# Timing analysis with the help of SPPAS and TGA tools

*Jue Yu*

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

## Abstract

This paper is trying to solve the big problems facing phoneticians and linguists in the study of duration, timing and speech rhythm, that is, heavy manual work during the annotating process, and how to generate more accurate and objective analysis results based on a large speech database. Two newly-developed speech tools are discussed: SPPAS, a tool for automatic phonetic segmentation of speech and TGA, a tool for automatic timing analysis. A case study was carried out to demonstrate that, with suitable models and tools for processing speech corpora, (1) the time required to transcribe speech data can be reduced with the help of SPPAS to about 33% of the manual annotation time, and (2) analysis of speech timing in annotations can be facilitated by using TGA.

**Index Terms**: SPPAS, TGA, Time Group Analysis, Time Trees, Chinese

## 1. Introduction

Speech timing is always a hot issue in phonetics, phonology, psychology and speech engineering. In order to study the relation between speech timing patterns and linguistic structures in Chinese dialects and in Chinese L2 speakers of English, a new approach is taken: Time Trees [1] are constructed from syllable annotations of speech recordings, and correspondences between their smallest constituents and language units are examined. This approach differs from more traditional studies in terms of "speech rhythm" or in terms of duration variation using models of duration difference averages, or in terms of different timing models, from the single-level duration models to multilevel modeling approaches and to studies of the multiple factors underlying durations.

So far, studies in speech timing show that an approach based on large corpora is necessary both for the study of speech production and for speech synthesis with reasonable quality. For example, Dellwo et al. [2] pointed out that rhythm studies really require the analysis of longer sequences of speech data, otherwise artifacts may appear in the results. Sagisaka et al. [3] state that fine control of segmental duration based on a large corpus has been proved to be essential in synthesizing speech with natural rhythm and tempo. Bigi & Hirst [4] came to the more general conclusion that today it is becoming more and more expected for linguists to take into account large quantities of empirical data, often including several hours of recorded speech.

Thanks to technological progress, a number of graphical software tools for creating annotated audio and/or video recordings of speech have become available such as Praat [5], Transcriber [6] and WaveSurfer [7], Anvil [8], Elan [9]. These tools are basically intended for manual annotation. But the present problem is the production of large numbers of annotation and their analysis. Large quantities of data require many hours of manual work, which is time-consuming (and can be really frustrating) and therefore imposes a severe restriction on the amount of data which can be used. A better solution for this problem is to borrow methods from speech engineering and use an automatic time-aligned phonetic transcription tool [10] [11]. The second problem is how to analyze speech timing using the large quantities of annotated data. For this purpose a tool designed for automatic timing analysis [12] is available.

This present paper is concerned with these 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, thus presenting a relatively efficient process for reducing the time required to transcribe speech data and for speech timing analysis.

## 2. Tools used in speech timing

### 2.1. SPPAS: automatic phonetic segmentation

SPPAS, Speech Phonetization Alignment and Syllabification, is a tool designed by Laboratorire Parole et Langage, Aix-en-Provence, France, to automatically produce annotations which include utterance, word, syllable and phoneme segmentations and their transcriptions from recorded speech. Currently it is implemented for four languages: French, English, Mandarin Chinese and Italian. It is said that adding other languages requires a a very simple procedure [4] [11].

SPPAS has a) a phonetician-friendly interface; b) a high rate of correct alignment (correct phoneme alignment rate: 88%; correct word alignment rate: 97.6%) [4]; c) generation of files in the TextGrid format, which can be easily analyzed in detail with the widely used Praat software workbench [3]; d) free web-based software and e) constant improvement [4].

SPPAS generates six TextGrid outputs, four of which are relevant here: (i) utterance segmentation, (ii) word segmentation, (iii) syllable segmentation and (iv) phoneme segmentation. In Chinese, simple words are monosyllabic, so (ii) and (iii) are the same units, but (ii) is orthographic and (iii) is phonetic. The output is illustrated as a screenshot in Figure 1, with TextGrid files merged in Praat.

