity of the syllabic cycle and the phrase stress cycle can vary. Duration patterns observed in speech data depend on the strength of coupling exerted by one oscillator on the other. This means that a model such as the one by O'Dell and Nieminen (2009) can express degrees of stress- and syllable-timing that often coexist in languages depending on, e.g. speech style as well as differences between languages themselves. The method is also able to express hierarchical relationships between relevant rhythmical units, unlike most linear "rhythm metrics" (cf. Asu and Nolan 2006).

An oscillator corresponds to a limit cycle attractor and as such has a natural frequency and natural phase. O'Dell and Nieminen (1999, 2009) recommend a method where working with natural frequencies can be avoided and the variables are reduced to phase only. This is done in order to arrive at simplified generalisations about the behaviour of oscillator collections, even without knowing the exact parameters governing the behaviour of the component oscillators (O'Dell and Nieminen 2009).

Limit cycle dynamics (cf. 3.1.2) give rise to a single phase variable that increases at a constant rate, so that the phase derivative (phase rate of change) equals the oscillator's natural frequency. To approximate the coupling function between

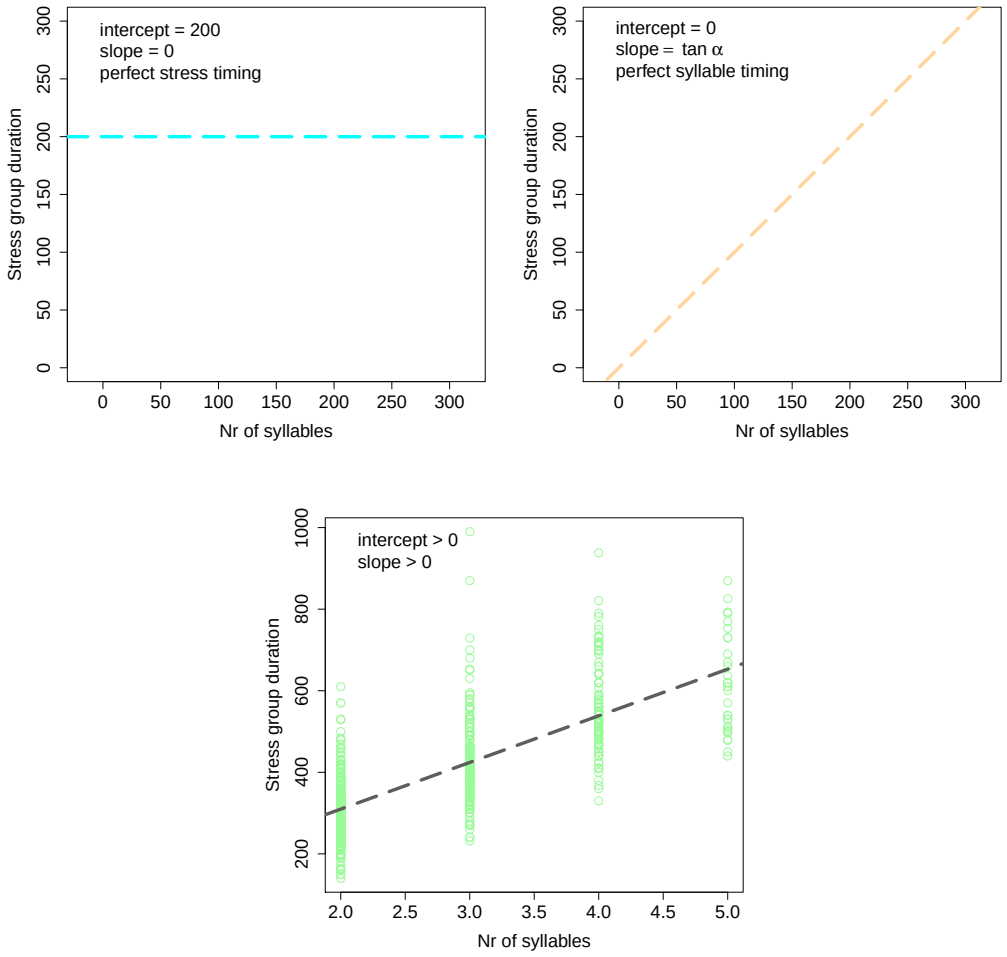


Figure 3.5: Hypothetical models of canonical rhythmic strategies as expressed by the variability of the stress group and the number of syllables contained in it, adapted from Barbosa (2002). The top left panel shows a model for perfect stress timing and the top right panel shows a model for perfect syllable timing. The bottom panel reflects a more realistic model with a non-zero intercept, as discussed by Eriksson (1991) and Beckman (1992).

the syllable and phrase stress oscillators, and so their influence on each other's frequency and phase, O'Dell and Nieminen (1999, 2009) use Average Phase Difference (APD) theory (Kopell 1988). The coupling between the oscillators is approximated by a function based on the difference between the two phases, averaged over a whole cycle of each oscillator.

Next, O'Dell and Nieminen (1999, 2009) mathematically model the hierarchical coupling of the stress cycle over the syllable cycle observed as the "rhythmic gradation" phenomenon, discussed in Section 3.1.1. In this case, stress group oscillator and a syllable oscillator are coupled by a function that depends on *n*, the number of syllables per stress group. The strength of coupling is captured by the constant *r*. The period of the stress group oscillator (T) at an equilibrium can be calculated as a linear function of *n* of the form found in Eriksson (1991), i.e. in Equation 3.2, with coupling strength (r), phase difference (ϕ) and frequency (ω) as variables:

$$T_1(n) = \frac{1}{\omega_1 + H(\phi_n)} = \frac{r}{r\omega_1 + \omega_2} + \frac{1}{r\omega_1 + \omega_2}n \tag{3.3}$$

As a result, regression coefficients a (the intercept), b (the slope) obtained empirically by (Eriksson 1991) are linked, via APD theory, to the relative coupling strength parameter r in the following way:

$$r = a/b \tag{3.4}$$

The ratio between the intercept to the slope presented in Figure 3.5 is consequently expressed in terms of the coupling strength *r*. This way, the coupled oscillator model demonstrates that rhythmic gradation is the result of the hierarchical coupling of two cycles (O'Dell and Nieminen 2009). Or, as it was put by Saltzman et al. (2008) in their discussion of the model:

> The key theoretical result was that the behavior of foot duration as a function of number of syllables depended on the degree of asymmetry of the coupling forces between the syllable and foot oscillators. (...) For English, the coupling from foot to syllable dominated the coupling from syllable to foot, and the ratio of coupling strengths could be specified as a function of the regression parameters in Eriksson's analyses. (Saltzman et al. 2008: 180)

The qualitative interpretation of the relative coupling strength values are given in Section 3.3.1.6.

3.2 A coupled oscillator model of speech rate-differentiated data in Polish

A model of spontaneous Polish data is provided using the O'Dell and Nieminen (1999) coupled oscillator model. The analysis aims to shed light on the rhythmic variability and speech rate issues in Polish on the basis of an annotated corpus of spontaneous speech in a task-oriented dialogue.

First of all, the objective of this experiment is to tackle a typological problem, especially relevant for Polish, in the light of the inconclusive results in the literature regarding its rhythm type, as discussed in Section 1.4.1 and in Chapter 2. No English data was analysed, since in all paradigms discussed so far, there is an agreement concerning the main rhythmic strategy employed in English. However, where possible, contrastive comparison of results with accounts on English in literature are given.

In particular, a specific oscillatory model by O'Dell and Nieminen (1999, 2009), introduced above, is used with Polish data. As discussed previously, periodicity on any single level (isochrony), syllable or stress interval based, cannot be postulated to characterise speech rhythm types. However, due to communicative task demands, certain timing strategies, understood as coordinative structures, are employed. These timing strategies are manifest in the interaction of the syllabic oscillator and the stress oscillator. In this sense, syllable-timing and stress-timing is used strictly to denote rhythmic strategies, understood as approximate states located on a continuum of relative coupling strength between two oscillators. The relative coupling strength is seen here as a *collective variable* (see Section 3.1.2 for a definition). Such a conceptualisation shifts the attention away from particular degrees of freedom, e.g. the syllable or the stress cycles, as the units of control, in favour of one coordination variable, relative coupling strength, expressing the interaction between these subsystems.

Also, as discussed in Chapter 2, global and local timing patterns are heavily influenced by speech tempo. The used model includes the speech rate parameter in a systematic way. Speech rate and a correlated variable "communicative

task" (roughly, speech style), are the parameters under influence of which the timing strategies reveal themselves, as employed in a given language. Speech rate is used as an *order parameter* (see Section 3.1.2 for a definition) along which the states of the collective variable can be observed. This view fundamentally assumes that it is possible to observe different rhythmic strategies within one language as a function of speech rate.

Additionally, a statistical multiple regression model of Rhythmic Prominence Interval duration as predicted by the number of syllables and speech rate will be calculated. This is done to provide a single duration model for prediction of RPI duration that might become useful in speech synthesis or other applications. The model is not linked to production or perception constraints.

The proper characterisation of relative prominence in the annotated data, as discussed in Section 3.1.1, is crucial for the adequate modeling of speech rhythm in a hierarchical perspective, as pursued here.

3.3 Experiment 1: coupling strength between rhythmic levels in a Polish dialogue corpus[8]

3.3.1 Material and annotation

DiaGest2, a Polish multimodal corpus of task-oriented dialogues (Karpiński et al. 2008a), was used. The data used in the present study come from eight speakers of standard Polish (four female and four male undergraduates) whose task was to instruct a dialogue partner in a paper folding task. The speakers had an origami like paper structure in front of them and were instructed to guide their dialogue partner, from whom the structure was concealed, towards constructing the same structure using plain sheets of paper. The task completion time was limited to approximately five minutes. The corpus was recorded audio-visually in a sound-treated room. The participants had no reported speech or hearing impairments.

[8]Preliminary results of this experiment were reported in Malisz (2011)

3.3.1.1 The phonetic syllable

In the present approach it is crucial segment and analyse units that delimit the intervals on each level of structure strictly relevant for the interaction between levels. In other words, as regards segment-prosody interaction, higher levels of timing hierarchy (be it the syllable, foot, phrase) require specific points to serve as anchors for the metrical beat, yield themselves to the influence of top-down rhythmical effects, and will also have their own bottom-up impact. The mapping of metrical beats onto the phonetic level proceeds via p-centres (Morton et al. 1976). The location of p-centres depends on the spectral and durational characteristics of the CV transition. According to Barbosa et al. (2005), the differences found in the vowel onset location relative to the syllable's perceptual centre in synchronisation tasks are due to differences in energy rises caused by onset consonants. Barbosa et al. (2005) showed that in case of the voiceless stop the steep energy rise allows for a closer synchronization of the p-centre with its abstract target: the vowel. In case of the fricative, the p-centre is located earlier, away from the vowel and towards the sibilant due to the high frequency noise. However, they also postulate that the underlying correspondence to vowel onsets is abstractly maintained. This way, the link between the perceptual and articulatory vocalic cycle is established. The vocalic identity of the linguistic beat was also confirmed in many other studies, see Dziubalska-Kołaczyk (2002) for review and discussion and the introduction to Chapter 4 for the articulatory rationale behind the vocalic cycle.

In the present model, it is enough to characterise the phonetic syllable in the form of the number of vowels within a Rhythmic Prominence Interval. The vowel counts were extracted using scripts in Praat (Boersma and Weenink 2012) from a syllabic segmentation that existed in the corpus[9], instead of delimiting vocalic intervals or vowel onsets.

[9]Syllabic boundaries were marked according to sonority principles. The Maximal Onset Principle was not used resulting in the closing of syllables in case of medial clusters such as in: "miasto" → mias.to, "mokry" → mok.ry. There are several problems with segmenting Polish syllables. Examples include cases such as proclitics plus nouns: "z okna" (out of the window), "w wodzie" (in the water). In the present corpus they were segmented as "zok.na" and "wwo.dzie". Symmetrical cases such as "oko" (eye) were labeled as "o.ko".

3.3.1.2 Phrase selection

Rhythmic prominences and phrasal structure were annotated using the Rhythm and Pitch system (henceforth RaP) (Breen et al. 2010). The advantage of the system over, e.g. ToBI is that RaP is largely theory-independent and based on perceptual judgments of native speakers. Also, no ToBI for Polish exists so far. Minor and major phrasal boundaries were delimited. The RaP minor phrase boundary is defined as a minimally perceptible disjuncture. It approximately corresponds to the ToBI break index 3 (Breen et al. 2010). Subsequently, the phrases were inspected in Praat (Boersma and Weenink 2012) in order to select fluent and coherent utterances for rhythmic analysis and modeling: Phrases with false starts, hesitation markers, hesitation lengthening, unintelligible speech portions, overlapping laughter, etc. were omitted. Overall, 411 phrases were selected.

3.3.1.3 Rhythmic prominence intervals

There is no consensus as to acoustic correlates of stress in Polish, as summarised in Section 1.4.2. Polish is generally considered to have perceptually weak stress that is mainly based on pitch accents and intensity peaks rather than duration, similarly to Czech, Finnish and Estonian (Jassem 1962; Klessa 2006; Lehiste 1970). Very often phonological expectations play a great role in the perception of stress in Polish. Lexical stress is placed on the penultimate syllable with few exceptions. Acoustic correlates of prominence are also not clear, as explained in 1.4.3. Therefore, it is rather difficult to delimit interstress intervals using objective methods based on e.g. acoustics as the first automatic step aiding a more time-efficient subjective annotation of rhythmic intervals. However, it was observed that prominence labels in the present annotation largely correlated with the main pitch accented syllable in a phrase in case of strong beats and lexical stress in case of weak beats.

It should be noted that Rhythmic Prominence Intervals (henceforth RPI) are not to be confused with Abercrombian feet. The purpose of the RPI is to reflect a composite of stress and accentual factors that are manifested in perceived prominence and are language specific. As discussed above, the perception of prominence does not always involve clear acoustic parameters and very often can

be defined as an influential expectation. These expectations may be phonological as is the case in Polish (Domahs et al. 2012) with its almost exceptionless penultimate stress or metrical exepectations (Wagner 2005), i.e. relating to universal tendencies for alternation of stresses and consequently, stress shifts and deletions.

Two native experts trained in RaP identified all rhythmically prominent syllables on two prominence levels: prominent and non-prominent. The annotation was based on perceptual judgments of the signal, i.e.: a prominent syllable was marked when a "beat" on a given syllable was actually perceived and not when phonological rules dictated lexical or sentence stress placement. Two labels were used, denoting perceptually strong and weak prominences. Phrases with at least two prominences and therefore at least one full inter-prominence interval were considered. In utterances where a prominent syllable was non-initial, the anacrusis was necessarily omitted. The two experts checked each other's annotation for obvious errors[10].

Rhythmic prominence intervals (RPI) in the pre-selected phrases were extracted from the RaP annotations, that is, by recording durations between one syllable marked as prominent and the next. The intervals were extracted only from within fluent stretches of speech, excluding pauses, in the selected phrases, as described above. Additionally, prominences marked on phrase-final syllables were excluded from the analysis to avoid boundary lengthening phenomena interfering with the main utterance rhythm. Polish uses lengthening liberally to mark phrasal boundaries, in some varieties multiplying the duration of an average syllable by a factor of five (Karpiński et al. 2008b). This way, durational effects of final lengthening were avoided.

Also, Kim and Cole (2005) found that in American English, correlations between the duration of the foot and the number of component syllables are strongest within an intermediate intonational phrase, rather than across pauses and intonational boundaries, where additional variability is introduced. They suggest that "the foot within the ip [intonational phrase - ZM] is a timing unit where

[10]Preliminary results on inter-rater reliability for selected dialogues between the annotation completed by the author of this thesis and a naive Polish native speaker was 60%. This rather low result shows that more work is needed on the subjectivity of prominence perception in Polish and clear instructions should be given to the annotators. However, for the purposes of this work, it is assumed that the annotation cross-checked by two experts is at least consistent.

a certain level of rhythmic stability exitsts" (Kim and Cole 2005: 2368).

3.3.1.4 Speech rate

Speech rate information was expressed in syllables per second, i.e. mean syllable length in each phrase per second was calculated. No anacruses or final lengthening were discarded for speech rate estimation. Four tempo categories were defined according to quartile ranges of phrase rate: up to six syllables per second (Tempo 1), between six and seven (Tempo 2), between seven and eight (Tempo 3) and from eight up to twelve syllables per second (Tempo 4). All phrases were grouped according to the four tempo categories, coupling strength values were calculated for these tempo groups (Malisz 2011).

3.3.1.5 Speech rate estimation for the analysis of relative coupling strength

The above method proved to be an erroneous approach to speech rate estimation (Michael O'Dell, pers. comm, see also Appendix A) when the tempo classes are used as a basis for coupling strength estimation across rates. By binning data in the aforementioned way, first of all, speech rate and the number of syllables in a foot are directly correlated. Because of that, the intercept does not change, however slope values do change, in the expected direction, down with increasing tempo. Nonetheless, because of the incorrect values of the constant, the coupling strength estimate is biased towards zero, with the slope values in the denominator determining the bias. For this reason, the relative coupling strength results in Malisz (2011) indicated an increasing RPI oscillator dominance with increasing tempo.

Michael O'Dell advised an adequate method of speech rate estimation for the purposes of the coupled oscillators model. The method as originally described by M. O'Dell (unpublished) is included, with permission of the author, in Appendix A. This approach was implemented in this work in the following steps:

a) each Rhythmic Prominence Interval size, from two to five syllables long was inspected separately first,

b) the duration values delimiting each interquartile range per given RPI size were recorded,

c) each RPI size was divided into tempo classes independently, depending on the values delimiting the interquartile ranges,

d) the RPI of different sizes with proportionally determined tempo class factor were then integrated into one dataset for further analysis.

This technique ensures that the changes in RPI duration, as categorised in the tempo class factor, will actually be correlated with increase and decrease in duration determined by speech rate and not with duration changes associated with the number of syllables. This approach of dividing the RPI size into tempo classes proportionally is consistent under one assumption: that for each RPI size (syllable count) slower tempo always means greater duration within each RPI size.

3.3.1.6 Relative coupling strength

The relative coupling strength parameter *r* expresses the interaction between the two coupled oscillators (O'Dell and Nieminen 2009). Relative coupling strength between the stress and syllabic oscillator was estimated empirically by:

a) measuring the durations of Rhythmic Prominence Intervals,

b) counting the number of phonetic syllables comprising the RPIs,

c) estimating intercept and slope coefficients by means of linear regression with the number of syllables as predictors of RPI duration,

d) calculating the relative coupling strength as the ratio:

$$r = a/b \tag{3.5}$$

where *a* is the intercept and *b* is the slope coefficient (Eriksson 1991; O'Dell and Nieminen 2009).

The *r* parameter is interpreted as follows: for values increasing over one the RPI oscillator exerts relatively more pressure on the syllable oscillator, i.e.

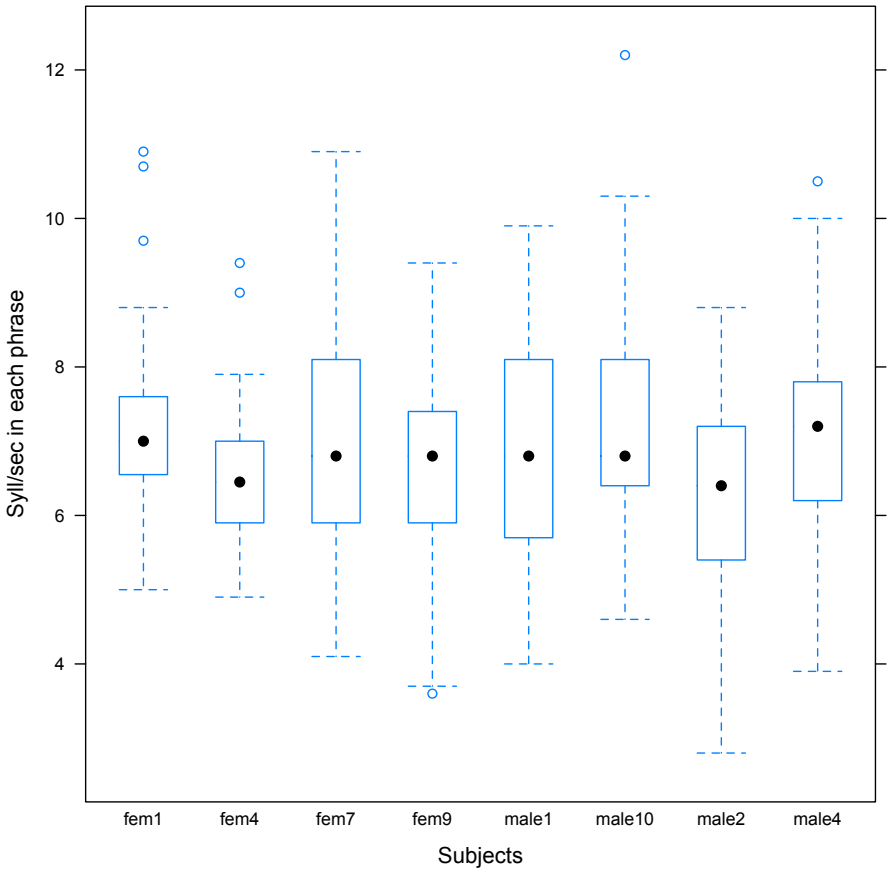


Figure 3.6: Distributions of speech rate in syllables per second for each subject. Dots denote distribution medians.

there is a tendency for the RPIs to equalise their periods and hence the syllable duration has to adapt. For values decreasing below one, the syllable period is relatively more "influential", i.e. the coupling is determined by the syllable oscillator. In this case, RPI duration variability depends more on the syllable count within an RPI; it increases vis-à-vis syllable count more cumulatively. Since relative coupling strength is a ratio, the value of one is not a strict cut-off but indicates a situation where a language or style is neither 'syllable'- nor 'stress'-timed. The slope is dependent on speech rate; the higher the slope the slower the tempo.

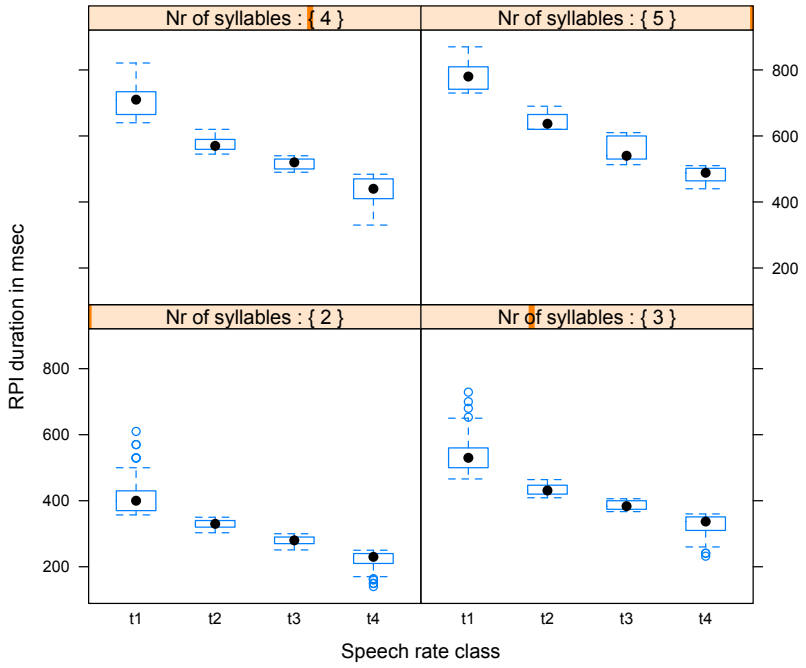


Figure 3.7: Distributions of Rhythmic Prominence Interval durations for four syllable sizes: from 2 to 5 and split into speech rate classes (estimated proportionally to syllable size).

3.3.2 Results

3.3.2.1 General rate effects

Speech rate distributions in syllables per second, normalised within each selected phrase, are shown in Figure 3.6 for each subject. It can be seen that some subjects (e.g.: *male10*) have occasionally reached speech rates as high as 12 syll/sec. Rates around 10 syll/sec were not uncommon. The minimum rate reached was 2.8 syll/sec. However, speech rates between approx. 6 and 8 syll/sec denote the range between the first and third quartiles of the data, with the overall mean at 6.9 syll/sec (median = 6.8, Std. Dev. = 1.34). It can be assumed this range defines the "normal" preferred rate for the studied speakers of Polish. The results are supported by individual speaker rate means that deviate from the overall mean by only 0.4.

