Looking for Rhythm in Speech

FRED CUMMINS
School of Computer Science and Informatics
University College Dublin

ABSTRACT: A brief review is provided of the study of rhythm in speech. Much of that activity has focused on looking for empirical measures that would support the categorization of languages into discrete rhythm 'types'. That activity has had little success, and has used the term 'rhythm' in increasingly unmusical and unintuitive ways. Recent approaches to conversation that regard speech as a whole-body activity are found to provide considerations of rhythm that are closer to the central, musical, sense of the term.

Submitted 2010 January 6; accepted 2012 July 13.

KEYWORDS: *rhythm, isochrony, speech rhythm*

MUCH like God, Tidiness, and Having a Good Time™, the concept of rhythm means many things to many people. As long as we are not forced to pin the notion down with an unforgiving definition, we can readily agree that the eloquent rhetoric of Martin Luther King is more rhythmic than the mumblings of an inarticulate teenager, that the rhythm of Japanese is qualitatively different from the rhythm of English, and that the verse of Gerald Manly Hopkins makes use of rhythm in ways not to be found in the insipid prose of a technical manual. Each of these observations might be backed up with intuitively convincing examples. Alas, examples and intuition do not suffice to ground either a rigorous empirical description, or a useful theory.

We will restrict our attention here to the empirical study of rhythm in speech. In so doing, we neglect many topics that might be of interest in teasing out parallels between the treatment of rhythm in language and in music. In particular, the musical notion of meter as an organizing scaffold upon which a specific sequence of events is hung in strong and weak alternation will be all but passed over. This finds treatment in the specialized field of metrical phonology (Liberman & Prince, 1977); there, formal devices such as grids and trees are used in a manner similar to the formal notation of note durations found in scores. But just as the score stands at some remove from the richness of a specific real-time performance, so these theories remain aloof from the speech signal and the articulatory movements that bring it forth. Both the archaic art of rhetoric, and the technical domain of poetics (Abercrombie, 1965) also use notions of rhythm that bear relation to music, perhaps in terms best mapped to composition, rather than performance. But these, too, we must pass over in the interest of conciseness.

It is with the messy business of performance that we start. As far back as 1939, Classé used a kymograph to study the succession of syllable onsets evident in the speech waveform, see Figure 1 (Classé, 1939). The speech employed was read English, and Classé wanted to inquire whether the impression of rhythmic regularity in the sequence of syllables encountered in prose could find empirical validation in isochronous interval measurements. Subjects read formal texts ranging from highly poetic (The Song of Songs) to informal prose (taken from Daniel Jones' transcriptions). They also tapped at points they considered rhythmically salient. This latter intervention is interesting, as it has the side effect of making the intervals between taps more regular than the corresponding speech intervals spoken freely, and thus will tend to favour the production of evenly spaced rhythmic beats.

FIG. 6. *harm*

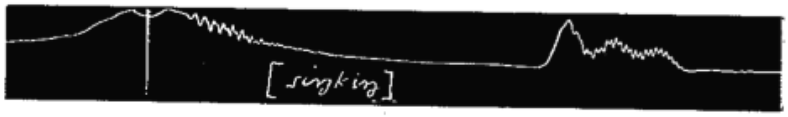

FIG. 7. *sinking*

Fig. 1: Kymograph tracings from Classé (1939) showing both pressure variation and the moment of manual tapping.

Classé's findings were not terrifically surprising. Even spacing between successive stressed syllables emerged as a tendency in the recordings—a tendency greatly encouraged when the lexical material was written with an ear to rhythm, when successive intervals contained phonetically matched segments and syllables, and when they had relatively similar grammatical construction. Any such tendency was disrupted by inter-sentence breaks. That, one might think, was that.

