CRITERIA FOR DATABASE AND TOOL DESIGN FOR SPEECH TIMING ANALYSIS WITH SPECIAL REFERENCE TO MANDARIN

Yu Jue and Dafydd Gibbon
School of Humanities, Zhejiang University, Hangzhou, China
Faculty of Linguistics and Literary Studies, Bielefeld University, Germany

ABSTRACT

This position paper investigates some of the problems in modelling speech timing for the design of speech databases and corpus analysis tools for phonetics and speech technology. First we examine a selection of phonetic approaches to speech timing analysis, the so-called 'rhythm metrics', and focus on explaining (1) inconsistencies (varying results for the same language) and (2) the failure to model rhythmic alternation. To overcome these problems we present a new perspective on the phonetic identification of rhythm patterns as a special case of duration modelling, including the additional criterion of alternation. We describe the Rhythm Parser, a tool for identifying hierarchical alternating patterns, and discuss results from applying it.

Index terms *– speech corpus, speech timing, rhythm metric, timing hierarchy, bottom-up analysis, peak unit*

1. OBJECTIVES AND OVERVIEW

This position paper is concerned with the requirements imposed on speech database analysis by the study of duration, timing and speech rhythm, and with suitable models and tools for processing speech corpora.

Speech unit durations, timing patterns and speech rhythm have interested linguists and phoneticians for decades (Campbell [7]), and are of increasing interest in speech technology. Some aspects, particularly rhythm, are controversial and have been examined from a number of points of view: as phonological structure (in autosegmental-metrical phonology in terms of tree structures and grids/histograms); as quantitative 'rhythm metrics' for phonetic patterning (with variability and smoothness measures based on descriptive statistics, Low et al. [24]); as dynamic cognitive and computational processes based on interacting oscillators (Barbosa [4]); as a by-product of sound structure without independent existence (Gut [20]).

We focus on formal and empirical properties of descriptive statistical studies (cf. also Gibbon [16]). We share the empirically critical approaches of Arvaniti [2] and Gut [20], but go further by showing formal flaws in the metrics leading to arbitrary numbers of false positives. Crucially, the models fail to account for *regularly iterated alternation* in rhythm. It is not sufficient only to criticise: we claim that duration is still a valid factor, and propose a new bottom-up model, differing from previous aproaches by incorporating alternation and hierarchy into the model, and by providing a proof-of-concept study with Mandarin data.

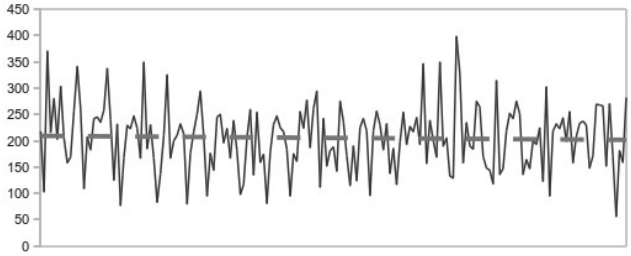

Figure 1: Syllable durations in a read Mandarin text. The slope of the dotted regression line models a slightly accelerating change in speech rate.

Figure 1 illustrates the issue with a sequence of syllable durations from a reading of the standard IPA text *The North Wind and the Sun* in Mandarin (with pauses removed). There are clear alternations at three main levels: long stretches (peaks above about 250ms), shorter units between longer peaks (around 210ms), and syllables. Durations decrease across longer stretches of the utterance, a global acceleration trend, and re-start to constitute new groups.

2. DEVELOPMENT OF RHYTHM METRICS

Traditionally speech rhythm has been regarded as regular recurrence in time (*isochrony*) of some given speech unit (Roach, [27]), and languages are assigned to two main rhythm classes: 'stress-timed' and 'syllable-timed', e.g. Abercrombie [1], sometimes also 'mora-timed'.