Table 3.2: Means and standard deviations of Rhythmic Prominence Interval durations for four syllable sizes: from two to five and split into speech rate classes (estimated proportionally to syllable size).

Speech rate class	Syllable number	mean	Std.Dev.
Tempo 1	*2*	412.1	52.9
	3	538.3	58.7
	4	709.0	50.5
	5	782.75	48.2
Tempo 2	*2*	329.7	13.0
	3	433.7	17.05
	4	576.4	25.6
	5	645.3	28.3
Tempo 3	*2*	279.0	14.1
	3	385.7	12.4
	4	515.3	15.8
	5	560.2	38.0
Tempo 4	*2*	221.5	25.3
	3	325.4	33.05
	4	433.9	41.7
	5	482.4	25.7

3.3.2.2 Rhythmic gradation and speech rate

This analysis uses the method of speech rate estimation that is proportional to the sizes of RPI in syllables, as described in Section 3.3.1.5. First, descriptive statistics of the Rhythmic Prominence Intervals are presented in Figure 3.7 and Table 3.2. It is evident that RPI duration increases with syllable size (there is a positive correlation, cf. Section 3.1.1) and decreases with speech rate.

Next, we analyse the simple linear models of RPI duration as a function of syllable number for each speech tempo group. Model equations, formulae in R (R Development Core Team 2011) and model estimates with correlation coefficients and relative coupling strength values are summarised in Table 3.3. Figure 3.8 presents the results graphically.

The analysis of rhythmic gradation across tempos suggests that Polish has an overall tendency to employ the syllable dominated rhythmic strategy, also across speech rates. Correlation coefficients for the tempo-differentiated analyses are high; given the fully spontaneous data. The best fit is obtained for moderate speech rates. The slope coefficient decreases with increasing speech rate as

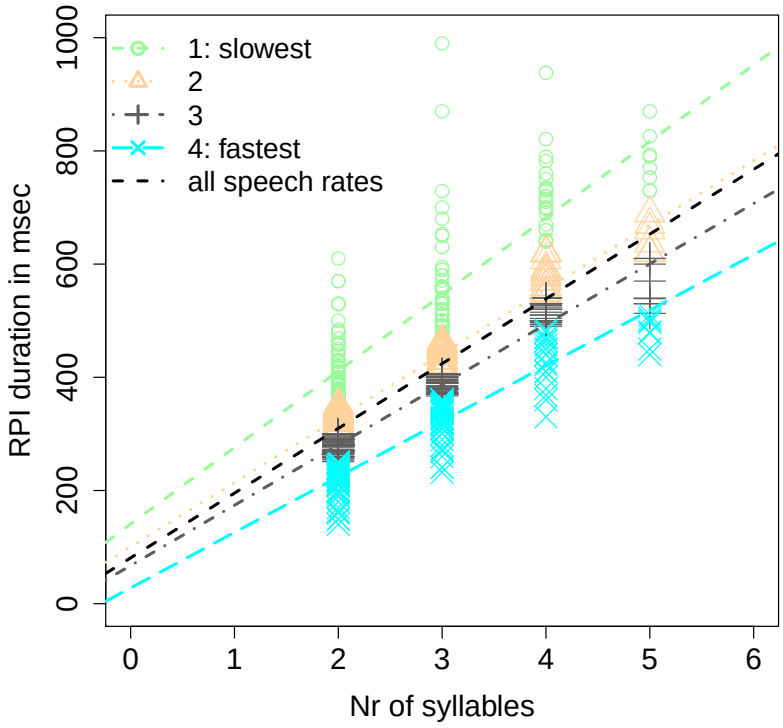


Figure 3.8: Regression results for particular tempo groups (see legend). The black dashed line denotes the linear regression model for all tempos.

assumed by O'Dell and Nieminen (1999). All relative coupling strength values resulting from the regressions on tempo differentiated data stay in the syllable-dominated range. The non-differentiated value in the model for "All" data in Table 3.3 and Figure 3.8 can be taken to identify the language typical strategy and equals $r = 0.73$, slightly higher than the overall values for Spanish ($r = 0.64$), as given by O'Dell and Nieminen (1999).

For the slowest tempo (Tempo 1), we see that the intercept value is closer to the stress timed values, when compared with the values that Eriksson (1991) found for stress timed languages in Table 3.1, i.e. the constant is higher. This means that some syllables in the slowest RPI might receive double the duration than in normal tempos. A much steeper slope is also observed here. There is an about three times higher difference in slopes (b difference equals 25.5 msec)

Table 3.3: Simple linear models of Rhythmic Prominence Interval duration as a function of the number of syllables, for each speech rate separately.

Speech rate class	Regression equation	Coupling strength	Corr. coeff.
Tempo 1: slowest	$I = 141 + 135n$	$r = 1$	$adj.R^2 = 0.81$
Tempo 2	$I = 101 + 113n$	$r = 0.9$	$adj.R^2 = 0.96$
Tempo 3	$I = 70 + 105n$	$r = 0.66$	$adj.R^2 = 0.96$
Tempo 4: fastest	$I = 28 + 98n$	$r = 0.28$	$adj.R^2 = 0.87$
All	$I = 83 + 113n$	$r = 0.73$	$adj.R^2 = 0.55$

Table 3.4: Regression on slopes and intercepts resulting from speech rate differentiated models in Table 3.3.

Coefficients	Estimate	Std. Err.	*t*-value	*p*-value
(Intercept)	438.04	50.84	8.62	$< .01$
c(1/b)	-39242.94	5613.83	-6.99	$< .01$
Adjusted R-squared: 0.941				

between the slowest speech Tempo 1 and speech Tempo 2 than the difference between the other slope values for Tempo 2 and Tempo 3 (b difference equals 7 msec) and between Tempo 3 and Tempo 4 (b difference equals 8.45 msec). The RPI duration is more greatly increased with each syllable added to the RPI in the slowest tempo than in other tempos.

The intercept in the fastest Tempo 4 is very low ($a = 28$) and overall, the intercept estimates fall quite rapidly with increasing speech rate. Relative coupling strength r is decreasing with tempo, indicating that the syllabic oscillator gains strength over the RPI oscillator with increasing rate. In order to confirm this result a further regression on slopes and intercepts was performed, as described in the first procedure in Appendix A. The coefficients *a* and *b* from Table 3.3 were entered into a simple regression where the intercepts were predicted by $(1/b)$. The relationship between tempo and r is estimated by the coefficient c. When the value of c is negative, coupling strength significantly decreases with increasing tempo, if c is positive, r increases. As seen in Table 3.4, the coefficient is significant and negative (p-value$= 0 < .01$). The syllable oscillator increases its strength over the RPI oscillator as speech rate increases.

3.3.2.3 A Rhythmic Prominence Interval duration model

Models with an interaction, with a quadratic term and with both an interaction and a quadratic term were compared to a simple additive model by means of ANOVA. The model including both higher order variables turned out to have a significantly better fit to the data (p-value$< .001$). Equation 3.6 presents the predicted RPI duration regression model as a weighted sum of predictor variables for a given syllable number in an RPI and a given speech rate class where:

a) b is the syllable number coefficient

b) c is the speech rate coefficient

c) d is the quadratic term coefficient

d) e is the speech rate coefficient in an interaction with syllable number

e) a expresses the intercept

$$\hat{I_n} = a + b(n) + c(speechrateclass) + (d(n)^2) + n(speechrateclass(e)) \quad (3.6)$$

By plugging in coefficient values from Table 3.5 a prediction of RPI duration estimates is possible. The model can be used to provide weights of the various factors that influence Rhythmic Prominence Intervals and can be tested in speech synthesis systems with a prominence component (Windmann et al. 2011).

3.3.3 Discussion

The mean value for speech tempo in Polish found in the present experiment, approx. 7 syll/sec corresponds to what was empirically found by Dellwo and Wagner (2003) to be the normal speaking rate for French subjects (normal range between approx. 6.3 and 7.2, mean = 7.3 syll/sec). Dellwo and Wagner (2003) explain that crosslinguistic differences in the normal speaking rates observed in their study (with English at mean = 5.9 and German mean = 5.6) might depend on the complexity of phonotactic structure. It is certainly plausible that the simpler syllable

Table 3.5: A multiple regression model of Rhythmic Prominence Interval duration as a function of the number of syllables and speech rate. Model Equation 3.6. Reference level for Tempo: "Tempo 1".

Variables	Estimate	Std. Err.	*t*-value	*p*-value
(Intercept)	98.090	17.194	5.705	$< .001$
Syllable number	164.388	10.762	15.275	$< .001$
Tempo 2	-38.685	12.491	-3.097	0.002
Tempo 3	-70.767	12.257	-5.773	$< .001$
Tempo 4	-112.515	12.258	-9.179	$< .001$
$I(Syllablenumber^2)$	-4.714	1.637	-2.879	0.004
Syllable number*Tempo 2	-21.783	4.354	-5.003	$< .001$
Syllable number*Tempo 3	-29.226	4.276	-6.836	$< .001$
Syllable number*Tempo 4	-36.641	4.288	-8.545	$< .001$

structure of French allows for a more liberal management of articulation rate than e.g. English. Consequently, Polish complex syllable types and lack of vowel reduction could place it in the less tempo "flexible" group of languages, with English or German. In fact, frequencies of complex syllable types in Polish are rather low: simple syllables (CV, CCV, CVC, CCVC) predominated in a large corpus analysed by Klessa (2006). Gibbon et al. (2007) suspected that the regularity of syllabic intervals in their Polish corpus (nPVI = 38) could be accounted for by a Zipf effect: the longest clusters are rare.

The properties of the model in the slowest tempo and a higher intercept value may come from the effect of accentuation on stressed syllable duration in Polish. As Malisz and Wagner (2012) showed, duration significantly manifests itself only on prominence level two in i.e. "strong prominence" (Section 1.4.3). Strong prominence and accentuation structure only significantly appear as correlates within the duration domain in slower, more elaborate tempos. Phrasing in slower tempos imparts a more temporally varied delivery that seems to increase the coupling strength towards the RPI oscillator. Apart from the behaviour in the slowest rate, Polish speakers in task-oriented dialogue exhibit an increasingly syllable-dominated timing with rising speech tempo demands.

The above explanation is compatible with the hypothesised task-dependent nature of rhythmic patterning proposed in Section 3.2. The proposal states that the correlated factors of speech rate and speech style are a manifestation of an order parameter able to reveal rhythmic strategies, expressed by coupling strenght, as

a continuum within languages. Fundamentally, such treatment implies that both "syllable timing" and "stress timing", conceptualised as rhythmic strategies, can be evidenced in languages when these parameters are manipulated.

Malisz et al. (2013) provide further evidence for this claim. They compare results obtained from the present corpus of spontaneous dialogues in Polish to models calculated for prepared speeches of several prominent Polish figures and estimate the relative coupling strength for these speeches. The r values calculated for the particular speakers are then used as a variable predicting the rate of glottalisation in Polish, among other factors. The same analysis is done, on similarly stylistically diverse material for German, a language described as stress-timed, typologically uncontroversial and similar in rhythmic structure to English. The resulting variation in the coupling strength between the RPI cycle and the phonetic syllable cycle approximates the complexity of rhythmic variation between languages and speech styles in a complete and continuous way. In particular, the results obtained by Malisz et al. (2013) indicate that Polish speakers and prominent figures, taken together, straddle both sides of the continuum determined by coupling strength, while German speakers stay in the stress oscillator dominance territory. The Polish data in fact presents a case where the use of both timing strategies, is able reveal a rhythm effect on glottal marking, apparently not evident in German, that exhibits rather uniform stress timing across styles (but not a "fixed" timing). The difference in r values between speech styles is significant in Polish. The distribution of r values for Polish in dialogue reported in Malisz et al. (2013), analysed for four speakers of the present corpus (here: eight speakers), ranged from extremely syllable oscillator dominated (r = 0.3) over to 'indeterminate' (values around 1) to somewhat stress-timed ($r = 1.25$). In prepared speeches the values ranged from $r = 0.85$ to $r = 2.8$. This means, that in prepared speeches, the RPI oscillator dominates over the syllabic one to a considerable degree.

The reported analysis can be summarised as follows: the rhythmic variable r approximates the relative coupling strength between the phonetic syllable and the Rhythmical Prominence Interval. These two intervals are nested, interacting rhythmic cycles that are assumed to express two coupled oscillators relevant for the rhythm of Polish. Relative coupling strength provides a single parameter that

quantifies the relevant rhythmic variability. The parameter values indicate that in Polish, the syllable oscillator dominated over the RPI oscillator and that its influence increases with speech tempo. In very slow tempos, the RPI cycle starts to prevail. In English, as O'Dell and Nieminen (2009) report, the relative coupling strength testifies to a stress cycle dominated strategy with $r = 2$. Additional evidence suggests that the stress-timing strategy in Polish might be characteristic of formal and slow styles of speaking. However, the relationships between speech style, speech rate and relative coupling strength demand further analysis.

Chapter 4: Rhythmic constituency and segmental duration

4.1 Introduction

Tuller et al. (1983) suggested that a vowel-to-vowel articulatory period can be posited for English both on a higher and lower timing level. Dynamical models of rhythm based on coupled syllable and stress oscillators (Barbosa 2006; O'Dell and Nieminen 1999) model the hypothesised rhythmical function of the vowel cycle in rhythm production. Evidence for articulatory timing based on the vowel-to-vowel cycle has been found by several phonetic studies. Spectrographic evidence was provided by Öhman (1966) where changes in $F2$ transitions of intervocalic consonants were accounted for by the spectral influence of the flanking vowels. However, what is quite important for the analysis of a Slavic language such as Polish, the coarticulatory effect across stops in Öhman (1966) was found for Swedish, English but not Russian. X-ray data from Browman and Goldstein (1990) showed a superimposition of consonants on vowels in simple monosyllables. Functional separation of vowels and consonants was suggested by kymographic evidence in Stetson (1951). An explicit model of vowel-to-vowel timing can be found in Fowler (1983) where the abstract use of vowel onsets as perceptual centre (p-centre) targets was first proposed.

The speech rhythm model in Barbosa (2002) and Barbosa (2006) is specified by the coupling of syllable-sized and phrase stress oscillator, where the first provides regularity and the other structure. Drawing on the above evidence on vocalic cycle articulation, the syllabic oscillator in Barbosa's rhythm model is implemented by vowel-to-vowel units. The model assumes there is a tendency to regularise the recurrence of vowel onsets in speech. Duration compensation phe-

nomena, such as, e.g. the voicing effect inside the vowel-to-vowel unit support the hypothesis of a relatively stable intervocalic onset period. The model suggests therefore that longer vs. shorter consonants within a vowel-to-vowel frame act as perturbing factors observed in the measured vocalic durations. These paradigmatic perturbations are corrected at least as an observable tendency by compensation phenomena.

Some effects of the superior unit on the subconstituents in Polish are subject of investigation in in this chapter. The voicing effect is tested first, as a manifestation of a rhythmic tendency to regularise the vocalic cycle. The next experiment looks at the behaviour of subconstituents in a context of a long consonant. In both cases it is hypothesised that a degree of duration compensation on the part of the vowel will occur, balancing the overall VC group duration. This hypothesis refers to the view that there are periodic control structures in speech dicussed in 3.1.3 on the level of the syllable-sized oscillator (Barbosa 2006).

4.2 The voicing effect

The voicing effect, i.e. the lengthening of a vowel preceding a voiced consonant, is a very well documented phenomenon in many languages (Chen 1970), especially in the Germanic family (Kohler 1977; Port 1981; Port and Dalby 1982; Port and O'Dell 1985). As Keating notes (Keating 1985), the phenomenon is a near phonetic universal but its mechanism is not well understood. In English we find the ratio of vowel preceding a voiceless consonant to preceding a voiced consonant in a range from 0.89 (sentence contexts) (Port 1977) to 0.69 (word lists) (House and Fairbanks 1953). An "exaggeration" of the phonetic universal was mentioned for English by e.g.: Keating (1985); Ohala and Ohala (1992) and Mitleb (1984), that is believed to result in a phonological rule (de Jong and Zawaydeh 2002). In Polish, no phonological vowel lengthening before voiced consonants was reported. In fact, Polish appears not to show vowel duration difference in the stop voicing context at all (Keating 1979, 1985), defying also the universal phonetic tendency for at least a 10% difference (Chen 1970; Keating 1985).

Given the above, a general issue arises whether the effect is dependent solely on physiological aspects of articulation and phonation (Kohler 1977; Klatt

1976), i.e. is it phonetic and universal, or is it specified in the phonology (Mitleb 1984). The former case is unlikely, given its apparent lack in Polish and other data, e.g. from Arabic (Port et al. 1980; Mitleb 1984; Flege and Port 1981) and Czech (Keating 1985). In the latter case, both the "exaggeration" (English) and the apparent "suppression" (Polish) of the effect would have to be implemented as part of the phonology.

In summary, we find a range of patterns with, e.g. Polish at one extreme and English at the other, plus, the majority of languages espousing the phonetic effect, located in the middle. What we do not find is a language that shows, e.g. a lengthening in the context of a voiceless consonant (Keating 1985). Consequently, in all existing scenarios, universal phonetic facts (the "default" pattern) interact with language specific timing constraints, as suggested by Keating (1985) and Port et al. (1980).

Other questions can be raised. Can the specific pattern depend on the function of the effect in a given language? Are the functions correlated systematically or just coincide? What structural level(s) do they operate on? What are the exact physiological aspects that underlie the phonetic "default" and how can they be "suppressed"? Both the physiological aspects as well as the specific role of the effect have not been conclusively explained.

There is an important relation that is especially relevant for the present thesis. The (lack of) difference in preceding vowel duration as a function of consonant voicing, a subsyllabic phenomenon, nonetheless participates in what Port et al. (1980) call "temporal microstructure". A possible explanation of the voicing effect patterning in these terms has been suggested by Keating (1985), others explicitely related it to the higher levels than segmental (Port et al. 1980; Lehiste 1977; Barbosa 2006). Therefore, a few possible functions of the voicing effect on the subsyllabic level will be discussed below. But a special focus will be placed on its hypothesised involvement in rhythmic structuring. After providing the background for an experiment on the voicing effect in Polish, results on vowel and consonant durations and ratios in voicing contexts will be presented and implications for both timing and rhythmic processes will be discussed.

4.2.1 Preceding vowel duration as a consonant voicing cue

The inverse relationship of duration between vowels and consonants often functions as a cue to the following consonant voicing (Lisker 1986; Port and Dalby 1982). Apart from the duration of the preceding vowel, voicing contrasts may depend on a number of other temporal parameters manifested in the acoustics. As many as 16 other cues have been suggested by Lisker (1986). The most important parameters are listed below:

a) consonant closure duration, where voiced obstruents are known to be shorter than voiceless (Luce and Charles-Luce 1985; Ohala 1983);

b) post-release voice onset time (VOT) and pre-release voice offset time (in pre-aspirated consonants) (Lisker 1978; Pind 1995);

c) the duration of voicing during the closure (Lisker 1986) (the duration of the "voice bar" relative to closure duration).

There have been discussions concerning the primacy of cues in language specific phonologies (Braunschweiler 1997; Port and Dalby 1982): are they combined into a percept of voicing or is there a primary cue? Port and Dalby (1982) have suggested that the C/V duration ratio allows for the number of temporal parameters to be reduced and is stable across speech rates.

According to Keating (1979), in Polish, voicing contrasts are primarily expressed by voicing throughout the closure with leading VOT in voiced obstruents vs. short-lag VOTs and little aspiration noise in voiceless obstruents, both in perception and production. Closure duration was found to overlap between /t/ and /d/ analysed in her study in medial stop voicing contrasts, even in read minimal pairs. However, the mean voiceless closure duration was significantly longer than the voiced closure duration (130.1 msec for /t/ and 91.5 msec for /d/) in the same study.

In summary, Polish and English respectively show a prevoiced vs. short lag VOT patterns for the voicing cues, short lag vs. long lag VOT patterns for the voicelessness cues, overlap vs. no overlap in closure durations between voicing categories and presence vs. absence of a consonant voicing effect on preceding

vowel duration. Additionally, in English, as mentioned above, the vowel duration cue to voicing is linguistically specified (de Jong and Zawaydeh 2002).

Kohler (1977, 2007) argues that a force feature differentiates between voiced and voiceless plosive classes in German and English where a greater articulatory force is required to produce voiceless consonants (*fortis*) than voiced (*lenis*). Additionally, voiceless stops induce a fast closing movement causing the preceding vocalic portion to close faster, i.e. the vowel is effectively shorter before a voiceless rather than before a voiced plosive. Klatt (1976) also proposed that an early glottal opening for a postvocalic voiceless stop is made ensuring that no low-frequency voicing cue is generated during an obstruent. The above would explain the physiology of the effect that takes place in English and German[1]. Keating's proposal above indicates that prosody potentially atenuates the voicing effect in Polish.

However, as noted earlier, the effect could not be claimed to be a phonetic universal. Polish and Arabic have been generally reported not to exhibit the effect. There are conflicting accounts however, concerning both these languages. Port et al. (1980) found that vowels lengthened in front of voiced stops in Arabic, in contrast to fricatives, according to Mitleb (1984) who found no such variation. Keating (1979) studied consonant voicing categorisation in Polish and as mentioned above, found that the phonemic voicing effect not only does not occur before stops but also in Czech (Keating 1985). Even though the difference of length between the consonants was significant, the longer voiceless plosives did not trigger compensation.

The function of the effect can also reside in a broader perceptual contrast as suggested by Kluender et al. (1988). The inverse relationship between vowel and consonant duration in voicing contrasts has been explained with a "principle of duration contrast" notably by Kluender et al. (1988). The principle states that segments following longer segments sound shorter than following shorter segments. In case of the voicing contrast, the principle would serve to enhance the differ-

[1]In German the voicing effect occurs even in case of final voicing neutralisation. In Polish the only study that reported results on a vowel effect in the context of final voicing neutralisation, and consequently claimed it is not a full neutralisation, is Slowiaczek and Dinnsen (1985). The study has been criticised for its methodology: using word lists where ortography might have led the participants to exaggerate the underlying contrast (Jassem and Richter 1989).

ences between voiced and voiceless closure durations. Fowler (1989) challenged this view claiming that in fact no such principle is evidenced in auditory processing, especially in speech, and showed experimental results where a reverse effect to the durational contrast effect was reported: closures following longer vowels were judged as longer (cf. "time shrinking", Section 1.2).

4.2.2 Syllable duration balance as a micro-prosodic function of the voicing effect

The function of the voicing effect has also been defined as a duration compensation between constituent segments. Voiceless consonants are usually longer than voiced (Ohala 1983) and the inverse relationship of vowel and consonant duration suggests a compensatory effect on the part of the vowel. The compensatory effect was suggested to be a phonetic representation of temporal planning and signaled to be relevant for the balance of syllable duration (Keating 1985; Lehiste 1977). Interpreted this way, the voicing effect "parameter" would be subject to prosodic constraints that are either universal, e.g. following the eurythmic principle, or are specific to the phonology. In the former case, depending on rhythmic type, syllable duration might be subject to stronger or weaker compensatory effects. Keating (1985) uses the contrasting examples of Polish and English to explain the notion:

> In English and presumably in other languages with vowel lengthening, the two ratios, vowel and closure, essentially balance each other, so that the syllable duration is relatively constant. (...) Polish, like English, has longer closure durations for voiceless stops. Because Polish shows the closure but not the vowel effect, its syllable durations are not balanced. (...) language-specific prosodic factors like stress or rhythm could make it desirable to balance intrinsic syllable durations. This factor may operate more powerfully in a language like English, with variable stress and vowel reduction, than in a language like Polish, with fixed stress. (Keating 1985: 122)

Tuller, Kelso, Harris (1983) suggested that in compensatory and coarticulation phenomena the consonants shorten the observed vocalic durations by being overlaid on a continuous vowel cycle, masking the vocalic intervals in the ongoing cycle but not interrupting their production. Compensation phenomena such as the voicing effect would then be a manifestation of co-articulatory tendencies on the

acoustic surface and, as a tendency, in fact happen in the majority of languages (Chen 1970).