THE GREAT ISOCHRONY SAFARI

Classé has the distinction of providing the first empirical test of an intuition that had found expression as far back as 1775, in Joshua Steele's *An Essay toward Establishing the Melody and Measure of Speech.* The intuition is that stresses in English have a tendency to form a regular series of beats or pulses. This informal intuition seems to have a degree of persuasiveness for many people, as it has popped up regularly in the interim. Daniel Jones (1960/1918) had expressed this succinctly when he said:

> ...there is a general tendency to make the 'stress-points' of stressed syllables follow each other at equal intervals of time, but ... this general tendency is constantly interfered with by the variations in the number and nature of the sounds between successive stress-points... (Jones, 1918)

This presumed characteristic of English is often contrasted with the perceived rhythm of other languages, which, being non-English, are found, unsurprisingly, to be different. But here common sense leaves the discussion, and the great Isochrony Hunt begins. The source of this unfortunate quest is a claim made by David Abercrombie, a highly respected phonetician, who unwisely asserted:

> As far as is known, every language in the world is spoken with one kind of rhythm or with the other. In the one kind, known as a *syllable-timed* rhythm, the periodic recurrence of movement is supplied by the syllable-producing process: the chest-pulses, and hence the syllables, recur at equal intervals of time—they are *isochronous*. French, Telugu, Yoruba illustrate this mode of co-ordinating the two pulse systems: they are syllable-timed languages. In the other kind, known as a *stress-timed* rhythm, the periodic recurrence of movement is supplied by the stress-producing process: the stress-pulses, and hence the stressed syllables, are isochronous. English, Russian, Arabic illustrate this other mode: they are stress-timed languages. (Abercrombie, 1967: 97)

Would that we could pass over this wild and wholly inaccurate assertion in silence, but we can not, for it spawned a veritable industry, peopled on the one hand by aspiring phoneticians who wanted to be the first to capture the isochronous beast in the wild, and, on the other, by well-meaning applied types who wanted to make use of this chimerical classification in ordering, teaching, and sorting the languages of the world.

The distinction between (perceptually) even timing of syllables and (perceptually) even timing of stresses can be traced to Lloyd-James (1940) (cited in Abercrombie, 1967, p.171), who coined the evocative terms "machine-gun rhythm" and "morse-code rhythm" for them respectively. Lloyd James was not talking about differences between languages though. He was referring to specific transient

patterns that might arise in the speech of an individual. Those familiar with Martin Luther King's "I have a dream" speech can find reasonably clear examples of each of these in the two phrases "[will be able to] SPEED UP THAT DAY" (syllable timing) and "BLACK men and WHITE men, JEWS and GENtiles, PROTestants and CATHolics" (stress timing). Kenneth Pike (1945) recast these as the now-familiar terms "syllable-timed" and "stress-timed", but it was Abercrombie who took these relatively inoffensive and informal observations about perceived regularity and turned them into dogma and into a strong claim about language typology.

The list of studies that have sought to underpin this distinction by measuring intervals in one language or the other is dispiritingly long. We might mention Nakatani (1981), Crystal and House (1990), Shen and Peterson (1962), Bolinger (1965), O'Connor (1968), Lehiste (1977), and others. But pride of place surely goes to Rebecca Dauer (1983), who systematically decimated any hopes anyone might have had that isochrony of inter-stress intervals was a characteristic of English, or that isochrony of syllables was characteristic of French, or that two classes of languages might be identified that were distinguished by any rhythmic characteristic as simplistic as stress-timing versus syllable-timing.

Dauer provided a list of potential properties that might collectively underpin people's intuition that languages differ in their characteristic rhythmic patterning. These included variability in syllable structure and complexity, differences in the degree to which vowels become shorter and more central in unstressed syllables, and the presence or absence of stress as a contrastive phenomenon. This opened up a whole new way of thinking about differences between languages, suggesting that they might vary along a number of dimensions. And so another hunt was on, this time not seeking the elusive isochronous monster, but rather looking for some (any) other empirical quantity or metric that might serve a similar role. The role in question had developed from an initial set of questions about speech in specific utterances, to a very different focus on validating a presumed language typology.