Isochrony is an elusive concept. Categorial distinctions between stressed-timed and syllable-timed languages have therefore been replaced by speech rhythm as a multidimensional percept covering several phonetic properties (e.g. syllable structure, vowel reduction), and prosodic properties (e.g. pitch accent and other markers of prominent and less prominent speech units). Dauer [10, 11] suggested that languages are not classified into distinct rhythmic classes but are located along a continuum from syllable-timed to stress-timed.

Many studies have introduced variability metrics which extend Dauer's approach, proposing specific quantitative

measures. Ramus et al. [26] proposed that differences in rhythm type could be accounted for by a set of variables derived solely from the acoustic duration of vocalic intervals: *%V* refers to the vowel and consonant sequence duration ratio; $\triangle V$ denotes standard deviation of durations for vowel sequences, $\triangle C$ for consonant sequences. The Ramus et al. approach ignores speech rate variation (Barry [5]; Dellwo [13]), however. Dellwo et al. [13] therefore introduced normalised measures, *VarcoC* and *VarcoV*, (percent standard deviation of consonantal and vocalic durations, each normalised by dividing by the mean).

Low et al. [24] proposed the *normalised Pairwise Variability Index* (*nPVI*) to exploit the duration difference between pairs of successive syllables, e.g. stressed and unstressed vowels, which tends to be much greater in stress-timed languages. They proposed a non-normalized *raw PVI* (*rPVI*) for the more stable consonantal intervals.

Several studies used pairs of these variability metrics to define the distribution of languages in a 2-dimensional space, showing that many languages, e.g. Greek, Malay, Romanian, Catalan and Polish, do not fit easily into the traditional syllable vs. stress timing dichotomy.

3. EMPIRICAL ISSUES

3.1 DATA ISSUES

Ramus et al. [26] started their rhythm studies based on 4 speakers per language and 5 sentences per speaker with careful control of speech rate, i.e. 20 sentences. Low et al. [24] made duration measurements on comparable passages of speech from eighteen languages (one speaker per language), i.e. 18 passages; this is the first study in which a fairly large number of languages was examined.

Dellwo et al. [12] pointed out that the use of a limited amount of data, especially in rhythm studies, which really require the analysis of longer sequences, may lead to artefacts in the results. The technological problems of creating large speech databases were a considerable obstacle in the early days of speech rhythm analysis, but technological progress has enabled many studies based on large corpora to be made in the meantime.

Rhythm metrics calculate fine-grained durational differences between vowels, consonants or syllables, but the segmentation procedures for these units vary widely across studies. Thomas et al. [29] counted formant transitions for obstruents as vocalic but Gut [19] counted them as parts of consonants. Postvocalic glides were counted as vocalic by Ramus et al. [26], but as consonants by Arvaniti [2] if they showed frication. For comparability, segmentation criteria must be explicitly justified and defined.

3.2 RELIABILITY

It has emerged in recent studies that the rhythm metrics do not yield similar results even for the same language (Arvaniti [2], Gut [20]). These inconsistencies can be demonstrated with results for Mandarin (cf. Table 1).

Table 1: Different metric values for Mandarin, found in the literature.

Sources:	*%V*	ΔV	ΔC	*nPVI-V*	*rPVI-C*
Low et al. [24]	55.80%	36.2	44.1	27	52
Mok et al. [25]	54.40%	53.1	45.40	51.65	56.2
Shao [28]	51.60%	55	38.3	48.7	42.9
He [21]	74.60%	36.8	34.6	39.5	41.7

The *%V* values range from 51.6% to 74.6%, ΔV ranges from 36.2 to 55, and the *nPVI-V* ranges from 27 to 51.65. These large differences make the metrics meaningless from the point of view of language typology, as results vary more in speech styles within languages than between languages.

