Praat on the Web: An Upgrade of Praat for Semi-Automatic Speech Annotation

Mónica Domínguez, Iván Latorre,
Mireia Farrús, Joan Codina-Filbà
Universitat Pompeu Fabra,
C. Roc Boronat, 138
08018 Barcelona, Spain
monica.dominguez|ivan.latorre|
mireia.farrus|joan.codina@upf.edu

Leo Wanner
ICREA and Universitat Pompeu Fabra,
Barcelona, Spain
leo.wanner@upf.edu

Abstract

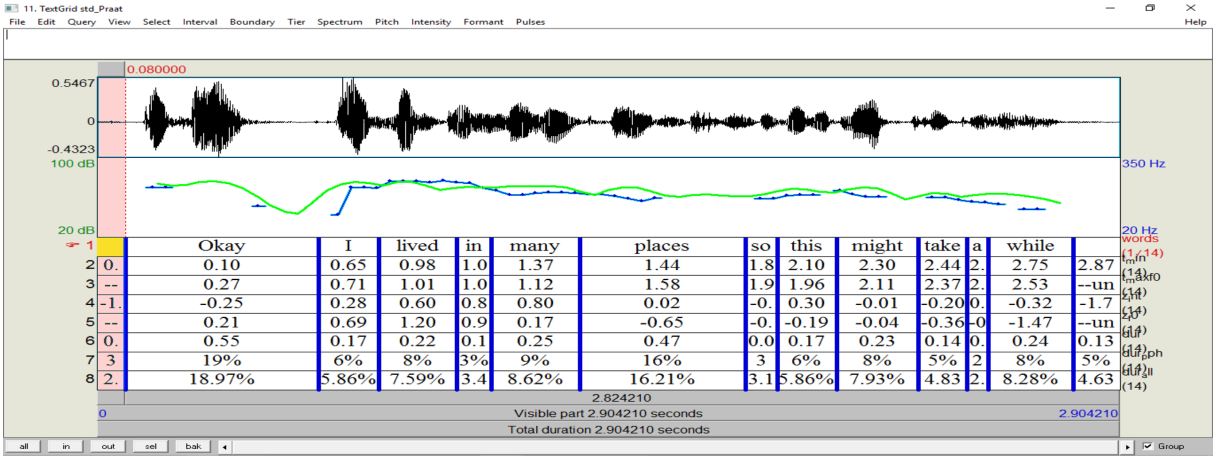
This paper presents an implementation of the widely used speech analysis tool Praat as a web application with an extended functionality for feature annotation. In particular, Praat on the Web addresses some of the central limitations of the original Praat tool and provides (i) enhanced visualization of annotations in a dedicated window for feature annotation at interval and point segments, (ii) a dynamic scripting composition exemplified with a modular prosody tagger, and (iii) portability and an operational web interface. Speech annotation tools with such a functionality are key for exploring large corpora and designing modular pipelines.

1 Motivation and Background

Automatic annotation of speech often involves dealing with linguistic and acoustic information that needs to be conveniently organized at different levels of segmentation (i.e., phonemes, syllables, words, phrases, sentences, etc.). Even though laboratory experiments on speech are controlled to a certain extent (e.g., minimal word pairs, short sentences, read speech) and are usually annotated manually, the increasing trend to analyze spontaneous speech, especially in human-machine interaction, requires tools to facilitate semi-automatic annotation tasks with a compact visualization for manual revision, presentation of results and versatile scripting capabilities.

The Praat software (Boersma, 2001) is one of the most widely used open-source tools for audio signal processing and annotation in the speech community. Praat has a dedicated text format called *TextGrid*, where stackable lines, called *tiers*, are mapped to the whole time-stamp of the associated sound file (cf. Figure 1). Accordingly, tiers account for the temporal nature of speech and take one compulsory parameter: the time-stamp of the *segments*, which are the smallest unit in a TextGrid. A time-stamp can be of two kinds: an interval (specifying the beginning and end time of each segment) or a point in time. This sequence of time-stamps is encoded in tiers as consecutive segments. Once (interval or point) segments are marked, they can take an optional string parameter, called *label*.

While suitable for a coarse-grained glance at the acoustic profile of speech, Praat shows two major limitations when it comes to more detailed annotation that also involves linguistic information. Firstly, Praat's segment annotations are opaque blocks of strings, and there is no function for a linguistic analysis of the labels. For instance, if an interval segment for the word *places* (as in the example shown in Figure 1) includes morphological information within the same label (e.g., "places: noun = plural"), there is no function in Praat that would allow the division of the string *places: noun = plural* into tokens of any kind, for example, *places — noun — plural* . Secondly, Praat is not modular, i.e., all automatic routines a user is interested in (e.g., detection of silent and voiced parts, annotation of intensity peaks and valleys, computing relative values, etc.) must be programmed together in a single script. No user need-driven composition of stand-alone off-the-shelf scripts for dedicated subroutines is possible, which implies that for any new constellation of the subroutines a new script must be programmed.



