Annotation Pro + TGA: automation of speech timing analysis

Katarzyna Klessa[1], Dafydd Gibbon[2]
[1]The Institute of Linguistics, Adam Mickiewicz University in Poznań,
[2]Fakultät für Linguistik und Literaturwissenschaft, Bielefeld University
[1]Poznań, Poland, [2]Bielefeld, Germany
E-mail: [1]klessa@amu.edu.pl, [2]gibbon@uni-bielefeld.de

Abstract

This paper reports on two tools for the automatic statistical analysis of selected properties of speech timing on the basis of speech annotation files. The tools, one online (*TGA*, Time Group Analyser) and one offline (*Annotation Pro+TGA*), are intended to support the rapid analysis of speech timing data without the need to create specific scripts or spreadsheet functions for this purpose. The software calculates, inter alia, *mean*, *median*, *rPVI*, *nPVI*, *slope* and *intercept* functions within interpausal groups, provides visualisations of timing patterns, as well as correlations between these, and parses interpausal groups into hierarchies based on duration relations. Although many studies, especially in speech technology, use computational means, enquiries have shown that a large number of phoneticians and phonetics students do not have script creation skills and therefore use traditional copy+spreadsheet techniques, which are slow, preclude the analysis of large data sets, and are prone to inconsistencies. The present tools have been tested in a number of studies on English, Mandarin and Polish, and are introduced here with reference to results from these studies.

Keywords: Annotation mining, Time Group Analysis, TGA, Annotation Pro+TGA, online tools, speech timing, interpausal group

1. Introduction: speech timing

Except for work on oscillator models (Barbosa, 2009, Inden et al., 2012) and in speech technology, which analyse the speech signal itself, timing properties of speech units such as phones, syllables and words or feet are often investigated indirectly in phonetics by analysing the time-stamps in speech annotations, made either automatically or manually with speech workbench tools such as *Praat* (Boersma, 2001), *WaveSurfer* (Sjölander, Beskow, 2000), *Elan* (Sloetjes, Wittenburg, 2008). The annotation analysis itself is often done either automatically with Praat, Unix shell, Perl, or Python scripts (e.g. Buschmeier et al. 2013, Gibbon 2013).

However, enquiries among colleagues have revealed that many do not have the computing skills or cooperations needed for this, and therefore laboriously copy time-stamps from the screen or from annotation files into a spreadsheet, then defining and adapting functions to analyse timing. This method is slow, permits only relatively small data sets, and is prone to inconsistency.

The present paper has two goals: first, to report on two facilitating tools, one online, one offline, for supporting researchers and students who are in this position, by replacing crucial manual annotation analysis steps by automated annotation mining of large annotated data sets, and, second, to demonstrate use of the tools in case studies on timing in English and Polish. Additionally, we will illustrate the tools' performance by referring to selected examples of the results obtained with these tools in other studies on timing in English, Polish and Mandarin Chinese.

Typical quantitative models which have been investigated by the copy+spreadsheet method are *PIM* (*Pairwise Irregularity Model*; Scott et al., 1996), *PFD* (*Pairwise Foot Deviation*; Roach, 1982), and *nPVI* (*normalised Pairwise Variability Index*; Low et al., 2001). A few tools and pre-prepared spreadsheets for calculating the values of some of these models are available at a number of places on the Internet. However, these appear to be more suitable for the analysis of individual examples and for teaching contexts than for the analysis of larger quantities of data. We propose a more efficient automatic approach to the analysis and modelling of speech timing in order to create comprehensive statistical analyses and visualisations for significant quantities of annotated speech data, and report on two tools which address this issue, *Annotation Pro* (Klessa et al., 2013a) and *TGA* (*Time Group Analyser*) (Gibbon, 2013), and on an integration strategy for these tools.

A comprehensive computational tool has the advantage of being more efficient in terms of user time and effort and faster on large data sets, and (after appropriate evaluation) of being more consistent and less prone to human error. The integrated *Annotation Pro + TGA* tool incorporates some *TGA* features and is intended to support the development of more robust and versatile timing models for a greater variety of data.

The structure of the present paper is as follows: Section 2 describes *Annotation Pro* and *TGA*. Section 3 reports on the integration of *TGA* functions into *Annotation Pro* and provides examples of use. Section 3 provides a conclusion and outline of ongoing work on modelling duration in large speech corpora and on comparing models for different corpora.