It is also difficult to maintain the validity of rhythm metrics in L2 speech (Arvaniti [2], White et al. [30]): most metrics failed to distinguish between native and non-native speech rhythm or to yield significantly different values for L2 learners at different proficiency levels. The metrics are clearly influenced by choice of speakers, materials and speaking style (Gut [20]). Barry [6] suggested that the rhythm in the speech of L2 learners is just the phonetic by-product of phonological processes. Thus, on empirical grounds the conclusion can be drawn that these metrics do not adequately model rhythm (Gibbon et al. [18]).

3.3 FACTORS INVOLVED IN SPEECH TIMING

Speech unit timing is influenced by phonetic features such as differences of manner, voicing and place of articulation, and by higher-level factors such as prosody and utterance rhythm, syntax, semantics, lexicon, state of the speaker, speech style, and quality of the material selection (Easterday et al. [14]). To avoid compromising the reliability of results, these factors are to be controlled for in all phases of preparing a speech database: pre-recording phase, recording phase, post-recording phase.

Differences of manner, voicing and place of articulation condition the lengths of segments. Tense vowels tend to be longer than lax vowels, and vowel durations vary inversely with vowel height for extremes of height (Crystal et al. [9]). Diphthongs are inherently longer than monophthongs, voiced plosive releases are generally shorter than voiceless plosive releases, and velar plosive releases are longer than labial or dental plosive releases. Mandarin tone 3 (fall-rise with creaky phonation) in isolation is longer than the unidirectional tones, maybe because tone direction change together with phonation type shift requires more time.

Segmental context can also influence segmental duration. In many languages segments are shortened in consonant-consonant sequences across word but not phrase boundaries (Klatt [23]). Vowels are longer before nasal followed by voiced plosive and tend to be longer before bilabials than before alveolars or velars. Vowels next to [b, d, g] tend to be longer than vowels next to [p, t, k]. Vowels

after plosives lengthen more than those after continuants. Sounds in clusters are shorter than singletons: [s] and [p] are shorter in [sp] clusters than as singletons.

Parts-of-speech, lexical and phrasal stress, information structure relate in complex ways to duration. Mandarin has no lexical stress like English, but Mandarin content words are more likely to have phrasal stress than functional words, thus increasing duration. According to the model of Coker et al. [8], stress is determined by newness of information, in this order: new nouns > prepositions as complements > new infrequent verbs, adjectives, adverbs, repeated nouns > repeated infrequent verbs > interrogatives, quantitatives > frequent verbs > less frequent function words > ordinary function words > schwa function words. Much variation is therefore grammatically conditioned. Grammar also influences segmental and syllable duration on different hierarchical levels. Environments such as the degree of word and sentence stress, the degree of finality, and position in the word, foot, and phrase etc., constrain the duration of segment and syllable on a higher level, to different extents in different languages. For example, vowels are longer in phrase-final words and before voiced consonants prepausally. Unstressed segments are shorter in duration and considered more compressible than stressed segments.

Campbell [7] defines a tripartite model with control of rate, prominence and boundary marking at the syllable level and above, which effectively predicts segmental durations within the syllable. Higher level rhythmic effects control speech timing to some extent, rather than being a function of other timing factors. At foot level syllable durations tend to be shorter as the number of syllables in the foot is increased. This compensatory effect is sensitive to syllable type (stressed-unstressed) and foot type, either headed with a stressed syllable, or anacrustic in phrase-initial position and lacking a stressed initial syllable (Campbell [7]). Huggins [22] used just-noticeable differences (JND) of phonetic segments to show that listeners are particularly sensitive to the rhythmical aspect of sentence timing: subjects are less sensitive to timing changes of adjacent syllables than rhythm changes between stressed syllables.

4. FORMAL PROBLEMS

A rhythm is a repeated, temporally regular iteration of alternating values of observable parameters, e.g. strong-weak, light-dark, loud-soft, consonant-vowel, hand raised vs. hand lowered (Gibbon [17]), over sequences (not necessarily binary) of structural units (foot, syllable, etc.). A formal model of speech rhythm should therefore capture not only *variability of durations* but also *iterated alternation* in duration patterning.