METRICS, METRICS, METRICS

In 1999, Ramus, Nespor & Mehler presented some novel phonetic measures that they thought might justify a presumed classification of languages into stress-timed and syllable-timed families. The authors were heavily committed to the two-way classification, and they had shortly before demonstrated that French newborn infants could discriminate between low-pass filtered speech in Japanese and English, but not between Dutch and English. They could also discriminate between the sets {English, Dutch} and {Spanish, Italian}, but not between the sets {English, Spanish} and {Dutch, Italian}. Of course, these discrimination results in no way confirm that languages fall into two groups, but they are certainly compatible with such a hypothesis, if it were to be established on independent grounds[1]. They arrived at two (correlated) variables, defined over an utterance: the proportion of vocalic intervals (%V) and the standard deviation of the duration of consonantal intervals (ΔC).

Results from nine languages are shown in Figure 2. These stem from 4 speakers per language, reading 5 short declarative sentences each. At first glance, there appear to be two distinct clusters, and one outlier. The clusters group languages claimed to be stress-timed (English, Dutch, Polish) together, while the so-called syllable-timed languages (French, Spanish, Italian, Catalan) form a second group. Japanese had long since been claimed to represent a third class: the mora-timed languages (Hoequist, 1983; Port et al., 1987).

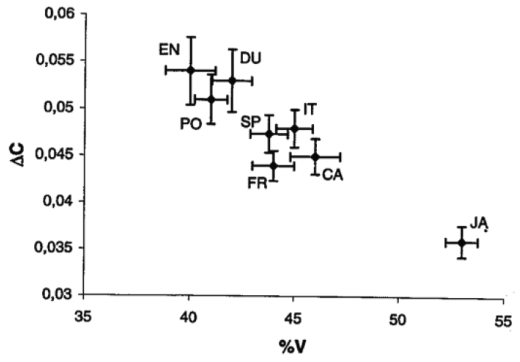

Fig. 2: Distribution of 9 languages over the (%V, ΔC) plane. Error bars show ±1 standard error. From Ramus et al. (1999).

With similar motivation, Grabe and Low (2002) employed a measure of local timing variability, the *Pairwise Variability Index*, or PVI, that quantifies the degree to which successive units (often, but not necessarily, syllables) differ in duration. Two variants were employed: the raw index (rPVI):

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

and a normalized form, that uses the average interval length within each pair as a normalization factor:

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

where *m* is the number of items contained in an utterance, and *d* is the duration of the *k*th item. The nPVI measure was applied to vowel durations, and the rPVI to the intervals between vowel onsets.

Figure 3 shows comprehensive results for 18 languages, with data from a single speaker for each language reading set texts in a recording booth. One can read what one likes into the resulting distribution. The authors claimed that the data "support a weak categorical distinction between stress-timing and syllable-timing ... [but] ... there is considerable overlap between the stress-timed and the syllable-timed group and hitherto unclassified languages" (Grabe and Low, 2002, p. 538). Nolan, from whom the PVI originally stems, has recently applied the measure at both syllable and foot level for four languages (Estonian, English, Mexican Spanish and Castilian Spanish) (Nolan and Asu, 2009). Five speakers of each read a short text to provide the data. There were serious methodological problems in defining units, especially the foot, in comparable fashion across language. Despite these, the author argued that syllable-timing and stress-timing were orthogonal dimension, such that a given language might exhibit characteristics of either, both, or neither.