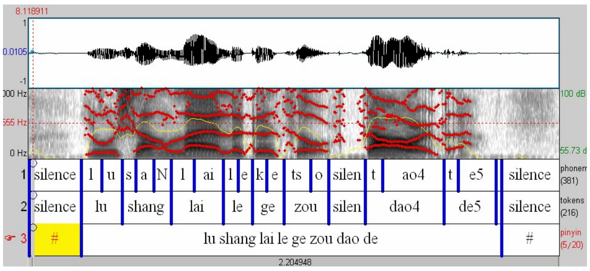
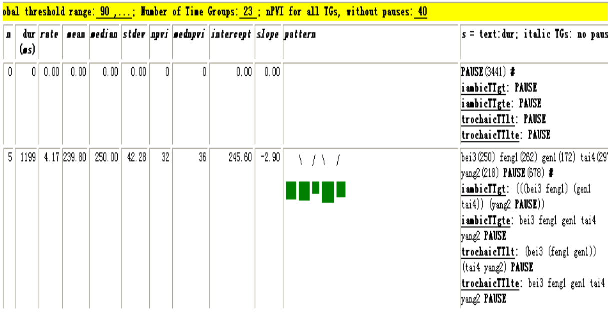

8.118911
1
0.0105
-1
100 Hz
555 Hz
0 Hz
100 dB
55.73 d
1 silence l u s a N l ai l e k e ts o silen t ao4 t e5 silence
phonemes (381)
2 silence lu shang lai le ge zou silen dao4 de5 silence
tokens (216)
3 # lu shang lai le ge zou dao de #
pinyin (5/20)
2.204948

Figure 1: *SPPAS output example for the Mandarin Chinese utterance “lu4 shang5 lai2 le5 ge4 zou3 daor4 de5” in Pinyin (On the street came a traveller ).*

## 2.2. TGA: automatic speech timing analysis

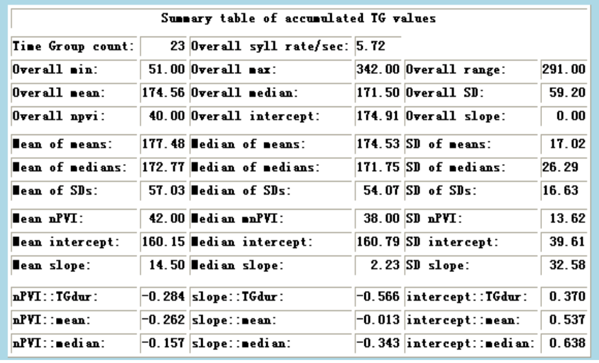
TGA, Time Group Analyzer [13] [14], is a tool for the automatic parsing of syllable sequences in speech annotations into Time Groups (TG), that is, inter-pausal groups, or into units based on deceleration models (consistent slowing down) or acceleration models (consistent speeding up). It captures not only the global timing patterns from the input speech annotation in TextGrid format, but also analyzes local patterns based on different duration thresholds (minimal duration difference between adjacent syllables). Most importantly for the present study, it also directly generates ‘Time Trees’ [1] using the local pattern, which are then available for analyzing the relationship between timing properties of the phonetic realizations and the underlying language categories.

TGA is also a web-based tool intended to facilitate the analysis work of phoneticians and linguists. TGA has been applied mainly to Mandarin and Hangzhou Chinese (a dialect, which is very different from Mandarin but shares the same typological structure) and English.


obal threshold range: 90 ,...; Number of Time Groups: 23 ; nPVI for all TGs, without pauses: 40

| n | dur (ms) | rate | mean | median | stdev | npvi | mednpvi | intercept | slope | pattern | s = text:dur; italic TGs: no paus |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.00 | 0.00 | 0.00 | 0.00 | 0 | 0 | 0.00 | 0.00 | | PAUSE (3441) #<br>iambicTTgt: PAUSE<br>iambicTTgte: PAUSE<br>trochaicTTlt: PAUSE<br>trochaicTTlte: PAUSE |
| 5 | 1199 | 4.17 | 239.80 | 250.00 | 42.28 | 32 | 36 | 245.60 | -2.90 | \ / \ / | bei3 (250) feng1 (262) gen1 (172) tai4 (29 yang2 (218) PAUSE (678) #<br>iambicTTgt: (((bei3 feng1) (gen1 tai4)) (yang2 PAUSE))<br>iambicTTgte: bei3 feng1 gen1 tai4 yang2 PAUSE<br>trochaicTTlt: (bei3 (feng1 gen1)) (tai4 yang2) PAUSE<br>trochaicTTlte: bei3 feng1 gen1 tai4 yang2 PAUSE |

Figure 2: *Local patterns of TGs for the Mandarin utterance“bei3 feng1 gen1 tai4 yang2”in Pinyin (the north wind and the sun).*


