Time Group Types in Mandarin Syllable Annotations*

Jue Yu
School of Foreign Languages
Tongji University
Shanghai, China
erinyu@126.com

Dafydd Gibbon
Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld
Bielefeld, Germany
gibbon@uni-bielefeld.de

***Abstract*—We describe an inductive procedure for automatically parsing syllable sequences in speech annotations into Time Groups (TGs) based on deceleration and acceleration measures, rather than predefined units from the prosodic hierarchy. The effect of different minimal duration difference thresholds (MDDTs) on the size and relation of these TGs to different grammatical units, e.g. 'word', 'phrase', 'sentence', is examined. The relative syllable isochrony in TGs is also determined with the *nPVI* measure. We found several non-trivial effects, demonstrating the plausibility of the methodology: (a) a stepwise, non-continuous relation between MDDT increases and number of induced TGs; (b) correspondence of induced TGs to sizes and ranks of grammatical units; (c) better correspondence with shorter phrasal units in the acceleration condition and longer discourse units in the deceleration condition; (d) comparability of relative isochrony using syllabic *nPVI* with the vocalic *nPVI* results from previous studies.**

Keywords—Mandarin; speech timing; Time Group; rhythm; acceleration; deceleration; relative isochrony

I. Introduction

Speech timing has been extensively studied in several disciplines, including phonetics and experimental phonology, speech engineering, psychoacoustics and psycholinguistics, and the sociolinguistics of discourse. The present study combines linguistic phonetic and computational methods in designing an inductive procedure for the automatic parsing of syllable sequences, Time Groups (TGs), in speech annotations. The procedure is implemented with a novel interactive web tool, the Time Group Analyzer (TGA), designed for linguistically oriented phonetic analysis across languages, with a view to applications in areas such as language testing and speech synthesis.

Analysis proceeds inductively from the syllable stream in the annotations without assumptions about phonological 'foot' or 'phrase' units. The criteria for sequence identification are deceleration (consistent slowing down; longest last; positive duration slope) and acceleration (speeding up; longest first; negative duration slope), a technique originally used in [1] (for pitch patterning, not syllable durations). Minimum Duration Difference Thresholds (MDDTs) are used to determine the minimal absolute duration difference which counts as a difference for purposes of TG parsing. The primary objective is to find TGs which relate to language categories: analysis results are compared with grammatical categories, and the dependence of TG parses on the MDDT, and of the size of associated grammatical units on the size of TGs, are investigated.

The analysis procedure also applies the *nPVI* measure of variability and 'evenness' of duration or relative isochrony to syllable sequences in TGs. Relative isochrony is a necessary (but not sufficient) criterion for rhythmic timing. However, the study is not directly concerned with speech rhythm as such; there are too many unresolved issues in this field to attempt full rhythm modelling, and distinctions between relative syllable and stress/foot isochrony [2] have turned out to be too simple or to rely on prior assumptions about linguistic structure (cf. critique and modifications in [3], [4], [5]): (a) overgeneralisation from specific speech styles to entire languages; (b) incorrect assumption that rhythm patterns are binary; (c) lack of consideration of interdependencies with segmental phonetic, pitch variation and higher level linguistic factors; (d) use of *global* variation metrics such as standard deviation, consonant-vowel timing ratios [6] or the *normalised Pairwise Variability Index* (*nPVI*) [7], which ignore essential *local* rhythmic properties such as alternation [8] and iteration or oscillation (as isochronous iteration) [9], though they are useful for modelling global relative isochrony. Overviews of these issues are given by [10], [11] and [12].

In the following section the choice of the read-aloud narrative genre is justified and the Mandarin data are characterised. In Section 3 the criteria for TG identification are discussed and the differential model of timing patterns in terms of acceleration and deceleration is defined. Section 4 is concerned with the results of processing the data with the TGA. In Section 5 the implications of the results for identifying linguistic constituents using TGs derived from the differential model are discussed. In Section 6, the final section, prospects for further research and application are outlined.