2. Custom speech timing tools

2.1. Annotation Pro

Annotation Pro is a new interactive annotation workbench[1] for:

1. multi-layer annotation of spoken and written (eg. morphological glossing) resources;
2. conducting perceptual tests.

An innovative feature, compared with other annotation tools, is the user-definable graphical representation of the feature space (see an example workspace in Figure 1). The idea underlying the design and implementation of the graphical component was to include the option of annotations using both categorial and dimensional rating scales (cf. also *Feeltrace* tool for the analysis of emotional speech using a two-dimensional space graphic representation developed by Cowie et al. (2000)).

The current implementation of the tool is in C# under Windows; porting to other operating systems is planned. The tool is freely available for research purposes.

Two further design requirements were flexibility for different use cases, and format compatibility.

The flexibility requirement is intended to enable extension and adaptation of the tool to project-specific needs, and for this purpose a plugin technology was designed and implemented, in which the plugins, also currently in C#, are compiled and launched at runtime. The plugin option makes it possible for users to add new options to the menu and to extend or automatize the built-in functionalities. Any plugin can be run either for a single annotation file or any desired number of files included in the program's workspace (cf. Karpiński et al. (2014) for a description and use cases of two *Annotation Pro* plugins calculating moving average of speech rate and *nPVI* as a means to investigate interlocutors' convergence in the temporal domain).

The format compatibility requirement covers a range of import/export options to and from other tools, the ones implemented so far include for example: Praat (Boersma, 2001), WaveSurfer (Sjölander, Beskow, 2000), Transcriber (Barras et al., 2001), as well as plain text or CSV database formats.

The perceptual test options enable the use of the annotation interface as a testing environment: it is for example possible to store the annotators' personal information and actions to a special output file, to blind the file names in the workspace, to set the number of allowed re-plays of speech signals within the workspace etc. The results of the perception tests can then saved as annotation tiers and used as a part of the multi-tier annotations, thus facilitating further uses of the perception data together with other types of available data (see eg. Wagner, 2012, for an example of a production and perception study of emotional speech in Polish)

The *Annotation Pro* tool is currently used in annotation projects based on Polish language resources for the analysis of timing phenomena in speech (see more detailed examples of use cases in Section 3 below), for speech technology as well as for the needs of language and speech documentation resources for minority languages spoken (at present or in the past) in the region of Poland (e.g. Halcnovian, Latgalian or Polish Yiddish, cf. Klessa, Wicherkiewicz, 2014). Advantages often