Summary table of accumulated TG values

| | | | | | |
|---|---|---|---|---|---|
| Time Group count: | 23 | Overall syll rate/sec: | 5.72 | | |
| Overall min: | 51.00 | Overall max: | 342.00 | Overall range: | 291.00 |
| Overall mean: | 174.56 | Overall median: | 171.50 | Overall SD: | 59.20 |
| Overall npvi: | 40.00 | Overall intercept: | 174.91 | Overall slope: | 0.00 |
| Mean of means: | 177.48 | Median of means: | 174.53 | SD of means: | 17.02 |
| Mean of medians: | 172.77 | Median of medians: | 171.75 | SD of medians: | 26.29 |
| Mean of SDs: | 57.03 | Median of SDs: | 54.07 | SD of SDs: | 16.63 |
| Mean nPVI: | 42.00 | Median mnPVI: | 38.00 | SD nPVI: | 13.62 |
| Mean intercept: | 160.15 | Median intercept: | 160.79 | SD intercept: | 39.61 |
| Mean slope: | 14.50 | Median slope: | 2.23 | SD slope: | 32.58 |
| nPVI::TGdur: | -0.284 | slope::TGdur: | -0.566 | intercept::TGdur: | 0.370 |
| nPVI::mean: | -0.262 | slope::mean: | -0.013 | intercept::mean: | 0.537 |
| nPVI::median: | -0.157 | slope::median: | -0.343 | intercept::median: | 0.638 |

*Figure 3: Quantative information of TGA output example for the whole Mandarin IPA text “the north wind and the sun”.*

The results generated by TGA are very informative, including (1) the corresponding text for each TG; (2) threshold information, (3) global and local timing patterns; (4) many kinds of quantitative information for each TG and for all TGs, such as nPVI, speech rate, slope etc. An output sample of local timing patterns of TGs and relative quantitive information are illustrated as a Screenshot in Figures 2 and 3.

# 3. A case study in Hangzhou-accented Mandarin and Standard Mandarin

## 3.1. Objective: empirical evaluation

The objective is to apply these tools, SSPAS and TGA, together, in order to test whether the tools can facilitate the whole annotating process and whether the combination can generate more satisfactory results than manual annotation and analysis.

## 3.2. Data

For the study, recordings of six speakers reading the same material, a well-known coherent story (the classic Aesop fable and IPA standard text. ‘The North Wind and the Sun’), were used, in a Mandarin translation. 3 subjects are from Hangzhou and 3 are native Beijing Mandarin speakers. The data is from the CASS corpus [15].

## 3.3. Procedure

The sound file in WAV format is opened in SPPAS with a transcription of the prompt text, and an IPU (Inter-Pausal Unit) segmented TextGrid file is created. The file is then opened and the segmentation breaks in the input text file are corrected manually, if necessary. Then the prompt text is corrected if for example there are any repeated words or if the speaker does not read exactly the text as it was written. The text is corrected so that it corresponds to what the speaker actually says. Then, the other functions are applied, Phonetization and Alignment, and a merged TextGrid file is generated. Finally the output is checked manually.

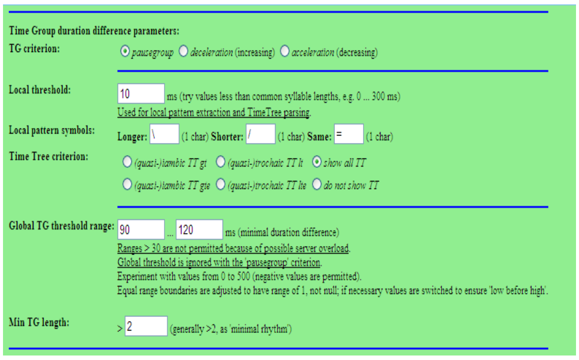
In order to quantify the efficiency gained by applying the above procedure, 3 recordings of Hangzhou speakers were transcribed following the procedure described above, and the other 3 recordings of Mandarin speakers were totally manually worked without the aid of SSPAS. The result shows that the procedure using SPPAS reduces annotation time to 33% of the time required for wholely manual annotation.

Once the annotated TextGrid file is ready, timing analysis with the help of TGA is performed. This tool is designed to handle interval tiers. The procedure is: using the web interface, input the TextGrid file, choose the target interval tier, set analysis options. The analysis is performed automatically. There are options for global and local timing patterns.

