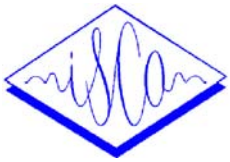

Speech Prosody 2002
Aix-en-Provence, France
April 11-13, 2002

Speech Rhythm and Rhythmic Taxonomy

Fred Cummins

Department of Computer Science
University College Dublin
also
Media Lab Europe
Dublin 8
fred.cummins@ucd.ie

Abstract

Of all prosodic variables used to classify languages, rhythm has proved most problematic. Recent attempts to classify languages based on the relative proportion of vowels or obstruents have had some success, but these seem only indirectly related to perceived rhythm. Coupling between nested prosodic units is identified as an additional source of rhythmic patterning in speech, and this coupling is claimed to be gradient and highly variable, dependent on speaker characteristics and text properties. Experimental results which illustrate several degrees of coupling between different prosodic levels are presented, both from previous work within the Speech Cycling paradigm, and from new data. A satisfactory account of speech rhythm will have to take both language-specific phonological properties and utterance-specific coupling among nested production units into account.

1. On Classification and Taxonomy

Taxonomy involves the determination of discrete classes. In its classical manifestation, living forms are divided into discrete groups (species, genera, families, etc), and criteria are established which help to decide which taxon a given exemplar should be assigned to. A basic assumption is that discrete classes exist underlyingly, and that a strict classification is, in principle, possible. In this regard it differs from the more general practice of biosystematics, which considers any and all relationships which exist among organisms.

The data on which a classification is made may, of course, be insufficient to allow unambiguous classification of a given exemplar. By way of a simple example, we might consider a simple racially homogeneous population of men and women, in which mens' heights are normally distributed around a given mean (say 2m) with a certain standard deviation (say 0.5m), while womens' heights are similarly distributed around a different mean (say 1.8m). Based only on a measure of height from an individual, we can only provide a probabilistic classification. Nonetheless, there is assumed to be a underlying discrete difference between the classes.

There are many forms of linguistic taxonomy, most of which have the property that we have strong reason to suspect a discrete difference in some formal feature between the languages. For example, some languages have a basic word order in which the subject is ordered before the verb, which in turn precedes the object, while others order these three elements differently. Taxonomic licence is granted because of the discrete nature of the elements involved.

2. Prosody as a Basis for Taxonomy

Prosody has often been used as a basis for classifying languages. The grab bag of phenomena which can be linked under the label "prosody" leaves considerable scope for creative classification. Attempts have been made to classify languages based on stress, accent, intonation, lexical and morphological tone, and, of course, rhythm. However, it has not always been possible to unambiguously identify discrete elements corresponding to each of these dimensions with the same robustness as in the segmental, morphological or lexical domains.

Distinctions based on syllable structure have been fairly uncontroversial, as a segmental inventory is relatively easy to obtain for a given language, and the principles of syllable structure have shown considerable generality. Linguistic theories such as Autosegmental Phonology or Optimality Theory have provided well-founded and empirically supported theories of underlying discrete structures which permit classifications within and across languages.

Distinctions based on fundamental frequency have had mixed success. On the one hand, one can identify languages which make use of lexical tone (e.g. Mandarin) and others which do not (e.g. English). Intermediate cases do exist (e.g. some dialects of Korean), but these are usually considered to represent transitional states of the language from one class to the other. The morphological use of tone familiar from the Niger-Congo languages of Africa represents another well-defined class.

On the other hand, phenomena related to phrasal accents and phrasal intonation have proved less obviously amenable to a conventional linguistic treatment. To be sure, there are several theories of phrasal intonation which relate observed pitch contours to a discrete set of underlying linguistic elements [16], however agreement among theories as to the nature and count of such elements has been hard to arrive at. The situation is further complicated by the many non-linguistic roles of intonation, such as in adding emphasis or expressive variation. Several studies have demonstrated gradient rather than categorical phenomena here [11, 10].