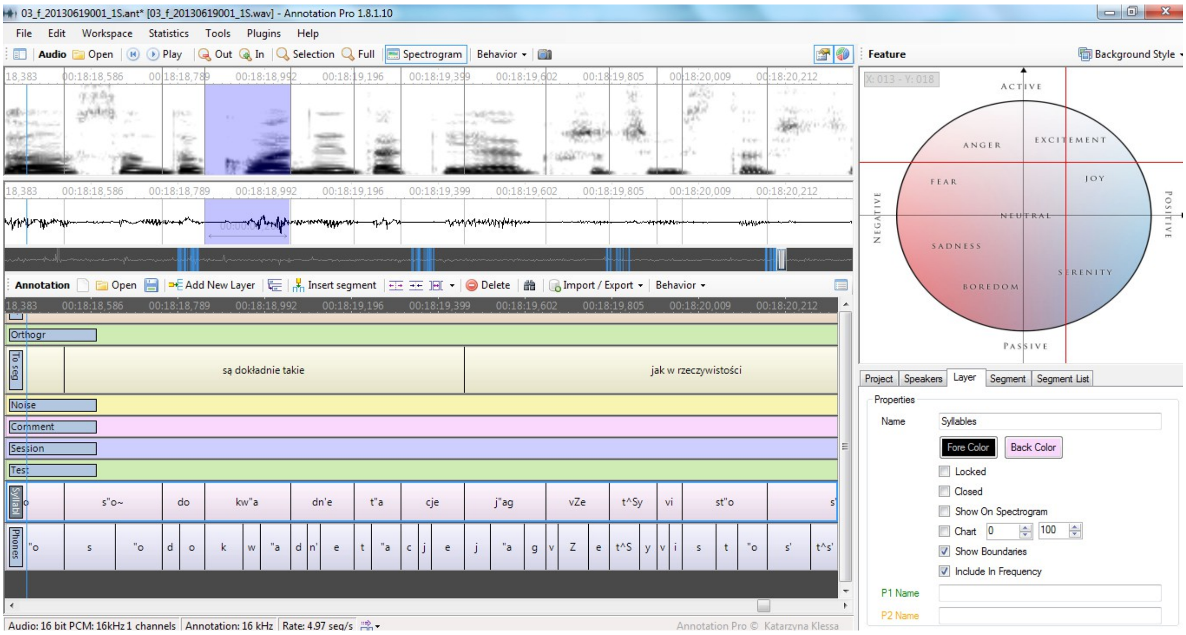

Figure 1: Annotation Pro - multilayer annotation workspace with a graphical representation of the feature space (here: activation-valence space for annotation of emotions, top-right corner).

1 http://annotationpro.org/download/

reported by the users are a high level of flexibility and user-friendliness of the interface.

2.2. TGA

The *TGA* is an online batch processing tool[2] which provides a parametrised mapping from time-stamps in speech annotation files in various formats to a detailed analysis report with statistics and visualisations. A comprehensive facility of this sort was not previously available.

The tool, which is implemented in Python using standard libraries and server-side CGI interaction, contains the following modules which are transparent to the user:

1. I/O and format import;
2. text extraction;
3. global basic descriptive statistics for all elements of the specified tier;
4. segmentation of the tier into interpausal *Time Groups* with
   a) statistics for individual *Time Groups*, and
   b) three new visualisation techniques for duration patterns: duration difference tokens, duration column charts, and *Time Trees*.

So far, the tool has mainly been used in analyses of Mandarin and English (Yu & al. 2013; Yu, 2013; Gibbon, 2013), with significant gains in time and consistency over the traditional copy+spreadsheet approaches.

A sample of a statistical report for all interpausal units in a file, from a case study in syllable duration alternation (Gibbon et al., 2013) is shown in Figure 2.

REPORT FOR GLOBAL AND ACCUMULATED INTERPAUSAL DURATION RELATIONS

Time Group criterion: pausegroup, local threshold: 10, Min valid TG length: 2
Only inter-pause intervals measured; pauses not included

Overall duration:	48504	Overall raw longer, ms:	15401	Overall raw shorter, ms:	14521
Overall min:	20.00	Overall max:	990.00	Overall range:	970.00
Valid Time Groups:	34	Overall rate/sec:	5.67		
Components: global tendencies					
Overall mean:	176.38	Overall median:	150.00	Overall SD:	113.58
Overall npvi:	62.00	Overall intercept:	156.12	Overall slope:	0.15
Mean of means:	182.18	Median of means:	176.70	SD of means:	34.75
Mean of medians:	168.68	Median of medians:	160.00	SD of medians:	40.88
Mean of SDs:	90.02	Median of SDs:	86.16	SD of SDs:	39.87
Mean of nPVIs:	60.00	Median of mnPVIs:	51.00	SD of nPVIs:	17.91
Mean of intercepts:	143.59	Median of intercepts:	130.80	SD of intercepts:	71.16
Mean of slopes:	10.65	Median of slopes:	11.86	SD of slopes:	41.10
Components: correlations					
mean::TGdur:	-0.190	median::TGdur:	-0.427	SD::TGdur:	0.230
nPVI::TGdur:	0.097	slope::TGdur:	0.061	intercept::TGdur:	-0.178
nPVI::mean:	0.128	slope::mean:	0.028	intercept::mean:	0.503
nPVI::median:	0.026	slope::median:	0.005	intercept::median:	0.310
nPVI::SD:	0.383	slope::SD:	0.051	intercept::SD:	0.229

Figure 2: Screenshot of summary of collated Time Group properties and correlations. Data: Aix-MARSEC corpus (Hirst et al., 2009), A series files.