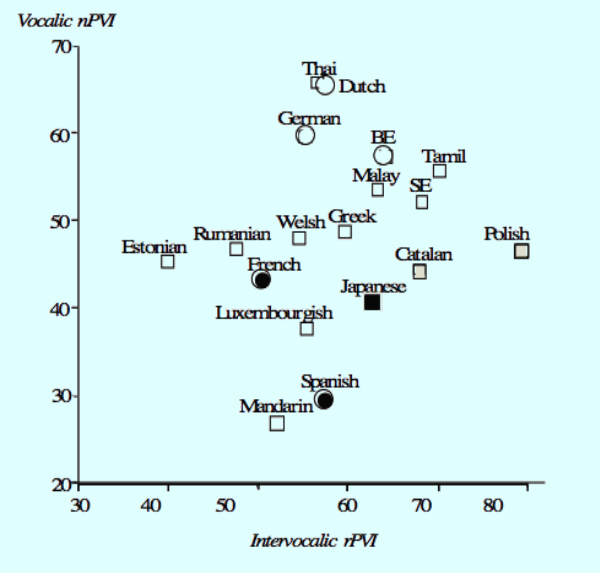

Fig. 3: Distribution of 18 languages over the (rPVI, nPVI) plane. Open circles: prototypically stress-timed languages, filled circles: prototypically syllable timed languages, open squares: mixed or unclassified. BE = British English, SE = Singapore English. From Grabe and Low (2002).

Several related metrics have subsequently been proposed, any of which might serve to locate languages in a low-dimensional 'rhythm-space', in the vain hope that the old and, it is surely now established, misguided classification into discrete rhythm types, might yet find some validation. Galves et al. (2002), proposed a sonority-based measure that obviated the need for manual annotation of the speech material. Gibbon and Gut (2001) contributed another, and Wagner and Dellwo (2004) provided yet another variant on the PVI in the service of the same highly questionable goals.

All the above work might be summarized as phoneticians employing their talents in the search for linguistic taxonomic goals that lie far removed from the reality of speech. In each case, a tiny amount of data is used to act as representative material for a given language. There is a yawning gulf between the immediate perception of temporal patterning that underlies the use of the term 'rhythm' in speech research and the abstract classification of languages that is sought through the use of metrics. The goal of constructing a taxonomy of languages based on a single concept of 'rhythm' has long ago

been revealed to be Quixotic. There is certainly nothing here that might hearken back to potential commonalities between musical rhythm and rhythm in speech.

EMBODIED RHYTHM

Let us return to rhythm in something approaching its core sense. When the public speaker pounds the lectern with his fist at every emphasis, we are made aware of rhythmic organization. When speech is turned into chant, rap, verse, or song, we find a marriage of rhythm in the musical and phonetic senses. When children in the playground turn a single exclamation or taunt into a repeated mantra, its rhythmic potential is unleashed. Here, we see there is still much room for study of the commonalities between musical rhythm and the rhythm of the voice. The voice is, after all, the most human of musical instruments.

Producing speech using the speech articulators may seem to be a highly specialized task. Masses and forces involved are small, movement targets are very precise and are intricately sequenced in time, and interaction between the moving parts and the rest of the world is minimal (for non-pipe smokers, at least). However, when speech movements are repeated cyclically, it has been possible to show that rhythmic constraints are apparent that are of a kind with rhythmic constraints on cyclic movement of the limbs, e.g. in juggling, walking, or dancing. In the speech cycling experimental paradigm, a short phrase is repeated in time with an auditory metronome. A canonical example is the targeted speech cycling reported in Cummins and Port (1998), where a short phrase, such as "big for a duck" is repeated along with a series of alternating high and low tones. The high tone sequence cues phrase onset, while the low tone provides a temporal target for the onset of the final stressed syllable ("duck"). It is quickly apparent that cyclic repetition like this is highly constrained, and the constraint lies in the temporal relationship between the sequence of syllables, and their organization into larger units, here the foot and the phrase. When the phase (relative time) of the low tone is varied from trial to trial, it becomes clear that some positions of the stressed syllable onset within the repeating phrase cycle are relatively natural, and can be maintained in a stable fashion, while others can not be so produced.

Figure 4 (left) shows data from 4 subjects, each of whom attempted to match the relative timing of a sequence of low and high tones, where the phase of the low tone within the High–High cycle was drawn, on each trial, from a uniform distribution in the range [0.3–0.7]. It is clear from the histograms of phases produced that subjects have a very strong propensity to produce one of three discrete patterns, corresponding, roughly, to the musical patterns shown on the right side of the figure. This constraint on the temporal organization within speech resembles the modes of cyclic organization found in limb movement. For example, all legged animals have at their disposal a finite, typically small, number of discrete gaits, each of which is characterized by stable phase relations among the legs.