4.1 THE RAMUS MODEL

The core issue is: do %V, ΔV and ΔC (Ramus et al. [26]) measure rhythm?

First, the *%V* metric sets the sum of the durations of all vowels over the whole sentence in relation to the duration of the sentence (pauses excluded). But high *%V* may indicate several properties, e.g. presence of both long and short vowels, or no syllable reduction. Low %V may result from long consonant clusters, as in English (with vowel length contrast) or Polish (without vowel length contrast).

Second, the ΔV metric refers to the standard deviation of vowel duration. High ΔV means there is more variation in vocalic length, which may be due to any of the following three factors: phonemic vowel length contrasts will tend to generate high ΔV (like *%V*); contrastive stress, emphasis and emotional expression will tend to involve vowel lengthening and consequently a higher ΔV; vowel reduction will also tend to have the same effect.

Third, the ΔC metric refers to the standard deviation of consonant duration. High ΔC may indicate simply that the language has complex consonant clusters such as CV, CCV, CCCV, ... CCCVCCC in English.

Each of the three variability metrics of Ramus et al., *%V*, ΔV and ΔC, is thus potentially affected by a number of factors which have little to do with the definition of rhythm as regularly iterated alternation. Further, the metrics miss the alternation component of rhythm because they average globally across whole utterances: in principle, durations of different lengths can be ordered randomly, longest-to-shortest or shortest-to-longest, and still have the same variability index as durations in a genuine alternating rhythmical sequence. So the variability metrics do contribute to showing one part of the definition of rhythm, position on a scale from 'smoothness' or regularity towards randomness, but not alternation. The formal properties of the model thus permit many different kinds of false positive inclusion of non-rhythmic properties as rhythm.

4.2 THE LOW AND GRABE MODEL

Low and Grabe (2002) developed a model of local binary duration relations, the *normalised Pairwise Variability Index* (n*PVI*): 100 multiplied by the average of all differences between neighbouring duration pairs, each difference divided by the average duration of the pair. They used *nPVI* for vocalic intervals, and *rPVI* ('raw', i.e. without normalisation by duration pair mean) for consonantal intervals. The minimum *nPVI* is 0 (full isochrony), normalisation by duration mean yields a maximum *nPVI* of 200 (normalisation by sum would yield 100). The maximum value is never reached but is approached asymptotically.

As a rhythm model, the *nPVI* is flawed. First, the *nPVI* (and the *rPVI*) compare durations pairwise, and thus presuppose that the rhythmic alternation is binary. However, that is not always the case: cf. *THAT is not ALways the CASE*. Second, neither the *nPVI* nor the *rPVI* can discriminate between short-long sequences and long-short sequences because they both take the absolute value of the difference within the pair: this destroys possible alternating

properties. Third, also because of the absolute operation, alternations, geometrical series (increasing or decreasing), and any mix of these can yield the same *nPVI* or *rPVI*. For example, the sequences 1 2 1 2 1 2, 1 2 4 8 16 and 1 2 1 2 1 2 4 8 16 32 generate the same *nPVI*: 66.66'.

The consequence of this critical discussion is that while a low *nPVI* indicates near isochrony of the speech unit concerned (e.g. the syllable), a high *nPVI* means simply that durations are subject to near-random variation. Thus the *nPVI* and the *rPVI* measure not rhythm but simply an overall 'smoothness', like other variability measures, and are open to the same false positives criticism as these.

5. A NEW APPROACH: THE RHYTHM PARSER

5.1 TIMING BEYOND THE SENTENCE

More important in the long run than the analysis of short data items is the development of a model of temporal organisation in longer stretches of connected speech.

It has frequently been noted that an important factor to be controlled for is speech rate. Figure 1 shows the sequences of syllable durations in a reading of the standard IPA text *The North Wind and the Sun* in Beijing Mandarin, with pauses removed. Initial inspection shows that medium length syllable durations cluster around 210 ms, with less frequent relatively regularly distributed longer and shorter syllables. The regression line of Figure 1 shows minimal acceleration of global speech rate.