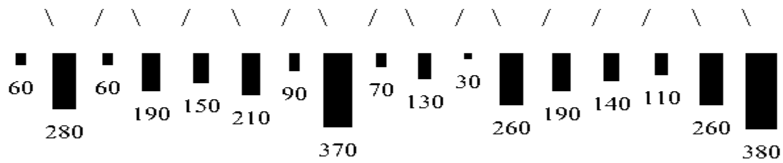

Figure 3: Top, duration difference token pattern. Bottom, top-suspended duration columns, duration scaled for width and length (bottom) ; Time Group from Aix-MARSEC data set (Hirst et al., 2009).

Detailed reports for individual Time Groups are also generated, with details of the *mean*, *median*, *standard deviation*, *nPVI*, *PIM*, *PFD*, as well as *slope* (representing acceleration and deceleration) and *intercept* linear regression values. Linear regression is selected mainly for the *slope* function, as a first approximation to examining acceleration and deceleration over large data sets.

Each individual Time Group is also provided with three visualisations of the Time Group structure:

2 http://wwwhomes.uni-bielefeld.de/gibbon/TGA/

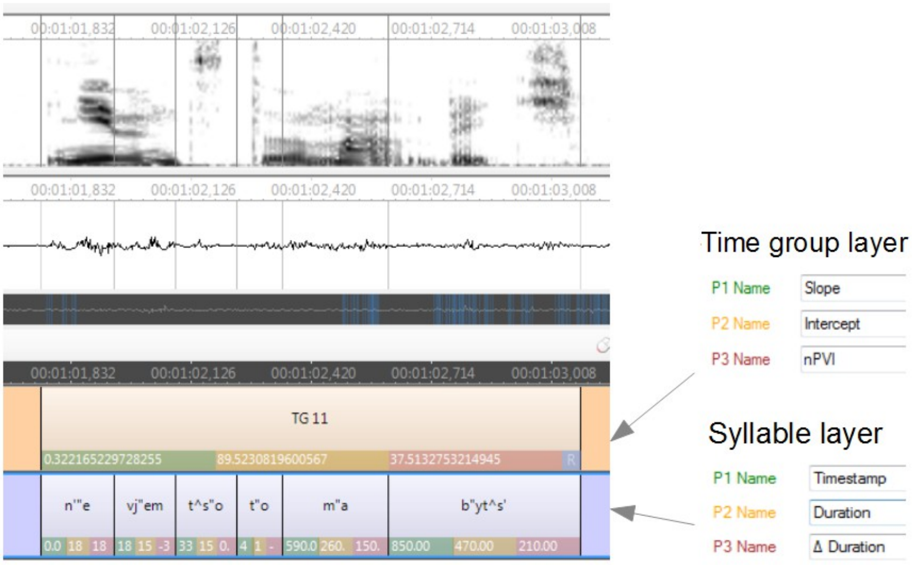

Figure 4: Example *TGA* results as displayed in the *Annotation Pro* interface for a Polish utterance: *Nie wiem co to ma być*. The obtained values are visible as segment parameters (P1, P2, P3) with coloured background at the bottom of each segment (a time group segment or a syllable segment respectively).

1. a visualisation of Δ-duration (duration difference) patterns as token icons (with '/', '=' and '\' for increasing, equal and decreasing duration, respectively);
2. a chart of Δ-duration values;
3. a visualisation of the duration relations as a parse tree graph whfich groups shorter and longer durations into a hierarchy of timing groups (Yu 2013).

The duration and Δ-duration visualisations provide a convenient immediate 'eyeball impression' of the rhythmical vs. irregular timing properties of interpausal groups as a guide to further analysis strategies.

An extensive case study (Yu, J., et al. 2014) on English syllable duration alternation revealed that Δ-duration below a threshold of approximately 50msec is unimportant for identifying rhythmic properties of utterances. This result means that raw duration data, such as that currently used in models like the *nPVI*, are likely to contain spurious Δ-duration values.