But nowhere has the effort at establishing and defending a prosodic taxonomy had a harder time than in the domain of 'rhythm'. Without doubt, much of this lack of progress can be traced to differing interpretations of the term 'rhythm'. It will be a contention of this paper that at least two independent dimensions have been called to service in characterizing rhythm. One of these is related to syllable structure and segmental inventories, and may therefore offer the basis for a taxon-

omy. The other relates to a gradient phenomenon, not yet well understood, which mediates the role of syllables in determining macroscopic timing patterns. Its gradient nature precludes it from supporting a classification among languages. Furthermore, it will be claimed, pre-theoretical perceptions of rhythm (whether characteristic of a speaker or a language) are derived from an interplay between the discrete and the gradient phenomena.

3. Where is Rhythm in Speech?

3.1. Rhythm across languages

Our formal approaches to characterizing rhythm in speech are grounded in a pre-theoretical perception of a patterning in time which speech and music have, to some degree, in common. We become aware of something like rhythmic properties in speech when we contrast speech in different languages, and this is presumably the reason why rhythm has so-often been called upon to support language classification. The ability to distinguish among languages based on a signal which preserves low frequency information has been documented in infants [13], while Ramus demonstrated a similar ability in adults using resynthesized speech in which segments were stripped of their identity, but not their broad phonetic class [17]. Many attempts have been made to identify a basis for this apparent perception of a rhythmic difference among languages. Simplistic notions based on isochronous units have been uniformly rejected [5].

Two current influential models [18, 9] take up a suggestion by Dauer [5] that languages may lie along a continuum (or in a continuous space), certain points of which have previously been identified with rhythmic classes (syllable-, stress- and mora-timed languages). They each develop continuous measures which can support clustering of languages in accordance with older taxonomic divisions. Since the introduction of the notion of gradient rhythmic qualities, it is no longer entirely clear that a taxonomy is being sought, as opposed to a more general systematic description of variation among languages.

Ramus et al. [18] arrive 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). Both of these measures will be directly influenced by the segmental inventory and the phonotactic regularities of a specific language. That is, any classification based on these variables can be related to an underlying discrete system, and so true classification is, in principle, possible.

Grabe and Low [9] relate rhythmic diversity to serial variability in (a) the inter-vowel-onset interval and (b) the interval between one vowel offset and the following onset. As with the previous measures, these two variables are not entirely independent, and their distributions will be dictated largely by the segmental inventory and phonotactics of a given language. Similar results have recently been suggested based on a sonority measure which captures the degree of obstruency in the signal [8]. Collectively these variables may be compared to alternative measures on our hypothetical population from Section 1: had we measured weight, or hair length, instead of height, we would likewise have found a bi-modal distribution, with the same underlying cause.

3.2. Rhythm within speaker

There is another, distinct, sense in which speech is rhythmical, and this is related to fluency. As we speak, the fluency with which speech is generated varies continually. We are all familiar with both the ease with which fluent speech flows, and the debilitating effect of its opposite, the dysfluent event. This type of rhythm is considerably harder to quantify, as it can vary substantially within a single utterance, and is apparently subject to the vagaries of expression and rhetorical force as much as to language-specific constraints[1].

Let the sentence presented by Abercrombie [1] as 'unambiguously' illustrating the stress-timed nature of English serve as an example: "Which is the Train for Crewe please". Abercrombie's suggestion was that the reader tap along with the stresses while saying the sentence, and indeed, it is not difficult to speak this sentence with 4 roughly isochronous beats on the stressed syllables. However, any naturalistic rendition without the associated tapping will depart substantially from this regular pattern. Furthermore, a syllable-based timing can likewise be imposed on this sentence (think "angry, seething, passenger faced with unhelpful guides"). Depending on the communicative situation, the rate of speech, the degree of expression, etc, rather different timing patterns can overlay one and the same utterance, for a single speaker. Some of these are regular enough that we would want our definition of speech rhythm to extend to them and their like. However, these patterns will clearly not be of much help in establishing a cross-language taxonomy.