In Barbosa (2006) study of Brazilian Portuguese, interpreted for the purposes of the rhythm model, the relative imbalance in duration between segments in the voicing effect environments, tended to trigger temporal adaptation on the part of the vowel. Moreover, in case of combinations of voicing, manner and place in the following context, reflecting significant and sizeable durational differences between consonants, as in: /CVrV/ vs. /CVSV/, vowel duration compensation occurred as well.

Barbosa notes that there exist languages that do not exhibit the voicing effect and therefore questions about how durations within the V-to-V unit are balanced can be asked. The V-to-V unit can be defined as a "phonetic syllable" and so both Barbosa (2006) and Keating (1979) essentially suggest a similar role of compensation phenomena as having a "syllable balancing" function preparing for the structure imparted by stress effects on duration to arise.

Barbosa raised objections to Keating's work on Polish (Barbosa 2006: 49) suggesting that a reading of isolated words in her experimental design might have influenced the results, e.g. with emphasis effects. A verification of Keating's results is undertaken in this thesis in order to establish how the assumed "inflexibility" of the Polish vowel-to-vowel units in this context might affect Barbosa's model. Stimuli containing stops and fricatives are also used to investigate the interactions between manner and voicing that could potentially affect variation, as indicated for e.g. Arabic.

The need to further test the mechanism in Polish is also highlighted by other studies that claim the presence of the effect in the language. Richter (1973) found it existed for vowels in isolated nonsense words. Similarly, Imiołczyk et al. (1994). Klessa (2006) however notes that vowel durations are shorter for continuous speech than in the logatomes studied in Richter (1973) indicating that the voicing effect may be an effect of hypercorrection (cf. Slowiaczek and Dinnsen 1985; Jassem and Richter 1989).

Recent Polish corpus data results (Breuer et al. 2006; Nowak 2006b) confirm only some contextual voicing effects on Polish mean vowel duration. In a study on Polish segmental duration (Breuer et al. 2006; Klessa et al. 2007) for

speech synthesis purposes over fifty features were tested both from the segmental and suprasegmental levels with a Classification And Regression Tree algorithm to verify the correlation with phone duration. The influence of the right context was rated high in the rankings obtained from three large corpora for two feature vectors. The identity of the sound directly following the sound in question appeared to be one of the two most important features within the feature vector, and manner of articulation was one of the first ten most important features. More complex interactions between the other fifty features were not discussed but are certainly relevant.

The author is not aware of any previous studies investigating mutual contextual duration effects of consonants and vowels in Polish with a focus on temporal compensation from a prosodic point of view. A recent study (Machač and Skarnitzl 2007) on Czech VC and CV sequences reports some temporal adaptation inside the Czech VC and CV. However, the Czech material was not controlled for lexical stress, emphasis and other factors, since it was based on data extracted from a corpus. This fact may explain the results that testified to compensation within CV syllables, as they might have been stressed. The present study on the duration effects within the Polish VC is more strictly controlled (see below).

4.3 Experiment 2: the voicing effect in Polish[2]

4.3.1 Data and methods

The dataset used to verify the hypothesis was a set of phrases produced by native speakers of Polish with target items in controlled positions. In the dataset the durational compensation effects or the lack thereof in VC groups within V-to-V units was studied. All analysed data were labeled according to prosodic annotation rules (cf. 4.3.1.1).

Eight speakers (four male and four female; 21-30 years old) of standard Polish were asked to repeat stimuli around 20 times (around two thousand tokens were recorded and annotated). In this experiment, the influence of Polish alveolar, retroflex and palatal fricatives /s, z, S, Z, s', z'/ on preceding vowel duration was investigated. The particular set of consonants was chosen because these particular

[2]Preliminary results of this experiment were published in Malisz and Klessa (2008)

fricatives are known to differ more than stops in their inherent durations within the voicing contrast in Polish (Klessa 2006). Only medial VC units were considered since Polish consonants do not contrast in voicing in final position: all voiced obstruents are devoiced word finally. Contexts with stops /p, b, t, d/ were included as well, for comparison with Keating's original work, which investigated /raCa/ type of stimuli with stops contrasting in voicing. Target words of the form kaCa were used, where C was one of the six fricatives or four stops under study. Usually z-score normalisation is used in measurements of vowel duration in materials and corpora containing different vowel phonemes. The data in the present study was non-normalised because there was only one vowel phoneme type studied: /a/. Raw or log-transformed, when indicated, durations were compared: log-normal transformation was shown to deal with positive skewness often encountered in the distribution of interval data (Rosen 2005).

The stimuli were presented on randomised strips of paper in the following form: "To nie jest kasa, to kasa", "To nie jest kaza, to kaza" The words to be filled into the gaps were given on separate strips of paper: "dobra", "tania" ("good", "cheap") for the first condition containing meaningful target words (e.g.: "kasa", "rasa", "Kasia") and "lośna", "dąpna" (no meaning) for the second condition containing nonsense target words ("kaza"). The two pairs of additional stimuli functioned as masking words. The nonsense target words are phonotactically legal sequences in Polish. In a few preparatory runs the speakers did not have problems with incorporating the nonsense target words into a meaningful carrier sentence, producing fluent utterances. The stimulus design placed target words in a carrier sentence in order to elicit a more natural speaking style as opposed to a word list used in Keating (1979). The main task was constructed so that the target word would not attract prominent focus, assigned to other potential locations in the frame. The "new information" status and position in the phrase attracted phrasal prominence to the masking words rather than to the target word. This way emphasis was controlled pragmatically. Gibbon et al. (2007) suggest that the lack of contrastive length in Polish might provide an extra degree of freedom for use in emphasis. Observationally, emphasis by lengthening is liberally used in Polish, an observation confirmed by the analysis in Malisz and Wagner (2012) discussed in Section 1.4.3 where phrasal prominence was correlated with

significant differences in duration of prominent syllables relative to others while lexical level prominences did not. The stimuli design and presentation control for the potential phrasal emphasis effects on duration.

The recordings were made in a sound-treated room using an MXL-700 cardioid condenser microphone connected to a PC running Windows XP Professional via an Edirol UA-25 USB audio interface. They were digitised at a sampling frequency of 44.1 kHz and bit depth of 16 bits in Audacity.

In summary, the experiment was controlled for the following factors that influence segment duration: vowel quality, syllable shape and count, position in the phrase, lexical stress, emphasis. It was not directly controlled for rate, however ratio measurements and mixed effects modeling are used further for analysis, methods that, respectively, normalise for the possible rate effects or correct individual extreme variation in the random effects structure.

4.3.1.1 Annotation and measurement

In the controlled speech data the signal was segmented manually using Praat (Boersma and Weenink 2012) speech analysis software and annotation tools. Oral constriction criteria were applied as guidelines for prosodic annotation as described in Turk et al. (2006). Oral constriction criteria correspond better to the objective of investigating temporal relations as they more closely correspond to the motor tasks that a speaker must dynamically tackle in the production of VC groups. The annotation procedure in this case is different from segmentation based on only acoustic cues, e.g., in the following cases:

a) a silent transition interval following voiceless fricative noise is included in the following vowel interval,

b) a silent interval occurring after a vowel, before a voiceless fricative noise sets in, is counted as belonging to the vocalic interval.

It is necessary to add, as target words beginning with /k/ were used, that the annotation standard used here sets the final boundary for /k/ at the consonant release rather than at the beginning of voicing. Any short transition period, not more than

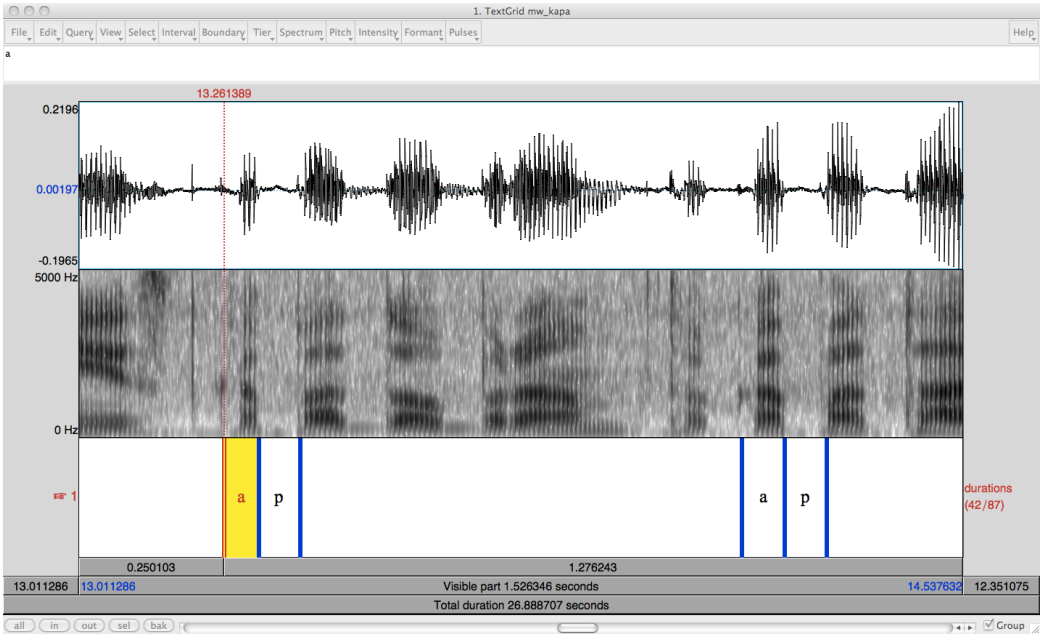


Figure 4.1: An example annotation of two repetitions of the "kapa" stimulus in a carrier phrase.

10 ms on average though, was included in the following vocalic interval. Screenshots in Figure 4.1 and Figure 4.2 show examples of Praat annotations of the recorded stimuli following the criteria described above. In general two intervals were marked in each recorded stimulus word:

a) the vocalic portion, from the initial consonant release (first boundary) to the beginning of the closure of the second consonant (second boundary),

b) the closure and release of the second consonant in the word. In case of stops at the end of the burst (third boundary) and in case of fricatives at the end of turbulent noise.

The durations of the intervals and their labels were extracted using a Praat script. The resulting experimental database was analysed using the R statistics software (R Development Core Team 2011).

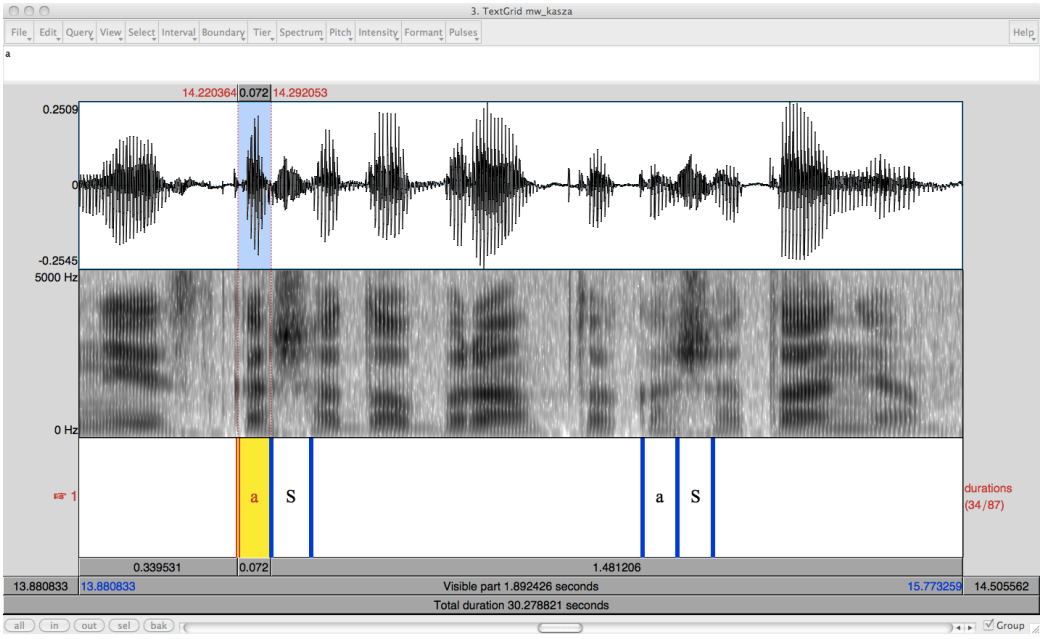


Figure 4.2: An example annotation of two repetitions of the "kaSa" stimulus in a carrier phrase.

4.3.2 Results

4.3.2.1 The voicing effect preceding fricative consonants

Figure 4.3 presents the consonant-consonant and vowel-vowel duration ratios for the kaCa stimuli in the experiment across speakers. The ratios were calculated by taking the mean duration of the vowels preceding voiced and voiceless consonants respectively and then forming the ratio (not by averaging individual ratios). The segments in contrast pairs, e.g. "kaza" vs. "kasa" (/z/ vs. /s/ and the corresponding preceding vowels), are of equal length when their ratio is 1. In case the vowel lengthens in front of a voiced consonant, then voicing effect exists and the vowel-vowel ratio amounts to more than 1.

In the fricative dataset 1445 samples were analysed. On average, the studied voiced fricative consonants were shorter than the voiceless ones by approx. 34% (ratio mean for consonants $= 0.66$, Std.Dev. $= 0.055$). The duration of the preceding vowels lengthened by 14% (ratio mean for preceding vowels $= 1.14$, Std.Dev. $= 0.07$, see below for statistical tests). Compare this to a 20% lengthening of vowels preceding voiced *stops* in German as found by Braunschweiler

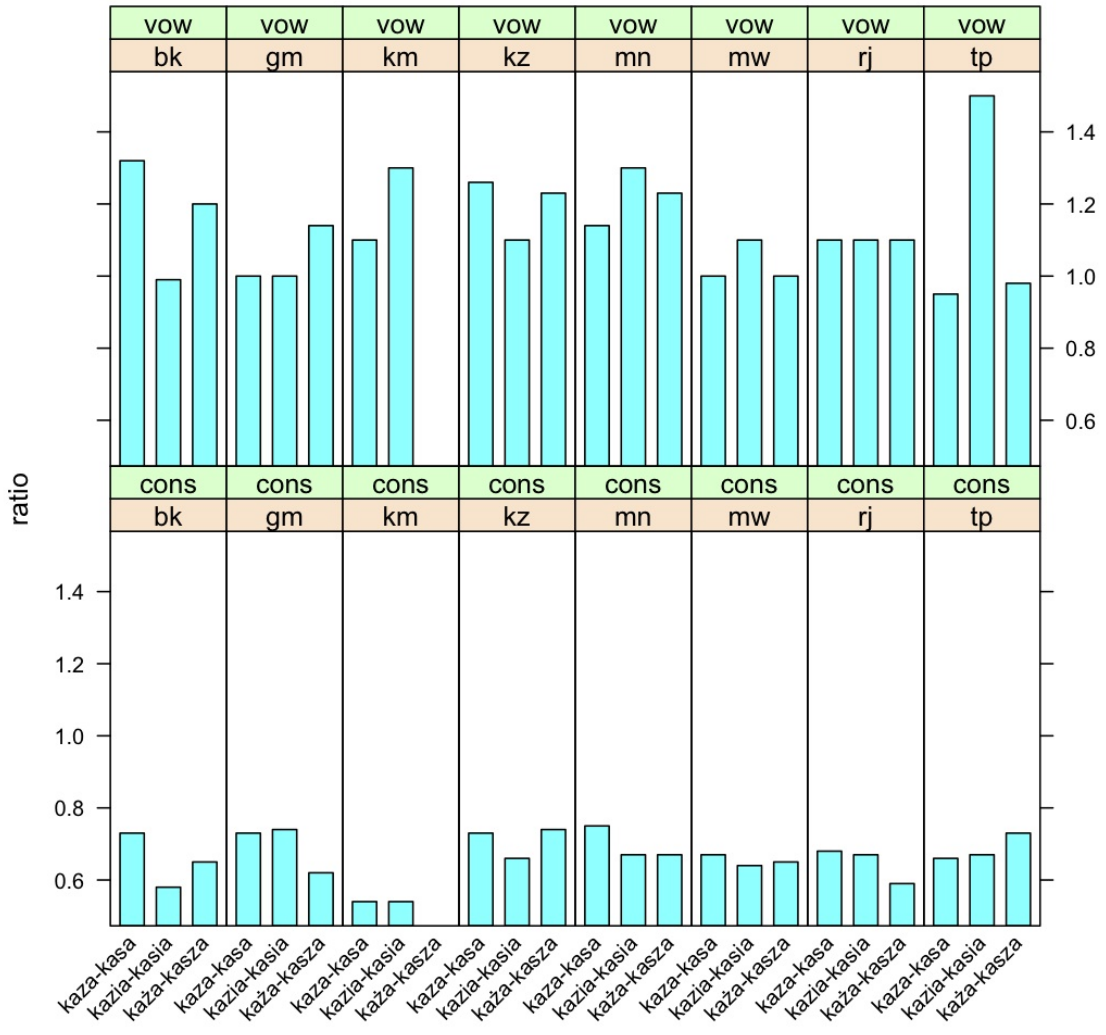


Figure 4.3: The ratio between the mean values for the vowels preceding a voiced and voiceless fricative in the kaCa stimuli (top panel); ratio between mean values of the voiced and voiceless fricative consonants in the kaCa stimuli (bottom panel) for each speaker.

Table 4.1: Means and standard deviations of vowel durations in miliseconds for each speaker in the fricative voicing condition.

	kaza		kasa		kazia		kasia		każa		kasza	
Subj	mean	St.Dev.	mean	St.Dev.	mean	St.Dev.	mean	St.Dev.	mean	St.Dev.	mean	St.Dev.
bk	71.4	20.4	54.3	12.7	65.5	19.2	65.9	21.8	74.2	24.2	61.9	21.0
gm	89.3	30.8	85.4	27.2	100.3	34.0	95.5	31.8	94.6	29.0	82.7	33.2
km	107.5	11.2	96.8	10.4	124.5	13.7	97.7	10.2	112.0	10.6	NA	NA
kz	147.3	45.2	116.9	52.2	137.6	33.8	124.9	46.9	149.8	53.0	121.4	36.2
mn	85.0	8.9	74.5	10.0	97.5	13.3	75.5	8.2	85.7	8.1	69.5	8.25
mw	77.5	12.5	75.7	6.9	84.4	7.5	76.9	7.9	76.8	7.8	73.2	7.8
rj	85.0	9.5	80.0	10.2	94.2	11.7	85.8	9.0	84.5	9.4	78.5	13.5
tp	95.8	8.4	101.0	19.8	133.8	32.8	90.0	24.7	90.0	9.4	91.35	19.7
All	101.2	38.9	86.6	31.3	109.4	34.8	91.3	32.2	98.6	37.8	85.8	32.0

(1997). However, the V/V ratio in the pre-voiced/pre-voiceless context respectively, displays speaker dependent behaviour where some speakers display values closer to 1 in some contrasts.

Table 4.1 presents the means in miliseconds for the vowels in the fricative kaCa set. The durations had a non-normal distribution (after log-transformation, one sample Kolmogorov-Smirnov test, $p < .001$) therefore a non-parametric Wilcoxon rank test was used to establish whether the means are the same. In all voicing contrast comparisons the means were significantly different at $p < .001$.

Individual timing strategies of the different speakers need to be regarded more closely. As can be seen in Figure 4.4, some speakers (*bk, kz, mn, km*) exhibited a tendency to differentiate vowel durations in the fricative task. Some speakers (*rj, mw*) did not show a great tendency to differentiate durations in the given contexts at all. Speakers *rj, mn* and *mw* showed also a low variability in the productions. The sentence realizations of these speakers were most uniform in terms of pitch changes and phrasing. Despite efforts to control for higher level prosodic factors such as emphasis, as it happens, speaker *tp* sometimes varied her productions more from sentence to sentence, in a less repetitive, more "illustrative" style.

In order to assess how much of the variance can be attributed to the main effect of consonant voicing while leaving out the idiosyncrasies of the speakers, linear mixed effects models (LMEM) were used (instead of ANOVA) to analyse variance. LMEMs are able to handle unbalanced datasets e.g.: with missing data

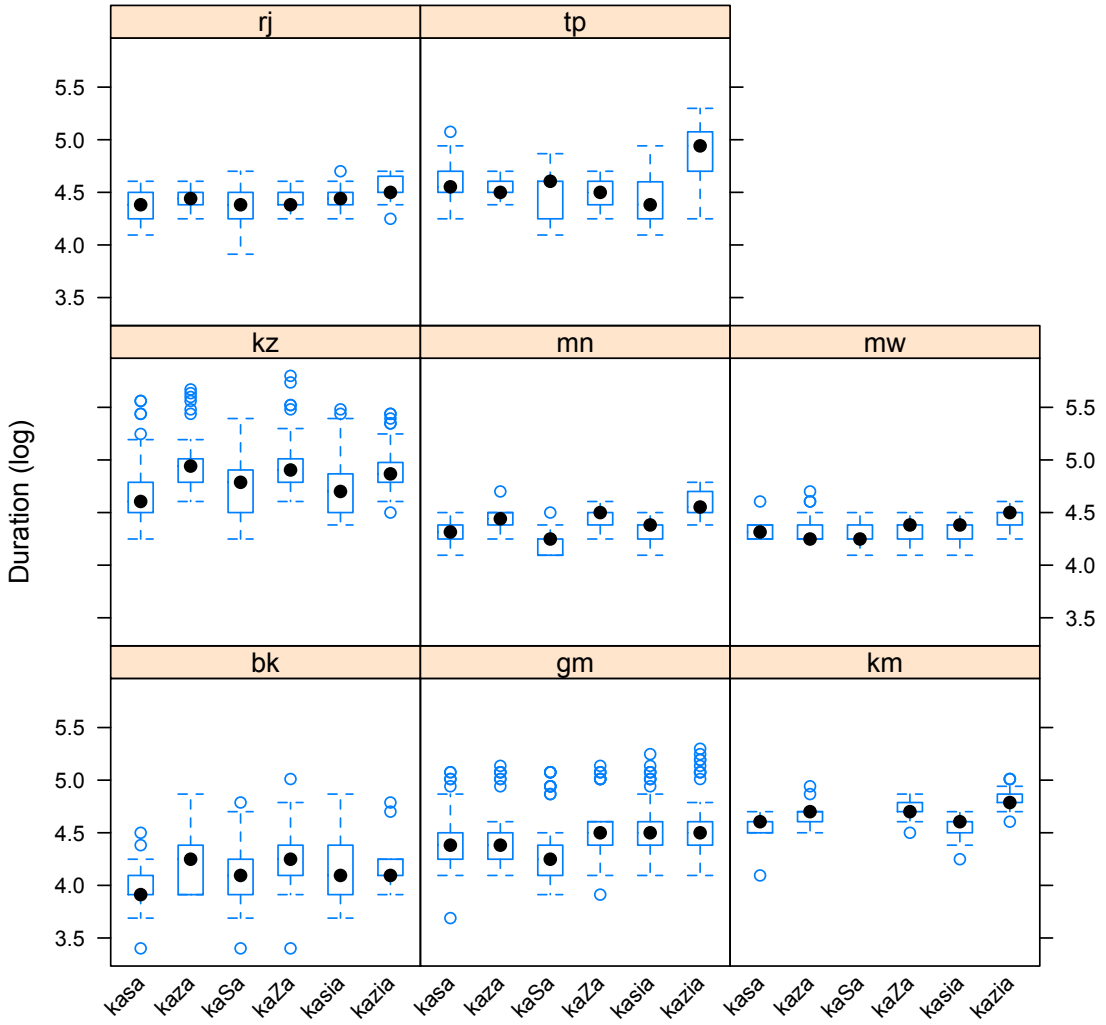


Figure 4.4: The distributions of vowel duration (log-transformed) preceding a voiced or voiceless fricative consonant per speaker.

Table 4.2: Parameter estimates of the linear mixed effects model for the fricative voicing condition in kaCa words. Model formula in R:
$Duration \sim Consonant\,voicing + Consonant\,place + (1|Speaker)$. Reference level for Consonant place: "alveolar".

Fixed effects	Estimate	Std. Err.	*t*-value	*p*-value
(Intercept)	84.6171	7.4771	11.317	$< .001$
Consonant voicing	12.9267	1.4443	8.95	$< .001$
Consonant place: "retroflex"	-0.4861	1.7772	-0.274	$= 0.78$
Consonant place: "palatal"	6.4103	1.7538	3.655	$< .001$
Random effects				
Residual variance: Var = 752.62, Std.Err. = 27.4				

and are flexible with respect to the number and variety of predictors entered into the model (categorical, continuous) (Baayen et al. 2008).