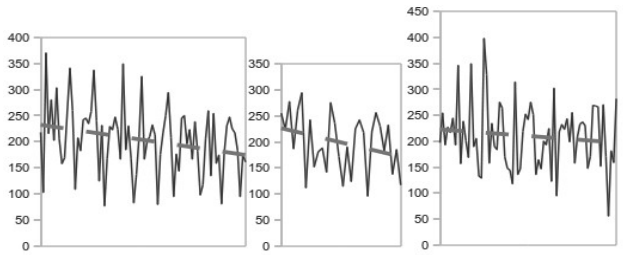

Figure 2: Three discourse segments of the reading of *The North Wind and the Sun* in Mandarin.

But closer examination of the data shows that the global speech rate measurement is misleading. Figure 2 shows optimal regression lines separately fitting and thus defining initial, medial and final parts of the data, which also correspond (within one or two syllables) with semantically distinct episodes in the narrative.

Figure 2 shows a clear acceleration slope for speech rate in each part, with syllables becoming shorter and shorter (except for final lengthening): the initial part shows clear acceleration, the medial part shows a even more acceleration, and the final part has the flattest slope.

The consequence of this illustration is that in speech database processing the hierarchical long-term temporal structure of the data must be investigated, and tools designed accordingly, both in phonetics and in speech technology. In this case, comparison of Figure 1 and Figure 2 clearly indicates that measures must be normalised for speech rate, and that the normalisation domain may need to be adjusted for different sections of the data.

5.2 TIMING BEYOND THE METRICS

The definition of rhythm involves not simply overall 'smoothness', but also the alternation of higher and lower values of speech parameters such as duration, pitch or other indicators of prominence. At the syllable level, sonority alternation between vowels and consonants may be taken as the alternation criterion. A sonority measure which captures the degree of obstruency in the signal for this purpose was introduced by Galves et al. [15].

Over longer stretches more complex prominence measures are needed (Asu et al. [3]). However, units such as the stress group or foot are linguistically defined, and thus represent *a priori* assumptions about the acoustic structure of speech. We propose a bottom-up acoustic measure for parsing into larger speech units (*Peak Unit*s), and have implemented a tool for examining the duration properties of these larger groups. The *Peak Units* are shallow or minimal *Time Trees* of depth 2, in the sense of Gibbon [16]. The data flow design is shown in Figure 3.

After annotation of the relevant speech segments, e.g. syllables, with Praat (or similar speech annotation tool), durations are automatically extracted from the annotation file. In previous rhythm metrics, durations were passed to a regularity processor ('rhythm metric') along route A in Figure 3, generating a *Smoothness Index* (*SI*) such as *%V* or *nPVI*. In our measure, in addition to *SI*, an *Alternation Index* is generated along route B by using a preprocessor parser to identify *Peak Units* on acoustic grounds alone.

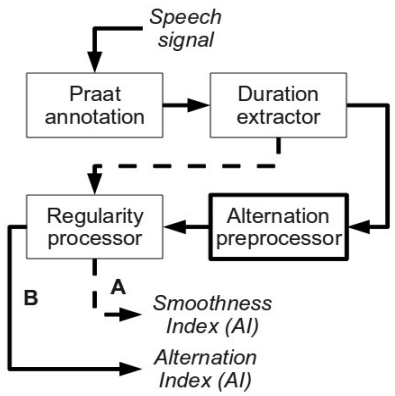

Figure 3: Data flow for handling alternation and smoothness in duration sequences.