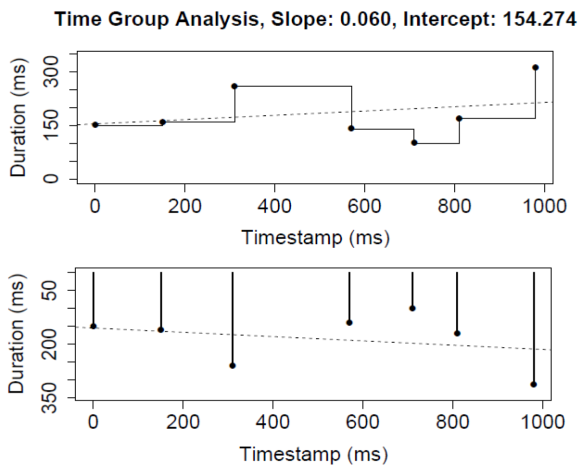

Figure 5: TGA results for a single Time Group with increasing (deceleration) duration tendencies shown as a duration differences chart (top) and a top-suspended bar chart (bottom).

3. TGA features in Annotation Pro

In order to make it possible to conduct Time Group Analysis directly in *Annotation Pro*, the *Annotation Pro* plugin facility was used to create a new plugin based on *TGA* functions. Additionally, the same functions were integrated natively. A further decision was made in order to enhance interactivity: *TGA* provides batch processing of an entire annotation tier, while two options have been implemented for the *Annotation Pro* interactive mode:

a) analysis of only the segment sequences selected by the user,
b) analysis of the entire tier (as in the on-line *TGA*).

The first *TGA* functions chosen for implementation and evaluation with *Annotation Pro* were the linear regression functions for duration patterns and also *nPVI*, based on the segmentation unit used in particular annotations (in the examples discussed below in this paper we use syllable as a basis but technically any other type of segment could be used depending on the level of segmentation). Other regression functions are planned.

The linear regression function yields local and overall values of *slope*, which expresses an approximation to tempo acceleration and deceleration, and *intercept*, together with relevant visualisations.

The values are calculated using syllable time-stamps as values for the *X* variable and syllable durations for the *Y* variable, and the results are stored in newly created annotation layers as the parameters of the relevant time group or syllable segments (as shown in Figure 4)

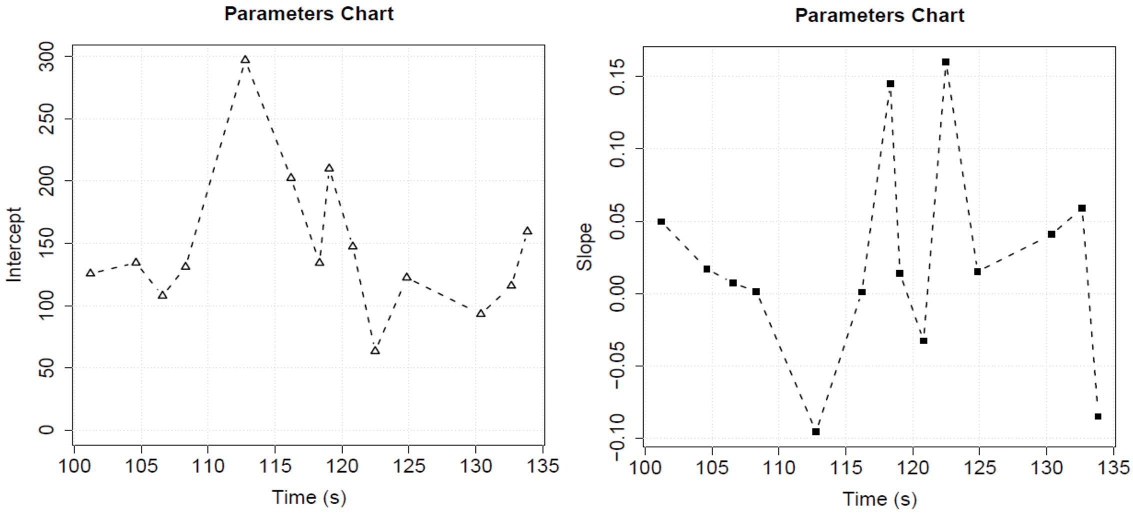

Figure 6: Example TGA results shown as a point chart of the variability of *intercept* (left) and of *slope* (right) for 14 subsequent phrases in a Polish dialogue utterance.

and may be plotted using an *R* (RDevelopment Core, 2013) script embedded in *Annotation Pro*.

Three main types of results visualisation have been implemented so far. The first two are shown in Figure 5: a Δ-duration chart (*x*: syllable duration time-stamps; *y*: syllable duration), and a top-suspended chart (*x*: syllable duration time-stamps, *y*: inverted axis of syllable duration, following the approach used in the original *TGA* tool). The graph renderings in Figure 5 differ superficially from the original *TGA* visualisations, but represent the same information.

