On the analysis of speech rhythm through acoustic parameters

Ingmar Steiner
Department of Phonetics, Saarland University
steiner@coli.uni-sb.de

March 30, 2005

Abstract

Several recent studies have rekindled interest in the traditional distinction of languages into rhythmic classes (e.g. stress-timed, syllable-timed) and presented evidence in the form of durational acoustic parameters in favor of such an interpretation. The present paper discusses some of the methods and data used in such work, and attempts to refine them in an independent study investigating the homogeneity of what has been described as acoustic correlates of rhythm. Adopting a variable consonant/vowel distinction during data preparation, this study shows that certain consonant classes are significantly more efficient than others in influencing the distribution of individual languages along durational parameters. These results support the view that rhythmic language classification is strongly determined by phonotactic factors.

Zusammenfassung

Mehrere Studien haben in den letzten Jahren das Interesse an der traditionellen Unterscheidung von Sprachen in Rhythmusklassen (z.B. akzentzählend, silbenzählend) neu entfacht und in Form von akustischen Dauerparametern Belege für eine solche Interpretation geliefert. Der vorliegende Artikel diskutiert einige der Methoden und Daten, die in diesen Arbeiten Verwendung gefunden haben und versucht, sie in einer unabhängigen Studie zu verfeinern, um die Homogenität dessen zu untersuchen, was als akustische Korrelate von Rhythmus beschrieben wurde. Indem eine variable Konsonant/Vokal-Unterscheidung bei der Datenaufbereitung zum Einsatz kommt, zeigt diese Studie, daß einige Konsonantenklassen signifikant höhere Effizienz aufweisen als andere, wenn es um ihren Einfluß geht, die Verteilung einzelner Sprachen entlang einiger Dauerparameter zu beeinflussen. Diese Ergebnisse stützen die Sichtweise, daß rhythmische Sprachklassifikation wesentlich durch phonotaktische Faktoren bestimmt wird.

1 Introduction

Recent studies into the acoustic correlates of speech rhythm have offered new evidence towards a weak form of the Rhythm Class Hypothesis. Traditionally, the Rhythm Class Hypothesis has differentiated between so-called *stress-timed* (e.g. English, Russian, Arabic), *syllable-timed* (e.g. French, Telugu, Yoruba) and *mora-timed* languages (e.g. Japanese) (Pike 1945, Abercrombie 1967). The strong form of the Rhythm Class Hypothesis, founded in a notion of foot, syllable or mora isochrony, respectively, has been repeatedly experimentally disproved (e.g. Ladefoged 1967, Ohala et al. 1979, Roach 1982). However, a weak form of the Rhythm Class Hypothesis, based principally on syllable structure and vowel reduction, has been found to offer a more plausible interpretation of acoustically measurable features of speech (Dasher and Bolinger 1982, Dauer 1983, 1987).

In the last few years, a number of studies, beginning with Ramus et al. (1999), have segmented speech data into vocalic and intervocalic (i.e. consonantal) intervals, measuring their duration and extracting parameters from these durations based on statistical analysis of varying complexity (e.g. Grabe and Low 2002, Duarte et al. 2001, Galves et al. 2002b, Cassandro et al. 2002, Ramus 2002). The speech data these studies are based on vary considerably in nature, as do the methods used for C/V segmentation and statistical analysis. However, the findings of these studies all seem to support the Rhythm Class Hypothesis, as speech data from languages analyzed appear to cluster in accordance with traditional rhythmic classification.

2 Recent studies inspected

There are several remarks to be made concerning the results of such studies, and they can be grouped into three categories, all of them influencing in their own way how the results can be interpreted. First, the speech data must be controlled with respect to several factors; second, data preparation such as C/V segmentation decisively determines the results to be expected; and third, the actual extraction of durational parameters depends naturally on the algorithms used.

It should be kept in mind that a study of rhythmic typology seeks to expose factors which influence linguistic rhythm, while normalizing for factors which do not.

2.1 Considerations in data selection

Turning to the first point, what factors can influence the acoustic correlates of speech rhythm at the stage of data collection or corpus selection? Some may not have a significant effect, while others strongly alter the results to be expected of such a study. Several factors have been discussed or come to mind in this regard: number of speakers, speaker age, speaker gender, number of languages, speech material (read or spontaneous), speech continuity, speech rate, number of recordings per speaker, semantic differences, etc. (this list could easily be expanded).