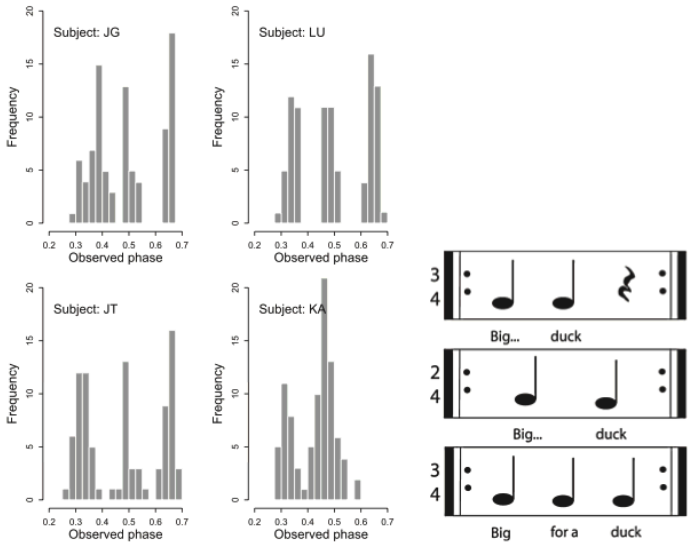

Fig. 4: Left: Histograms of the relative time (phase) of the stressed syllable onset within the overall phrase repetition cycle for four subjects. Target phases were uniformly distributed between 0.3 and 0.7. Right: Musical interpretation of the three dominant patterns observed.

In producing music, a musician's whole body comes into play. Breathing, posture, foot movements, torso sway, all conspire in the enactment of a musical performance. Speech movement too goes beyond the vocal tract, and can involve the whole body. Gestures have been extensively studied as they co-occur with speech. We gesture in almost all speaking situations, even when the conversational partner is not spatially present (Goldin-Meadow, 1999). While most gestures are devoid of overt linguistic content, they are still integrated into the temporal unfolding of speech in ways that are still being uncovered (Cassell et al., 1999; Leonard & Cummins, 2011; Wachsmuth, 1999). Speakers and listeners coordinate their eye movements (Richardson et al., 2007) and even their posture (Shockley et al., 2003). Even the brain activity of speakers and listeners during spoken communication has been shown to exhibit mutual entrainment or coupling (Stephens et al., 2010).

It would be a grave mistake, therefore, to consider speech as a highly modular, encapsulated form of activity, in which all movement subserves the goal of producing a sequence of sounds. Rather, it is a whole-body activity, in which the intentions, emotions, and engagement of the individual in the joint domain of conversational interaction are signaled using a multiplicity of cues tightly interwoven in time. In this respect, it bears strong resemblance to an improvisatory musical performance. I have elsewhere argued that one way of understanding the role of rhythm in speech, and in music, is to see rhythm as a social affordance that allows multiple individuals to entrain their movements (Cummins, 2009a; Cummins, 2009b). This is a plausible way of describing the role of rhythm in sustaining dance activity, or even ensemble playing. It may also allow us to see commonalities between these group activities and choral or group speaking. Indeed, in a laboratory situation, it is found that speakers can synchronize their speech very tightly indeed (Cummins, 2009b), even though, as noted at length above, there is no regular isochronous pulse to scaffold their performance. Rather, it seems, common knowledge of what it is to speak, together with a shared set of goals in the form of an agreed text, suffice to allow a remarkable and sustained coupling in time. Perhaps there are lessons to be learned here that can feed back from the study of rhythm in speech to the domain of music. If asked what the most plausible continuation is for a sequence of notes C-D-E-F-G-A-B, I suspect a psychologist would say C′, while a musician would say "anything except C′". Likewise, the essence of rhythm may lie, not in a soulless, but predictable, isochrony, but in the intuitive knowledge of what it is to move together.

END NOTES

[1] Among the Romance languages often claimed to be syllable-timed, French plays an odd role. While it is sometimes unthinkingly lumped together with Spanish and Italian (which themselves are interestingly different), its prosody is markedly different, and at a first pass, is distinguished primarily by intonational characteristics rather than timing or stress.