This variability raises the question of whether the kind of index proposed by Ramus, Grabe and others can meaningfully be said to capture anything about *rhythm* in speech. The discrete basis for the suggested taxonomy can be argued to be grounded in segmental inventories and syllabic phonotactics, and can therefore be accounted for without reference to anything resembling the pre-theoretical notion of rhythm described at the start of this section. More succinctly, where is the bom-di-bom-bom in %V?

The argument to be developed here is that there are indeed two distinct phenomena here, which interact to provide a perception of rhythm in speech. On the one hand, there are linguistic units which vary discretely across languages. Thus English has its heavy and light syllables, stresses, feet etc, while Japanese has its Morae, perhaps a bi-moraic foot, and so on. These are symbolic, linguistic entities familiar from phonology, and language taxa can be constructed on foot[2] thereof. To some extent these alone dictate the alternation of light and heavy elements in spoken language, and so they contribute to the rhythmic signature of a language.

These units also serve as participants in hierarchical timing relationships, in which smaller prosodic units are nested within larger units, and the degree of coupling between levels varies in gradient fashion, as dictated by fluency, conversational intent, urgency, etc. As coupling varies continually, so too does the perceived rhythmicity of speech, and, perhaps, perceived fluency, though this direct association has yet to be tested.

The gradient coupling between prosodic levels (syllables within feet, feet within phrase, etc) has been identified and modelled before [15]. It has also been observed experimentally in the Speech Cycling paradigm [4, 19], in which subjects repeat a short phrase in time with an external metronome. Results from Speech Cycling experiments with English and Japanese speakers will now briefly be reviewed to see if they can illuminate the relationship between these two interacting sources of "rhythm".

[1] Examples of particularly fluent speech exhibiting syllable-timed and stress-timed characteristics within an utterance by a single speaker are given at http://cspeech.ucd.ie/~fred/speechrhythm/speechrhythm.html.

[2] sorry.

4. Speech Cycling Results

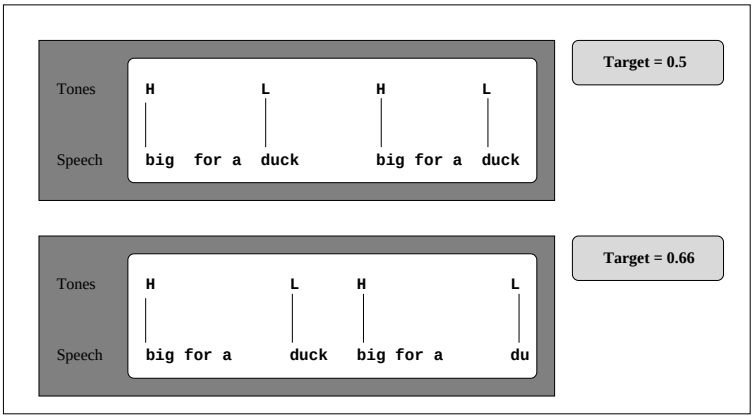

Figure 1: *Targeted Speech Cycling task, as used with English speaking subjects (reported in [4]). 'Target' refers to the phase of the L tone within the H-H cycle.*