While age and gender of speakers may not make a significant contribution to the study of speech rhythm in most cases, it would be advisable to include both genders and a variety of age groups. In fact, as with all empirical studies, a larger sample allows more powerful interpretations, so the overall number of speakers should be reasonably large.

Since spontaneous speech tends to exhibit phenomena such as hesitations, pauses, stammering, false starts and repetitions, there seems to be a general agreement that read speech is less difficult to analyze for purposes of rhythmic studies, as such discontinuities can be significantly reduced in experimental conditions with prepared texts.

Speech rate, however, has been identified as an uncontrollable source of potentially strong artifacts in measuring speech rhythm, especially since the parameters recent studies have extracted and used for analysis are durational in nature. As we shall see, this problem is not insurmountable and can be addressed at different stages of empirical study.

It is crucial to avoid conflating inter-speaker and inter-language variation. Therefore, data is required from more than one speaker per language analyzed. This point has been neglected in one otherwise promising study (Grabe and Low 2002), rendering its results somewhat limited in applicability.

The choice and number of languages to analyze is often narrowed by access to native speaker informants of the respective languages. Nevertheless, all of the recent studies have included English data, as well as at least a few other languages previously classified as stress-timed or syllable-timed. Where these languages cluster, they serve as reference for the interpretation of the rhythmic properties of languages hitherto unclassified with respect to the Rhythm Class Hypothesis.

2.2 Considerations in data preparation

All recent studies towards acoustic correlates of speech rhythm share a fundamental segmentation of sampled speech data into two interval categories. Almost all of them agree that these should be vocalic and consonantal intervals and thus perform a C/V segmentation during data preparation.

This view, however, tacitly presupposes both that C/V segmentation is a necessary prerequisite for the extraction of durational parameters and subsequent rhythmic analysis, and that it is easily performed. Unfortunately, these presumptions oversimplify an important point:

Assuming that the acoustic correlates of linguistic rhythm in speech are in fact based on the complexity of syllable structure and the presence or absence of vowel reduction, as proposed by Dauer, how is the boundary between what counts as a vocalic interval and what is marked as consonantal defined? Ramus et al. acknowledge the problematic nature of this question in passing and leave it to subsequent studies to elaborate on the issue, but so far, it has not been explicitly addressed. The reasons that the C/V distinction is not a trivial problem are twofold.

First of all, it must be decided whether the basis for the distinction between vocalic and consonantal intervals is to be a phonological one, based on the phonemic system of the language in question, or a phonetic division, based on acoustic cues. While the domain of linguistic rhythm seems to entail a preference for the former, this also renders the comparison of languages more difficult whose phonology handles the C/V distinction differently. What counts as a vowel in one language may not always be a vowel in another, and vice versa. Certain segment types, such as approximants, also receive unequal treatment, depending on where in the syllable they occur. On the other hand, the studies discussed in this paper attempt to capture acoustic correlates of rhythm, so that a purely phonetic distinction may be preferable. This in turn can pose problems on how to classify intervals exhibiting such phenomena as vowel devoicing, syllabic consonants and so forth, phenomena which surface regularly in languages such as the Tokyo accent of Japanese (Laver 1994) and Tashlhiyt Berber (Coleman 1999), respectively, as well as most other languages. Grabe and Low's study, for instance, opts to classify devoiced vowels as consonantal intervals, thereby eliminating the carrier syllables from the analysis.

This last point is worth elaborating on. Numerous languages allow syllable nuclei to consist not just of vowels, but alternatively of sonorants such as nasals and laterals, and even fricatives, as defined by the concept of the *sonority hierarchy*. Those languages that formally prohibit nonvocalic nuclei still exhibit them as phenomena of reduced speech. Are such segments to be labeled as vocalic in the context of C/V segmentation? A negative answer would result in consonantal intervals containing more than one syllable, a prospect at best unattractive for studies devoted to rhythmic analysis.

An alternative circumventing this question could be to abandon the notion of C/V segmentation altogether, instead falling back on a segmentation into syllable nuclei and inter-nuclear intervals. However, such a prosodic segmentation entails its own set of problems.

The second issue brought about by C/V segmentation is its rigidity. Depending on a phonological or phonetic approach, the criteria for C/V distinction may be language-dependent or universal and acoustically motivated. Once the C/V distinction has been established and encoded into the data by annotation, there is virtually no way of modifying this class distinction short of relabeling the data. Galves et al. (2002b) are the notable exception to this problem, as they automate the C/V distinction by computing a sonority measure. Data presented in this paper, however, suggest that varying the C/V distinction can have a revealing impact on language clustering in the parameter space proposed by recent studies on acoustic correlates of rhythm.