* This paper is supported by the National Social Science Foundation of China (Grant No. 15BYY183).

978-1-4673-8279-3/15/$31.00 ©2015 IEEE

II. MANDARIN DATA AND ANNOTATION

The present study is concerned with timing in Mandarin Chinese. Timing is influenced by many phonetic factors, such as prosody and rhythm, syntax and semantics, including the linguistic prosodic hierarchy and deep semantic hierarchies (see Fig. 1). Mandarin Chinese has sometimes been classified as a syllabic language with reference to orthography, morphology, syllabic lexical tone and speech timing. In addition to factors involved in stress-intonation languages such as English, Mandarin also has tone as a syllable timing factor, with tone 3 (falling-rising) being tendentially longer than tones 1, 2 and 4.

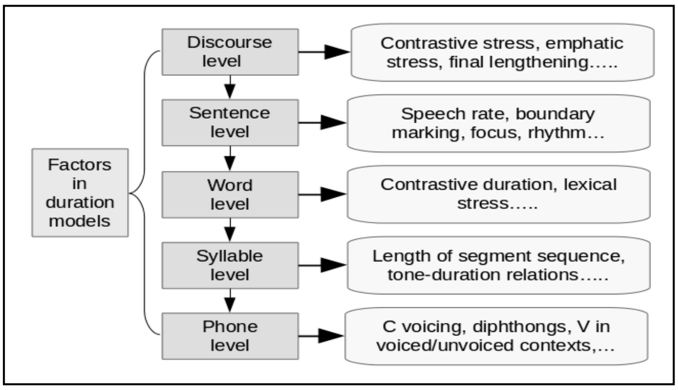

Fig. 1. Factors in duration patterning.

In [13] 190 million Chinese syllables were analysed in different phonetic contexts, with the result that prosodic boundaries, syllable structure, tones and location in words can influence syllable timing. The role of the hierarchical organisation of multiple factors in syllable timing in spontaneous speech is shown in [14], semantic influence being the most important. In [15] systematic contributions of phonological timing and multiple layers of discourse units to the dynamics of output speech rate are discussed.

The study examines properties of the syllable as a whole, rather than fine control in terms of subsyllabic factors or tone, which are not included at this stage of the research. The study is also concerned with multiple layers, but using an inductive method based on acoustic information from annotated syllable time-stamps, rather than a deductive method starting with predefined larger linguistic units.

The focus in the present contribution is on the timing of syllables in formal speech. The Mandarin Chinese speech data are from the read-aloud narrative genre, which is rather formal and relevant in many application fields from language comparison to speech synthesis. Read speech (a) can provide a clear case of a well-defined speech style; (b) is intrinsically more homogeneous than spontaneous speech because of the impact of fewer discourse factors; (c) has fewer reduction phenomena, so the smallest unit for analyzing is syllable rather than foot; (d) is also a natural speech activity in many practical contexts such as spontaneous or practised reading aloud of narratives, in instructional speech and in many speech synthesis applications (note that features of spontaneous speech such as disfluencies are not always excluded).

The production task is the reading of a well-known coherent story, the classic Aesop fable and IPA standard text ‘The North Wind and the Sun’, in Mandarin translation, with no specific preparation or practice. Data from productions of ten speakers are used, five male and five female and the speakers are native Beijing Mandarin speakers. The data are from the CASS corpus [16].

The recordings were annotated at the syllable level using the Praat speech workbench [17]. Mandarin syllable structure is (O)(G)V(N), (Obstruent, Glide, Vowel, Nasal), traditionally represented as a binary structure, either O+GVN or OG+VN. Syllable boundaries were identified as follows: (1) utterance initial voiceless obstruents were assigned a 50ms closure; (2) in adjacent syllables ...V+V … a perceptual criterion is used; (3) if adjacent V are identical, or if a sequence Nasal+Nasal occurs, each segment is assigned 50% of the combined duration.