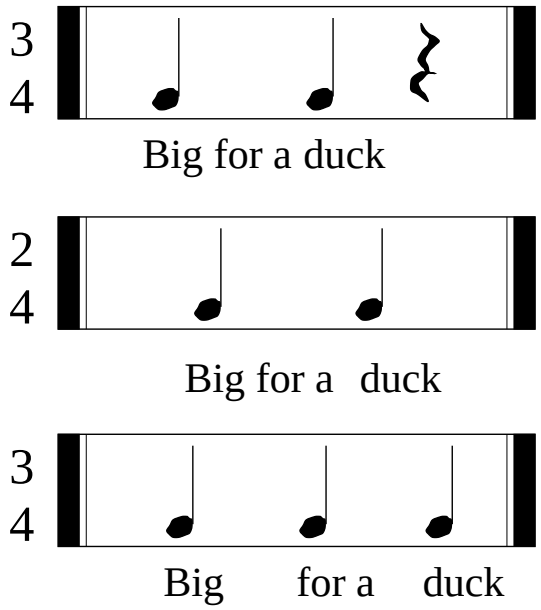

Figure 2: *Rhythmic patterns produced by English speakers in [4].*

In [4], English speaking subjects repeated short phrases such as "big for a duck" in time with a two-tone metronome. The phrases were always of the form "X for a Y", and their stated goal was to align the onset of "X" with the first, higher, tone, and the onset of "Y" with the second, lower, tone. The relative timing of the two tones was varied systematically to see what ways the stressed foot could be accommodated within the repeating Phrase Repetition Cycle (PRC). The task is illustrated in Figure 1. The results were unambiguous and readily interpretable. Under these conditions, subjects could produce only three patterns reliably. These patterns are illustrated in Figure 2. Each of these patterns can be understood as the strict nesting of one unit (the stress foot) within a larger unit (the PRC). For the third pattern, this requires introducing a nonce stress on the content word *for*, and indeed we found that some subjects did not produce this pattern, as they did not discover this strategy.

In related work, Tajima had both English and Japanese speakers repeat short phrases in time with a repeating metronome [19]. The metronome here consisted only of a single repeating tone, and subjects were instructed to align the onset of the phrase with this tone. The texts used contained carefully controlled segmental material which tested the relative stability of syllable and mora durations at a range of prosodic positions. The similarities and differences found across languages are illuminating. Firstly, both languages showed preferences for prominent syllables (stressed in English, pitch accented in Japanese) to fall at easily predictable points within the PRC (one half, two thirds, etc.). Evidence for temporal stability of a foot-like unit was found. In English, this is the conventional stress-foot, delimited by the onsets of successive stressed vowels. In Japanese, there was some evidence for a bi-moraic foot, within which individual morae were nested. (Independent evidence from morphology for the bi-moraic foot had hitherto lacked any supporting phonetic evidence.) The strategies employed by individual speakers in adhering to the set task constraints varied much more across Japanese speakers than among English speakers. Some Japanese speakers appeared to make use of a bi-moraic foot, while others showed no evidence of such a construct. All English speakers (in [4] and [19]) showed clear evidence of using the stress foot as a production unit in satisfying the given task demands.

The speech cycling task(s) represent an extreme case of rhythmic organization, where the only stable way to satisfy task demands appears to be production of a hierarchical rhythmic structure, in which one phonological unit is nested within the other. The nature of the phonological unit which is available to solve the problem appears to vary across languages, and may in fact support a discrete classification among languages. Under speech cycling conditions, where a practiced phrase is being repeated, cognitive load is minimal, and upcoming production demands are maximally predictable. Under these circumstances, there appears to be no impediment to the tight coupling between distinct levels in a timing hierarchy.

Further circumstantial evidence for the language-specific nature of the discrete units which constitute levels in a timing hierarchy comes from attempts by the present author to extend the methods of [4] to speakers of Italian and Spanish. Unlike Japanese, both of these languages have lexical stress, and so it was possible to devise text sets with stress patterns comparable to English phrases (e.g. Eng: MANning the MIDdle/It: MUNGo la MUCca/Sp: BUSca la MOto). Subjects could thus be asked to align the first stressed syllable with a high tone, and the second with a low tone, as before. However, after obtaining data from 4 speakers of each language, it became obvious that the targeted speech cycling task, which had been relatively easy to conduct with English speakers, was extremely problematic for speakers of these other two languages. Whereas English speakers typically required about 5 minutes instruction before the experiment could begin, speakers of Italian and Spanish were unable to attempt the task without at least 30 minutes of intensive practice, and they remained very uncomfortable with the task thereafter. Analysis of their data revealed either extreme

variability, or production of a single, simple rhythmic pattern, with the second stress located half way between phrase onsets. The unexpected difficulty and high variability of the data precluded statistical analysis, but the obvious inference to be drawn was that the stress foot, which enables English speakers to coordinate the relative timing of stresses within the PRC, was simply not available to these speakers as a unit, despite the existence of lexical stress in their language.