The handling of pauses is another matter entirely, but there seems to be a broad consensus to simply exclude pauses (silent or filled) from all further processing.

2.3 Considerations in parameter extraction

Presupposing a satisfactory C/V segmentation, the main issue in which recent studies have parted directions is that of parameter extraction. The acoustic correlates of rhythm are unanimously interpreted as statistical measures computed from the duration of vocalic and consonantal intervals in the speech signal, but the fundamental difference between the individual studies lies in the algorithms used for the computation of these durational parameters.

Ramus et al. propose the percentage of total vocalic interval duration in the overall duration of the utterance ($\%V$) and the standard deviation of consonantal interval duration (ΔC) as rhythmically classifying parameters. While the simplicity of these measures may seem naïve at first, their motivation is founded in two of the structural features of the weak Rhythm Class Hypothesis as proposed by Dauer: vowel reduction and syllable complexity. As such, $\%V$ and ΔC are remarkably efficient in discriminating languages by rhythmic class, as attested by the clustering of the language mean values in the $\%V/\Delta C$ plane. This is illustrated in Figure 1.

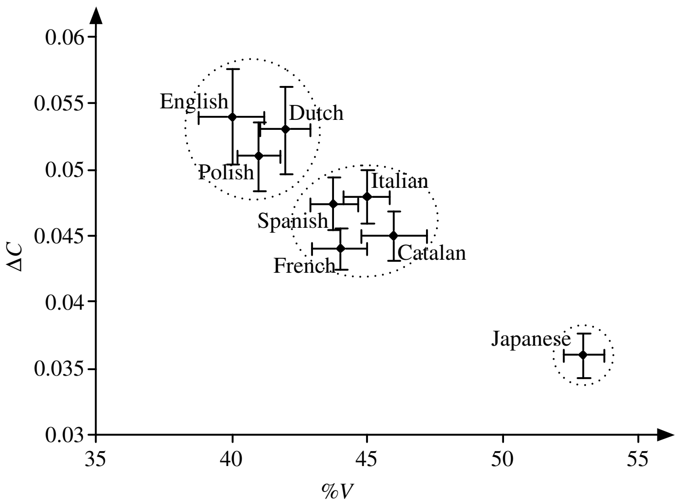

Figure 1: Language distribution in the $\%V/\Delta C$ plane. Error bars represent ± 1 standard error (from Ramus et al. 1999).

Other studies attempt to duplicate and refine these results by devising more elaborate statistical measures to extract from interval durations. Grabe and Low (2002) use a "pairwise variability index", both raw ($rPVI$) and normalized ($nPVI$), to function as ΔC and $\%V$, respectively. Galves et al. (2002b) define the sample mean of a sonority function to play the role of $\%V$, and δS for ΔC. While all of these studies succeed in achieving results comparable to Ramus et al. (1999), details will not be discussed here for reasons of brevity, and the reader is referred to the respective publications.

The problem of speech rate control was one of the principal motivations for the independent study by Grabe and Low. They attempt to normalize for local speech rate variation through statistical means at the level of parameter extraction. While such an approach is potentially fruitful, Galves et al. (2002a) present evidence to the contrary. Additionally, Dellwo and Wagner (2003) show that speech rate does not necessarily influence the distribution of $\%V$ and ΔC values strongly enough to obliterate Rhythm Class clustering.

To summarize this section, recent studies have taken speech data from a number of rhythmically distinct languages, segmented these data into consonantal and vocalic intervals and computed statistical measures from the duration of these intervals. Differences arise mainly in the first and third stages, more precisely in the sample of languages, the number of speakers per language, and especially the methods used to compute the statistical measures, which are interpreted as the acoustic correlates of rhythm.

3 Evidence for a refined analysis

The aim of the present study is to show that to a considerable extent, studies analyzing acoustic correlates of rhythm according to the methods outlined in the previous section have their outcome determined by the circumstances of

C/V segmentation. This step, intermediate between data selection and parameter extraction, is therefore in focus in the analysis described in this section. To this end, measurements were carried out on speech rate controlled data and analyzed with a variable C/V distinction. Finally, it is proposed that specific parameters other than $\%V$ and ΔC are most efficient at discriminating the languages analyzed by rhythmic class. The results support the interpretation that linguistic rhythm is substantially dependent on language specific phonotactic characteristics.