The syllable tier time stamps and text in Praat TextGrid attribute-value format are automatically converted by the TGA into fixed-position tabular CSV vector format with the syllable record structure quadruple <*start*, *end*, *duration*, *label*> for further processing (timestamps re-formatted to ms), e.g. the Praat format

```
intervals [2]:
   xmin = 3.4419
   xmax = 3.6912
   text = "bei3"
```

is initially converted to

[344.19, 369.12, 24.93, 'bei3']

before further processing.

III. METHOD

A. Time Group types

The analysis algorithm represents a function $f : A \rightarrow TG^+$, i.e. a mapping from the annotation representation A to a sequence of at least one TG. The TG output is a tuple of MDDT, TG count, TG variation index, and a two-level sub-tuple, with the following structure:

$$\langle T, TGcount, TGvar, \langle\langle nPVI, int, slope\rangle\ syllseq\rangle\rangle$$

Two possible Time Group types are defined as syllable sequences which decelerate (duration increasing) or accelerate (duration decreasing):

$$TG_{dec} = syll_1, ..., syll_n, \Delta t(syll_i\text{-}1, syll_i) + threshold$$

$$TG_{acc} = syll_1, ..., syll_n, -\Delta t(syll_i\text{-}1, syll_i) + threshold$$

Raw duration differences are not considered, only differences which exceed a specifiable threshold, the MDDT. The variable *threshold* is introduced in order to define and traverse a search space for possible TGs in terms of minimum differences between syllable durations: different MDDTs not only handle ‘duration noise’ fluctuation, but also determine different sizes of TG. The different TG sizes are predicted to relate to different sizes of grammatical constituent. The MDDT is applied locally to neighbouring syllables, and therefore normalises for local absolute duration variability 'noise', so a sequence with many local decelerations may turn out to be globally accelerating, and vice versa.

In addition to information for individual TGs, overall TG variability in the search space for a given MDDT is calculated, as well as mean syllable variability (here: *nPVI*) over the entire set of outputs, separately for deceleration and acceleration. The various options for measuring syllable duration variability were mentioned previously. It is known that empirically there are small differences with the same data between measures such as standard deviation and the *nPVI* [18], and both are used here. Syllable durations are locally normalised with the *nPVI*; global *z*-score normalisation over syllable types is not feasible since few syllables occur sufficiently frequently for this to make sense; segment durations are not considered.

As already noted, metrics like standard deviation and the *nPVI* do not take alternation into account and therefore do not model rhythm. Nevertheless the measures are still valid as measures of overall regularity. Lower values of the variability measures mean greater relative isochrony of syllables, but higher variability values unfortunately have no unambiguous meaning as they can be alternating or not, and are certainly not interpretable as a tendency to foot or stress timing: the meaning is simply 'less isochronous syllables' (or whatever unit is being measured). Larger units require independent measurement of their alternation.

The dynamically defined TG is a potential candidate for higher level rhythm analysis, and the status of speech as having relative syllable timing versus TG timing is modelled as a ratio $nPVI_{syll} : nPVI_{TG}$ (cf. a related approach for Finnish [19], but for syllable-foot relations, not syllable-TG relations). Alternation of strong-weak sequences is given at the syllable level as C-V alternation. Within TGs, alternation holds between distinctly longer and shorter syllables. In TG sequences, alternation of monotonic deceleration and acceleration is the criterion.

B. Implementation

The TGA parsing model was implemented as a web tool in order to make timing analysis in terms of linguistic units as simple, efficient, accessible and collaborative as possible. The implementation of the TG_{dec} and TG_{acc} definitions in a segmentation algorithm which cycles through the specified range of MDDTs is straightforward (ignoring minor technical details):

```
initialise TG syllable list
for each Δthresh level
 select {deceleration | acceleration}
  for each syllable
   if difference + threshold detected
     add syllable to list
    else
     calculate diff index
     emit list w Δthresh info
     start new list
     calculate and emit descr stats
```

The TGA system is implemented in Python as an interactive CGI application with a web form interface. The application takes as inputs Praat format annotations with user-specified syllable tier, and user-selected parameters including MDDT range. The syllable tier annotation information is extracted from the Praat TextGrid file; and Praat attribute-value notation is converted to a plain tuple format <*start*, *end*, *duration*, *label*> for processing.