Vowel durations in the fricative voicing condition were analysed by formulating a host of LMEMs in the using the lme4 package (Bates et al. 2011) in R[3]. The variables "Consonant Voicing" and "Consonant Place" were entered in the models as fixed factors. Speaker IDs and experimental Stimulus were entered as random factors (Clark 1973). Models with simple fixed effects and their interactions were compared for the best fit using ANOVA. Log-likelihood tests revealed that a model with simple main effects and one random effect (Speaker) provides the best fit to the data. Parameter estimates are reported in Table 4.2. p-values were calculated by means of Markov chain Monte Carlo (MCMC) sampling.

The model shows that also after including other effects in the model the main effect of consonant voicing in the fricative context remains statistically significant. The intercept (grand mean) of the preceding vowel under consonant voicing increases by 13 msec. Palatal place of articulation has also a significant lengthening effect on the preceding vowel by 6.4 msec.

4.3.2.2 The voicing effect preceding stop consonants

Neither in the labial stop contrast pair ("kapa" vs. "kaba") nor in the dental pair ("kata" vs. "kada") was the voicing effect on vowel duration found. A Wilcoxon test returned respectively the values $W = 18754$ (p-value$= 0.6948$) and $W = 15678$ (p-value$= 0.06375$). A total of 719 samples were analysed in

[3]Version 0.999375-42 was used

Table 4.3: Means and standard deviations of vowel durations in miliseconds for each speaker in the stop voicing condition.

	kapa		kaba		kata		kada	
Speaker	mean	Std.Dev.	mean	Std.Dev.	mean	Std.Dev.	mean	Std.Dev.
bk	75.5	12.8	71.7	11.5	64.4	9.0	64.4	9.0
gm	82.0	25.25	79.3	36.4	95.7	31.25	106.3	51.9
km	89.2	8.3	86.0	11.0	89.5	7.05	96.4	10.0
kz	113.9	25.1	125.6	45.4	111.5	32.2	146.25	49.7
mn	79.5	6.9	82.0	7.7	68.7	7.6	69.4	9.7
mw	65.45	9.1	64.6	10.0	68.4	9.6	67.4	8.1
rj	72.2	6.5	77.5	9.7	72.1	6.3	67.5	11.6
tp	94.0	18.2	108.9	17.45	100.0	8.2	120.8	25.4
All	85.3	22.8	86.5	30.8	84.4	26.1	96.0	41.35

the stop context. Keating's results on stops (Keating 1979) are this way replicated using different stimuli: carrier phrases rather than lists of words. Keating (1979) reports on mean vowel durations of 167.4 msec before /t/ and 169.5 msec before /d/ resulting with the ratio of 0.99. In the present data the ratio for the labial voicing context is 0.985 and the dental context 0.88. The mean values for the consonant duration are presented in Table 4.4.

The lack of consistent differences can be also seen in Figure 4.5 where the log-transformed distributions of vowel duration per each speaker in this condition are presented. Some speakers exhibited a great variability in their productions (notably *gm, kz* and *tp*) others on the other hand were consistent but the lack of the effect is apparent. The mean raw durations in miliseconds and the standard deviation values for the stop voicing effect are shown in Table 4.3.

4.3.2.3 Consonant duration differences

All fricative pairs differed significantly in length (difference significant at $p < .001$, Wilcoxon rank sum test, see also Table 4.4) Fricative duration across the voicing distinction in Polish is clear cut. The density plots and logarithmic means of the fricative pairs are shown in Figure 4.6.

Stop durations also produced significant differences: the "kapa" vs. "kaba" pair and "kata" vs. "kada" both at $p < .001$. The differences within the voicing condition in these stops can be assessed in Figure 4.7. It is however evident that the distributions of these consonants overlap greatly.

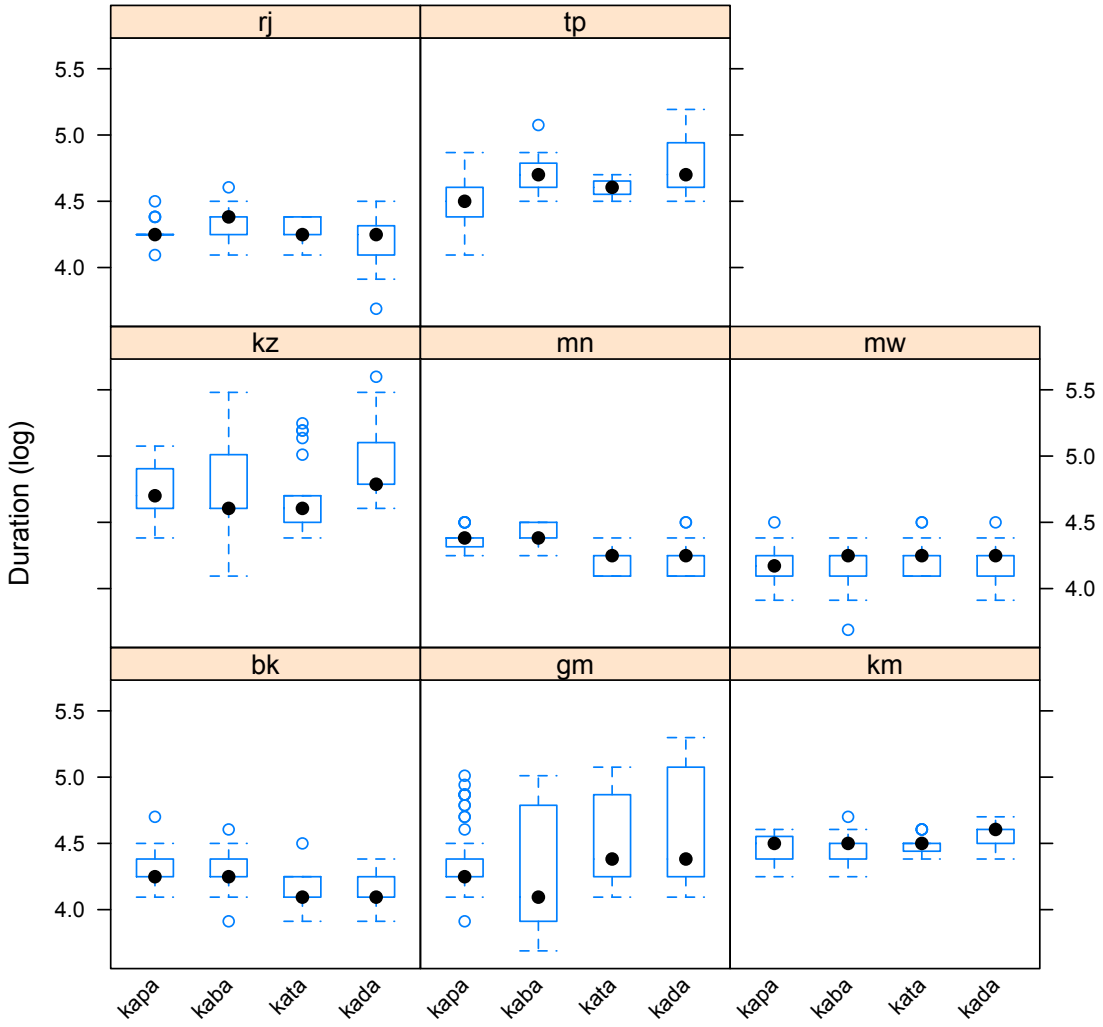


Figure 4.5: The distributions of vowel duration (log-transformed) preceding a voiced or voiceless stop consonant per each speaker.

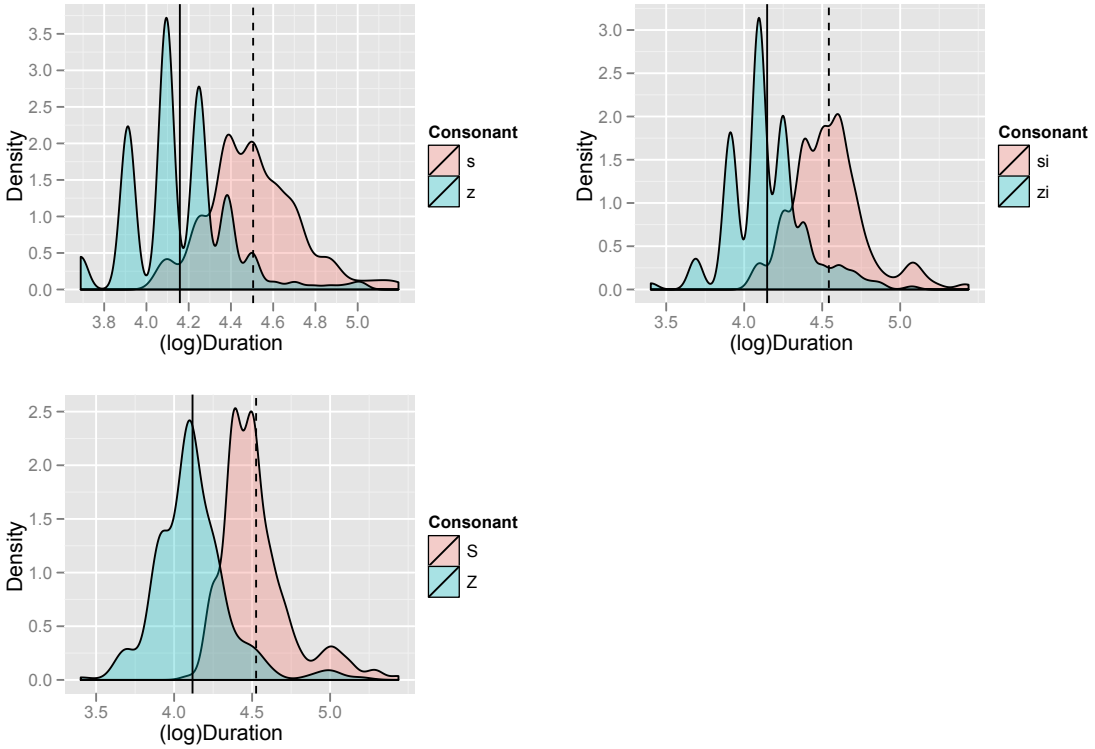


Figure 4.6: Estimated density plots for alveolar, retroflex and palatal fricative (log)duration within the voicing contrast. Vertical lines denote distribution means.

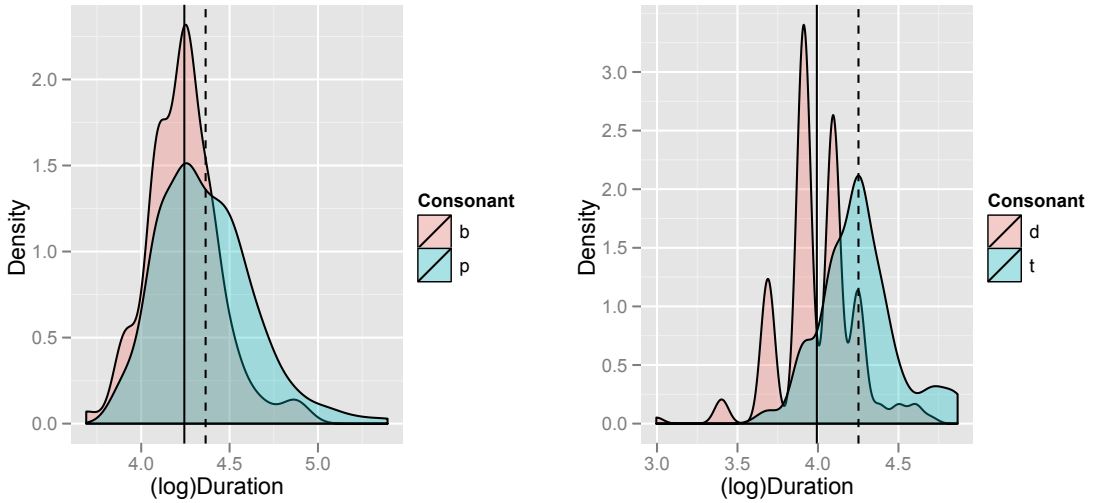


Figure 4.7: Estimated density plots for labial and dental stop (log)duration within the voicing contrast. Vertical lines denote distribution means.

Table 4.4: Means and standard deviations of stop and fricative durations in miliseconds in the voicing condition, kaCa target words.

Consonant	mean	Std.Dev.
/p/	81.6	24.8
/b/	71.2	15.9
/t/	72.35	18.8
/d/	55.75	13.3
/s/	92.7	21.5
/S/	95.0	25.5
/s'/	96.8	26.0
/z/	65.7	16.8
/Z/	63.4	18.3
/zi/	65.2	17.3

4.3.2.4 Is there temporal compensation within VC groups in the voicing contexts?

On the basis of the fact that only the fricative set induces preceding vowel lengthening, it can be suggested that it is not the voicing contrast that requires it but the duration contrast significantly present in the fricative set. It is clear that on the microtemporal level, there is an interaction of manner of articulation effect and voicing on preceding vowel duration. However it seems that the evidence of a consistent duration difference within fricatives points to a possibility that it has to be planned. Consequently, the lengthening of vowel duration preceding e.g. a very short voiced fricative is coordinated as well.

However, in order to judge, e.g. the temporal predictions of rhythm and timing models, described in the Introduction, that postulate tendencies towards vocalic cycle regularity, it is necessary to see which consonants in this dataset could introduce the greatest duration "perturbations" into the vowel-to-vowel frame. And secondly, to observe if a vocalic adaptation, or at least a tendency for it exists and whether it is based only on temporal constraints and none other. Figure 4.8 presents the log-transformed distributions of consonant durations sorted from the median shortest to the longest in the present dataset. The mean and standard deviation of consonant duration is also presented in Table 4.4.

The mean duration of /p/, the longest stop in this dataset, was compared to several voiceless fricatives. The mean duration of the labial voiceless unaspirated /p/ in the material was 82 msec; the mean duration of the voiceless fricatives /s/,

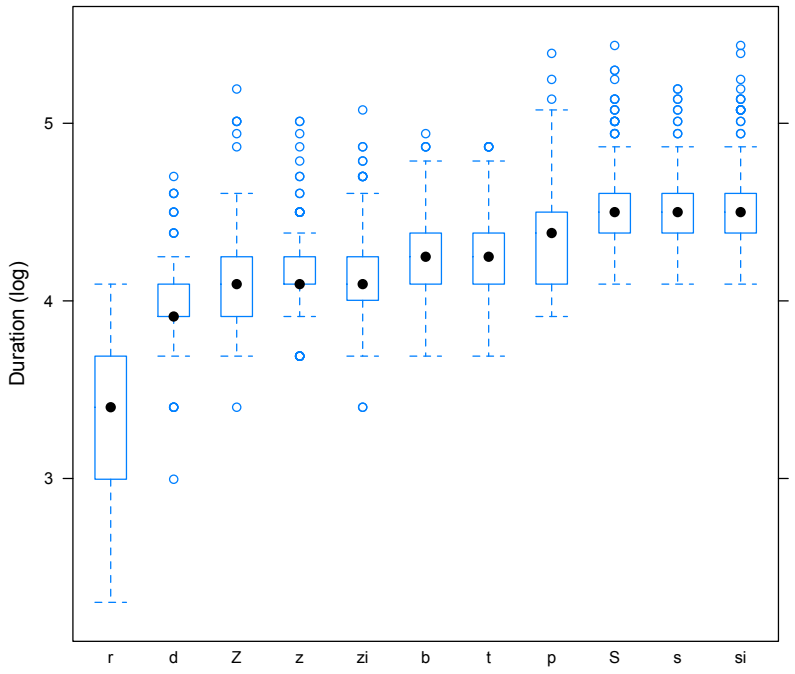


Figure 4.8: The distributions of consonant durations (log-transformed) sorted from the median shortest to the longest in the present dataset.

Table 4.5: Means and standard deviations of the consonant /r/ and the preceding vowel durations in miliseconds for each speaker producing the /kara/ target word.

	/a/		/r/	
Speaker	mean	Std. Dev.	mean	Std. Dev.
km	111.1	14.5	21.6	7.05
mn	95.0	11.9	30.5	6.04
mw	86.0	10.9	31	7.6
rj	92.3	9.7	36.4	9.02
All	96.07	14.7	30.12	8.84

/S/ and /s'/ were 93 msec, 95msec and 97 msec respectively. The differences between /p/ and /S/ (W = 14433.5, p-value< 0.001) as well as /p/ and /s'/ were all significant (W = 14559.5, p-value< 0.001). Also the shortest segments were compared: consonant duration in the "kada" target words vs. "kaZa", "kazia" and "kaza". All differences were significant at p-value$< .001$.

The dataset also involved target words containing a consonant that is usually produced as an alveolar trill or tap in Polish: /r/. The mean duration of /r/ (30 msec, Std.Dev$= 8.8$) was shorter than both the shortest stop (by 37%) and fricative (by 31%) studied here. Four subjects (*km, mn, mw, rj*) pronounced the "kara" target words in the same sentence frames as described in Section 4.3.1. The mean and standard deviation of the resulting durations for this target word are summarised in Table 4.5.

Since an alveolar tap is a very short stop, a LMEM was formulated including all stops studied in Section 4.3.1 plus the /kara/ data. The tap was characterised in the Consonant Place factor as an "alveolar" and in the Consonant Voicing factor as "voiced". The results are presented in Table 4.6. The alveolar tap context has a significant lengthening effect on the preceding vowel, by 19 msec. Compare this to the 13 msec lengthening in front of voiced fricatives that are on average also 30 msec longer than the average tap. A significant voicing effect is only evident with the inclusion of the sonorant alveolar tap together with the stop data.

The above results suggest that vowels preceding consonants that are markedly longer than most other consonants, shorten significantly. At the same time vowels preceding consonants that are markedly shorter than most other consonants lengthen significantly. However the degree of stricture and voicing effects are dif-

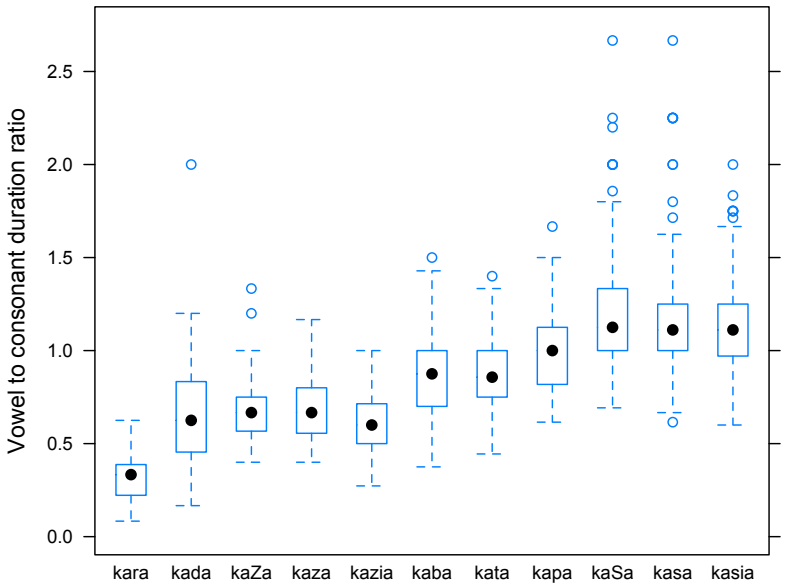


Figure 4.9: The distributions of the vowel to consonant duration ratios sorted from the median shortest to the longest participating consonant in the kaCa dataset.

ficult to separate from the potential independent effect of duration since they are positively correlated: the lower the degree of stricture, the shorter the consonant, especially if it is a sonorant or a voiced consonant.

Figure 4.9 presents the ratio between the consonant and the vowel i.e. the relative duration within a VC group in the kaCa target words. The figure represents the level of durational balance within a VC group related to the increasing (left to right on the horizontal axis) median consonant duration. The duration ratio is given in msec rather than log-transformed for reasons of clarity (ratio = 1 denotes equal durations between the two segments). This way, the impact of inherent consonant duration on the vowel-to-vowel period as well as correlated factors such as consonant manner and voicing can be assessed. The ratio range for most consonants is around 0.6 to 1.1 with the exception of target words with /r/ (0.35). It appears that the vocalic subunit does not contribute to the balancing of the VC group. The VC group is shorter with shorter consonants yielding the ratios lower than unity and it is longer with longer consonants for ratios higher

Table 4.6: Parameter estimates of the linear mixed effects model for the stop voicing condition including the /kara/ stimuli. Model formula in R:
$Duration \sim Consonant\,voicing + Consonant\,place + (1|Speaker) + (1|Stimulus)$.
Reference level for Consonant place: "labial".

Fixed effects	Estimate	Std. Err.	*t*-value	*p*-value
(Intercept)	81.4	7.6	10.7	$< .001$
Consonant voicing	6.4	3.2	2.018	$< .05$
Consonant place: "alveolar"	18.95	4.7	3.99	$< .001$
Consonant place: "dental"	4.48	3.18	1.41	$= .16$
Random effects				
Residual variance: Var = 552.7, Std.Err. = 23.5				

than unity.

4.3.3 Discussion

Keating (1985) discussed the possibility that the voicing effect is a phonetic universal. However, given the lack of a systematic effect in Polish, Czech and Arabic, a physiological explanation has not been so far straightforwardly possible. The question of the universal applicability of the voicing effect is also crucial to the understanding of grammars (Keating 1985; Kawahara 2011). As Kawahara (2011) notes, the lack of the effect in Polish bears on the issue of automatic applicability of phonetic rules, since it seems that the degree of lengthening before voiced consonants is language specific. The rhythmic-compensatory aspect of the voicing effect cannot be claimed to apply universally either as it has been shown not to function in Arabic in Polish (Port et al. 1980), so Kawahara (2011). An attempt to replicate the results on Polish (Keating 1979) was undertaken in the present work with special focus on the rhythmic-compensatory aspect of the voicing effect.

Keating (1979) reported some closure duration overlap between homorganic stops in Polish, suggesting they do not always clearly fall into two groups (Keating 1979: 177). The present study found greatly overlapping consonant duration distributions for stops but not for fricatives. Stop and fricative pairs both showed statistically significant mean duration differences within the voicing contrast.

Keating (1979) also for the first time showed that there is no voicing effect of stops on preceding vowel duration in the language. Similar results were ob-

tained in the present work regarding the differences between stop consonants and preceding vowels, using target words in carrier sentences rather than word lists, as in (Keating 1979). It was confirmed that Polish deviates from a major tendency concerning temporal relations: it does not systematically observe the shortening of vowels in front of voiceless consonants, relatively longer than voiced, all other things being equal, in target words of the kaCa form. However, the present experiment showed a voicing effect in front of fricatives where the vowels lengthened in the context of voiced fricatives by 14 msec. Certainly, the pre-fricative effect is not as "exaggerated" as it is in English. In English we find vowel-to-vowel ratios in similar environments ranging from 1.6 (Lisker 1957) to 1.35 (Port 1977), depending on a study.

The results also suggest that segment timing in voicing effect environments in Polish is to some extent speaker dependent. The rather large interspeaker variability found in both consonant tasks corresponds to conflicting reports on the matter in other languages. Barbosa (2006) reports on the basis of similar Brazilian Portuguese data that stops do not always show significant differences for the voicing effect for all speakers, which is agreed to exist in Brazilian Portuguese as a phonological cue to voicing.

Some authors have suggested that weakening or absence of the effect may have sources on the rhythmical level (Port and Dalby 1982), depending on the rhythm type. Moreover, Keating suggested that the syllable duration balance achieved via the voicing effect can be explained by rhythmic characteristics of a language, e.g. fixed vs. mobile stress and presence or absence of vowel reduction. In English mobile lexical stress and large vocalic variability due to stress might require the balancing of the syllable in voicing contrasts. Polish with fixed lexical stress and a tendency towards full unr such as the voicing effect to support regular re-occurrence of vowels in production. Regarding the interactions of microtemporal structure and rhythmic constituency, Polish does not seem to exhibit a straightforward compensation mechanism balancing V-to-V stretches. The results for VC groups involving fricatives however show a tendency to balance the V-to-V segment timing relations, as predicted by the model in Barbosa (2006). There is also a significant duration effect on preceding vowels in case of /r/.