3.1 Data selection

To avoid some of the potential artifact sources identified in the previous section, speech data for the present study was taken from the BonnTempo Corpus, a collection of annotated recordings designed specifically with rhythmic analysis in mind (Dellwo et al. 2004). The corpus consists of read speech, with a translation of the same text in each of several languages, numerous speakers per language, and 5 recordings of controlled varying speech rate per speaker, roughly averaging 100 syllables per recording. The data is manually C/V annotated according to phonological criteria, as well as syllable segmented.

From this corpus, a number of speakers were selected, including 4 German speakers (2 male, 2 female), 5 English (2m, 3f), 5 French (2m, 3f) and 2 Italian (1m, 1f), amounting to 80 recordings.

3.2 Data preparation

To allow a flexible C/V distinction along the sonority scale, all recordings were manually annotated on the basis of the BonnTempo Corpus C/V segmentation. According to visual and auditory cues in spectrographic analysis, intervals were segmented into vowels (labeled as v), approximants (a), laterals (l), nasals (n), fricatives (f) and stops (s). Additionally, to preserve syllabic structure, syllabic laterals were labeled as L and syllabic nasals as N.

A preliminary analysis duplicating the methods of Ramus et al. (1999) displayed comparable clustering of languages into stress-timed (German and English) and syllable-timed languages (French and Italian). The variation of speech rate is reflected in a pronounced deviation along the ΔC axis, as shown in Dellwo and Wagner (2003).

3.3 Parameter extraction

By varying along the sonority scale the boundary between intervals classified as "vocalic" and "consonantal", it becomes possible to subsequently include approximants, laterals, and so on in $\%V$, removing them from ΔC. Since this step undermines the reality of $\%V$ being the percentage of vocalic intervals, $\%V$ and ΔC were replaced by the abstract measures $\%X$ and ΔY, respectively, with various configurations assigning interval types to the sets X and Y by sonority. For example, the configuration $\%vaLN/\Delta lnfs$ plots the percentage of the total duration of all vocalic, approximant, syllabic lateral and syllabic nasal intervals against the standard deviation of all clusters of non-syllabic lateral, non-syllabic nasal, fricative and stop intervals.

4 Results

As several configurations of sonority class assignments were analyzed in this manner, it became evident that when $\%X$ contains nasals and laterals, the distribution of languages in the $\%X/\Delta Y$ plane converges, as individual languages and rhythmic classes become merged.

To isolate the durational parameters principally responsible for the initial distribution of languages, 262 individual parameters were extracted for each of the 80 recordings by computing $\%X$ and ΔY for all possible combinations of the 8 interval types.[1] Subsequent discriminant analysis reinforced the observation identifying nasals and laterals as significant contributors to language distribution and provided the compelling result that not $\%V$ and ΔC (or its equivalent, $\Delta alLnNfs$) provide maximal separation of rhythmic classes, but rather $\%l$ and $\%n$, the percentages of lateral and nasal intervals of overall utterance duration (Figure 2). For the data analyzed, the canonical correlation of the discriminant function containing $\%l$ and $\%n$ (between stress-timed and syllable-timed language groups) was 0.902, compared to only 0.822 for $\%V$ and ΔC. For a detailed description of this analysis, the reader is referred to Steiner (2004).

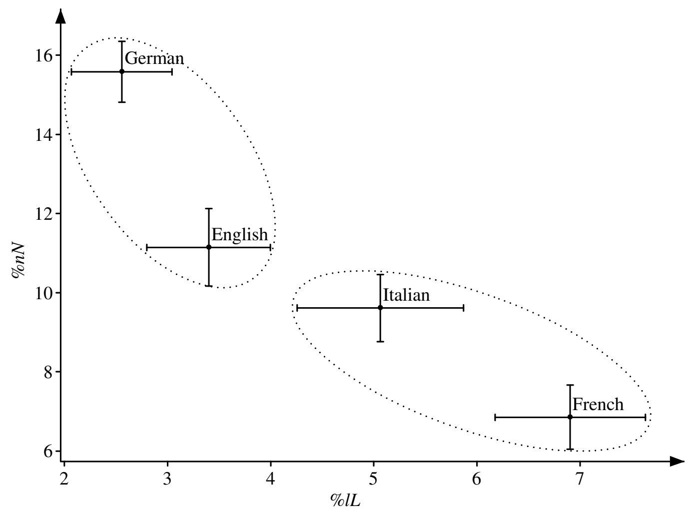

Figure 2: Language distribution in the $\%lL/\Delta nN$ plane. Error bars represent standard deviation.

5 Discussion

While it is possible that the results presented above apply only to the data analyzed, several generalizations are nevertheless applicable.