The output contains basic quantitative information about syllables and syllable sequences, including information on relative syllable isochrony in the entire narrative. General properties of the syllables are given, followed by two large tables, one for the deceleration condition and one for the acceleration condition. Each row represents values for one MDDT range in the representation of the tuple which was introduced previously (cf. Fig. 2).

Direction: deceleration

Threshold	PU count	PU s3t-nPVI	PU lists, format = s3t-nPVI (intercept,slope) sylltext(sylldur)* # (PU with no pause: italic, pause: bold, s3t-nPVI without pauses)
90	33	72	0 (0.00,0.00) **PAUSE**(3441) # 32 (245.60,-2.90) bei3(250) feng1(262) gen1(172) tai4(297) yang2(218) **PAUSE**(678) # 47 (76.50,100.50) you3(110) yi4(110) hui2(311) **PAUSE**(427) # 0 (198.00,44.00) *bei3(198) feng1(242)* # 30 (170.50,3.17) *gen1(151) tai4(223) yang2(184) zai4(113) nar4(188) zheng1(192) lun4(237) shui2(165)* # 62 (81.20,49.70) de5(71) ben3(175) shi5(123) da4(254) **PAUSE**(553) # 32 (138.70,38.70) *zheng1(171) lai2(132) zheng1(210) qu4(274)* # 32 (110.33,32.00) *jiu4(126) shi4(111) fen1(190)* # 51 (82.20,36.60) bu4(55) chu1(155) gao1(185) di1(133) lai2(249) **PAUSE**(571) # 36 (156.83,30.50) *zhe4(179) shi2(143) hou5(240)* # 31 (114.16,12.80) lu4(146) shang5(161) lai2(144) le5(86) ge4(76) zou3(188) daor4(258) de5(213) **PAUSE**(357) # 25 (135.93,9.68) ta1(154) shen1(184) shang5(177) chuan1(164) zhe5(89) jian4(134) hou4(173) da4(201) yi1(296) **PAUSE**(484) # 0 (181.00,0.00) *ta1(181)* # 50 (127.04,16.11) men5(83) lia3(206) jiu4(116) shuo1(200) hao3(241) le5(158) **PAUSE**(549) # 27 (155.08,0.01) *shui2(152) neng2(150) xian1(211) jiao4(170) zhe4(94) ge5(102) zou3(158) daor4(204)* # 37 (110.27,10.40) de5(68) tuo1(193) xia4(134) ta1(120) de5(76) hou4(154) da4(214) yi1(270) **PAUSE**(205) jiu4(181) suan4(198) shui2(177) # 69 (86.90,39.90) de5(64) ben3(187) shi5(115) da4(221) **PAUSE**(835) # 0 (199.00,90.00) *bei3(199) feng1(289)* # 38 (184.16,-7.47) jiu4(181) shi3(217) jinr4(157) de5(75) gua1(201) qi3(164) lai2(153) le5(116) **PAUSE**(489) # 46 (174.00,-6.20) bu2(125) guo4(246) **PAUSE**(159) ta1(182) yue4(120) shi4(91) gua1(187) # 45 (167.38,1.45) de5(77) li4(194) hai5(298) **PAUSE**(239) na4(165) ge5(89) zou3(168) daor4(263) de5(191) **PAUSE**(205) ba3(144) da4(216) yi1(134) guo3(173) de5(120) yue4(162) jin3(269) **PAUSE**(674) # 27 (149.57,13.57) hou4(117) lai2(159) bei3(222) feng1(229) mei2(152) far3(240) le5(213) **PAUSE**(442) # 40 (178.10,19.10) *zhi3(213) hao3(186) jiu4(134) suan4(294)* # 0 (108.00,0.00) le5(108) **PAUSE**(631) # 33 (148.71,4.85) guo4(162) le5(92) yi2(88) huir4(299) **PAUSE**(285) tai4(199) yang2(166) chu1(158) lai5(188) le5(161) **PAUSE**(457) # 42 (191.14,0.28) ta1(280) huo3(195) la4(127) la4(169) de5(123) yi2(108) shai4(342) **PAUSE**(390) # 30 (129.17,21.89) *na4(169) ge5(84) zou3(174) daor4(222) de5(214) ma3(255) shang4(246)* # 15 (126.81,5.30) jiu4(94) ba3(100) na4(150) jian4(155) hou4(162) da4(193) yi1(188) tuo1(199) xia4(137) lai2(166) le5(143) **PAUSE**(585) # 18 (150.70,29.20) *zhe4(172) xiar4(171) bei3(163) feng1(272)* # 17 (143.00,34.50) zhi3(144) hao3(175) cheng2(214) ren4(246) **PAUSE**(406) # 52 (100.40,48.50) *ta1(169) men5(80) lia3(171) dang1(231) zhong1(336)* # 0 (222.00,0.00) *hai2(222)* # 71 (106.50,13.21) shi5(101) tai4(197) yang2(119) de5(51) ben3(182) shi5(115) da4(258) **PAUSE**(1227) # **MEAN s3t-nPVI = 38, for PU with no pause and length =>3 (deceleration)**