REFERENCES

Abercrombie, D. (1965). A phonetician's view of verse structure. In *Studies in Phonetics and Linguistics*. London: Oxford University Press, pp. 16–25.

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Bolinger, D. (1965). Pitch accent and sentence rhythm. In Abe, I. and Kanekiyo, T. (Eds.), *Forms of English: Accent, Morpheme, Order*. Cambridge, MA: Harvard University Press, pp. 139–180.

Cassell, J., McNeill, D., & McCullough, K. (1999). Speech-gesture mismatches: evidence for one underlying representation of linguistic and nonlinguistic information. *Pragmatics and Cognition*, Vol. 7, No. 1, p.1.

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

Crystal, T. H. & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *Journal of the Acoustical Society of America*, Vol. 88, No. 11, pp. 101–112.

Cummins, F. (2009a). Rhythm as an affordance for the entrainment of movement. *Phonetica*, Vol. 66, Nos. 1–2, pp. 15–28.

Cummins, F. (2009b). Rhythm as entrainment: The case of synchronous speech. *Journal of Phonetics*, Vol. 37, No. 11, pp.16–28.

Cummins, F. & Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, Vol. 26, No. 2, pp. 145–171.

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

Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). Sonority as a basis for rhythmic class discrimination. In *Proceedings of Prosody 2002*.

Gibbon, D. & Gut, U. (2001). Measuring speech rhythm. In *Seventh European Conference on Speech Communication and Technology*. ISCA.

Goldin-Meadow, S. (1999). The role of gesture in communication and thinking. *Trends in Cognitive Sciences*, Vol. 3, No. 11, pp. 419–429.

Grabe, E. & Low, E. (2002). Durational Variability in Speech and the Rhythm Class Hypothesis. *Laboratory Phonology,* Vol. 7, pp. 515–546.

Hoequist, Jr., C. (1983). Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica*, Vol. 40, pp. 203–237.

Jones, D. (1967). *An Outline of English Phonetics*. Cambridge, UK: Heffner and Sons Ltd, 9th edition. 1st edition published 1918.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, Vol. 5, pp. 253–263.

Leonard, T. & Cummins, F. (2011). The temporal relation between beat gestures and speech. *Language and Cognitive Processes*, Vol. 26, No. 10, pp. 1457-1471.

Liberman, M. & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, Vol. 8, pp. 249–336.

Nakatani, L. H., O'Connor, K. D., & Aston, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica*, Vol. 38, pp. 84–106.

Nolan, F. & Asu, E. (2009). The pairwise variability index and coexisting rhythms in language. *Phonetica*, Vol. 66, Nos. 1-2, pp. 64–77.

O'Connor, J. D. (1968). The duration of the foot in relation to the number of component sound-segments. Technical Report Progress Report 3, Phonetics Laboratory, University College, London.

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

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *Journal of the Acoustical Society of America*, Vol. 81, No. 5, pp. 1574–1585.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, Vol. 73, No. 3, pp. 265–292.

Richardson, D., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination. *Psychological Science,* Vol. 18, No. 5, pp. 407.

Shen, Y. and Peterson, G. G. (1962). Isochronism in English. In *Studies in Linguistics, Occasional Papers 9*, pp. 1–36. University of Buffalo.

Shockley, K., Santana, M., & Fowler, C. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology-Human Perception and Performance*, Vol. 29, No. 2, pp. 326–332.

Stephens, G., Silbert, L., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences,* Vol. 107, No. 32, pp. 14425-14430.

Wachsmuth, I. (1999). Communicative rhythm in gesture and speech. In *Proceedings of the International Gesture Workshop on Gesture-Based Communication in Human-Computer Interaction*. London, UK: Springer-Verlag, pp. 277–289.

Wagner, P. & Dellwo, V. (2004). Introducing YARD (Yet Another Rhythm Determination) and re-introducing isochrony to rhythm research. In *Speech Prosody 2004, International Conference*. ISCA.