1. It should have become clear that in studies basing their findings on some sort of C/V distinction, the resulting distribution is asymmetric and that

[1] Only 8 of these parameters were of the $\%X$ kind, since all combinations would have yielded only dependent parameters (i.e. $\%v + \%a = \%va$).

the "functional load" of individual consonant classes is not homogeneous. This means that some consonant classes play a more important role than others in influencing any analysis grouping all consonants together.

2. Assuming the acoustic correlates of rhythm can be captured by methods such as those used in recent studies towards this end, and the position of individual languages in figures plotting durational measures based on a C/V distinction (e.g. the $\%V/\Delta C$ plane) assigns those languages to rhythmic classes, then the nature of the C/V distinction determines this assignment.

3. By an extension combining these two points, the statistical distribution of certain significant consonant classes in individual languages determines these languages' membership in rhythmic classes. This means that language assignment to rhythmic classes is determined by language internal phonotactics, as well as principles of syllabification. This provides new acoustic evidence for the validity of Dauer's proposals regarding the reformulation of the Rhythm Class Hypothesis.

4. If phonotactic characteristics determine an individual language's position relative to other languages with respect to certain parameters, this "phonotactic profile" could serve as a rhythmic feature complex in its own right, opening a new rhythmic typology based on phonotactics.

While these points are open to discussion in light of the evidence presented in this paper, it should also be taken into consideration that speech rhythm and rhythmic typology cannot be adequately captured by durational measures. Consider the position of Polish and Catalan in Ramus et al.'s study. While these languages have not been conclusively assigned to a traditional Rhythm Class, their distribution in the $\%V/\Delta C$ plane seems to suggest their classification as stress-timed and syllable-timed, respectively, which indicates that syllable complexity (a phonotactic factor), not vowel reduction, is the significant parameter. On the other hand, the ΔV distribution presents a different picture, which suggests that simple durational statistics of consonantal and vocalic intervals explain some, but not all rhythmic variability.

It may be in order to re-examine the acoustic correlates of rhythm altogether. If speech rhythm is regarded as the temporal arrangement of prominent and non-prominent syllables in a certain pattern, there are other factors at work through which syllable prominence is achieved, alongside duration, such as pitch, loudness, and segmental quality. After all, Dauer (1987) proposes 8 independent parameters of rhythmic distinction, of which only two have been taken into account in recent studies (and vowel reduction is in fact not even a purely durational dimension). An admittance of these other factors into an acoustic analysis of rhythm may grant new insight into the parameters governing rhythmic similarity.

Also, it should be kept in mind that the first formulations of the Rhythm Class Hypothesis were perceptual, if not downright introspective, in nature. Therefore, perceptual experiments present themselves as a convenient way of correlating the acoustic with perceptual correlates of rhythm. Some work has already been done in this direction (Ramus et al. 1999, 2003), but it would be interesting to investigate directly the influence of phonotactic factors on the perception of speech rhythm.

5.1 Applications

Reliable acoustic correlates of speech rhythm could prove an invaluable asset in foreign language acquisition. Most pronunciation instruction focuses on segmental articulation or word stress at best, ignoring speech rhythm characteristic for the target language. Identifying the essence of speech rhythm could help reduce the effects of native rhythm in a foreign language, which accounts of much of the native-speaker perceived “strangeness”, even when segmental pronunciation is flawless.

Infants learn to discriminate languages based on prosodic cues long before they acquire knowledge pertaining to the phonotactic properties of their mother tongue. The psychological reality of Rhythm Classes in this context is demonstrated by Ramus et al. (2003), who nevertheless employ their usual C/V distinction for their experimental data.

One relatively obvious application of acoustic correlates of speech rhythm lies in the domain of automatic spoken language identification. In situations where a multilingual speech user interface has to determine the language of discourse, identification of the user’s spoken language by prosody presents an option preferable, or at least complementary, to lexical matching or similar approaches which require speech recognition on the segmental level. In fact, Farinas and Pellegrino (2001) present results which show that rhythmic features alone can allow a Gaussian mixture model to identify languages with a reliability greater than 80%.

Finally, the notions of syllable isochrony and acoustic correlates of speech rhythm could provide new insight to durational modeling in speech synthesis. The author is currently engaged in research investigating such possibilities, among other things, to improve the perceived naturalness of synthetic speech rhythm by refining the durational models used.

References

Abercrombie, D. (1967). *Elements of General Phonetics*. University of Edinburgh Press, Edinburgh.