Fig. 2. Extract from web tool TGA output ('PU' , is replaced by 'TG' in the present study).

IV. RESULTS

MDDTs turn out to influence the number of TGs which are parsed, i.e. identified. The relation is summarised in Fig. 3 for male speakers under deceleration and acceleration TGs. Zero MDDT means that raw duration differences are used.

The lowest MDDTs which make a difference are around 90ms. As MDDTs are increased, fewer and larger Time Groups are identified. From around 250ms the influence of pauses, which are in general longer than syllables, becomes dominant. Horizontal lines indicate MDDT ranges which generate a given TG parse. Vertical lines indicate steps between different TG parses.

The scale of MDDTs appears to show abrupt changes rather than a smooth continuum: there are conspicuous steps between certain MDDTs. These steps may tend to show a relation of different MDDTs to different sizes of linguistic units (see Section 5).

The relation between speakers varies from being roughly linear to a function describing smaller changes in TG number as MDDTs increase.

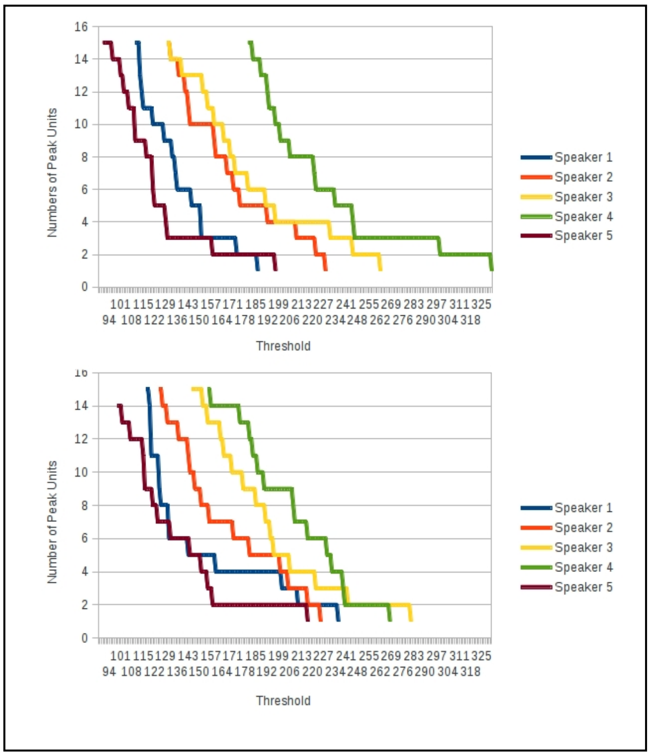

Fig. 3. TG:Duration difference threshold ratios, 5 male Mandarin speakers, read-aloud narrative (note that 'Peak Units' is to be read as 'TGs' here, and above is deceleration model, acceleration below).