We examine (1) two models for the phonetic definition of Peak Units: internal *acceleration* (syllables get shorter) and internal *deceleration* (syllables get longer); (2) we define a threshold parameter which currently is varied

empirically; this will be replaced in due course by a regression based criterion. The data visualised in Figure 1 are examined using an implementation of the model, the Rhythm Parser. Syllables are inherently alternating (vowels alternate with consonants), and the Rhythm Parser model adds an alternation criterion for larger units: the *Alternation Index*, the ratio of syllable *nPVI* to *Peak Unit nPVI*

6. RESULTS

The *Acceleration* condition appears to perform better with smaller units (partly due to the presence of shorter enclitic function words), while the *Deceleration* condition yields Peak Units which relate well to sentences (smaller threshold) and paragraphs (larger threshold).

Table 2: Peak Unit parse output from Rhythm Parser.

#	|PU|	Size of PU & syllable sequence, with lengths.
1	17369	84: you3(218) yi4(103) hui2(370) bei3(216) feng1(280) gen1(203) tai4(303) yang2(205) zai4(158) nar4(169) zheng1(257) lun4(341) shui2(260) de5(109) ben3(207) shi5(183) da4(242) zheng1(245) lai2(235) zheng1(258) qu4(337) jiu4(240) shi4(125) fen1(231) bu4(77) chu1(167) gao1(229) di1(223) lai2(247) zhe4(224) shi2(167) hou5(349) lu4(185) shang5(230) lai2(167) le5(83) ge4(137) zou3(209) daor4(325) de5(167) ta1(200) shen1(210) shang5(232) chuan1(213) zhe5(80) jian4(176) hou4(217) da4(250) yi1(294) ta1(208) men5(95) lia3(176) jiu4(144) shuo1(245) hao3(250) le5(196) shui2(223) neng2(167) xian1(238) jiao4(181) zhe4(98) ge5(117) zou3(205) daor4(259) de5(135) tuo1(254) xia4(159) ta1(174) de5(81) hou4(178) da4(231) yi1(247) jiu4(225) suan4(217) shui2(184) de5(95) ben3(175) shi5(161) da4(255) bei3(224) feng1(277) jiu4(187) shi3(261) jinr4(294)
2	6341	32: de5(112) gua1(242) qi3(152) lai2(181) le5(188) bu2(142) guo4(275) ta1(236) yue4(172) shi4(115) gua1(190) de5(124) li4(225) hai5(242) na4(218) ge5(96) zou3(219) dao4(256) de5(230) ba3(183) da4(232) yi1(138) guo3(185) de5(117) yue4(197) jin3(254) hou4(193) lai2(227) bei3(217) feng1(244) mei2(193) far3(346)
3	7707	37: le5(157) zhi3(238) hao3(201) jiu4(169) suan4(349) le5(190) guo4(205) le5(133) yi4(129) hueir4(398) tai4(328) yang2(158) chu1(234) lai2(190) le5(184) ta1(275) huo3(264) la4(170) la4(148) de5(144) yi2(118) shai4(314) na4(136) ge5(147) zou3(219) daor4(252) de5(242) ma3(275) shang4(251) jiu4(136) ba3(164) na4(147) jian4(200) hou4(193) da4(224) yi1(123) tuo1(302)
4	4924	24: le5(95) xia4(218) lai2(232) zhe4(223) xia4(243) bei3(199) feng1(255) zhi3(158) hao3(208) cheng2(233) ren4(237) ta1(230) men5(148) lia3(171) dang1(269) zhong1(268) hai2(266) shi5(152) tai4(270) yang2(171) de5(56) ben3(181) shi5(159) da4(282)

A sample of Rhythm Parser output (condition *Deceleration*, threshold 180) is shown in Table 2, yielding four Peak Units, Peak Unit lengths, Syllable counts, Syllable lengths. The example shows that Peak Units at certain threshold levels correspond well with semantic text units. Tentative error rates for comparisons of the three threshold conditions are shown in Table 3. A plausible interpretation for Table 3 is that different hierarchical levels of Peak Units are captured at different threshold levels. Space does not permit detailed discussion; application to extensive data is in progress.