5. Where else to look?

The work of Grabe and Ramus and colleagues [9, 18] constitutes strong prima faciae evidence for categorical distinctions among languages based on the kind of linguistic unit on which rhythm is "hung". Evidence from Speech Cycling illustrates how, under rather extreme elicitation conditions, entrainment of one prosodic unit within another can be induced. Speech Cycling alone will not suffice to make the case that there is a continually varying level of entrainment between units at one level (syllables, perhaps feet) and prosodic units at a higher level (feet, perhaps phrases), as suggested by O'Dell and Nieminen [15] and Barbosa [2].

The claim being made here is that there is such entrainment, and that the degree of entrainment varies within speaker and across utterances. Because of this high degree of variability, the resulting rhythmic forms are not stable enough to support a rhythmic taxonomy. However, the sort of forms that can emerge are dictated largely by the discrete categories mentioned above, and so we will expect language-specific manifestations of entrainment between prosodic levels.

The evidence for temporal entrainment among prosodic units at distinct timescales under more natural speaking conditions is not uncontroversial. Attempts to identify compensatory shortening within the foot as unstressed syllables are added yielded negative results [12]. Some studies have produced weak evidence of compensatory durational adjustment toward weak isochrony [14, 7], but most such investigations have been fruitless [5]. However, none of these investigations have considered the degree of entrainment between prosodic levels, and hence the strength of rhythmic regularity, to be a continuously variable function. We have recently found some intriguing evidence for a demonstrable entrainment between prosodic levels in read speech, without metronomic influence. These experiments are as yet at an early stage, but they do suggest where we might continue to look in order to tease apart the gradient contribution to rhythmic patterning within a speaker's utterances.

6. Metrical Structure

Methods As part of a larger experiment still underway, speakers provided readings of word lists, where each list contained 8 trochaic forms (e.g. "tango, lighter, daddy, wiper, pony, cutter, pinky, mango"). A total of 54 readers each read 6 such lists in "as regular a form as possible". That is, they were instructed to produce something akin to an isochronous series. From each reading, P-centers, corresponding roughly to vowel onsets, were obtained by semi-automatic means (following the method of [4]), and the first six inter P-center intervals were plotted in several ways. (The final two intervals are not shown, as the last one lacks a measurable right edge.)

Results Two illuminating plots are shown in Fig 3. In the top panel, the first six inter-onset intervals have been computed, and each divided by the mean inter-onset interval. The median and IQR of each is shown (n=318), and the only interval which

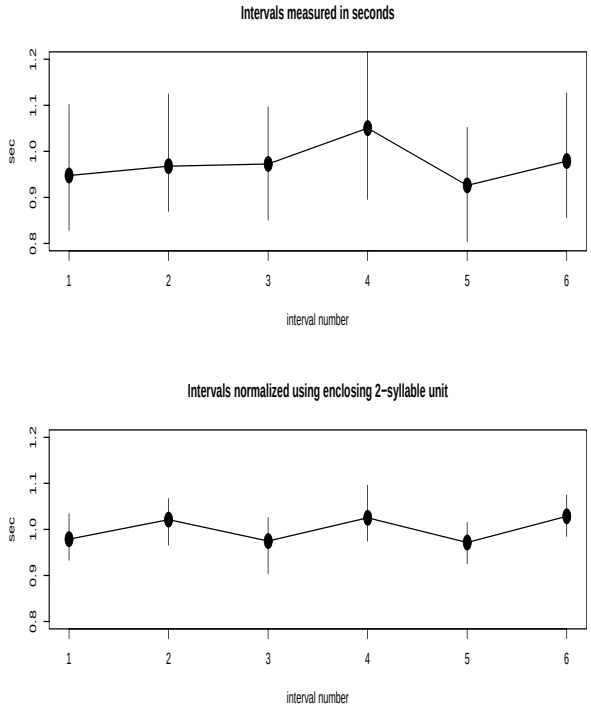

Figure 3: *Median and IQR of intervals from trochaic list reading task.*

stands out is the fourth, separating the first group of four from the second. This interval is longer and more variable than all the others.