Manner of articulation effects may override the voicing effect before stops

and support it before sibilant fricatives studied here. In English, vowels preceding fricatives tend to be longer than those preceding stops, and vowels preceding voiced fricatives tend to be even longer (Klatt 1976): an apparent compounding of the voicing and manner effects, also observed in the present Polish data, to an extent. Indeed, an extensive study of a Polish corpus in Klessa (2006) demonstrated that vowels are longer if the degree of stricture of the following consonant is lower. This finding could explain what appears to be a general damping of compensatory effects in the stop effect environment. Consequently, there might be a physiological effect connected to manner of articulation happening in front of fricatives that triggers the voicing effect in these environments, common to both Polish and English. It is possible, that physiological aspects of the manner of articulation influence the magnitude of the effect. This possibility needs to be looked into in the future, if a compensatory interpretation of the voicing effect is to be found valid for Polish.

What seems an interesting candidate however, is the preservation of contrast between the voiced and voiceless three-way sibilant series in Polish. Nowak (2006a) investigated the perceptual categorisation of Polish sibilants and quotes Bladon et al. (1987) statement on the perception of Shona fricatives that "it is certainly not safe to assume, in a language [...] with a three-way place distinction among sibilants, that the sibilants can be identified reliably from spectral features alone" (Bladon et al. 1987: 63). Nowak suggests that such an assumption also holds for the Polish sibilant contrast. Nowak (2006a) found that formant transition cues are necessary for Polish speakers in sibilant identification tasks in VCV words. He also found that the postconsonantal vowel has a higher cue value than the preceding vowel. However, preceding vowel duration was not studied in Nowak (2006a). It is possibly worth to speculate if preceding vowel duration is also one of the phonetic perceptual cues to fricative voicing in Polish. In the present thesis however any questions concerning perceptual categorisation of following consonants with the help of vocalic duration cues cannot be answered and will be addressed in future work.

In terms of indirect evidence for the lack of voicing effect in Polish, Rojczyk (2010) found that Polish learners of English as a foreign language need to acquire the systematic voicing effect on the duration of English vowels. We have

no data whether Polish learners behave differently in voicing contrasts involving fricatives. Given these facts, Keating's suggestions that it is part of Polish phonology not to employ and/or exaggerate the effect for high level linguistic processes (be it prosodic or phonological) is plausible.

However to complete the picture, the effects of very long and very short consonants on preceding vowel duration were assessed. All these consonants were significantly different in duration, but only fricatives and the very short /r/ caused vocalic temporal adaptation. It is possible therefore that the temporal compensation within a vowel-to-vowel unit in Polish is being "switched on" once a certain duration threshold is reached. Judging by the general uniformity of durations of the vowel /a/ in the contexts studied here, it is the very short voiced fricatives and /r/ that trigger a degree of duration compensation.

The above does not apply to target words containing /d/ that nonetheless contain the shortest stop in the studied words. It could be argued that the lack of consistent duration cue to voicing in the two alveolar stop contrast could need a "supporting" cue from the preceding vowel; similarly to the needs for a place contrast between the three-way fricative contrast as suggested above. However it seems that the voicing of the closure is a sufficient cue to this homorganic stop contrast Keating (1979). Certainly the temporal compensation effect suffers in the face of the lack of it in preceding the very short /d/.

The existence of compensation especially in case of voiced fricatives and /r/ needs further explanation. Port et al. (1980) report on a similar experiment for Arabic. Port et al. (1980) looked at several variables using carrier sentences with stimuli containing the consonants /t/, /d/ and /r/, and phonologically long and short /a/ vowels that preceded them. They concluded that there is "minimal evidence" (Port et al. 1980: 240), for temporal compensation in Arabic. Voicing significantly influenced the duration of stops but did not influence the duration of the preceding vowels. However, similarly to the results in the present study, the Arabic tap /r/, an extremely short segment, caused a lengthening of the preceding vowels by 10 msec (in the present data by 19 msec). Mitleb (1984) however, contrary to Port et al. (1980) found that vowel duration does not vary preceding voiced vs. voiceless fricatives in Arabic but does in front of stops, a mirror image of what has been found for Polish in this work. As Ham (2001) observes commenting on

Mitleb (1984): “while it is possible that fricatives and stops behave differently in this regard in Arabic, there is no obvious reason why this should be the case”.

4.4 The geminate effect

The often reported[4] inverse effect of geminates on vowel duration presents an example of a possible temporal micro- and macrostructure interaction, similar to the voicing effect. It is therefore of similar relevance for the syllabic oscillator periodicity hypothesis found in dynamical rhythm models. The following laboratory experiment with Polish native speakers investigates the behaviour of vocalic subconstituents under the duration “pressure” of the geminate within VC groups. Polish, contrary to English, exhibits eligible consonant length distinctions. Additionally, there is no data on the potential presence of the inverse duration relationship between geminates and preceding vowels in Polish. We also find hypotheses concerning the influence of both the canonical syllable and the vowel-to-vowel cycle on segment timing in this context, as well as proposals relating the effect to rhythmic types; the literature will be reported on below.

Smith (1995) conducted a kinematic study of words containing geminate consonants in Italian and Japanese. She connected the timing predictions of her study to rhythmic typologies by relating established rhythmic types with two models of segment timing coordination. Vowel-to-vowel timing was correlated with syllable and stress-timed languages in her study and consonant-and-vowel timing with mora-timed languages. Temporal processes in the organisation of gestures were analysed in the Articulatory Phonology framework in Smith (1995). Their temporal structure was specified in terms of phasing relations.

Concerning the temporal structure of CVCV and CVC:V sequences that are of interest here, in case of the consonant-vowel timing model, Smith’s prediction was that: “if the duration of the consonant were to increase, the time between the two phases of the consonant [closure and release], and hence between the vow-

[4]Myodynamic, aerodynamic and acoustic evidence can be found in: Al-Tamimi (2004) for Arabic, Tserdanelis and Arvaniti (2001) for Cypriot Greek, Esposito and di Benedetto (1999) for Italian, Delattre (1971) for English, German, Spanish and French, Maddieson (1984) also lists Kannada, Tamil, Telugu, Hausa, Icelandic, Norwegian, Finnish, Hungarian and Amharic, among others.

els, might be expected to increase" (Smith 1995: 207). In case of vowel-to-vowel timing where consonants are imposed on a continuous vocalic train, the vowels should remain relatively unaffected, *in articulatory terms*, by adding consonants but, *superficially*, manifest compensation phenomena ("observations on the output" (Browman and Goldstein 1990)). Similar duration trade-off relations between consonants and vowels in such structures are assumed in the rhythm model by Barbosa (2006) which was developed for syllable- and stress-timed languages.

Rhythmic type was also suggested to correlate with different patterns of the geminate effect on preceding vowel duration. Ham (2001) observed that syllable-timed languages such as Italian exhibit a shortening effect of geminates on the preceding vowels, resulting in a durational inverse or compensation. Several researchers (Campbell 1999; Idemaru and Guion 2008) have shown that Japanese, a mora-timed language, lacks the durational inverse in the context of geminates; Japanese vowels are in fact slightly longer in the pre-geminate position and shorter in the post-geminate position. The patterns mentioned are also correlated with the amount of overlap in the singleton-geminate length distributions: Italian shows a large degree of overlap (Payne 2005) whereas Japanese tends to have robust durational differences.

In the present and former experiment we focus on the vocalic cycle as the rhythmic constituent however, questions of a possible syllabification effect in the context of geminates and preceding vowels cannot be ignored. Maddieson (1984) discusses various effects on vowel duration as evidence for canonical syllable constituency by focusing on phenomena found universally. He concentrates on a) closed syllable vowel shortening in general and b) geminate vs. singleton effects on preceding vowel duration in particular. He suggests that the word-internal intervocalic singleton vs. geminate contrast offers a straightforward test for syllabification effects on vowel duration. This contrast is also studied here in the differently syllabified minimal pairs: pa.pa vs. pap.pa. According to Maddieson (1984), the vowel duration difference that relates to the syllabification of the following consonant (Maddieson 1984: 89) manifests in the vowel shortening within the closed syllable preceding the geminate. Importantly, he reports on evidence that this tendency is found in different languages "with and without vowel length contrast, at different speech rates and under different prosodic conditions" (Mad-

dieson 1984: 91). Maddieson (1984) discusses Japanese as the apparent counterexample to this tendency: vowels lengthen in front of geminates and shorten after geminates. He argues that the first part of the geminate in Japanese is not the coda of the first syllable but a syllabic consonant resulting in divisions such as /ka.n.na/.

4.4.1 Intervocalic geminates in Polish

Polish geminates are either underlying or derived through gemination across morpheme and clitic boundaries (Thurgood and Demenko 2003; Pająk and Bakovic 2010). It can be argued whether Polish possesses true geminates; depending on which view on phonological representation of geminates is adopted. The phonological status of gemination has been disputed. According to Delattre (1971) geminates are two identical consonants where the first occupies the coda of the first syllable and the other the onset of the next syllable. In this view, geminates can be seen as a special case of clusters that consist of two identical re-articulated consonants that occupy two skeletal slots. According to Ladefoged (1971) on the other hand, geminates are long consonants that occupy a single timing slot in the syllable structure. More recent autosegmental accounts of phonological representation of geminates such as McCarthy (1986), suggest that the so called "true" geminates represent one feature bundle but occupy two timing slots while "fake" geminates have two separate feature bundles and timing slots.

English is a good example of a language with "fake", i.e. concatenated geminates. Concatenated geminates arise as identical consonants follow each other across morpheme or word boundaries, e.g. "fun name", "un-name" (Oh and Redford 2012). Most intervocalic geminates in Polish, arise as a result of morpheme concatenation e.g. "lekka", "miękki". Despite spanning a morpheme boundary, the consonant length does provide the lexical contrast between, e.g. "lekki" (adj. masc. *light*) and "leki" (pl.n. *medicine*), contrary to English. Therefore, the above are examples of true geminates. Non-derived and non-concatenated gemination also occurs in Polish, however only in loanwords such as "fontanna" (fountain) or "ballada" (ballade) (Pająk and Bakovic 2010). Neither Thurgood and Demenko (2003) nor Pająk (2010) when discussing different aspects of geminate

production and perception in Polish, commit themselves to solving the phonological status of Polish intervocalic geminates.

Moreover, even in languages that possess many tautomorphemic geminates, such as Italian, the status of the phonological representation ("are they heterosyllabic or not?") has been disputed and not settled (Zmarich et al. 2011). It is however agreed, that for all contexts and all languages it is the closure duration (in case of stops) or consonant duration (in case of fricatives) that constitutes the main temporal feature of the singleton-geminate contrast (Ridouane 2010).

As in the case of the phonetic voicing effect studied in Experiment 1, a hypothesis will be tested that a phonetic, universal tendency to equalise the duration of the intervocalic period is found in the present geminate-singleton contrast. The period "correction" in the presence of a geminate should also correspond to the hypothesised gesture timing relations modeled by Smith (1995) for the rhythmic types most likely applicable to Polish: either syllable- or stress- timed ("vowel/stress based" rather than "mora counting"). Consonant duration distinguishes between words in Polish and so this context is permitted to be taken into consideration in the present study.

4.5 Experiment 3: the geminate effect in Polish[5]

In Experiment 1, an effect of voicing on vowel duration was only found for preceding fricatives. Moreover, no systematic evidence in favour of the vocalic cycle regularisation hypothesis was found. An experiment on the geminate effect on preceding vowel duration is conducted to clarify the problems posed to all models of rhythm mentioned above and to the rhythm class hypothesis in general. In the design of the experiment and in the results no claim is made as to the phonological status of vocalic duration variability as a potential additional cue to geminate categorisation. Results are interpreted only in terms of timing relations between segments. Questions of categorisation would require a greater number of speakers and additional perceptual experiments.

[5]Preliminary results of this experiment were reported on in Malisz (2009)

4.5.1 Data and methods

Six speakers of standard Polish (three male and three female; 20-30 years old) from the Great Poland (Wielkopolska) region were asked to repeat stimuli around 10 times (2126 tokens were recorded and annotated). Target words of the form / paCa/ and / paC:a/ were used as stimuli (where C was one of the eight fricatives or four stops under study). The stop target words involved the bilabials /p/, /b/, and the dentals /t/, /d/. The fricative target words involved the alveolar voiceless /s/, retroflex /S/ and /Z/ and the alveolo-palatal voiceless /s'/ contrasting in length, word medially. The majority of resulting target words are nonsense words (with the exception of "papa", "passa" and "paSa"). The presentation of the stimuli followed the same procedure as in the previous experiment described in Section 4.3.1.

4.5.1.1 Annotation and measurement

The data were transcribed manually using Praat (Boersma and Weenink 2012) according to oral constriction criteria for prosodic annotation as defined in Section 4.3.1.1.

Stop geminates were defined as a consonant with a prolonged closure and a final release burst. Fricative geminates were defined as a consonant with a prolonged sustained frication noise, evident as energy in the higher frequency bands visible on the spectrogram, and as an aperiodic waveform. Some speakers occasionally produced two released consonants in the stop geminate trials, e.g. two doubly articulated stops in "pap.pa". vs. a geminate stop in "pappa" with a prolonged closure. Doubly released stops and affricates are quite common in hyper-articulated Polish speech (Thurgood and Demenko 2003). All consonants with a double release and/or vocalic epenthesis (5%) were excluded from the analysis. Only medial VC-VC: sequences were considered.

Table 4.7: Means and standard deviations of stop and fricative durations in miliseconds in the geminate condition, paCa target words.

Singletons			Geminates		
Consonant	mean	Std.Dev.	Consonant	mean	Std.Dev.
/p/	79.6	12.65	*/pp/*	177.6	42.4
/b/	67.0	11.7	*/bb/*	154.9	34.0
/t/	70.2	9.0	*/tt/*	164.0	30.3
/d/	54.8	11.4	*/dd/*	153.0	38.0
/s/	95.65	9.7	*/ss/*	182.3	33.3
/S/	91.2	11.3	*/SS/*	200.9	48.4
/s'/	97.1	14.8	*/s's'/*	178.4	36.5
/Z/	61.9	8.0	*/ZZ/*	150.3	36.8

4.5.2 Results

4.5.2.1 Consonant length differences

All speakers produced significantly longer geminate consonants than singleton consonants (all measures p-value$<$ $.001$, Wilcoxon rank sum test). The mean and standard deviation of consonant duration in the singleton/geminate context is reported in Table 4.7. The mean ratio of geminate to singleton length was 2.4 for stops (2.3 for voiceless, 2.53 for voiced) and 2.1 for fricatives (1.96 for voiceless, 2.42 for voiced). Maddieson (1984) reports on geminate to singleton ratios for various languages ranging from 1.5 to 3. An important condition, i.e for a significant difference in duration between geminates and singletons is met, which, apart from its phonemic status, additionally testifies to the reality of the consonantal contrast in Polish.

4.5.2.2 Vowel duration differences

Figure 4.10 and Figure 4.11 present the results of the vowel duration study in pre-geminate and pre-singleton contexts, grouped into stops and fricatives, separately for each speaker. The interspeaker variability appears to be more pronounced than in the case of the voicing effect in Section 4.2 with greater standard deviations for each target word pair. However, an effect of consonant length on vowel duration can be observed for most speakers in both stop and fricative environments.

Figure 4.12 groups mean vowel durations by manner of articulation and voicing. A non-parametric test (Wilcoxon rank sum test) showed that vowels

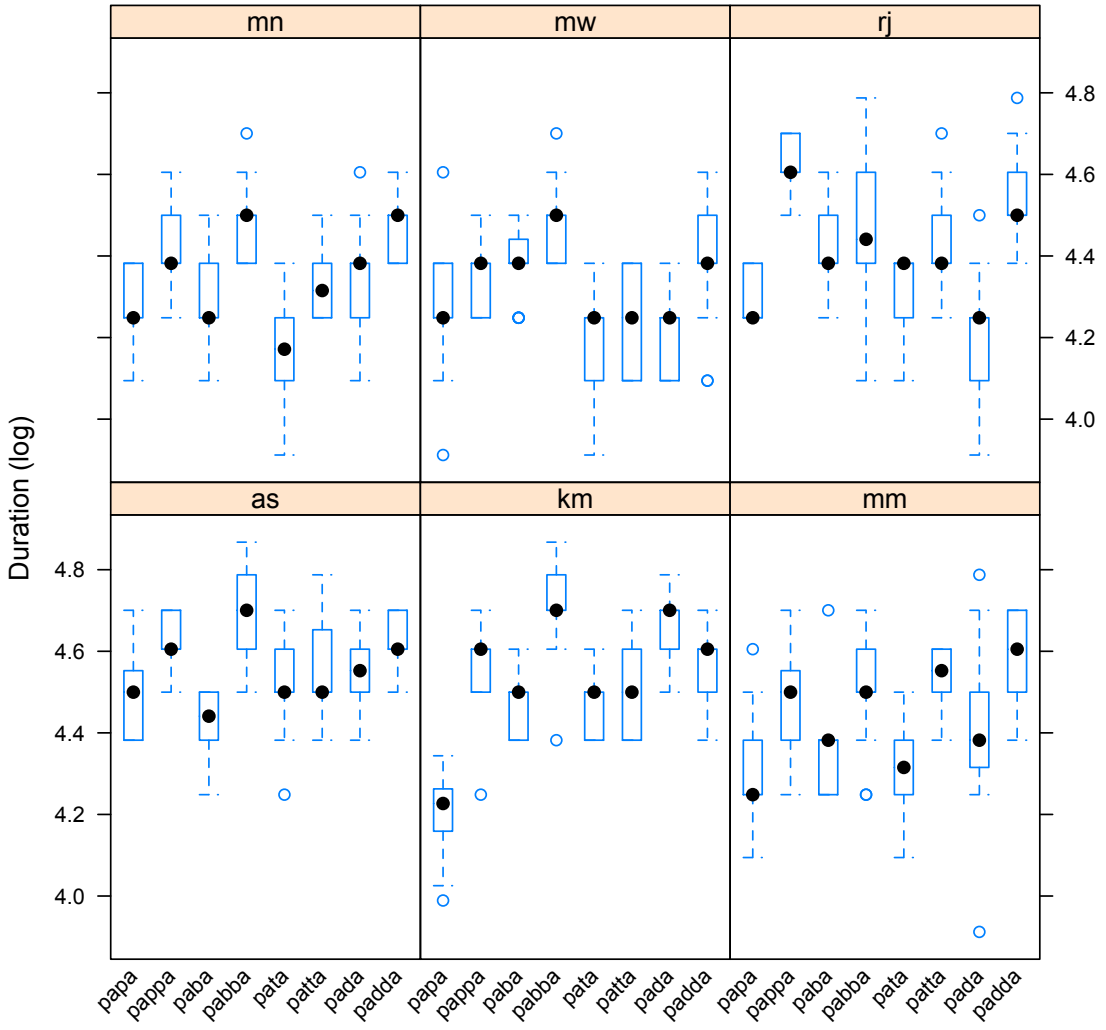


Figure 4.10: The distributions of vowel duration (log-transformed) preceding a singleton or geminate stop consonant per each speaker.

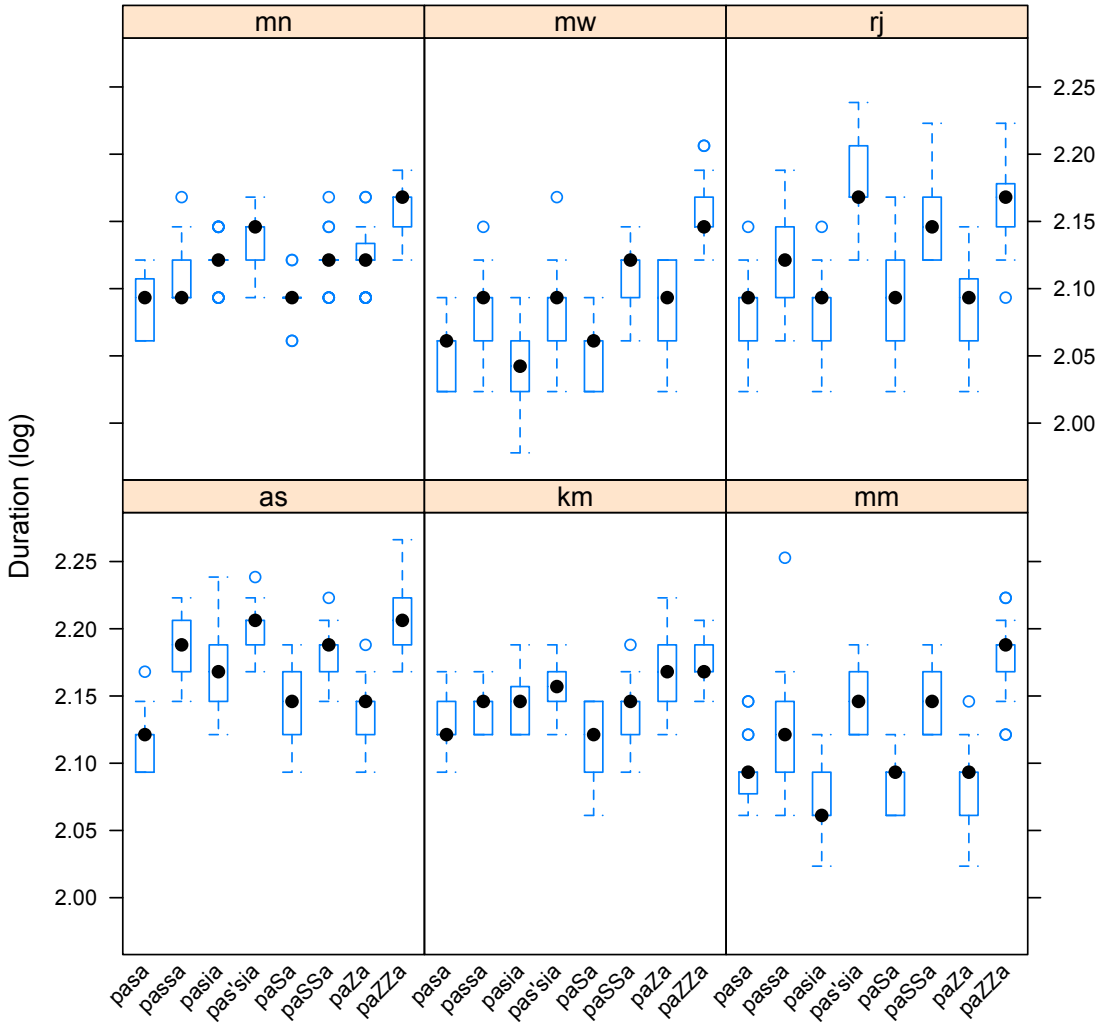


Figure 4.11: The distributions of vowel duration (log-transformed) preceding a singleton or geminate fricative consonant per each speaker.

lengthened in the context of a geminate by ca. 12 ms preceding stops and 17 msec preceding fricatives, the difference is significant at $p < .001$. Regarding the mean durations of vowels preceding singleton consonants, it can be observed that first of all, the difference in duration across consonant manners is small, with a tendency for a lengthening effect of fricatives. Second of all, as expected, the effect of voicing is small with singletons, a result that confirms the experimental findings on the voicing effect in the present thesis. The previous experiment showed no effect of voicing in front of stops, with significant lengthening influence of voiced fricatives. Nonetheless, a clear effect of consonant length can be observed in both manner and voicing conditions in Figure 4.12. Cumulatively, vowels should lengthen more before geminate fricatives than stops (mean $= 92.8$, Std.Dev. $= 19$ for fricatives, mean $= 84.6$, Std.Dev . $= 15$ for stops), especially before the voiced geminate fricative (mean $= 113.6$, Std.Dev. $= 16$ for voiced geminate fricative, mean $= 98.2$, Std.Dev. $= 18$ for voiceless geminate fricative).

Given the individual differences and a multitude of possible predictors of vowel duration in this context, linear mixed models were used for analysis as in Experiment 1, Section 4.3.2.1. Vowel durations in the context of both stops and fricatives were entered to LMEMs with Speaker and Stimulus as random variables. The fixed factors under consideration were Length (geminate, singleton), Manner (stop, fricative) and Voicing (voiced, voiceless). In general, factors other than Length served as controls for the main effect of Length. All possible interactions were checked, none were statistically significant. Consonant place was not entered due to an unsufficient number of contrasts. The p-values were obtained by means of MCMC sampling. The number of observations analysed was 2125. Table 4.8 presents the formula and estimates of the LMEM.