1		Okay	I	lived	in	many	places	so	this	might	take	a	while		words (1/14)
2	0.	0.10	0.65	0.98	1.0	1.37	1.44	1.8	2.10	2.30	2.44	2.	2.75	2.87	t_{min} (14)
3	--	0.27	0.71	1.01	1.0	1.12	1.58	1.9	1.96	2.11	2.37	2.	2.53	--un	t_{maxf0} (14)
4	-1.	-0.25	0.28	0.60	0.8	0.80	0.02	-0.	0.30	-0.01	-0.20	0.	-0.32	-1.7	z_{int} (14)
5	--	0.21	0.69	1.20	0.9	0.17	-0.65	-0.	-0.19	-0.04	-0.36	-0	-1.47	--un	z_{f0} (14)
6	0.	0.55	0.17	0.22	0.1	0.25	0.47	0.0	0.17	0.23	0.14	0.	0.24	0.13	dur (14)
7	3	19%	6%	8%	3%	9%	16%	3	6%	8%	5%	2	8%	5%	dur_{ph} (14)
8	2.	18.97%	5.86%	7.59%	3.4	8.62%	16.21%	3.1	5.86%	7.93%	4.83	2.	8.28%	4.63	dur_{all} (14)

2.824210

0 Visible part 2.904210 seconds 2.904210

Total duration 2.904210 seconds

Figure 1: Standard Praat visualization:Annotation using tiers.

In order to remedy these limitations, advanced users have found workarounds. Thus, the first limitation is remedied by either extracting information to an external file, as ProsodyPro (Xu, 2013) does, or by annotating in parallel tiers with cloned time segments and different labels, as shown in Figure 1. To circumvent the second limitation, experienced users tend to program in external platforms and call Praat for performing specific speech processing routines. For example, Praaline (Christodoulides, 2014) extracts acoustic information from Praat for analysis in the R statistic package (R Core Team, 2013) and visualization in the Sonic visualizer (Cannam et al., 2010). However, these workarounds make the use of Praat cumbersome.

The Praat on the Web tool presented in this paper aims to address the aforementioned Praat limitations. More precisely, it upgrades Praat along the lines observed in state-of-the-art natural language processing (NLP) annotation interfaces as encountered for SEMAFOR[1] (Tsatsaronis et al., 2012), Brat[2] (Stenetorp et al., 2012), or GATE[3] (Cunningham et al., 2011). Such an upgrade is instrumental for prosody studies, among other, which are described as a combination of features (not only acoustic, but also linguistic) and therefore benefit greatly from a versatile semi-automatic approach to annotation and a compact visualization of those features.

Praat on the Web involves three main technical aspects: (i) a multidimensional feature vector within segment labels (see Figure 2 for illustration), (ii) a web-based implementation, and (iii) an operational interface for modular script composition exemplified as a prosody tagger. Given that many Praat scripts are freely available and shared in the speech community for different specialized tasks, one of the advantages of modular scripting within the same platform is keeping a library of scripts for easy replacement of independent subtasks within a larger pipeline. The dynamic composition approach presented in this paper, thus, promotes tests on how different configurations affect the final output of the architecture, and positively impacts reproducibility of experiments in a user-friendly web environment.

Praat on the Web is available for extended feature annotation, but compatible with the original Praat format, as a web application[4] and as a local version;[5] source code and all scripts mentioned in this paper as well as a tutorial are available in a Github account.[6] and distributed under a GNU General Public Licence.[7]

[1] http://www.cs.cmu.edu/ ark/SEMAFOR/
[2] http://brat.nlplab.org/
[3] https://gate.ac.uk/
[4] http://kristina.taln.upf.edu/praatweb/
[5] implemented for Praat v.6.0.11
[6] https://github.com/monikaUPF
[7] http://www.gnu.org/licenses/