For the local patterns, values less than common interval lengths can be tried, while for the global patterns, based on deceleration and acceleration models, a wider range of thresholds can be used. The variable threshold is introduced in order to define and traverse search space for possible TGs in terms of minimum differences between syllable durations: different thresholds are relevant for different sizes of TG. A previous study of deceleration and acceleration relations [15] has shown that there are conspicuous steps between certain thresholds, possibly indicating a ‘quantum leap’ between different sizes of linguistic units relating to timing unit sizes.

Finally, local Time Trees can be generated for the TGs, using local quasi-iambic (deceleration) or quasi-trochaic (acceleration) conditions (Figure 4).

An advantage of the TGA tool is that, rather than measuring timing properties of 'a priori' linguistic units, such as 'foot' or 'phrase', or focusing primarily on rhythm, the tool applies an inductive procedure for the automatic parsing of a hierarchy of interval sequences, and then the generalised results can be compared in an independent step with linguistic units. Further quantitative properties of these interval sequences are also available, in particular variation and 'evenness' of interval durations in the sequences.



**Time Group duration difference parameters:**

**TG criterion:** ● *pausegroup* ○ *deceleration* (increasing) ○ *acceleration* (decreasing)

**Local threshold:** 10 ms (try values less than common syllable lengths, e.g. 0 ... 300 ms)
Used for local pattern extraction and TimeTree parsing.

**Local pattern symbols:** **Longer:** \ (1 char) **Shorter:** / (1 char) **Same:** = (1 char)

**Time Tree criterion:** ○ *(quasi-)iambic TT gt* ○ *(quasi-)trochaic TT lt* ● *show all TT*
○ *(quasi-)iambic TT gte* ○ *(quasi-)trochaic TT lte* ○ *do not show TT*

**Global TG threshold range:** 90 ... 120 ms (minimal duration difference)
Ranges > 30 are not permitted because of possible server overload.
Global threshold is ignored with the 'pausegroup' criterion.
Experiment with values from 0 to 500 (negative values are permitted).
Equal range boundaries are adjusted to have range of 1, not null; if necessary values are switched to ensure 'low before high'.

**Min TG length:** > 2 (generally >2, as 'minimal rhythm')

Figure 4: *Option interface for TG duration difference parameters in TGA.*

In the present study, the pause group condition was used. The global threshold is only required for determining deceleration and acceleration, and is ignored with the pause group criterion. In this study the local pattern and Time Trees are in the focus. Only the quasi-iambic Time Tree model is taken into account because initial inspection showed closer relationships for this model than for the quasi-trochaic model. Relations between Time Tree constituents and multisyllable words were investigated and the percentage of agreement in each TG was calculated. The following example shows a quasi-iambic Time Tree (using brackets, not a graph) of the Mandarin utterance "*zhe4 shi1hou5, lu4 shang5 lai2 le5 ge4 zou3 daor4 de5*" (*at that time, on the street came a traveller*), and a grammatical bracketing of the utterance:

quasi-iambic Time Tree:
(((zhe4 (shi2 hou5)) (((lu4 shang5) (lai2 (le5 (ge4 zou3)))) daor4)) (de5 **PAUSE**))
Grammatical bracketing:
((zhe (shi hou)), (lu shang) ((lai) (le) (ge) (zou daor de)))

The groups (shi2 hou5) and (lu4 shang5) correspond to words; (ge4 zou3) is not a grammatical constituent. Also, factoring out the effect of the pause, (lai2 le5 ge4 zou3 daor4 de5) corresponds to a grammatical constituent.

### 3.4. Results on speech timing

The results illustrated in Figure 5 show:

1. Iambic groups with pause excluded are more meaningful than those with pause included.
2. The graphs in Figure 5 show that the results in the percentage of correlation of the Time Tree components with multisyllable words in each TG for all the speakers are very similar until about 50ms. This is approximately the length of the shortest syllables. It is only when the threshold gets longer that interesting results start to appear. It seems that HZ-2 and HZ-3 speaker have better Mandarin-like timing, which corresponds to the evaluation results in Mandarin speech proficiency of the 3 Hangzhou speakers.
3. After 50ms, the percentage of correlation of the Time Tree components with multisyllable words increases rapidly and steadily until the increase stops at a certain threshold, varying with different speakers.
4. Quantitative properties show that the difference in speech rate between Hangzhou speakers and Mandarin speakers is not significant (F(1, 5) =0.04, p>0.05) and doesn't correlate with Mandarin speech proficiency ($R^2$=0.028, p>0.05). Neither do the nPVIs of syllables durations between the two (F(1, 5) =0.444, p>0.05 ).