Cassandro, M., Collet, P., Duarte, D., Galves, A., and Garcia, J. (2002). A universal linear correlation among acoustic correlates of rhythm. http://www.ime.usp.br/~tycho/prosody/sonority/linear/linear.pdf.

Coleman, J. (1999). The nature of vocoids associated with syllabic consonants in Tashlhiyt Berber. In *Proceedings of the 14th International Congress of Phonetic Sciences*, volume 1, pages 735–738, San Francisco. http://www.phon.ox.ac.uk/~jcoleman/ICPhS.ps.

Dasher, R. and Bolinger, D. L. (1982). On pre-accentual lengthening. *Journal of the International Phonetic Association*, 12:58–71.

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

Dauer, R. M. (1987). Phonetic and phonological components of language rhythm. In *Proceedings of the 11th International Congress of Phonetic Sciences*, volume 5, pages 447–450, Tallinn.

Dellwo, V., Steiner, I., Aschenberner, B., Dankovičová, J., and Wagner, P. (2004). BonnTempo Corpus and BonnTempo Tools: A database for the study of speech rhythm and rate. In *Proceedings of the 8th International Conference on Spoken Language Processing*, Jeju.

Dellwo, V. and Wagner, P. (2003). Relations between language rhythm and speech rate. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona.

Duarte, D., Galves, A., Garcia, N. L., and Maronna, R. (2001). The statistical analysis of acoustic correlates of speech rhythm. In *Proceedings of the Workshop on Rhythmic Patterns, Parameter Setting and Language Change*, Bielefeld. Zentrum für interdisziplinäre Forschung.
http://www.ime.usp.br/~tycho/zif/zif_papers/duarte.pdf.

Farinas, J. and Pellegrino, F. (2001). Automatic rhythm modeling for language identification. In *Proceedings of the 7th European Conference on Speech Communication and Technology*, Aalborg.
http://www.ddl.ish-lyon.cnrs.fr/PDF/pellegrino/FPEurospeech2001.pdf.

Galves, A., Garcia, J., Duarte, D., and Galves, C. (2002a). Appendix: Are the lengths of consonantal intervals correlated? In *Proceedings of Speech Prosody 2002*, Aix-en-Provence.
http://www.ime.usp.br/~tycho/prosody/sonority/appendix.pdf.

Galves, A., Garcia, J., Duarte, D., and Galves, C. (2002b). Sonority as a basis for rhythmic class discrimination. In *Proceedings of Speech Prosody 2002*, Aix-en-Provence.
http://www.ime.usp.br/~tycho/prosody/sonority/rev4.pdf.

Grabe, E. and Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. In Gussenhoven, C. and Warner, N., editors, *Papers in Laboratory Phonology 7*. Mouton de Gruyter, Berlin.
http://www.phon.ox.ac.uk/~esther/ivyweb/Grabe_Low.doc.

Ladefoged, P. (1967). *Three Areas of Experimental Phonetics*. Oxford University Press, London.

Laver, J. (1994). *Principles of Phonetics*. Cambridge University Press, Cambridge.

Ohala, J. J., Riordan, C. J., and Kawasaki, H. (1979). Investigation of pulmonic activity in speech. In *Proceedings of the 9th International Congress of Phonetic Sciences*, volume 1, page 205, Copenhagen.

Pike, K. L. (1945). *The Intonation of American English*. University of Michigan Press, Ann Arbor, MI.

Ramus, F. (2002). Acoustic correlates of linguistic rhythm: Perspectives. In *Proceedings of Speech Prosody 2002*, pages 115–120, Aix-en-Provence.
http://www.ehess.fr/centres/lscp/persons/ramus/ramus_sp02.pdf.

Ramus, F., Dupoux, E., and Mehler, J. (2003). The psychological reality of rhythm classes: Perceptual studies. In *Proceedings of the 15th International Congress of Phonetic Sciences*, pages 337–342, Barcelona. http://www.ehess.fr/centres/lscp/persons/ramus/ICPhS03.pdf.

Ramus, F., Nespor, M., and Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73:265–292. http://www.ehess.fr/centres/lscp/persons/ramus/Cognition99.pdf.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In Crystal, D., editor, *Linguistic Controversies*, pages 73–79. Edward Arnold, London. http://www.personal.rdg.ac.uk/∼llsroach/phon2/frp.pdf.

Steiner, I. (2004). Zur Rhythmusanalyse mittels akustischer Parameter. Master's thesis, Universität Bonn. http://www.coli.uni-saarland.de/∼steiner/files/pdf/MA-Arbeit.pdf.