Based on the parsed output, rhythm related regularity measurements can be made of syllable duration regularity and Peak Unit duration regularity, and the *Alternation Index* can be calculated. This index only makes sense for smaller Peak Units, not for the larger units under discussion here.

Table 3: Peak Unit – paragraph correspondence.

Threshold	PU#	*n*	*err*	Initial false		Final false	
				-	+	-	+
>=178 <=181	1	84	0.06	0	0	0	5
	2	32	0.34	5	0	6	0
	3	37	0.24	0	6	3	0
	4	24	0.13	0	3	0	0
>=182 <=188	1	116	0.52	0	0	0	6
	2	37	0.24	0	6	3	0
	3	24	0.13	0	3	0	0
>=189 <=206	1	153	0.02	0	0	3	0
	2	24	0.13	0	3	0	0

In Table 4, the syllable regularity for the *Deceleration*:180 condition is relatively high (*nPVI*=36), while the peak regularity of *nPVI*=52 is something of an artefact under this condition, with only 4 units to measure. For comparison, results for the condition *Acceleration*:90 are included, showing Peak Unit *nPVI* (57) for a much larger number of Peak Units, i.e. a slightly larger but more realistic result, and a smaller Alternation Index. The *Deceleration*:90 condition, on the other hand (not in the table), yields a worse Peak Unit *nPVI* of 67.

Table 4: Regularity values.

***Deceleration*:180**	
Threshold:	180
Number of PeakUnits:	4
Number of Syllables:	177
Mean Syllables per PeakUnit:	44
Syllable nPVI:	36
PeakUnit nPVI:	52
Syllable/PeakUnit npVI ratio:	0.69
***Acceleration*:90**	
Threshold:	90
Number of PeakUnits:	27
Number of Syllables:	177
Mean Syllables per PeakUnit:	6
Syllable nPVI:	36
PeakUnit nPVI:	57
Syllable/PeakUnit npVI ratio:	0.62

7. CONCLUSION AND OUTLOOK

In this position paper we have examined a number of models which claim to be 'rhythm metrics', and pointed out a number of empirical and formal problems which are inherent in these metrics as models of rhythm. We illustrated the empirical problems (inconsistency of results; speech rate issues) with reference to Mandarin. We also identified the main formal problem with the rhythm metrics: they do not identify the alternations which are characteristic of 'real rhythm'. However, we claim that duration is still a valid factor, and propose an extended model as a Rhythm Parser incorporating an Alternation Preprocessor which identifies Peak Units as a basis for determining temporal regularity values using the *nPVI*. We conducted preliminary tests on an extract from a corpus of Mandarin speech. The proof-of-concept analysis is based on a corpus fragment.

Current work is applying the Rhythm Parser to large corpora. The Rhythm Parser itself is being further developed to include automatic comparison between Peak Units and text units unit sizes determined by different thresholds. We predict that threshold regions can be identified in this way which discriminate systematically between categorially different levels of timing units. The current Python CGI implementation of the Rhythm Parser can be accessed at:

http://wwwhomes.uni-bielefeld.de/gibbon/rhythm.html

8. REFERENCES

[1] Abercrombie, D. *Studies in Phonetics and Linguistics*. London: Oxford University Press. 1965.

[2] Arvaniti, A. The usefulness of metrics in the quantification of speech rhythm. *Phonetica* 66:46-63. 2009.

[3] Asu E.L. and F. Nolan. Estonian rhythm and the pairwise variability index. *Fonetik* 2005, Department of Linguistics, Göteborg University, 29-32. 2005.

[4] Barbosa, P. A. Generating duration from a cognitively plausible model of rhythm production. In *Eurospeech 2001*, Aalborg, Denmark, September 3–7, (2):967–970. 2001.

[5] Barry, W. J., B. Andreeva, M. Russo, S. Dimitrova and T. Kostadinova. Do rhythm measures tell us anything about language type? In *ICPhS XV,* 2693–2696. Barcelona. 2003.