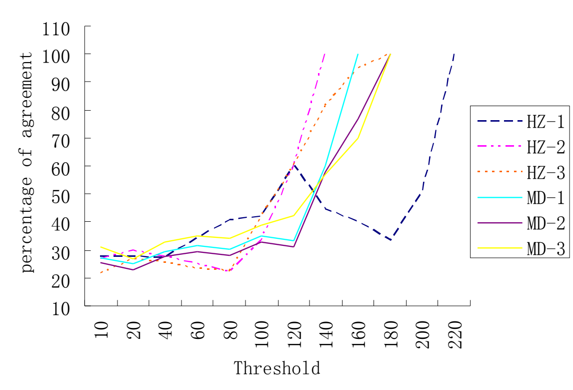


Figure 5: *Comparison of local iambic patterns in pause groups between Hangzhou (HZ) and Mandarin (MD) speakers.*

## 4. Conclusion

The study showed that SPPAS, an automatic annotation tool, applied in the procedures described in this paper, can reduce overall transcription time by about 33%. TGA, a tool for automatic timing parsing of interval sequences in speech annotations can be used, just as in the above case study, to investigate the relations between Time Tree constituents and multisyllable words, with comparison of quantitative properties; thus to distinguish native and non-native speech (though the data itself is not large enough). It can also facilitate research into areas such as cross-linguistic phonetic studies, into heuristics for using grammar-speech relations in speech technology, and into the provision of timing criteria for the evaluation of L2 speech proficiency. Both tools together allow phoneticians and linguists to work with larger corpora and to spend more of their time on analysis and less on manual tasks involved in transcription and calculation.

## 5. References

[1] Gibbon, D., "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. 281-209, 2006.

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

[3] Sagisaka, Y., Kato, H., Tsuzaki, M. and Nakamura, S., "Speech timing and cross-linguistic studies towards computational human modeling", In Proc. Oriental COCOSDA 2009, 1-8, Beijing, 2009.

[4] Bigi B., and Hirst D., "Speech phonetization alignment and syllabification (SPPAS): a tool for the automatic analysis of speech prosody", In Proc. of Speech Prosody 2012, Shanghai, 2012.

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

[6] Barras, C., Geoffrois, E., Wu, Z. and Liberman. M. "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", First International Conference on Language Resources and Evaluation (LREC), 1373-1376, May 1998, and http://transag.sourceforge.net/, 2011.

[7] Sjölander, K., Beskow, J., "Wavesurfer - an open source speech tool", in ICSLP/Interspeech, Beijing, China, October 16-20, 464-467, ISCA, 2000, http://www.speech.kth.se/wavesurfer/.

[8] Kipp, M., "Anvil, DFKI, German Research Center for Artificial Intelligence", http://www.anvil-software.de/, 2011.

[9] Sloetjes, H. and Wittenburg, P., "Annotation by category - ELAN and ISO DCR", LREC 6, 2008, http://www.latmpi.eu/tools/elan/.

[10] Serridge, B., and Castro, L., "Faster time-aligned phonetic transcriptions through partial automation", In Proceedings of ISCA Tutorial and Research Workshop on Experimental Linguistics, 189-192, Atenas, 2008.

[11] Bigi, B., "SPPAS: a tool for the phonetic segmentations of speech", LREC 8, 1748-1754, Istanbul, 2012.

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

[13] Gibbon, D., "TGA. A tool for automatic speech timing analysis", [Computer Software], Bielefeld: G. Dafydd, Universität Bielefeld, http://wwwhomes.uni-bielefeld.de/gibbon/tga-3.0.html, 2012.

[14] Gibbon, D. "TGA: a web tool for Time Group Analysis". Proc. Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop, Aix-en-Provence, forthcoming, August 2013.

[15] Li A., Zheng, F., Byrne, W., Fung, P., Kamm, T., Liu, Y., Song, Z., Ruhi, U., Venkataramani, V. and Chen, X., "CASS: A phonetically transcribed corpus of Mandarin spontaneous speech", in Proc. Interspeech 2000, 485-488, Beijing 2000.