As the model shows, Consonant Length has a significant main effect on preceding vowel duration at $p < .001$. The estimate value of 15 msec approximately reflects the mean for stop and fricative geminate effects presented in Figure 4.12, i.e. 12 msec for stops, 17 msec for fricatives. Somewhat unexpectedly, there is no significant interaction between manner and voicing. It appears that voicing affects vowels both in front of stops and fricatives in this environment, vowels are longer preceding voiced consonants, by 8.4 msec. As far as manner effects are concerned, vowels are shorter preceding stops by 10 msec, relative to preceding

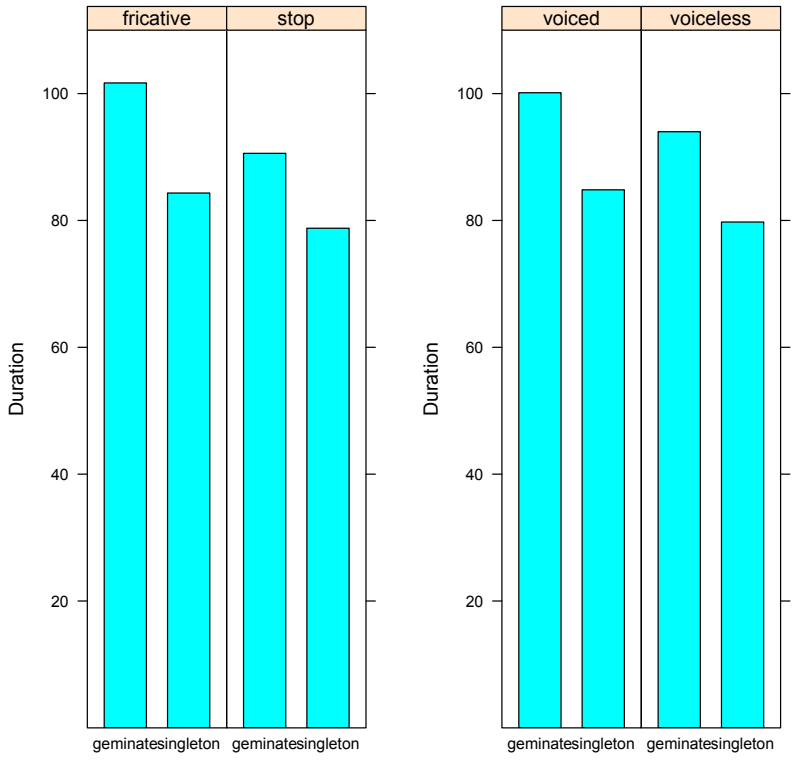


Figure 4.12: Absolute mean duration (in msec) of vowels preceding geminate and singleton consonants grouped by manner of articulation or voicing.

Table 4.8: Parameter estimates of the linear mixed effects model for the geminate condition in paCa words. Model formula in R: $Duration \sim Consonant\,voicing + Consonant\,manner + Consonant\,length + (1|Speaker) + (1|Stimulus)$. Reference level for Consonant manner: fricative, for Consonant length: singleton.

Fixed effects	Estimate	Std. Err.	*t*-value	*p*-value
(Intercept)	84.01	4.3	19.34	$< .001$
Consonant voicing	8.38	2.02	4.13	$< .002$
Consonant manner	-10.17	1.96	-5.19	$< .001$
Consonant length	14.98	1.89	7.91	$< .001$
Random effects				
Residual variance: Var = 128.13, Std.Dev. = 11.32				

fricatives. The additive effect of all these factors is that vowels are longest in front of voiced fricative geminates and shortest in front of voiceless stop singletons.

4.5.2.3 Is there temporal compensation within VC groups in the consonant length context?

Apart from the strengthened effect in case of fricatives, in general, vowels lengthen in front of geminates in Polish. However, the lengthening of vowels in Japanese entails also post-geminate shortening. If the same effect could be found for Polish, at least some adaptation to achieve relative stability of vowel-to-vowel onsets in words such as "passa", "pappa" would be found.

4.5.2.4 Following vowel duration in the geminate context

The responses of two speakers from the present dataset were additionally annotated for post-consonantal vocalic durations (henceforth: V2) in paC(C:)a stimuli. The mean realisations produced by the speakers for all three segments are presented in Figure 4.13. A first look at the correlation between consonant duration and V2 is also offered in Figure 4.14. Evidence for a negative correlation, i.e. as consonant duration increases V2 duration decreases (p-value$< .001$) is found. This result suggests duration patterns in this context as found in Japanese.

A LMEM was fitted to the V2 data. The number of observations was 765. The estimates of the LMEM can be found in Table 4.9. Only a significant lengthening effect of Voicing on the following vowel was found. Manner and

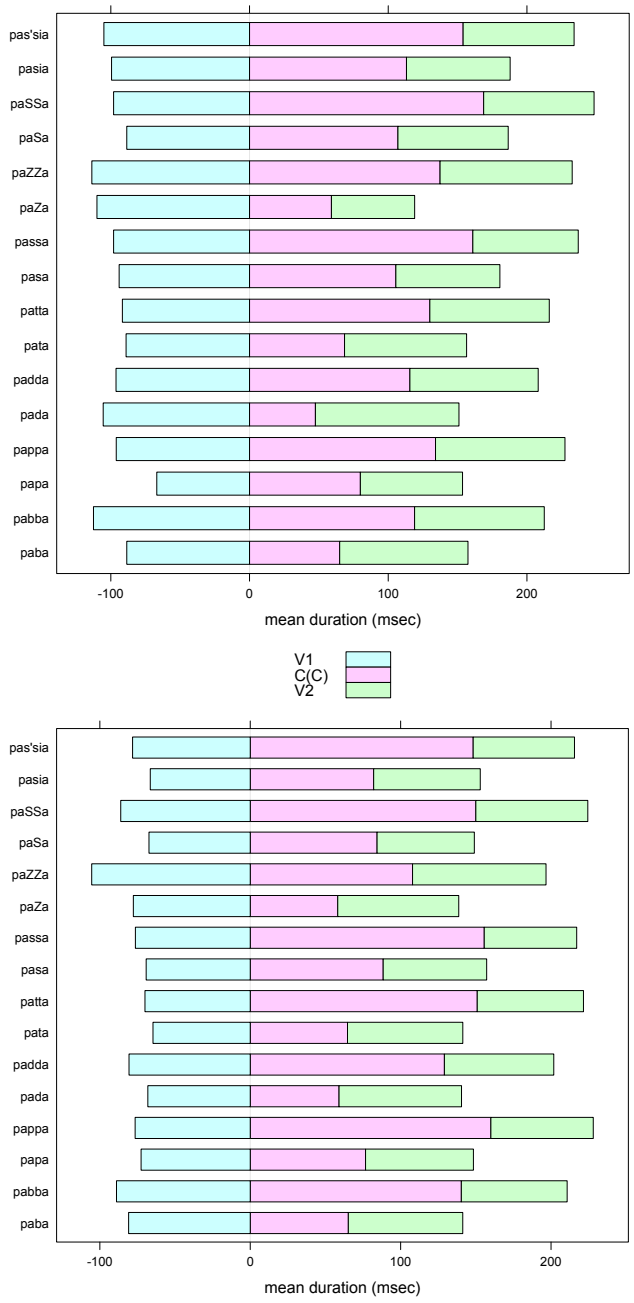


Figure 4.13: The mean durations of the first vowel (V1), the consonant (either a singleton C or a geminate CC) and the second vowel (V2) in the responses to paC(C)a stimuli. Speakers km (top panel) and mw (bottom panel).

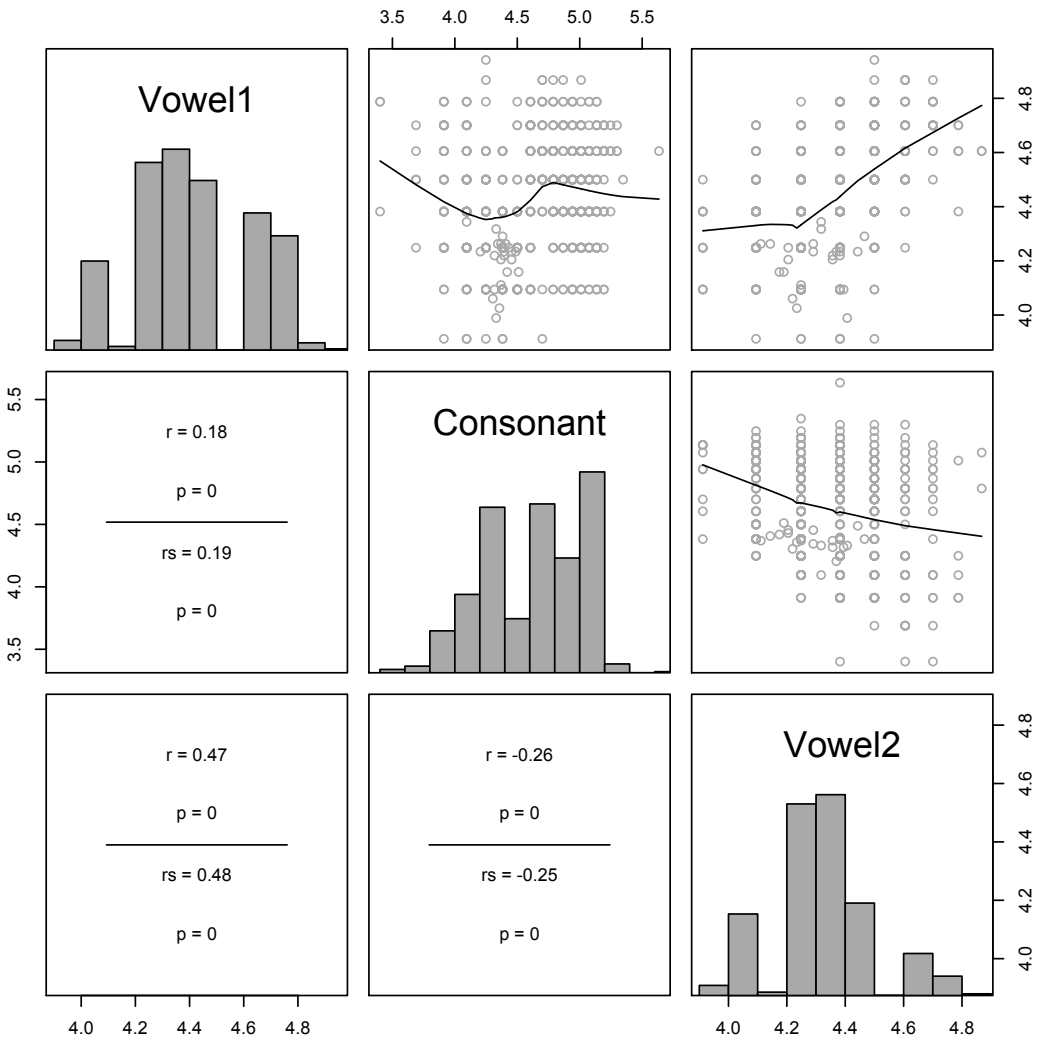


Figure 4.14: Correlation diagrammes and coefficients between Consonant duration, the preceding (Vowel1) and the following vowel (Vowel2). Durations in msec.

Table 4.9: Parameter estimates of the linear mixed effects model for the postconsonantal vowels in two speakers. Model formula in R:
$Duration \sim Consonant\,voicing + (1|Speaker) + (1|Stimulus)$.

Fixed effects	Estimate	Std. Err.	*t*-value	*p*-value
(Intercept)	75.05	6.5	11.5	$< .001$
Consonant voicing	10.78	2.33	4.63	$< .001$
Random effects				
Residual variance: Var = 85.23, Std.Dev. = 9.23				

Consonant length have not reached significance. An independent effect for prolonging the vowel following the geminate suggested in the correlations could not be confirmed.

4.5.3 Discussion

The results of Experiment 2 show that Polish follows Japanese in the timing pattern of pre-geminate vowel duration. In both languages consonant length differences show no overlap and vowels co-vary with long consonants that follow. Given the post-geminate shortening effect apparent in Japanese (Campbell 1999; Idemaru and Guion 2008), a level of temporal compensation interpretation can still be applied to the Japanese vowel-to-vowel unit in this context: where the vowel lengthens as a result of geminate presence, it also shortens post-consonantally. Such a pattern would testify to a different timing strategy handling the consonantal "perturbation", be it strategy based on a proposal by Smith (1995) for Japanese or by Maddieson (1984). Neither of the strategies discussed by these authors is the case in Polish: voiceless consonants seem to shorten the following vowels by approx. 11 msec, the duration of the consonant has no significant effect.

Given the discussion in Maddieson (1984) of the closed syllable vowel shortening apparent in many languages in the following geminate context it seems quite implausible that Polish should exhibit behaviour that is different both from the most universal pattern as well as the special case of Japanese. Even if "the Japanese pattern" for the geminate effect was accepted in case of Polish, the consequences of this timing strategy are unclear. Smith (1995) suggested that "the patterns of temporal organisation observed among articulatory gestures can vary among languages, but seem to vary in a way that corresponds to the traditional

descriptions of languages' rhythm, and can be described in terms of how different gestures are coordinated in time" (Smith 1995: 220). Does this effect mean that Polish is mora-timed in parts? Such a conclusion seems unlikely.

Nowak (2006b) found in a corpus study that following segment manner was ranked higher than voicing, as far as segmental features were concerned, as estimated from multiple regression on vowel duration. Vowels were significantly shortest in front of stops, then longer in front of nasals > affricates > fricatives > laterals > approximants > and longest in front of a trill. Nowak (2006b) did not have an explanation for the consonant manner effect, but he suspected, similarly to Klessa (2006), that Polish vowels are longer before segments with smaller degree of oral constriction. Given the results by Nowak (2006b) and Klessa (2006) the present results on the manner effect in the following direction: vowel is shortest before stop > fricative > longest before a tap/trill, correspond to what was found by these authors. Klessa (2006) reported, also on a corpus, that in a syllable of the CV type in Polish, the average vowel duration was ca. 82 msec > CVC $= 67$ msec > CCV $= 78$ msec > CCVC $= 59$ so there was an apparent shortening effect of a closed syllable on vowel duration, as predicted by Maddieson (1984). The geminate environment in Polish seems to place exceptional conditions on the preceding vowels, as is evident from the results in this section, since such a common effect was not found in the present work.

Given the implausibility of a solution based on timing at this moment, a " non-rhythmic interpretation of all lengthening effects in both experiments reported in the present chapter will be offered.

Evidence from Experiment 1 showed a mild voicing effect in the context of fricatives. This result was confirmed in the present experiment, with additional evidence of a significant voicing effect also for stops within the geminate-singleton contrast material. Additionally, a significant geminate effect was found, however no inverse compensation on the part of the preceding vowel ocurred. On the contrary, the vowels appear to lengthen in front of a geminate by a mean of 14 msec. An explanation for the maximum possible, additive lengthening effect on the preceding vowel, that of a fricative voiced geminate, might stem from constraints on the voicing of obstruents, in combination with other factors. As Ohala (1983) explains, the tendency for geminate obstruents to become voiceless is particularly

strong, since the longer the closure, the more likely it is to be devoiced.

Furthermore, to maintain voicing in fricatives, two conflicting forces are interacting: the oral pressure needs to be low for sustained voicing, however it should also be sufficiently high to cause high air velocity passing through the constriction. As a result of these aerodynamic constraints, to sustain voicing and frication, as well as the length contrast, a lot of articulatory effort is required. Potentially, the pre-geminate vowel lengthening is a case of *anticipatory effort*. In fact, vocalic durations seem to lengthen in accordance with the increased articulatory effort associated with sustaining consonant voicing in the following order of complexity: manner of articulation (singleton fricatives, not singleton stops) > consonant length (geminate stops and fricatives). Also, as suggested in the previous experiment, the necessity to keep the three-way place contrast between fricatives in Polish probably strengthens the increasing cost of producing a voiced, geminate fricative.

Conclusion

The present thesis considered two current approaches to speech rhythm: rhythm metrics and dynamical models. It was argued that some metrics model characteristics of rhythm that are necessary but insufficient: the linear, binary alternation of segments, very susceptible to speech rate variation, speaker variation and other factors. Therefore, one of the objectives of the dissertation, to elucidate the typological status of speech rhythm in Polish, could not be achieved using this popular methodology. Throughout this work, a perspective on speech rhythm derived from coordination dynamics and dynamical systems theory, was supported. In this perspective, rhythm is treated as a coordinative device that operates on multiple, interacting timescales. The characteristics of this interaction have been exploited over the years by phoneticians not directly involved in dynamical modeling. The thesis reviewed some of these studies such as Asu and Nolan (2006); Eriksson (1991); Jassem et al. (1984), demonstrating their greater accuracy and adequacy in representing speech rhythm variability, as compared to rhythm metrics.

A dynamical model by O'Dell and Nieminen (2009) was used to represent the coupling of Rhythmic Prominence Intervals and phonetic syllables in a corpus of Polish task-oriented dialogues. The results show a general tendency for syllable oscillator domination in Polish, especially with increasing speech rate. In the slowest tempo, due to accentual effects on duration in Polish that manifest in the regression models, the relative coupling strength value is higher, tending towards the stress oscillator dominated rhythmic strategy. The result supports approaches to speech rhythm that do not assume that the presence of several rhythmic strategies in one language is mutually exclusive. These strategies evidently co-exist and their interaction is dependent on style, speech rate and structural factors. The results in this part of the thesis also contain a duration model determining fac-

tors influencing RPI duration such as tempo, number of syllables etc. The model therefore enables prediction of RPI duration.

The second and third experiments deal with local and detailed predictions made by dynamical rhythm models. The models (Barbosa 2006; O'Dell and Nieminen 2009) postulate an influence of higher level oscillators on the duration of their subconstituents, in this case, segments. The laboratory experiments tested the hypothesis that the syllabic oscillator has an impact on the subordinate segments. According to the literature, a duration balancing tendency should be observed, regularising the duration of the phonetic syllable in contexts that contain particularly long constituents such as, e.g. geminates. The existence of such an effect in Polish would support the hypothesis that the basic characteristic of the syllabic oscillator is its tendency towards periodicity.

In the second experiment an analysis of the voicing effect on the preceding vowel was presented in the context of the abovementioned hypothesis, but also in order to conflicting accounts in the literature concerning the presence of the voicing effect in Polish. The results show that the voicing effect is unsystematic and limited to contrasts involving fricatives. Fricatives are also found to be significantly longer than stops in the studied material. However, a clear confirmation of the rhythmic hypothesis based on this context cannot be made.

The third experiment analysed the effect of a long consonant (a geminate consonant) on the preceding vowel. It is the first such study on Polish known to the author. A controlled experiment was conducted and a significant lengthening effect of the geminate on the preceding vowel was observed. The result, uncompatible with predictions found in the literature for any canonical rhythm type, as well as unexpected with regards to the rhythmic hypotheses tested directly in this experiment, complicates the interpretation of this context in terms of timing, both global and local.

In summary of Chapter 4, it was shown that the contexts previously found to incur vowel compensation in other languages either do not fully incur the same process in Polish (Experiment 1) and/or presents a case that does not comply with generalisations regarding rhythmic types (Experiment 2). To explain both results concerning segments within the vowel-to-vowel frame (the phonetic syllable), an explanation was suggested: the increased duration of vowels is due to anticipa-

tory, articulatory effort associated with the necessity to sustain the voicing of the following consonant. The additive lengthening effects on the vowel reach their maximum in the context of a following voiced geminate fricatives in which, due to aerodynamic factors, maintaining voicing is particularly difficult.

The dissertation implemented a current methodology used to model speech rhythm variability with Polish data. It also offered a new perspective on the duration of syllables and Rhythmic Prominence Intervals in Polish. The results obtained in this work are going to facilitate the full explanation of the postulated "atypical" nature of Polish speech rhythm in the future, as well as support continuing studies on Polish voicing and geminates.

Abstract in Polish

Niniejsza praca przedstawia współczesne modele rytmu w mowie stosowane w typologii rytmicznej języków polskiego i angielskiego, oraz wyniki eksperymentów nad rytmem i iloczasem w języku polskim. Wszystkie omówienia oraz eksperymenty zawierają porównania z istniejącą szeroką literaturą i wynikami dotyczącymi badanych zjawisk w języku angielskim.

Praca zawiera omówienia tzw. "miar rytmicznych" (ΔV, %V, nPVI itp.) i rozważa ich wielorakie braki w charakterystyce wielopoziomowego ujęcia rytmu oraz inne słabości metodologiczne: wartości uzykane w literaturze w wyniku stosowania miar rytmicznych są niestabilne względem tempa mowy, różnic indywidualnych mówców, materiału tekstowego i innych zmiennych.

Następnie przedstawione są modele rytmu oparte na oscylatorach sprzężonych, które w zadowalający sposób wyrażają właściwości wzorców iloczasowych i akcentowych w mowie, np.: a) strukturę hierarchiczną, gdzie elementy hierarchii są zagnieżdżone w sobie według prostych stosunków 2:1, 3:1 i tak dalej, b) tempo mowy, jako parametr kontrolujący zmiany w strategiach rytmicznych tj. strategiach gdzie wpływ oscylatora sylabicznego silniej wpływa na akcentowy i odwrotnie.

Modele dynamiczne w obliczeniach strategii rytmicznych używają jednostek iloczasowych takich jak Rhythmic Prominence Interval (Przedział Prominencji Rytmicznej), wyznaczanych przez rodzimych mówców poprzez anotację mowy spontanicznej ocenianej percepcyjnie, oraz sylaby fonetycznej, tj. przedziału od samogłoski do samogłoski. Praca omawia hipotezy przemawiające za użyciem takich jednostek i za badaniem wzajemnych wpływów tych jednostek na ich przebiegi w czasie (częstotliwości) jako sposobu ocenienia rodzaju rytmu w mowie.

W pierwszym eksperymencie powyższe jednostki zostały wyznaczone w dialogu mowy polskiej oraz wykonano obliczenia parametru sprzężenia relatywnego jako funkcji tempa mowy. Wyniki pokazują ogólną tendencję dla dominacji oscylatora sylabicznego w polskim, zwłaszcza wraz z wzrostem tempa mowy. Wynik wspiera takie ujęcia rytmu w mowie, które nie zakładają wzajemnego wykluczania się tradycyjnych typów rytmicznych, ale ich współistnienie, zależne od stylu, tempa oraz czynników strukturalnych. Wyniki na tym etapie zawierają również statystyczny model iloczasowy określający czynniki, które wpływają na długość RPI, takie jak tempo, ilość sylab w RPI itd. Model umożliwia więc przewidywanie długości RPI.

Eksperyment drugi i trzeci zajmują sie lokalnymi, szczegółowymi aspektami przewidywań modeli dynamicznych rytmu w zakresie wpływu oscylatorów umiejscowionych na prozodycznie wyższych poziomach, na iloczas głosek niżej w hierarchii. Eksperymenty badają hipotezę wpływu oscylatora sylabicznego na podległe głoski w warunkach laboratoryjnych. Według literatury, oscylator ten powinien wykazywać tendencje harmonizujące, regulujące długość sylaby fonetycznej w kontekstach, w których występują szczególnie długie spółgłoski składowe, takie jak np.: geminaty. Tego typu efekt wspierałby hipotezę, że podstawową cechą oscylatora sylabicznego jest jego regularność (periodyczność) lub tendencja do regularności.

Eksperyment drugi bada kontekst wpływu dźwięcznych i bezdźwięcznych spółgłosek na iloczas poprzedzającej samogłoski w kontekście wyżej wymienionej hipotezy. Badanie wykonano także ze względu na literaturę podającą sprzeczne wyniki dotyczące istnienia wpływu dźwięczności na iloczas samogłoski poprzedzającej (“the voicing effect”) w języku polskim. Niniejsze wyniki pokazują, że wpływ ten występuje w niewielkim stopniu oraz niesystematycznie, ograniczony do kontrastów związanych ze spółgłoskami trącymi. Spółgłoski te, są również w większości dłuższe od wybuchowych w tym materiale, jednak w pracy nie stwierdza sie jednoznacznie, że znaleziono potwierdzenie hipotez rytmicznych przedstawionych powyżej, gdzie wyniki mozna uzasadnić odgórnymi, rytmicznie umotywowanymi wpływami.

Eksperyment trzeci bada kontekst geminaty i jej wpływ na poprzedzającą samogłoskę. Jest to pierwsze tego rodzaju badanie dla języka polskiego

znane autorce. W kontrolowanym eksperymencie, odkryto efekt wzdłużający poprzedzającą samogłoskę przez długą spółgłoskę. Wynik komplikuje jednoznaczne odrzucenie lub potwierdzenie hipotez rytmicznych.