In the lower panel of Fig 3, each interval has been normalized by a containing interval. For the first two intervals, the normalizing interval is the duration of the first two intervals, for intervals three and four, it is the sum of intervals three and four, and for five and six, it is the sum of intervals five and six. In order to make these measurements directly comparable with those of the top panel, all normalized intervals are again divided by the mean for the whole data set. This representation of interval duration tells a very different story. Now interval duration, expressed as a proportion of a containing two-interval unit, is much less variable. There is also a clear alternating pattern, where the first interval of each two-interval "foot" is shorter than the second.

A simple model which can account for these data would be one in which produced units are hierarchically organized, with a binary nesting of units at one level inside those at the next, and the further constraint that each unit at each level be subject to some degree of final lengthening. In this way, the inter-word intervals plotted here would be grouped into two-word "feet", with the second interval in each "foot" exhibiting some final lengthening. Each pair of two-word "feet" would again group into four-word units, of which there are two in each list. The additional lengthening arising from this grouping is visible in the top panel of Figure 3 as the long fourth interval. Interval durations expressed in milliseconds are highly variable, reflecting rate variation across list readings and from one speaker to the next. When each interval is re-expressed as a proportion of a containing interval, however, the data become much more coherent.

The task of reading a regular list of 8 trochees, while not as rhythmically constrained as speech cycling, is still carefully designed to elicit maximally rhythmical speech production[3]. Given speech material which lends itself to simple rhythmical grouping, speakers do indeed impose a rhythmic organization on their speech, resulting in durations which are interpretable in terms of simple meter. Not all speech is this regular, however. In the following section, we report some new data which provides tentative support for the hypothesis that hierarchical timing is imposed under much less stringent speaking conditions.

7. Temporal structure as Characteristic of an Individual Speaker

Methods In the course of a larger experiment, readings from 27 speaker pairs were obtained reading the first paragraph of the rainbow text. For each pair of speakers, A and B, a reading was first obtained from A, then A and B read together, attempting to remain in synchrony with one another, then Speaker B read the text. After some intervening practice at this, the process was repeated, with Speaker B starting, then A and B together, and finally Speaker A. From each recording, the final sentence ("When a man looks for something beyond his reach, his friends say he is looking for the pot of gold at the end of the rainbow") was excised, and 16 well defined points in the waveform were identified by hand. These points correspond to reliably recognizable events such as stop releases, vowel onsets etc, and together they divided the utterance into 15 sub-intervals of approximately 2–4 syllables each.

Results This sequence of 15 intervals can again be viewed in two ways. Firstly, we can consider the vector of 15 millisecond values, each expressing a well defined interval. We would naturally expect two utterances recorded in the synchronous condition to be fairly similar by this measure.

However, we can obtain a very crude representation of the rhythmical structure of an utterance by expressing each interval instead as a proportion of some larger containing interval. The above sentence is normally read as two intonational phrases (separated at the comma), so we can re-express the sequence of measurements such that each interval is now given as a proportion of the containing IP (or the measurement points most nearly located at the two ends of that IP). This is also a vector of intervals, but each is expressed as a function of the overall temporal organization of the phrase.

Something rather surprising happens when we consider the similarity of two utterances using these two measures. For each synchronous utterance, we computed the Euclidean distance between this utterance and all 163 other utterances for which all 15 interval measurements were available. We then ordered this list of 163 distances, and noted the index of the matched utterance in the ordered list. The matched utterance is that spoken by another speaker in synchrony with the present utterance. A low index means that the two utterances are similar by this measure. The top left panel of Figure 4 shows the distribution of this index for 92 synchronous utterances, and it can be seen that, in general, the index tends to be low in the ordered list of 163 distances, suggesting a reasonable temporal match between utterances.