The third graph type is based on the values of segment parameters calculated for Time Groups. The values of two *TGA* parameters were selected as an example which is especially useful for plotting: duration difference *slope*, expressing tempo acceleration and deceleration, and/or *intercept* (Figure 6). Apart from these two parameter charts, *Annotation Pro* enables plotting charts based on any other parameters available for segments in the annotation layers. For the *TGA* component, apart from the *slope* and *intercept* charts it is also possible to quickly generate plots of unit durations, Δ-durations, and the *nPVI*.

The *R* scripts used for plotting can be copied or modified either in an internal editor inside the *Annotation Pro* tool or directly in *R*. This way, the user can adjust the appearance and contents of the charts. The results (together with any desired data present in other annotation layer: transcriptions, segmental or suprasegmental annotations etc.) can also be easily exported to a CSV database format and subjected to further analysis as desired, either by custom scripts, in an external spreadsheet, with statistical software.

Figure 7 shows the values of *slope* for time groups (interpausal groups) produced by eight Polish speakers (4 females, 4 males) reading the same text in two target speaking rates: fast and normal rate (*Paralingua* data, Klessa et al., 2013b). It can be seen from this example data visualisation that *slope* variability is significantly higher at normal speaking rate (larger minimum-maximum differences overall and between particular time groups for particular speakers). The observation is a higher number of time groups realized in the normal rate for most speakers (6-13 in fast rate, and 7-18 in the normal rate).

Another example of using the *Annotation Pro* + *TGA* is an earlier study where the framework was applied to compare timing patterns in Polish speaking styles: read speech and dialogue recordings by investigating the *slope*, *intercept*, *nPVI* values and speaking rates. The results showed that the feature that distinguishes the styles most significantly was related to the *slope* values (Gibbon et al., 2013).

The *nPVI* formula (Low et al., 2001) has been implemented as one of the options in the *TGA* module in *Annotation Pro*. The base unit substituted in the *nPVI* formula depends on the segmentation level used in annotations, here we used syllable-based units (similarly as with the regression functions). As an example, Table 1 shows the means and standard deviations (SD) for *nPVI* and for the *slope* values for spoken and read utterances produced by 20 Polish speakers in read and conversational speech (Paralingua Dial sub-corpus, two reading tasks, one dialogue task, Klessa et al., 2013b). Altogether, 1185 interpausal time groups were analysed (675 conversational, 510 read utterances).

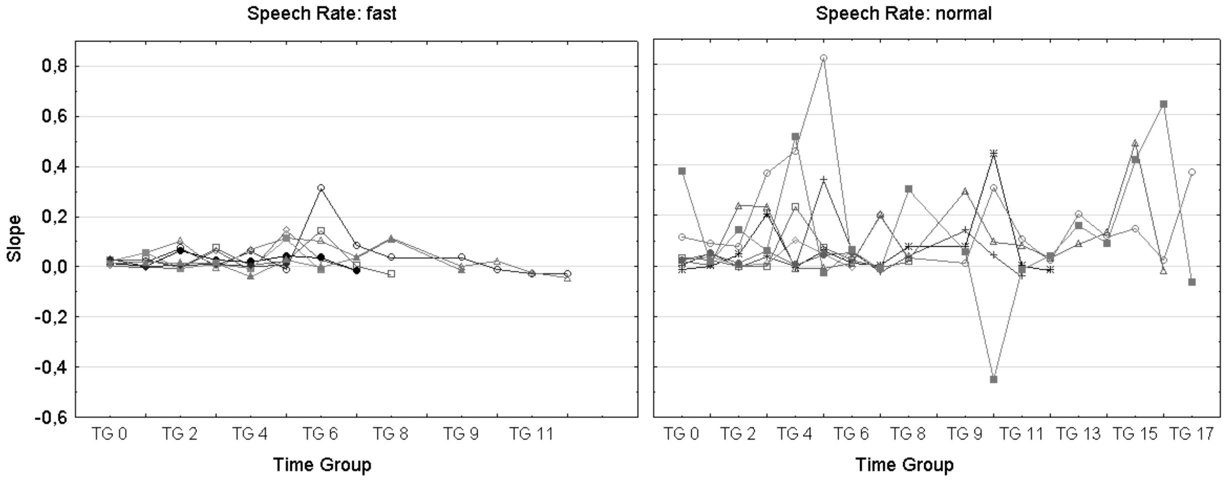

Figure 7: *Slope* variability in Polish read speech for 8 speakers in two tempi. Each line chart represents *slope* variability for interpausal groups realized by one speaker (Data: Paralingua, Emo sub-corpus; Klessa et al., 2013b).