Konteksty, w których, w wielu innych językach, stwierdzono kompensację iloczasów samogłosek pod wpływem następujących spółgłosek, albo nie występują w pełni w języku polskim (Eksperyment 2) lub przestawiają problem niezgodny z wyjaśnieniami takich efektów opartymi na typologii rytmicznej (Eksperyment 2). Dla obu eksperymentów dotyczących segmentów, zaproponowano inne wyjaśnienie wzdłużeń samogłosek. Mianowicie, jako efekt wzmożonego wysiłku związanego z fonacją spółgłoski następującej. Najsilniejszy taki efekt stwierdzono dla następujących głosek długich, dźwięcznych i trących, w przypadku których, ze względów aerodynamicznych, szczególnie trudno jest utrzymać fonację.

Praca wdraża nowatorską metodologię analizy rytmu oraz tempa w mowie oraz wnosi nowe spojrzenie na iloczas sylab i interwałów miedzyakcentowych w języku polskim. Wyniki osiągnięte w pracy mają w przyszłości znacznie ułatwić wyjasnienie czynników wpływających na, po pierwsze, postulowaną w literaturze "nietypowość" rytmu polskiego języka mówionego, a po drugie, kwestię stabilności iloczasu sylaby. Wyjaśnienie tych czynników przyczyni się również do udoskonalenia istniejących modeli rytmu oraz iloczasu w mowie.

References

Abercrombie, David. 1991. *Fifty years in phonetics*. Edinburgh: Edinburgh University Press.

Al-Tamimi, Feda. 2004. "An experimental phonetic study of intervocalic singleton and geminate sonorants in Jordanian Arabic", *Al-Arabiyya*, 29:37–52.

Allen, George D. 1975. "Speech rhythm: It's relation to performance universals and articulatory timing", *Journal of Phonetics*, 3, 75-86.

Arvaniti, Amalia. 2009. "Rhythm, timing and the timing of rhythm", *Phonetica*, 66, 1-2:46–63.

Asu, Eva Liina and Francis Nolan. 2005. "Estonian rhythm and the pairwise variability index", in: *FONETIK 2005*, Göteborg, Sweden: Göteborg University, 29–32.

Asu, Eva Liina and Francis Nolan. 2006. "Estonian and English rhythm: A two-dimensional quantification based on syllables and feet", in: *Proceedings of Speech Prosody 2006*, Dresden, 249–252.

Baayen, R. Harald, Doug J. Davidson and Douglas Bates. 2008. "Mixed-effects modeling with crossed random effects for subjects and items", *Journal of Memory and Language*, 59:390–412.

Barbosa, Plinio Almeida. 2000. "Syllable-timing in Brazilian Portuguese", *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 16:369 – 402.

Barbosa, Plinio Almeida. 2002. “Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production”, in: *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 163–166.

Barbosa, Plinio Almeida. 2006. *Incursões em torno do ritmo da fala* [Investigations of speech rhythm]. Campinas: Pontes.

Barbosa, Plinio Almeida. 2007. “From syntax to acoustic duration: A dynamical model of speech rhythm production”, *Speech Communication*, 49:725–742.

Barbosa, Plinio Almeida, Pablo Arantes, Alexsandro R. Meireles and Jussara M. Vieira. 2005. “Abstractness in speech-metronome synchronisation: P-centres as cyclic attractors”, in: *Proceedings of Interspeech 2005*, Lisbon, 1441–1444.

Barry, William J., Bistra Andreeva, Michaela Russo, Snezhina Dimitrova and Tanja Kostadinova. 2003. “Do rhythm measures tell us anything about language type?”, in: *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2693–2696.

Bates, Douglas, Martin Maechler and Ben Bolker. 2011. *lme4: Linear mixed-effects models using S4 classes*, URL `http://CRAN.R-project.org/package=lme4`, R package version 0.999375-39.

Beckman, Mary E. 1992. “Evidence for speech rhythms across languages”, in: Yoh’ichi Tohkura, Eric Vatikiotis-Bateson and Yoshinori Sagisaka (eds.), *Speech perception, production and linguistic structure*. Tokyo: OHM Publishing Co., 457–463.

Beckman, Mary E. and Jan R. Edwards. 1990. “Lengthenings and shortenings and the nature of prosodic constituency”, in: *Papers in laboratory phonology I: Between the grammar and the physics of speech*. Cambridge: Cambridge University Press, 152–178.

Bell-Berti, Fredericka and Katherine S. Harris. 1981. “A temporal model of speech production”, *Phonetica*, 38:9–20.

Bertinetto, Pier Marco. 1988. “Reflections on the dichotomy stress- vs. syllable-timing”, *Quaderni del Laboratorio di Linguistica, Scuola Normale Superiore, Pisa*, 2:59–85.

Bertinetto, Pier Marco and Chiara Bertini. 2007/2008. “Towards a unified predictive model of Natural Language Rhythm”, *Quaderni del Laboratorio di Linguistica*, 7.

Bladon, Anthony, Christopher Clark and Katrina Mickey. 1987. “Production and perception of sibilant fricatives: Shona data”, *Journal of the International Phonetic Association*, 17, 1:39–65.

Boersma, Paul and David Weenink. 2012. *Praat: Doing phonetics by computer. Version 5.3.04*, URL `http://www.praat.org/`.

Bouzon, Caroline and Daniel Hirst. 2004. “Isochrony and prosodic structure in British English”, in: *Proceedings of Speech Prosody 2004*, Nara, Japan, 223–226.

Braunschweiler, Norbert. 1997. “Integrated cues of voicing and vowel length in German: A production study”, *Language and Speech*, 40, 4:353–376.

Breen, Mara, Laura Dilley, John Kraemer and Edward Gibson. 2010. “Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)”, *Corpus Linguistics and Linguistic Theory*, 8:277–312.

Breuer, Stefan, Katarzyna Francuzik (Klessa) and Grażyna Demenko. 2006. “Analysis of Polish segmental duration with CART”, in: *Proceedings of Speech Prosody 2006*, Dresden, 137–140.

Browman, Catherine P. and Louis Goldstein. 1990. “Articulatory gestures as phonological units”, *Phonology*, 6:201–251.

Browman, Catherine P. and Louis M. Goldstein. 1992. “Articulatory phonology: An overview”, *Phonetica*, 49, 3-4:155–80.

Bunta, Ferenc and David Ingram. 2007. “The acquisition of speech rhythm by bilingual Spanish and English speaking 4- and 5-year-old children”, *Journal of Speech and Hearing Research*, 50, 4:999–1014.

Byrd, Dani and Elliot Saltzman. 2003. “The elastic phrase: Modeling the dynamics of boundary-adjacent lengthening”, *Journal of Phonetics*, 31:149–180.

Campbell, Nick. 1999. “A study of Japanese speech timing from the syllable perspective”, *Journal of the Acoustical Society of Japan*, 3, 2:29–39.

Chen, Matthew. 1970. “Vowel length variation as a function of the voicing of the consonant environment”, *Phonetica*, 22, 3:129–159.

Clark, Herbert H. 1973. “The language as-fixed-effect fallacy: A critique of language statistics in psychological research”, *Journal of Verbal Learning and Verbal Behavior*, 12:335–359.

Clements, George N. 2003. “Feature economy in sound systems”, *Phonology*, 20, 3:287–333.

Clements, George N. 2006. “Feature organization”, in: Keith Brown (ed.), *The Encyclopaedia of Language and Linguistics*. Oxford: Elsevier, vol. 4, 433–441.

Couper-Kuhlen, Elizabeth. 1993. *English speech rhythm: Form and function in everyday verbal interaction.* Amsterdam: Benjamins.

Crosswhite, Katherine. 2003. “Spectral tilt as a cue to stress in Polish, Macedonian and Bulgarian”, in: *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 767–770.

Crystal, Thomas H. and Arthur S. House. 1988. “Segmental duration in connected speech signals: Syllabic stress”, *Journal of the Acoustical Society of America*, 83, 4:1574–1585.

Cumming, Ruth E. 2011. “Perceptually informed quantification of speech rhythm in Pairwise Variability Indices”, *Phonetica*, 68, 4:256–277.

Cummins, Fred. 2002. “Speech rhythm and rhythmic taxonomy”, in: *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 121–126.

Cummins, Fred and Robert Port. 1998. “Rhythmic constraints on English stress timing”, *Journal of Phonetics*, 26, 2:145–171.

Dasher, Richard and Dwight L. Bolinger. 1982. “On pre-accentual lengthening”, *Journal of the International Phonetic Association*, 12:58–69.

Dauer, Rebecca M. 1983. “Stress-timing and syllable-timing re-analysed”, *Journal of Phonetics*, 11:51–62.

de Jong, Kenneth and Bushra Adnan Zawaydeh. 2002. “Comparing stress, lexical focus, and segmental focus: Patterns of variation in Arabic vowel duration”, *Journal of Phonetics*, 30:53–75.

Delattre, Pierre C. 1971. “Consonant gemination in four languages: An acoustic, perceptual and radiographic study (part I)”, *International Review of Applied Linguistics in Language Teaching*, 9, 1:31–53.

Dellwo, Volker. 2006. “Rhythm and speech rate: A variation coefficient for deltaC”, in: Pawel Karnowski and Imre Szigeti (eds.), *Language and language processing*. Frankfurt am Main: Peter Lang, 231–241.

Dellwo, Volker, Marie-José Kolly and Adrian Leemann. 2012. “Speaker identification based on speech temporal information: A forensic phonetic study of speech rhythm in the Zurich variety of Swiss German”, in: *Book of abstracts of the Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Santander, Spain.

Dellwo, Volker and Jacques Koreman. 2008. “How speaker idiosyncratic is measurable speech rhythm?”, in: *Book of abstracts of the Annual Conference of the International Association for Forensic Phonetics and Acoustics*, Lausanne, Switzerland.

Dellwo, Volker and Petra Wagner. 2003. “Relations between language rhythm and speech rate”, in: *15th International Congress of Phonetic Sciences*, Barcelona, 471–474.

Deterding, David. 1994. “The rhythm of Singapore English”, in: *5th Australian International Conference on Speech Science and Technology*, Perth, 316–321.

Deterding, David. 2001. “The measurement of rhythm: A comparison of Singapore English and British English”, *Journal of Phonetics*, 29, 2:217–230.

Dłuska, Maria. 1950. *Fonetyka polska* [The phonetics of Polish]. Warszawa: PWN Polskie Wydawnictwo Naukowe.

Dogil, Grzegorz. 1999. “The phonetic manifestation of word stress in Lithuanian, Polish and German and Spanish”, in: Harry van der Hulst (ed.), *Word prosodic systems in the languages of Europe*. Berlin: Mouton de Gruyter, 273–311.

Domahs, Ulrike, Johannes Knaus, Paula Orzechowska and Richard Wiese. 2012. “Stress deafness in a language with fixed word stress: An ERP study on Polish”, *Frontiers in Psychology*, 3, 439:1–15.

Dziubalska-Kołaczyk, Katarzyna. 2002. *Beats-and-binding phonology*. Frankfurt am Main: Peter Lang.

Eliasmith, Chris. 1995. *Mind as a dynamic system*, Master’s thesis, University of Waterloo, Waterloo, ON.

Elman, Jeffrey L. 1995. “Language as a dynamical system”, in: Robert Port and Timothy van Gelder (eds.), *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press, 195–226.

Eriksson, Anders. 1991. *Aspects of Swedish speech rhythm*, Ph.D. thesis, University of Göteborg, Göteborg.

Esposito, Anna and Maria Gabriella di Benedetto. 1999. “Acoustical and perceptual analysis of gemination”, *Journal of the Acoustical Society of America*, 106, 4:2051–2061.

Fastl, Hugo and Eberhard Zwicker. 2007. “Subjective duration”, in: *Psychoacoustics*. Springer Verlag, 265–269.

Ferragne, Emmanuel and François Pellegrino. 2004. “A comparative account of the suprasegmental and rhythmic features of British English dialects”, in: *Modelisations pour l’Identification des Langues*, Paris, France.

Ferragne, Emmanuel and François Pellegrino. 2007. “Automatic dialect identification: A study of British English”, in: *Speaker Classification II*. Berlin and Heidelberg: Springer Verlag, *Lecture Notes in Computer Science*, vol. 4441, 243–257.

Flege, James Emil and Robert Port. 1981. “Cross-language phonetic interference: Arabic to English”, *Language and Speech*, 24, 2:125–146.

Folkins, John W. and James H. Abbs. 1975. “Lip and jaw motor control during speech: Responses to resistive loading of the jaw”, *Journal of Speech and Hearing Research*, 18:207–220.

Fowler, Carol A. 1980. “Coarticulation and theories of extrinsic timing”, *Journal of Phonetics*, 8:113–133.

Fowler, Carol A. 1983. “Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet”, *Journal of Experimental Psychology*, 112, 3:386–412.

Fowler, Carol A. 1989. “Real objects of speech perception: A commentary on Diehl and Kluender”, *Ecological Psychology*, 1, 2:145–160.

Fraisse, Paul. 1963. *The psychology of time*. New York: Harper and Row.

Galves, Antonio, Jesus Garcia, Denise Duarte and Charlotte Galves. 2002. “Sonority as a basis for rhythmic class discrimination”, in: *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 11–13.

Gibbon, Dafydd. 2003. “Computational modelling of rhythm as alternation, iteration and hierarchy”, in: *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 2489–2492.

Gibbon, Dafydd. 2006. “Time types and time trees: Prosodic mining and alignment of temporally annotated data”, in: Stefan Sudhoff (ed.), *Methods in Empirical Prosody Research*. Walter de Gruyter, 281–209.

Gibbon, Dafydd, Jolanta Bachan and Grazyna Demenko. 2007. “Syllable timing patterns in Polish: Results from annotation mining”, in: *Proceedings of Interspeech/Eurospeech 2007*, Antwerp, 994–997.

Gibbon, Dafydd and Flaviane Romani Fernandes. 2005. “Annotation mining for rhythm model comparison in Brazilian Portuguese”, in: *Proceedings of Interspeech 2005*, Lisbon, 3289–3292.

Gibbon, Dafydd and Ulrike Gut. 2001. “Measuring speech rhythm”, in: *Proceedings of Eurospeech*, Aalborg, Denmark, 91–94.

Gibson, James J. 1975. “Events are perceivable but time is not”, in: J.T. Fraser and N. Lawrence (eds.), *The study of time*. New York: Springer Verlag, 295–301.

Grabe, Esther and Ee Ling Low. 2002. “Durational variability in speech and the rhythm class hypothesis”, in: Carlos Gussenhoven and Natasha Warner (eds.), *Papers in laboratory phonology VII*. Berlin and New York: Mouton de Gruyter, 515–546.

Gracco, Vincent. 1988. “Timing factors in the coordination of speech movements”, *The Journal of Neuroscience*, 8, 12:4628–4639.

Gray, Wayne D. 2012. “Great debate on the complex systems approach to cognitive science”, *Topics in Cognitive Science*, 4, 1:1–94.

Grondin, Simon. 2010. “Timing and time perception: A review of recent behavioral and neuroscience findings and theoretical directions”, *Attention, Perception and Psychophysics*, 72, 3:561–582.

Guaitella, Isabelle. 1999. “Rhythm in speech: What rhythmic organizations reveal about cognitive processes in spontaneous speech production versus reading aloud”, *Journal of Pragmatics*, 31:509–523.

Haken, Hermann. 1982. *Synergetik*. Berlin, Heidelberg, New York: Springer Verlag.

Haken, Hermann, J. A. Scott Kelso and Heinz Bunz. 1985. "A theoretical model of phase transitions in human hand movements", *Biological Cybernetics*, 51:347–356.

Ham, William H. 2001. *Phonetic and phonological aspects of geminate timing*. New York, London: Routledge.

Hayes, Bruce. 1984. "The phonology of rhythm in English", *Linguistic Inquiry*, 15:33–74.

Hayes, Bruce and Stanisław Puppel. 1985. "On the rhythm rule in Polish", in: Harry van der Hulst and Norval Smith (eds.), *Advances in Nonlinear Phonology*. Dordrecht: Foris, 59–81.

House, Arthur S. and Grant Fairbanks. 1953. "The influence of consonant environment upon the secondary acoustical characteristics of vowels", *Journal of the Acoustical Society of America*, 25, 1:105–113.

Idemaru, Kaori and Susan Guion. 2008. "Acoustic covariants of length contrast in Japanese stops", *Journal of the International Phonetic Association*, 38:167–186.

Imiołczyk, Janusz, Ignacy Nowak and Grażyna Demenko. 1994. "High intelligibility text-to-speech synthesis for Polish", *Archives of Acoustics*, 19, 2:161–172.

Jassem, Wiktor. 1962. *Akcent jezyka polskiego* [The accent in Polish]. Wrocław: Ossolineum.

Jassem, Wiktor, David R. Hill and Ian H. Witten. 1984. "Isochrony in English speech: Its statistical validity and linguistic relevance", in: Dafydd Gibbon and Helmut Richter (eds.), *Intonation, accent and rhythm: studies in discourse phonology.* Berlin: Walter de Gruyter, 203–225.

Jassem, Wiktor and Lutosława Richter. 1989. “Neutralization of voicing in Polish obstruents”, *Journal of Phonetics*, 17:317–325.

Jones, Mari Riess and Marilyn Boltz. 1989. “Dynamic attending and responses to time”, *Psychological Review*, 96:459–491.

Karpiński, Maciej, Ewa Jarmołowicz-Nowikow and Zofia Malisz. 2008a. “Aspects of gestural and prosodic structure of multimodal utterances in Polish task-oriented dialogue”, *Speech and Language Technology*, XI:113–122.

Karpiński, Maciej, Ewa Jarmołowicz-Nowikow, Zofia Malisz, Konrad Juszczyk and Michal Szczyszek. 2008b. “Rejestracja, transkrypcja i tagowanie mowy oraz gestów w narracji dzieci i dorosłych [The recording, transcription and annotation of speech and gesture by adults and children in a narration corpus]”, *Investigationes Linguisticae*, XVI:83–98.

Kawahara, Shigeto. 2011. “Experimental approaches in theoretical phonology”, in: Marc van Oostendorp, Colin J. Ewen, Elizabeth Hume and Keren Rice (eds.), *The Blackwell Companion to Phonology*. Blackwell Publishers Ltd, 2283–2303.

Keane, Elinor. 2006. “Rhythmic characteristics of colloquial and formal Tamil”, *Language and Speech*, 3, 49:299–332.

Keating, Patricia. 1979. *A phonetic study of a voicing contrast in Polish*, Ph.D. thesis, Brown University, Providence, RI.

Keating, Patricia. 1985. “Universal phonetics and the organization of grammars”, in: Victoria Fromkin (ed.), *Phonetic Linguistics*. Academic Press, 115–132.

Keating, Patricia. 1990. “Phonetic representations in a generative grammar”, *Journal of Phonetics*, 18:321–334.

Keller, Brigitte Zellner and Eric Keller. 2002. “Representing speech rhythm”, in: *Working Papers of COST*, vol. 258, 154–164.

Kelso, J. A. Scott. 1995. *Dynamic patterns: The self organization of brain and behavior*. MIT Press.

Kelso, J. A. Scott, Betty Tuller and Carol A. Fowler. 1972. “The functional specificity of articulatory control and coordination”, *Journal of the Acoustical Society of America*:S103.

Kim, Heejin and Jennifer Cole. 2005. “The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase”, in: *Proceedings of Interspeech 2005*, Lisbon, 2365–2368.

Klatt, Dennis H. 1976. “Linguistic uses of segmental duration in English: Acoustic and perceptual evidence”, *Journal of the Acoustic Society of America*, 59, 5:1208–1221.

Klessa, Katarzyna. 2006. *Modelowanie iloczasu głoskowego na potrzeby syntezy mowy polskiej* [Modelling segmental duration for Polish speech synthesis purposes], Ph.D. thesis, Adam Mickiewicz University, Poznań.

Klessa, Katarzyna, Stefan Breuer and Grażyna Demenko. 2007. “Optimization of Polish segmental duration prediction with CART”, in: *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, 77–80.

Kluender, Keith R., Randy L. Diehl and Beverly A. Wright. 1988. “Vowel length differences before voiced and voiceless consonants: an auditory explanation”, *Journal of Phonetics*, 16:153–169.

Kohler, Klaus. 1977. “The production of plosives”, *Arbeitsberichte des Instituts für Phonetik an der Universität Kiel (AIPUK)*, 8:30–110.

Kohler, Klaus. 2003. “Domains of temporal control in speech and language. From utterance to segment”, in: *15th International Congress of Phonetic Sciences*, Barcelona, 7–10.

Kohler, Klaus. 2007. “Beyond laboratory phonology: The phonetics of speech communication”, in: Maria-Josep Solé, Patrice Speeter Beddor and Manjari Ohala (eds.), *Experimental approaches to phonology*. Oxford University Press, 41–53.

Kopell, Nancy. 1988. “Toward a theory of modelling central pattern generators”, in: Avis H. Cohen, Serge Rossignol and Sten Grillner (eds.), *Neural control of rhythmic movements in vertebrates*. New York: John Wiley & Sons, 369–413.

Ladefoged, Peter. 1971. *Preliminaries to linguistic phonetics*. Chicago, IL: The University of Chicago Press.

Lee, Christopher S. and Neil P. Mcangus Todd. 2004. “Towards an auditory account of speech rhythm: Application ofa model of the auditory ’primal sketch’ to two multi-language corpora”, *Cognition*, 93:225–254.

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge, MA: MIT Press.

Lehiste, Ilse. 1977. “Isochrony reconsidered.”, *Journal of Phonetics*, 5, 3:253–263.

Lindblom, Björn. 1963. “Spectrographic study of vowel reduction”, *Journal of the Acoustical Society of America*, 35:517–525.

Lisker, Leigh. 1957. “Closure duration and the intervocalic voiced-voiceless distinction in English”, *Language*, 33:42–49.

Lisker, Leigh. 1978. “In qualified defense of VOT”, *Language and Speech*, 21, 4:375–383.

Lisker, Leigh. 1986. “’Voicing’ in English: A catalog fo acoustic features signaling /b/ versus /p/ in trochees”, *Language and Speech*, 29:3–11.

Liss, Julie M., Laurence White, Sven L. Mattys, Kaitlin Lansford, Andrew J. Lotto, Stephanie M. Spitzer and John N. Caviness. 2009. “Distinguishing dysarthrias using rhythm metrics”, *Journal of Speech and Hearing Research*, 52:1334–1352.

Löfqvist, Anders. 2010. “Theories and models of speech production”, in: William J. Hardcastle, John Laver and Fiona E. Gibbon (eds.), *Handbook of phonetic sciences*. John Wiley & Sons, 353–377.

Loukina, Anastassia, Greg Kochanski, Chilin Shih, Elinor Keane and Ian Watson. 2009. "Rhythm measures with language-independent segmentation", in: *Proceedings of Interspeech 2009*, 1531–1534.

Low, Ee Ling. 1994. *Intonation patterns in Singapore English*, Master's thesis, Cambridge University.

Low, Ee Ling, Esther Grabe and Francis Nolan. 2000. "Quantitative characterisations of speech rhythm: 'Syllable-timing' in Singapore English", *Language and Speech*, 43:377–401.

Luce, Paul A. and Jan Charles-Luce. 1985. "Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production", *Journal of the Acoustical Society of America*, 78:1949–1957.

Machač, Pavel and Radek Skarnitzl. 2007. "Temporal compensation in Czech?", in: *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 537–540.

Maddieson, Ian. 1984. "Phonetic cues to syllabification", *UCLA Working Papers in Phonetics*, 59:85–101.

Malisz, Zofia. 2004. *Speech Rhythm: Solutions for Temporal Phenomena in Natural Speech*, Master's thesis, School of English, Adam Mickiewicz University, Poznań.

Malisz, Zofia. 2005. "Speech cycling tasks for Polish", in: *Poznań Linguistic Meeting 2005*, Poznań.

Malisz, Zofia. 2006. "Segment-prosody interaction and phonetic models of speech rhythm and rhythmic typology", *Quaderni del Laboratorio di Linguistica, Scuola Normale Superiore, Pisa*, 6.

Malisz, Zofia. 2009. "Vowel duration in pre-geminate contexts in Polish", in: *Proceedings of INTERSPEECH 2009*, Brighton, UK, 1547–1550.