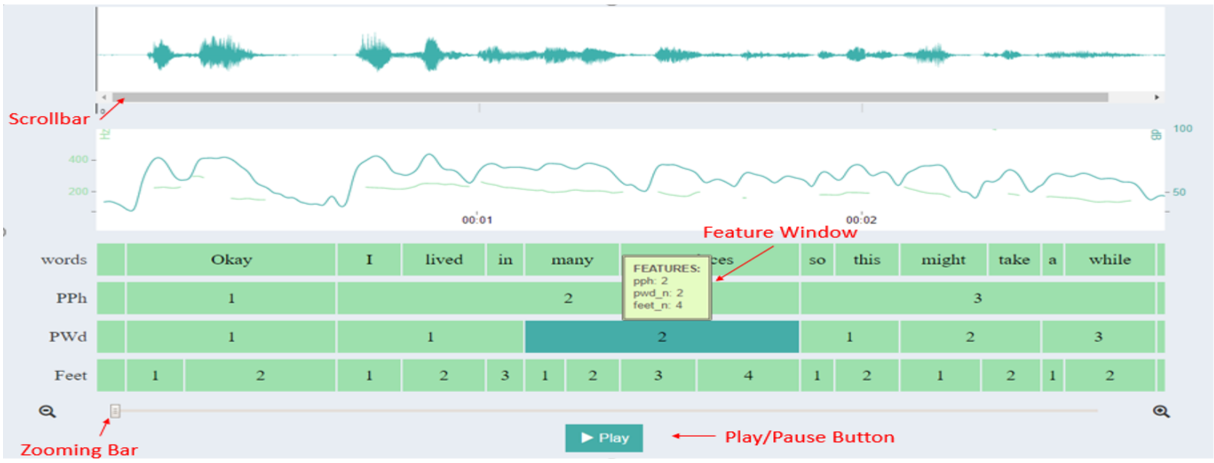


Figure 2: Praat on the Web's visual enhancement of the standard Praat.

2 Annotating in parallel tiers versus using features

Annotations in tiers are convenient for studying nested elements in the speech signal. For example, Selkirk (1984) proposes a hierarchical structure of intonation where smaller units (e.g., prosodic feet) are embedded in larger ones (e.g., prosodic words and prosodic phrases), as Figure 2 shows. However, if each layer needs to be annotated in stacked tiers with cloned times as previously shown in Figure 1, a long collection of repeated tiers for each new layer information blurs visual presentation and makes manual revision tasks harder.

Praat on the Web's main menu on our webpage includes a first demo (accessible by clicking on the button "Enter Demo 1"), where the user can upload their own audio and TextGrid files for visualization and playback. Sample files with feature annotations, which can serve as inspiration or examples, are also provided in the demo. Waveform, fundamental frequency (F0) and intensity curves are displayed on the screen together with the annotated tiers. There are some practical differences with respect to the standard Praat, which are summarized in Table 1. Whereas standard Praat uses keyboard commands to perform actions during annotation such as zooming and playback, Praat on the Web has dedicated buttons for these actions, as illustrated in Figure 2.

Action	Standard Praat	Praat on Web
Zooming	keyboard shortcuts (ctrl+i/o/n)	sliding bar signaled with amplifying glass symbol
Audio playback	shift button or segment + time bar click	play/pause button or segment + waveform click
Scroll left/right	scrollbar below TextGrid	scrollbar below waveform

Table 1: Comparison: actions in standard Prat and Praat on Web.

Further demonstration of visualization capabilities using automatic scripts for merging tiers and splitting features (Demos 3 and 4 respectively) are also available in the online demo webpage. Users can upload their own cloned TextGrids entering Demo 3 and click on the 'run' button to automatically annotate selected cloned tiers as features. In Demo 4, this action is reversed, i.e., feature vectors are converted to cloned tiers. All TextGrids generated in Praat on the Web are displayed in the browser and can also be downloaded for local use clicking on the "Download" button.

3 Dynamic Scripting Composition

Entering Demo 2 through the main menu of Praat on the Web, an example of dynamic scripting composition can be run on available samples or uploaded files. The configuration of the automatic prosody tagger[8] appears in the right part of the screen (see Figures 3 and 4). The pipeline varies depending on

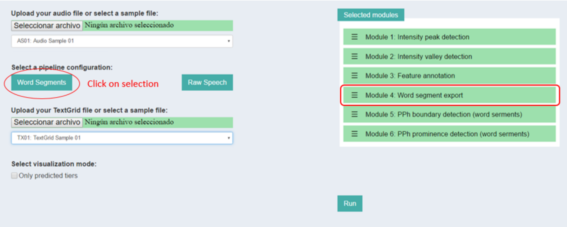
[8] Further information on the prosody tagger' methodology, technical specifications and evaluation is provided in Domínguez et al. (2016).

the selected configuration.

The prosody tagger is made up of a total of eight modules, three of which (from Module 1 to 3) are common for the two possible configurations:

1. Word segments (see Figure 3): when clicking on this button, six modules will appear in the "Selected modules" box. Modules 5 and 6 predict boundaries and prominence respectively on both acoustic information annotated in Modules 1 to 3 and word segments exported by Module 4. A TextGrid with the word alignment needs to be provided to run this configuration.

2. Raw speech (see Figure 4): when clicking on this button, five modules will appear in the "Selected modules" box. Prediction is performed on acoustic information and thus, Module 4 is not in the pipeline and alternative Modules 5 and 6 are chosen for this pipeline.