V. Discussion

Durations and duration differences have been investigated in many studies. However, the grammatical relevance of MDDTs and the TGs they determine has not been investigated before. The prediction of a relation between MDDTs and grammatical units is not trivial, as it is not obvious that MDDTs may relate to grammatical units.

The relation between TGs parsed at different MDDT levels and linguistic units, including words, phrases, clauses and sentences, was investigated, motivated by the informal observations about stepwise increases in threshold effects. It is clear that, if MDDTs relate to grammatical units at all, then smaller thresholds characterise smaller grammatical units and larger thresholds characterise larger grammatical units, from word size to sentence size and above. This hypothesis was examined.

Two conditions for the definition of TGs were distinguished: (1) TGs were defined as interpausal units, and their deceleration properties; (2) TGs were defined by deceleration properties only. Acceleration was not considered in this part of the study.

The first result is that the MDDT table, Table 1, shows three tendencies: (1) TGs as interpausal units show, as expected, much lower error rates at every threshold level than sequences where deceleration alone is the criterion; (2) nevertheless, TGs based on deceleration alone provide clear indications of timing domains for grammatical units, though not for specific types of grammatical unit; (3) longer TGs relate to fewer and fewer smaller and more and more larger grammatical units. These results are based on the deceleration model.

The data quantities in this investigation do not justify more detailed quantitative evaluation, but the tendencies are clear and motivate further work with this methodology.

A second result, based on initial results from TGs parsed by the acceleration criterion also indicate so far that acceleration TGs tend to be better at defining time domains of smaller units (e.g. words with suffixes, compound words, some phrases), while deceleration TGs with larger MDDTs tend to be better at defining time domains of larger units (phrases, clauses, sentences); this result is not entirely unexpected, except that the units in the present study are not units from a predefined prosodic hierarchy but parsed from syllable durations annotated from the speech signal on the basis of variable MDDTs.

TABLE I. Relation of decelerating TGs with or without final pause to grammatical categories (speaker JiaY, female).

Deceleration (pause-free TGs)				
Thresh	**TGs**	**Valid TGs**	**Syll. Err. rate**[a]	**Associated language unit types**
114	16	13	28%	phrase (3), compound phrase (1), clause (6), sentence (3)
122-123	11	11	36%	phrase (2), clause (5) sentence (4)
124-132	10	10	29%	phrase (2), clause (4) sentence (4)
133-150	9	9	33%	phrase, clause sentence
158	6	6	28%	phrase, clause sentence
189-206	1	1	20%	sentence
Deceleration (TGs with final pause)				
114	19	16	10%	phrase (8), clause (4), sentence (3)
122-123	19	19	6%	phrase (7), clause (5) sentence (3)
124-132	19	16	5%	phrase (6), clause (6) sentence (3)
133-150	19	16	4%	phrase (5), clause (6) sentence (4)
158	19	19	2%	phrase (3), clause (8), sentence (5)
189-206	19	19	0%	phrase (2), clause (10), sentence (6)

[a] Syllable error rate means that a TG has too many/few syllables for an exact match.

A third result is rather a spin-off of this study: the overall *nPVI* for syllables in TGs for the entire narrative is low (the range of the *nPVI* metric is from 0 to an asymptote of 200): the syllable *nPVI* is 38 in the deceleration condition, 39 in the acceleration condition. This compares with the average *nPVI* of 42 for vocalic spans, taken from the four studies reported in [8], though these values vary rather wildly, reflecting different data genres more than different languages. The standard deviation of duration differences is also small: typically about 12 for both acceleration and deceleration (durations in ms). The results show strong relative syllable isochrony.