When the intervals are expressed as proportions of their containing IPs, however, this similarity goes away. The bottom left panel of Fig 4 plots the same distribution, but this time

[3]The data collected also include somewhat irregular lists which are currently undergoing analysis.

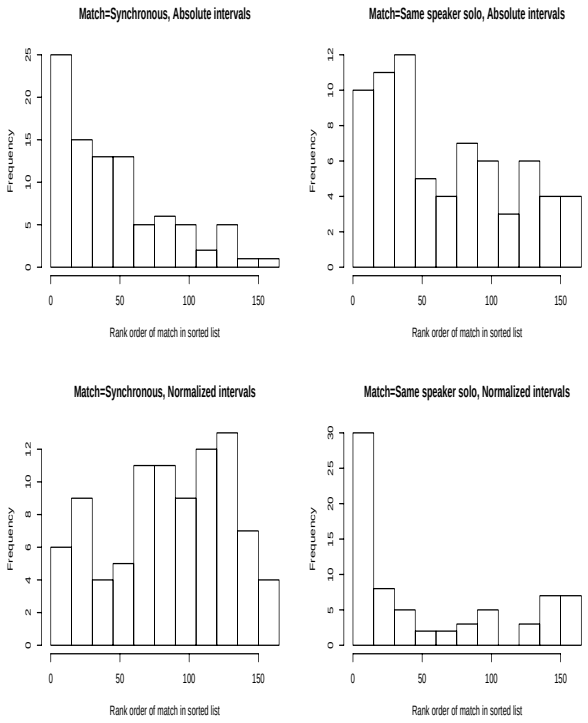

Figure 4: *Distributions of rank order of matched utterances. Details in text.*

using these proportional durations. This distribution no longer has the decaying exponential shape previously seen, and it is not clear that it is different from a uniform distribution, which is the expected distribution if the similarity measure were entirely worthless.

We can carry out the same procedure again, but this time we define the matching utterance to be the solo reading given by the same speaker immediately prior to or immediately after the synchronous reading. The top right panel of Fig 4 plots the distribution of indices so obtained (n=73). Not surprisingly, when we do this using intervals expressed as absolute values, the Euclidean distance between vectors does not do a very good job of picking out utterances by the same speaker. Finally, we can look for the matching utterance (by the same speaker) using normalized intervals (lower right panel). What emerges, quite remarkably, is that this measure does a very good job indeed at expressing similarity between two utterances by the same speaker, even though those utterances were elicited under quite distinct circumstances (reading alone and in synchrony with another speaker).

8. Discussion

Both the preceding experimental results illustrate the coordination of temporal intervals at one level with those at a higher level. In the word list example, metrical structure based on the hierarchical nesting of each word within a two-word unit was evident. In the preceding example, a sequence of temporal intervals in which each interval is expressed as a proportion of a larger interval was demonstrated to be characteristic of an individual speaker, and quite stable across different elicitation conditions. This accords with the finding that timing at both

phoneme and word level remains largely unaltered in speech produced by professional mimics, even though the resulting speech is perceived to be similar to the target voice [6, 20].

All of which brings us back to the subject of speech rhythm. The argument was made that a gradient phenomenon, not yet well understood, mediates the role of syllables in determining macroscopic timing patterns. Its gradient nature precludes it from supporting a classification among languages. Furthermore, it was claimed, pre-theoretical perceptions of rhythm (whether characteristic of a speaker or a language) are derived from an interplay between the discrete and the gradient phenomena. The intervals between stressed syllable onsets have long been held to be of singular importance in the perception of English speech rhythm.