Malisz, Zofia. 2011. “Tempo differentiated analyses of timing in Polish”, in: *Proceedings of the 17th International Congress of Phonetic Sciences*, Hong Kong, 1322–1325.

Malisz, Zofia and Katarzyna Klessa. 2008. “A preliminary study of temporal adaptation in Polish VC groups”, in: Plinio Almeida Barbosa and Cesar Reis (eds.), *Proceedings of Speech Prosody 2008*, Campinas: Editora RG/CNPq, 383–386.

Malisz, Zofia and Petra Wagner. 2012. “Acoustic-phonetic realisation of Polish syllable prominence: A corpus study”, *Speech and Language Technology*, 14/15:105–114.

Malisz, Zofia, Marzena Żygis and Bernd Pompino-Marschall. 2013. “Rhythmic structure effects on glottalisation: A study of different speech styles in Polish and German”, *Laboratory Phonology - Journal of the Association for Laboratory Phonology*, 4, 1:119–158.

Martin, James G. 1972. “Rhythmic (hierarchical) versus serial structure in speech and other behavior”, *Psychological Review*, 79:487–509.

Mary, Leena and Bayya Yegnanarayana. 2008. “Extraction and representation of prosodic features for language and speaker recognition”, *Speech Communication*, 50:782–796.

McAuley, J. Devin. 1995. *Perception of time as phase: Toward and adaptive-oscillator model of rhythmic pattern processing*, Ph.D. thesis, Indiana University, Bloomington, IN.

McCarthy, John J. 1986. “OCP effects: Gemination and antigemination”, *Linguistic Inquiry*, 17:237–293.

Meireles, Alexsandro R., João Paulo Tozetti and Rogério R. Borges. 2010. “Speech rate and rhythmic variation in Brazilian Portuguese”, in: *Proceedings of Speech Prosody 2010*, Chicago, IL, 1–4.

Mitleb, Fares M. 1984. “Voicing effect on vowel duration is not an absolute universal”, *Journal of Phonetics*, 12:23–27.

Morton, John, Steve Marcus and Clive Frankish. 1976. “Perceptual centers (p-centers)”, *Psychological Review*, 83, 5:405–408.

Nakajima, Yoshitaka, Gert ten Hoopen, Gaston Hilkhuysen and Takayuki Sasaki. 1992. “Time-shrinking: A discontinuity in the perception of auditory temporal intervals”, *Perception and Psychophysics*, 51, 5:504–507.

Nam, Hosung, Louis Goldstein and Elliot L. Saltzman. 2010. “Self-organization of syllable structure: A coupled oscillator model”, in: Francois Pellegrino, Egidio Marsico, Ioana Chitoran and Christophe Coupé (eds.), *Approaches to phonological complexity*. Berlin and New York: Mouton de Gruyter, 299–328.

Nazzi, Thierry, Josiane Bertoncini and Jacques Mehler. 1998. “Language discrimination by newborns: Towards an understanding of the role of rhythm”, *Journal of Experimental Psychology: Human Perception and Performance*, 24:756–766.

Nespor, Marina. 1990. “On the rhythm parameter in phonology”, in: Iggy Roca (ed.), *Logical issues in language acquisition*. Dordrecht: Foris, 157–175.

Newlin-Łukowicz, Luiza. 2012. “Polish stress: A phonetic investigation of phonological claims”, in: *The 36th Penn Linguistics Colloquium*, Philadelphia.

Norton, Alec. 1995. “Dynamics: An introduction”, in: Robert Port and Timothy van Gelder (eds.), *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press, 45–68.

Nowak, Paweł. 2006a. “The role of vowel transitions and frication noise in the perception of Polish sibilants”, *Journal of Phonetics*, 34, 2:139–152.

Nowak, Paweł. 2006b. *Vowel reduction in Polish*, Ph.D. thesis, University of California, Berkeley, CA.

O’Dell, Michael L. 2003. *Intrinsic timing and quantity in Finnish*, Ph.D. thesis, University of Tampere, Finland.

O'Dell, Michael L. and Tommi Nieminen. 1999. "Coupled oscillator model of speech rhythm", in: *14th International Congress of Phonetic Sciences*, San Francisco, 1075–1078.

O'Dell, Michael L. and Tommi Nieminen. 2009. "Coupled oscillator model for speech timing: Overview and examples", in: *Nordic Prosody: Proceedings of the 10th conference*, Helsinki, 179–190.

Ogden, Richard. 1996. "Where is timing? A response to Caroline Smith", in: Amalia Arvaniti and Bruce Connel (eds.), *Papers in laboratory phonology IV: Phonology and phonetic evidence*. Cambridge University Press, 223–234.

Oh, Grace E. and Melissa A. Redford. 2012. "The production and representation of fake geminates in English", *Journal of Phonetics*, 40, 1:82–91.

Ohala, John J. 1983. "The origin of sound patterns in vocal tract constraints", in: Peter F. MacNeilage (ed.), *The production of speech*. New York: Springer Verlag, 186–192.

Ohala, Manjari and John J. Ohala. 1992. "Phonetic universals and Hindi segment durations", in: J.J. Ohala, T. Nearey, B. Derwing, M. Hodge and G. Wiebe (eds.), *Proceedings of International Conference on Spoken Language Processing*, Banff: University of Alberta, Edmonton, 309–355.

Öhman, Sven E.G. 1966. "Coarticulation in VCV utterances: Spectrographic measurements", *Journal of the Acoustical Society of America*, 39, 1:151–168.

Ordin, Mikhail, Leona Polyanskaya and Christiane Ulbrich. 2011. "Acquisition of timing patterns in second language", in: *Proceedings of Interspeech 2011*, Florence, Italy, 1129–1132.

Pająk, Bożena. 2010. "Contextual constraints on geminates: The case of Polish.", in: *Proceedings of the 35th Annual Meeting of the Berkeley Linguistics Society*, Berkeley, CA: University of California, 269–280.

Pająk, Bożena and Eric Bakovic. 2010. "Assimilation, antigemination, and contingent optionality: The phonology of monoconsonantal proclitics in Polish", *Natural Language and Linguistic Theory*, 28:643–680.

Patel, Aniruddh D. 2010. *Music, language and the brain*. Oxford University Press.

Patel, Aniruddh D. and Joseph R. Daniele. 2003. "An empirical comparison of rhythm in language and music", *Cognition*, 87, 1:B35–B45.

Payne, Elinor. 2005. "Phonetic variation in Italian consonant gemination", *Journal of the International Phonetic Association*, 35, 2:153–181.

Perkell, Joseph S. and Dennis H. Klatt (eds.). 1986. *Invariance and variability in speech processes*. Hillsdale, NJ and London: Lawrence Erlbaum.

Pike, Kenneth. 1945. *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Pind, Jürgen. 1995. "Constancy and normalization in the perception of voice offset time as a cue for preaspiration", in: *Acta Psychologica*, vol. 89, 53–91.

Port, Robert. 1977. *The influence of speaking tempo on the duration of stressed vowel and medial stop in English trochee words*, Ph.D. thesis, University of Connecticut, Storrs, CI.

Port, Robert. 1981. "Linguistic timing factors in combination", *Journal of the Acoustical Society of America*, 69, 1:262–174.

Port, Robert. 2003. "Meter and speech", *Journal of Phonetics*, 31:599–611.

Port, Robert. 2013. "Coordinative structures for the control of speech production", webpage, last checked 1 April 2013, URL `www.cs.indiana.edu/ port/teach/641/coord.strctr.html`.

Port, Robert, Salman Al-Ani and Shosaku Maeda. 1980. "Temporal compensation and universal phonetics", *Phonetica*, 37:235–252.

Port, Robert and Jonathan Dalby. 1982. "C/V ratio as a cue for voicing in English", *Perception and Psychophysics*, 2:141–52.

Port, Robert and Adam P. Leary. 2000. "Speech timing in linguistics", Tech. rep., Department of Linguistics, Indiana University, Bloomington, IN.

Port, Robert and Michael O'Dell. 1985. "Neutralization of syllable-final voicing in German", *Journal of Phonetics*, 13:455–71.

Port, Robert, Keiichi Tajima and Fred Cummins. 1999. "Speech and rhythmic behavior", in: Geert J. P. Savelsburgh, Han van der Maas and Paul C. L. van Geert (eds.), *The non-linear analysis of developmental processes*. Amsterdam: Elsevier, 5–45.

Port, Robert and Timothy van Gelder (eds.). 1995. *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Prieto, Pilar, Maria del Mar Vanrell, Lluisa Astruc, Elinor Payne and Brechtje Post. 2012. "Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English and Spanish", *Speech Communication*, 54, 6:681–702.

Prince, Alan S. 1983. "Relating to the grid", *Linguistic Inquiry*, 14, 1:19–100.

Querleu, Denis, Xavier Renard, Fabienne Versyp, Laurence Paris-Delrue and Gilles Crèpin. 1988. "Fetal hearing", *European Journal of Obstetrics and Gynecology and Reproductive Biology*, 28, 3:191–212.

R Development Core Team. 2011. *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, URL `http://www.R-project.org/`, ISBN 3-900051-07-0.

Ramus, Franck. 2002. "Acoustic correlates of linguistic rhythm: Perspectives", in: *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 115–120.

Ramus, Franck, Emmanuel Dupoux and Jacques Mehler. 2003. "The psychological reality of rhythm classes: Perceptual studies", in: *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 337–342.

Ramus, Franck, Marina Nespor and Jacques Mehler. 1999. "Correlates of linguistic rhythm in the speech signal", *Cognition*, 73:265–292.

Richter, L. 1973. "The duration of Polish vowels", *Speech Analysis and Synthesis*, 3/1973:87–115.

Ridouane, Rachid. 2010. “Gemination at the junction of phonetics and phonology”, in: Cecile Fougeron, Barbara Kühnert, Mariapaola d’Imperio and Nathalie Vallee (eds.), *Papers in laboratory phonology 10: Variation, Detail and Representation*. de Gruyter Mouton, 61–91.

Roach, Peter. 1982. “On the distinction between ’stress-timed’ and ’syllable-timed’ languages”, in: David Crystal (ed.), *Linguistic controversies*. London: Edward Arnold, 73–79.

Rojczyk, Arkadiusz. 2010. *Temporal and spectral parameters in perception of the voicing contrast in English and Polish*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Rosen, Kristin M. 2005. “Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison”, *Journal of Phonetics*, 33, 4:411–426.

Rouas, Jean-Luc, Jerome Farinas, François Pellegrino and Regine Andre-Obrecht. 2005. “Rhythmic unit extraction and modelling for automatic language identification”, *Speech Communication*, 47:436–456.

Rubach, Jerzy. 1974. “Syllabic consonants in Polish”, *Journal of Phonetics*, 2:109–116.

Rubach, Jerzy and Geert Booij. 1985. “A grid theory of stress in Polish”, *Lingua*, 66:281–319.

Russo, Michaela and William J. Barry. 2008. “Isochrony reconsidered. Objectifying relations between rhythm measures and speech tempo”, in: *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 419–422.

Saltzman, Elliot L. and Dani Byrd. 2000. “Task-dynamics of gestural timing: Phase windows and multifrequency rhythms”, *Human Movement Science*, 19:499–526.

Saltzman, Elliot L., Anders Löfqvist and Subhobrata Mitra. 2000. “’Glue’ and ’clocks’: Intergestural cohesion and global timing”, in: Michael B. Broe and

Janet B. Pierrehumbert (eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*. London: Cambridge University Press, 88–101.

Saltzman, Elliot L. and Kevin G. Munhall. 1989. "A dynamical approach to gestural patterning in speech production", *Ecological Psychology*, 1, 4:333–382.

Saltzman, Elliot L., Hosung Nam, Jelena Krivokapić and Louis Goldstein. 2008. "A task-dynamic toolkit for modeling the effects of prosodic structure on articulation", in: Plinio Almeida Barbosa, Sandra Madureira and Cesar Reis (eds.), *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 175–184.

Sasaki, Takayuki, Daigoh Suetomi, Yoshitaka Nakajima and Gert ten Hoopen. 2002. "Time-shrinking, its propagation, and gestalt principles", *Perception and Psychophysics*, 64, 6:919–931.

Sawicka, Irena. 1995. "Fonologia [Phonology]", in: Henryk Wróbel (ed.), *Gramatyka współczesnego języka polskiego. Fonetyka i fonologia* [Modern Polish grammar. Phonetics and phonology]. Instytut Języka Polskiego, PAN, 7–111.

Scott, Donia R., Stephen D. Isard and Bénédicte de Boysson-Bardies. 1986. "On the measurement of rhythmic irregularity: A reply to Benguerel", *Journal of Phonetics*, 13:327–330.

Selkirk, Elizabeth. 1984. *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.

Shockley, Kevin, Daniel C. Richardson and Rick Dale. 2009. "Conversation and coordinative structures", *Topics in Cognitive Science*, 1:305–319.

Slowiaczek, Louisa M. and Daniel A. Dinnsen. 1985. "On the neutralizing status of Polish final devoicing", *Journal of Phonetics*, 13:325–341.

Smith, Caroline L. 1995. "Prosodic patterns in the coordination of vowel and consonant gestures", in: Bruce Connel and Amalia Arvaniti (eds.), *Phonology and phonetic evidence*. Cambridge University Press, *Papers in laboratory phonology*, vol. IV, 205–222.

Spence, Melanie J. and Mark S. Freeman. 1996. “Newborn infants prefer the maternal low-pass filtered voice, but not the maternal whispered voice”, *Infant Behavior and Development*, 19, 2:199–212.

Steffen-Batogowa, Maria. 2000. *Struktura akcentowa języka polskiego*. Warszawa: PWN Polskie Wydawnictwo Naukowe.

Stetson, Raymond Herbert. 1951. *Motor phonetics*. Amsterdam: North-Holland.

Tajima, Keiichi and Robert Port. 2003. “Speech rhythm in English and Japanese”, in: John Local, Richard Ogden and Rosalind Temple (eds.), *Papers in laboratory phonology VI: Phonetic interpretation*. Cambridge: Cambridge University Press, 322–339.

Tajima, Keiichi, Bushra Adnan Zawaydeh and Mafuyu Kitahara. 1999. “Cross-linguistic comparison of speech rhythm using a speech cycling task”, in: *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, CA, 285–288.

Tatham, Mark and Katherine Morton. 2002. “Computational modelling of speech production: English rhythm.”, in: Angelika Braun und Herbert R. Masthoff (ed.), *Festschrift für Jens-Peter Köster zu Ehren seines 60. Geburtstags*. Stuttgart: Franz Steiner, 383–405.

Thurgood, Elżbieta and Grażyna Demenko. 2003. “Phonetic realizations of Polish geminate affricates”, in: *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 1895–1898.

Tilsen, Sam. 2009. “Multitimescale dynamical interactions between speech rhythm and gesture.”, in: *Cognitive Science*, vol. 33, 839–879.

Tilsen, Sam. 2011. “Metrical regularity facilitates speech planning and production”, *Laboratory Phonology - Journal of the Association for Laboratory Phonology*, 2, 1:185–218.

Tserdanelis, Georgios and Amalia Arvaniti. 2001. “The acoustic characteristics of geminate consonants in Cypriot Greek”, in: *Proceedings of the 4th International Conference on Greek Linguistics*, Thessaloniki, Greece, 29–36.

Tuller, Betty and J. A. Scott Kelso. 1991. “The production and perception of syllable structure”, *Journal of Speech and Hearing Research*, 34:501–508.

Tuller, Betty, J. A. Scott Kelso and Katherine S. Harris. 1983. “Converging evidence for the role of relative timing in speech”, *Journal of Experimental Psychology*, 9, 5:829–833.

Turk, Alice, Satsuki Nakai and Mariko Sugahara. 2006. “Acoustic segment durations in prosodic research: A practical guide”, in: S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert and P. Augurzky et al. (eds.), *Methods in Empirical Prosody Resarch*. New York, Berlin: Mouton de Gruyter, 1–28.

Turk, Alice and Stefanie Shattuck-Hufnagel. 2013. “What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić and Goswami and Leong”, *Laboratory Phonology - Journal of the Association for Laboratory Phonology*, 4, 1:93–118.

Turvey, Michael T. 1990. “Coordination”, *American Psychologist*, 45, 8:938–953.

Vihman, Marilyn May, Satsuki Nakai and Rory DePaolis. 2006. “Getting the rhythm right: A cross-linguistic study of segmental duration in babbling and first words”, in: Louis M. Goldstein, Douglas H. Whalen and Catherine T. Best (eds.), *Laboratory Phonology VIII: Varieties of Phonological Competence*. Mouton de Gruyter, 341–366.

Wagner, Petra. 2005. “Great Expectations: Introspective vs. perceptual prominence ratings and their acoustic correlates”, in: *Proceedings of Interspeech 2005*, Lisbon, 2381–2384.

Wagner, Petra. 2007. “Visualizing levels of rhythmic organization”, in: *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, 1113–1116.

Wagner, Petra and Volker Dellwo. 2004. “Introducing YARD (yet another rhythm determination) and reintroducing isochrony to rhythm research”, in: *Proceedings of Speech Prosody 2004*, Nara, Japan, 227–230.

White, Laurence and Sven L. Mattys. 2007a. “Calibrating rhythm: First language and second language studies”, *Journal of Phonetics*, 35, 4:501–522.

White, Laurence and Sven L. Mattys. 2007b. “Rhythmic typology and variation in first and second languages”, in: Pilar Prieto, Joan Mascaró and Maria-Josep Solé (eds.), *Segmental and prosodic issues in Romance phonology*. Current issues in linguistic theory, Amsterdam: John Benjamins, 237–257.

Wiget, Lukas, Laurence White, Barbara Schuppler, Izabelle Grenon, Olesya Rauch and Sven L. Mattys. 2010. “How stable are acoustic metrics of contrastive speech rhythm?”, *Journal of the Acoustical Society of America*, 127, 3:1559–1569.

Windmann, Andreas, Igor Jauk, Fabio Tamburini and Petra Wagner. 2011. “Prominence-based prosody prediction for unit selection speech synthesis”, in: *Proceedings of Interspeech 2011*, Florence, Italy, 325–328.

Wittmann, Marc. 1999. “Time perception and temporal processing levels of the brain”, *Chronobiology International*, 16, 1:17–32.

Xiu, Yi. 2008. “Multi-dimensional information coding in speech”, in: *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 17–26.

Zmarich, Claudio, Barbara Gili Fivela, Pascal Perrier, Christophe Savariaux and Graziano Tisato. 2011. “Speech timing organization for the phonological length contrast in Italian consonants”, in: *Proceedings of Interspeech 2011*, Florence, Italy, 401–404.

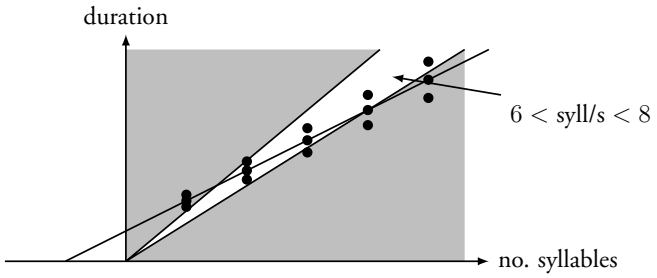
Appendix A: A method for coupling strength estimation at different rates, by Michael O'Dell

What follows is an unpublished manuscript written by Michael O'Dell and attached with permission of the author. The first procedure described in the following manuscript was implemented in Experiment 1 as described in Subsection 3.3.1.5 of this thesis.

Estimating coupling strength at different tempos

(cf. Bertinetto & Bertini 2010)

There is a serious problem with the heuristic of estimating speech tempo from data as syllables per second for the purpose of evaluating a possible correlation between tempo and coupling strength estimated using a regression. To see this, note that restricting cases to a certain range of syllables per second, say between 6 and 8, will make the regression line move closer to the origin, thus biasing the coupling strength estimate towards zero.



Is there a better way?

We can make the assumption that tempo distribution is independent of the number of syllables. If this assumption is false, then it will be next to impossible to assess the two effects separately.

The problem is to decide which data points for a given syllable count correspond to which data points for a different syllable count. For instance, how do we decide which measured durations of 2 syllables represent the same tempo as a given measured duration of 5 syllables? Whatever procedure is chosen, the assumption of independence provides a diagnostic indication of possible failure: If there is a systematic relation between syllable count and estimated tempo distribution, then the procedure has failed.[2] For instance, we expect the above procedure of grouping the data based on syllables per second to produce systematically faster tempo estimates as syllable count increases (cf. above diagram).

[2] Unfortunately, lack of an obvious systematic relation is no guarantee that the procedure has succeeded.

Call a procedure which does not fail this test (of empirical independence) *consistent*. Obviously any procedure which divides the data for each syllable count into proportional groups will be consistent. If we assume that slower tempo always means greater duration within a given syllable count, then tempo will be a monotonically increasing function of the cumulative distribution function (CDF) for each syllable count and therefore (with enough data) the inverse empirical CDF can be used as an estimate of tempo class which will at least be consistent.

A simple procedure based on standard (frequentist) regression analysis could be as follows: partition the sorted data points for each syllable count into groups based on various ranges of quantiles (identical for all syllable counts) and carry out a regression on each group. For instance, take the bottom 10 % of durations for each syllable count as roughly representing one tempo and perform a regression, then do the same for the next 10 %, etc. This would provide some indication as to how tempo affects the slopes and intercepts and therefore the relative coupling strength.

A further regression on the slopes (say b_i) and intercepts (say a_i) of the individual regression lines could then be carried out to estimate the relationship between tempo and relative coupling strength. If the coefficient c_1 in $(a_i/b_i) = c_0 + c_1(1/b_i)$ is negative, then increasing tempo decreases relative coupling strength; if positive, the opposite.

A cautionary note: As mentioned above, this procedure assumes that slower tempo always means greater duration once syllable count is fixed. However, additional variation in duration (due to segmental differences, measurement error, etc.) could undermine this assumption. Also, in an oscillator model including other oscillators (rhythmic levels), ignoring these levels will be introduce extra variation. In general, unexplained variation will bias the estimate of c_1 (see above) toward smaller values, possibly even changing its sign. Any additional information about the data (covariates such as syllable type, segmental content, etc.) could help to restore confidence in the assumption of a monotonic relation between tempo and duration, once the covariates are given.

A somewhat more robust procedure (but with more assumptions) could be as follows. Perform two regressions, one on mean durations, $\mu(n) = c_{\mu,0} + c_{\mu,1}n$ and one on standard deviations, $\sigma(n) = c_{\sigma,0} + c_{\sigma,1}n$ (where n is syllable count). Then we can solve for the value of n which gives zero σ: $n_0 = -c_{\sigma,0}/c_{\sigma,1}$, and the corresponding value of μ: $\mu_0 = c_{\mu,0} - c_{\mu,1}c_{\sigma,0}/c_{\sigma,1}$. Now a negative μ_0 would indicate that increasing tempo decreases relative coupling strength, whereas positive μ_0 would indicate the opposite. This again assumes variance is entirely due to tempo changes, which is not likely to be the case, but if we assume an independent (additive) error variance, it can be estimated at the same time. Instead of a linear regression

on standard deviations, we perform a quadratic regression on variances, $\sigma^2(n) = V(n) = c_{V,0} + c_{V,1}n + c_{V,2}n^2$. Now n_0 can be estimated as $n_0 = -c_{V,1}/(2c_{V,2})$, and plugging into the regression formula for the means gives $\mu_0 = c_{\mu,0} - c_{\mu,1}c_{V,1}/(2c_{V,2})$ (same interpretation as before).[3] In either case, the empirical regression might give uninterpretable results. In particular, n_0 should be restricted to $n_0 < 1$ regardless of the regression results, otherwise increasing tempo would increase duration for some n.

Naturally Bayesian counterparts to the above procedures could be carried out more flexibly and with the added advantage that instead of point estimates for parameters (such as μ_0) we have posterior distributions for them, which allows an assessment of the probability of interesting value ranges (say, negative vs. positive).

The problems in grouping data into tempo classes based on the data can be sidestepped if we have independent knowledge of tempo classification, for instance if subjects were instructed to speak at different rates. Another situation where these problems can be avoided is when tempo differences are very large and durations naturally fall into non-overlapping regions corresponding to discrete tempo classes.

[3] Error variance is then estimated as $\sigma_\epsilon^2 = c_{V,0} - c_{V,1}^2/(4c_{V,2})$.