[6] Barry,W. J. Rhythm as an L2 problem. How prosodic is it? In J. Trouvain and U. Gut, eds. *Non-native prosody: Bridging the Gap between Research and Teaching*. 97–120. 2007.

[7] Campbell. N. Multi-level Speech Timing Control. University of Sussex, doctoral dissertation. 1992.

[8] Coker. C., N. Umeda, and Browman. C. Automatic synthesis from ordinary English text. *IEEE Trans. Audio Electroacoust*, AU-21:293–297. 1973.

[9] Crystal. T. H. and House. A. S. Segmental durations in connected-speech signals: Currrent results. *Journal of the Acoustical Society of America* 83(4):1553–1573. 1988.

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

[11] Dauer, R. Phonetic and phonological components of language rhythm, *ICPhS XI*, Talinn, Estonia. 447–450. 1987.

[12] Dellwo, V., B. Aschenberner, J. Dancovicova and P. Wagner. The BonnTempo-Corpus and Tools: A database for the combined study of speech rhythm and rate. *Interspeech 2004 (ICSLP)*. Jeju Island, Korea, 777–780. 2004.

[13] Dellwo, V. and P. Wagner. Relations between language rhythm and speech rate. In *ICPhS XV*. 471–474. Barcelona. 2003.

[14] Easterday, S., J. Timm and I. Maddieson. The effects of phonological structure on the acoustic correlates of rhythm. *ICPhS XVII*, 623–626. 2011.

[15] Galves. A, J. Garcia, D. Duarte and C. Galves. Sonority as a basis for rhythmic class discrimination. In *Prosody* 2002. Aix-en-Provence, France. 2002.

[16] Gibbon, Dafydd. Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In Sudhoff, Stefan & al., eds. *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter. 281–209. 2006.

[17] Gibbon. D. Formal models of oscillation in rhythm, melody and harmony. *Speech and Language Technology* 14/15:35–44. 2012.

[18] Gibbon, D. and U. Gut. Measuring speech rhythm. In *Proceedings of Eurospeech*. 91–94. Aalborg. 2001.

[19] Gut, U. *Non-native Speech: a Corpus-based Analysisi of Phonological and Phonetic Properties of L2 English and German*. Frankfurt: Peter Lang. 2009.

[20] Gut. U. Rhythm in L2 speech. *Speech and Language Technology* 14/15:83–94. 2012.

[21] He, L. Interlanguage Rhythm. University of Edinburgh MA. thesis. 2010.

[22] Huggins. A. W. F. Just noticeable differences for segment duration in natural speech. *JAcSocAm.* 4(51):1270–1278. 1972.

[23] Klatt. D. H. Duration characteristics of pre-stressed word-initial consonant clusters in English. *Technical Report* QPR. 108, MIT, Cambridge MA. 1973.

[24] Low, E. L., E. Grabe and F. Nolan. Quantitative characterisations of speech rhythm: 'Syllable-timing' in Singapore English. *Language and Speech*, 43:377–401. 2000.

[25] Mok, P. K. and V. Dellwo. Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English, In *Proceedings of Speech Prosody 2008*, Campinas, Brazil, 423–426. 2008.

[26] Ramus, F., M. Nespor and J. Mehler. Correlates of linguistic rhythm in the speech signal. *Cognition*, 72:1–28. 1999.

[27] Roach, P. On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal, *Linguistic Controversies*. 73–79, London: Arnold. 1982.

[28] Shao P. F. The Comparison and the Evaluation in Prosody between Accented Mandarin and Standard Mandarin. Shangdong Normal University, MA. thesis. 2009.

[29] Thomas, E. R. and P. M. Carter Prosodic rhythm and African American English. *English World-Wide* 27:331–355. 2006.

[30] White, L. and S. L. Mattys. Calibrating rhythm: First Language and Second Language studies. *Journal of Phonetics* 35:501–522. 2007.