VI. Conclusion

Given a syllable segmentation, the TGA method parses syllable sequences into Time Groups based on deceleration or acceleration of syllable durations; when MDDTs, thresholds of minimal difference durations, are adjusted, different thresholds produce TGs of different lengths which tend to relate to grammatical units of different sizes. Applications are envisaged in the evaluation of L2 speaking proficiency and in natural speech synthesis.

In contrast to previous studies, the results show that not only interpausal unit TGs function in this way but also TGs parsed independently of units on the prosodic hierarchy, according to deceleration and acceleration in timing patterns. Whether the deceleration and acceleration relations have any general functional or typological linguistic significance beyond the present Mandarin data remains to be seen in future studies of TG hierarchies.

References

[1] D. Huber, Aspects of the communicative function of voice in text intonation-constancy and variability in Swedish fundamental frequency contours. Dissertation. Institutionen för Svenska Språket, Göteborgs Universitet, 1988.

[2] K. Pike, The Intonation of American English, Ann Arbor: University of Michigan Press, 1945.

[3] I. Lehiste, "Isochrony reconsidered," Journal of Phonetics, vol. 5, 1977, pp. 253–263.

[4] W. Jassem, D. R. Hill and I. H. Witten, "Isochrony in English speech: its statistical validity and linguistic relevance," in D. Gibbon and H. Richter [Eds], Intonation, Accent and Rhythm: Studies in Discourse Phonology, Berlin: Walter de Gruyter, 1984, pp. 203-225.

[5] R.M. Dauer, "Stress-timing and syllable-timing re-analysed," Journal of Phonetics, vol. 11, 1983, pp. 51–62.

[6] F. Ramus, M. Nespor and J. Mehler, "Correlates of linguistic rhythm in the speech signal," Cognition, vol. 72, 1999, pp. 1–28.

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

[8] J. Yu and D. Gibbon, "Criteria for database and tool design for speech timing analysis with special reference to Mandarin," O-COCOSDA 2012, University of Macau, 2012, pp. 41-46.

[9] F. Cummins and R. F. Port, "Rhythmic constraints on stress timing in English," Journal of Phonetics, vol. 26(2), 1998, pp. 145-171.

[10] P. A. Barbosa and S. Madureira, "Toward a hierarchical model of rhythm production: evidence from phrase stress domains in Brazilian Portuguese," in Proc. ICPhS 14, San Francisco, Vol. 1, 1999, pp. 297-300.

[11] U. Gut, "Rhythm in L2 speech," in D. Gibbon, D. Hirst and N. Campbell [eds], Rhythm, melody and harmony in speech. Studies in honour of Wiktor Jassem. Speech and Language Technology 14/15, Polish Phonetics Society, Poznań, Poland, 2012, pp. 83-94.

[12] D. Gibbon, "Time Types and Time Trees: Prosodic mining and alignment of temporally annotated data," S. Sudhoff et al., Methods in Empirical Prosody Research, Berlin: Walter de Gruyter, 2006, pp. 281-209.

[13] Feng, Y. Q., M. Chu, L. He and S. N. Lu, "Statistical analysis of syllable durations in Mandarin," 5th National Conference on Phonetics, Beijing, 2001, pp. 536-541.

[14] J. F. Cao, "Rhythm grouping and speech timing," In *Proceeding of ICSLP'2000*, Beijing, 2000, pp. 357-360.

[15] Q. Y. Tseng and C. Y. Su, "Dynamic discourse speech tempo and phonological timing," ICPhS XVII. Hongkong, 2011, pp. 2026-2029.

[16] A. Li et al., "CASS: A phonetically transcribed corpus of Mandarin spontaneous speech," Proc. Interspeech 2000, Beijing, 2000, pp. 485-488.

[17] P. Boersma, "Praat, a system for doing phonetics by computer," Glot International 5:9/10, 2001, pp. 341-345.

[18] D. Gibbon and F. R. Fernandes, "Annotation mining for rhythm model comparison in temporally annotated data," in Proc. Interspeech 2005, Lisbon, 2005, pp. 3289-3292.

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