As it can be seen, the mean values of *slope* differ clearly for the two speaking styles while for the *nPVI* the relative difference is also present but less significant (some speakers tended to produce almost exactly the same mean *nPVI* for the two types of speech). When looking at mean SDs, both measures exhibit significantly higher values in the dialogues as compared to read speech, especially the mean SD of slopes in conversational utterances shows great dispersion from the mean value, which might be explained by the high variation of speech tempi in dialogues, as well as by the specificity of guidelines accepted during annotation and segmentation of the data into interpausal time groups (the segmental content of the so-called 'filled pauses' or 'hesitation markers' was not treated as a pause, cf. Yu et al., 2014).

Measure/Task		**Dialogue**	**Reading**
nPVI	*Mean*	48,25	43,15
	Mean SD	19,66	9,19
Slope	*Mean*	0,23	0,08
	Mean SD	0,53	0,13

Table 1: Syllable *nPVI* and *slope* values in conversational and read Polish speech by 20 speakers (Data: Paralingua, Dial sub-corpus, Klessa et al., 2013b).

4. Conclusion and outlook

The TGA online tool was designed to support phoneticians in basic statistical analysis of annotated speech data. In practice, the tool provides not only rapid analyses but also the ability to handle larger data sets than can be handled manually.

The integration of *TGA* statistical and visualisation functions into *Annotation Pro+TGA* results in a powerful computational enhancement of the existing *Annotation Pro* phonetic workbench, for supporting experimental analysis and modelling of speech timing. Integrated interactive statistical analysis and the dimension of duration data pattern visualisation constitute a new magnitude of speech corpus analysis resource capability for the general phonetic user with few programming skills.

Another benefit of the integration is relatively easy inspection of the potential correlations or dependencies of timing information on other features observed and annotated for the analysed speech signals. Using the information available in the remaining time-aligned layers in *Annotation Pro* is especially interesting when the graphical representation of the feature space is used for annotation of non-categorial features of speech (e.g. paralinguistic features in some cases difficult to annotate with categorial scales only). Combining the integrated duration information with perception-based annotation of continuous speech features within one flexible workspace is expected to provide new insights into the relations between timing and other phenomena in speech.

Currently work is in progress on applying *TGA* and *Annotation Pro* to large corpora of Polish and English data, and on implementing automatic inter-corpus comparison methods for the outputs of the *TGA* functions. Work is ongoing on another languages. These results will be reported on, together with a further functional evaluation in practical annotation environments as a future work.

In addition to applications in phonetic duration modelling, the statistical and visualisation functions of the *Annotation Pro + TGA* tool are used to provide additional support for the development of corpora for unit selection and other corpus-based methods of speech synthesis, as well as for the development of heuristics for providing temporal prosodic information for automatic speech recognition systems, particularly for enhancing the phrasing capabilities of language models.

We predict that analysis and visualisation facilities of this kind will become increasingly important for phonetic and technological work with very large annotated speech corpora.

5. References

Barbosa, P. (2009). Measuring speech rhythm variation in an oscillator-based framework. In *Proceedings of Interspeech 2009*, Brighton, UK, pp. 1527-1530.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2001). Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Communication,* 33(1–2), pp. 5–22.

Buschmeier, H., Włodarczak, M. (2013). TextGridTools: A TextGrid Processing and Analysis Toolkit for Python. *Proceedings of 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV)*, pp. 152-157.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M. (2000). FEELTRACE: An instrument for recording perceived emotion in real time. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, Newcastle.

Gibbon, D. (2006). Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In Sudhoff, S. et al. (Eds.), *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter., pp. 281-209.

Gibbon, D. (2013). TGA: a web tool for Time Group Analysis. In D. Hirst & B. Bigi (Eds.) *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, pp. 66-69.