In the word list experiment, we saw that these intervals do in fact partake in a strictly metrical structure, demonstrable and measurable in real time, when the spoken material is sufficiently regular. The units (feet delimited by stressed syllables) are language specific (Japanese, for example, has no correlate of stress), but the participation of these units in genuinely rhythmical structures is dependent on the nature of the spoken utterance.

In the second experiment we saw that the entrainment among levels does exist in some form when the material is less regular. The resulting pattern is not perceived as being rhythmic in a musical sense, but in common with the simple metrical example, there is a demonstrable coupling between intervals at one prosodic level and those at a higher level.

Little is known about the nature or origin of these production constraints which impose hierarchical temporal structure upon an utterance. The similarity which can be observed between speech cycling patterns and patterns of coordination among the limbs [3] suggests that the origin is to be sought in the demands imposed by the finely tuned coordination of heterogeneous components in speech production, and is thus one aspect of motor control in speech. But the elements upon which these patterns are built are embedded in the phonological regularities which typify a given language. Progress in the study of speech rhythm will require taking both the linguistic units and their forms of coordination into account.

9. Acknowledgments

Keiichi Tajima (ATR) helped in preparation of the word lists. Work supported by a grant from the Irish Higher Education Authority.

10. References

[1] David Abercrombie. *Elements of general phonetics*. Aldine Pub. Co., Chicago, IL, 1967.

[2] Plínio Almeida Barbosa. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. In *Proceedings of Prosody 2002*. 2002. to appear.

[3] Fred Cummins and Robert F. Port. Rhythmic commonalities between hand gestures and speech. In *Proceedings of the Eighteenth Meeting of the Cognitive Science Society*, pages 415–419. Lawrence Erlbaum Associates, 1996.

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

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

[6] Anders Eriksson and Pär Wretling. How flexible is the human voice?–a case study of mimicry. In *Proceedings of EUROSPEECH*, volume 2, pages 1043–1046, Rhodes, Greece, 1997.

[7] Edda Farnetani and Shiro Kori. Effects of syllable and word structure on segmental durations in spoken Italian. *Speech Communication*, 5:17–34, 1986.

[8] A. Galves, J. Garcia, D. Duarte, and C. Galves. Sonority as a basis for rhythmic class discrimination. In *Proceedings of Prosody 2002*. 2002. to appear.

[9] Esther Grabe and Ee Ling Low. Durational variability in speech and the rhythm class hypothesis. In *Papers in Laboratory Phonology 7*. 2000. to appear.

[10] Carlos Gussenhoven. Discreteness and gradience in intonational contrasts. *Language and Speech*, 42(2–3):283–305, 1999.

[11] D. Robert Ladd and Rachel Morton. The perception of intonational emphasis: continuous or categorical? *Journal of Phonetics*, 25:313–342, 1997.

[12] Lloyd H. Nakatani, Kathleen D. O'Connor, and Carletta H. Aston. Prosodic aspects of American English speech rhythm. *Phonetica*, 38:84–106, 1981.

[13] T. Nazzi, J. Bertoncini, and J. Mehler. Language discrimination by newborns: towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24:756–766, 1998.

[14] Sieb G. Nooteboom. *Production and Perception of Vowel Duration*. PhD thesis, Utrecht, The Netherlands, 1972.

[15] Michael L. O'Dell and Tommi Nieminen. Coupled oscillator model of speech rhythm. In *Proceedings of the International Congress of Phonetic Sciences*, San Francisco, 1999.

[16] Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA, 1980. Reprinted by the Indiana University Linguistics Club.

[17] Franck Ramus and Jacques Mehler. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America*, 105(1):512–521, 1999.

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

[19] Keiichi Tajima. *Speech Rhythm in English and Japanese: Experiments in Speech Cycling*. PhD thesis, Indiana University, Bloomington, IN, 1998.

[20] Pär Wretling and Anders Eriksson. Is articulatory timing speaker specific? – evidence from imitated voices. In *Proc. FONETIK 98*, pages 48–52, 1998.