Gibbon, D., Klessa, K., and Bachan, J. (2013). Duration and speed in speech events. To appear in: Mikołajczak-Matyja, N., Karpiński, M. (Eds.), *Studies in Phonetics and Psycholinguistics,* Poznań.

Hirst, D. Auran, C. and Bouzon, C. (2009). The Aix-MARSEC database, 2002-2004. *Tech. Report, Equipe Prosodie et Représentation Formelle du Langage*, Laboratoire CNRS UMR 6057 Parole et Langage, Université de Provence, Aix-en-Provence.

Inden, B., Malisz, Z., Wagner, P., & Wachsmuth, I. (2012). Rapid entrainment to spontaneous speech: A comparison of oscillator models. In Miyake, N., Peebles, D., & Cooper, R. P., (Eds.), *Proceedings of 34th Annual Conference of the Cognitive Science Society*, Austin, TX: Cognitive Science Society.

Karpiński, M., Klessa, K., Czoska, A. (2013). Local and global alignment in the temporal domain in Polish task-oriented dialogue. To appear in: *Proceedings of 7th Speech Prosody Conference*, Dublin, 20-23 May 2014.

Klessa, K., Karpiński, M., Wagner, A. (2013a). Annotation Pro - a new software tool for annotation of linguistic and paralinguistic features. In D. Hirst & B. Bigi (Eds.) *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, pp. 51-54.

Klessa, K., Wagner, A., Oleśkowicz-Popiel, M., Karpiński, M. (2013b). "Paralingua" – a new speech corpus for the studies of paralinguistic features. In Vargas-Sierra, Ch. (Ed.), *Corpus Resources for Descriptive and Applied Studies. Current Challenges and Future Directions: Selected Papers from the 5th Int. Conf. on Corpus Linguistics, Procedia – Social and Behavioral Science*, Vol. 95, pp. 48-58.

Klessa, K., Wicherkiewicz, T. (2014). Design and Implementation of an On-line Database for Endangered Languages: Multilingual Legacy of Poland. To appear in: *Proceedings of 6th International Conference on Corpus Linguistics (CILC 6*), Las Palmas de Gran Canaria, 22-24 May 2014.

Low, E.L., Grabe, E., Nolan, F. (2001). Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English". *Language and Speech* 43 (4), pp. 377-401.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal, (Ed.), *Linguistic Controversies: Essays in Linguistic Theory and Practic*. London: Edward Arnold, pp. 73-79.

Scott, D. R., Isard, S. D. and de Boysson-Bardies, B. (1986). On the measurment of rhythmic irregularity: a reply to Benguerel. *Journal of Phonetics* 14, pp. 327–330.

Sjölander, K., Beskow, J. (2000). WaveSurfer – an Open Source Speech Tool. In *Proc. of 6th ICSLP Conference 2000*, Vol. 4, Beijing, pp. 464–467.

Sloetjes, H., Wittenburg, P. (2008). Annotation by category – ELAN and ISO DCR. In *Proceedings of 6th International Conference on Language Resources and Evaluation, LREC 2008*, Marrakech, pp. 816-820, on-line: http://www.lat-mpi.eu/tools/elan/

Team, RDevelopment Core (2013). R: A language and environment for statistical computing. ISBN 3-900051-07-0. R Foundation for Statistical Computing. Vienna, Austria.

Wagner, A. (2012). Emotional speech production and perception: A framework of analysis. *Speech and Language Technology*, vol. 14/15 (2011/12), PTFon, Poznań, pp. 163-183.

Yu, J. (2013). Timing analysis with the help of SPPAS and TGA tools. *Proceedings of the Tools and Resources for the Analysis of Speech Prosody (TRASP) Workshop*, Aix en Provence, pp. 70-73.

Yu, J., Gibbon, D. (2012). Criteria for database and tool design for speech timing analysis with special reference to Mandarin. In *Proceedings of Oriental COCOSDA 2012 (IEEEexplore Conf ID 21048)*, pp. 41-46.

Yu, J., Gibbon, D., Klessa, K. (2014). Computational annotation-mining of syllable durations in speech varieties. To appear in: *Proceedings of 7th Speech Prosody Conference*, 20-23 May 2014, Dublin.