Upload your audio file or select a sample file:
Seleccionar archivo Ningún archivo seleccionado
AS01: Audio Sample 01

Select a pipeline configuration:
Word Segments   Click on selection   Raw Speech

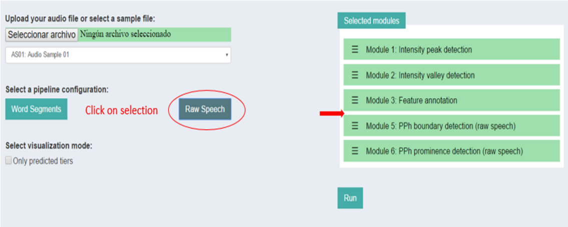
Upload your TextGrid file or select a sample file:
Seleccionar archivo Ningún archivo seleccionado
TX01: TextGrid Sample 01

Select visualization mode:
□ Only predicted tiers

Selected modules
☰ Module 1: Intensity peak detection
☰ Module 2: Intensity valley detection
☰ Module 3: Feature annotation
☰ Module 4: Word segment export
☰ Module 5: PPh boundary detection (word serments)
☰ Module 6: PPh prominence detection (word serments)

Run

Figure 3: Configuration with word segments.



Upload your audio file or select a sample file:
Seleccionar archivo Ningún archivo seleccionado
AS01: Audio Sample 01

Select a pipeline configuration:
Word Segments   Click on selection   Raw Speech

Select visualization mode:
□ Only predicted tiers

Selected modules
☰ Module 1: Intensity peak detection
☰ Module 2: Intensity valley detection
☰ Module 3: Feature annotation
☰ Module 5: PPh boundary detection (raw speech)
☰ Module 6: PPh prominence detection (raw speech)

Run

Figure 4: Configuration for raw speech.

The users can select in the web interface the output of the prosody tagger by ticking the option "only predicted tiers" displayed at the bottom left side of the screen. If that option is not ticked, all tiers generated by each module are shown. The output of the tagger (including annotated features of each segment) is displayed on screen in the browser; it can also be downloaded in TextGrid format for local use.

A further add-on of Praat on the Web is that includes a centralized repository of scripts and data. The action of selecting modules for the sample prosody tagger has been scripted in this demonstration to be automatically done, and the web interface allows moving around modules to prove that modules are also manually interchangeable.

4 Conclusions

We have presented the tool Praat on the Web, which aims to take speech annotations to meet the increasingly demanding requirements in the field of speech technologies. In such a scenario, user-friendly semi-automatic annotation tools within one versatile common platform are key to make steady progress in the study of complex events, like prosody, over large amounts of data. Praat on the Web shows several advantages over standard Praat in that it offers: (i) intuitive visualization of segment annotations using features displayed in a dedicated window; (ii) easy modularity of computational tasks within the same Praat platform; (iii) ready-to-use web environment with no pre-installation requirements for presentation of results. The two first characteristics are achieved including functionality for feature annotation. Consequently, the smallest unit in a Praat TextGrid is no longer an opaque string label, but a well-structured linguistic unit containing a *head*, a *feature name* and a *feature value*.

At the time of publication, Praat on the Web runs with sample or uploaded files for visualization, playback and automatic prediction of PPh boundaries and prominence. In the future, user account management will be introduced for researchers to upload their scripts and create their own pipeline configurations. The web interface is well-suited for annotation and demos (like this one) and teaching purposes; we also plan to extend it with online edition of manual annotations.

Praat on the Web is a first step in the transformation of speech annotation tools to meet the standards already set in other branches of computational linguistics. A move in this direction is especially needed for integrative research and reproducibility that require user-friendly tools for designing automatic processes with enhanced visualization capabilities.

Acknowledgements

This work is part of the KRISTINA project, which has received funding from the *European Unions Horizon 2020 Research and Innovation Programme* under the Grant Agreement number H2020-RIA-645012. It has been also partly supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). The third author is partially funded by the Spanish Ministry of Economy and Competitivity through the *Juan de la Cierva* program.

References

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

C. Cannam, C. Landone, and M. Sandler. 2010. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the ACM Multimedia 2010 International Conference*, pages 1467–1468, Firenze, Italy, October.

G. Christodoulides. 2014. Praaline: Integrating tools for speech corpus research. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*, Reykjavik, Iceland.

H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, v Petrak, Y. Li, and W. Peters. 2011. *Text Processing with GATE (Version 6).*

M. Domínguez, M. Farrús, and L. Wanner. 2016. An automatic prosody tagger for spontaneous speech. In *Proceedings of COLING*, Osaka, Japan.

R Core Team, 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

E. O. Selkirk. 1984. *Phonology and Syntax: The relation between sound and structure*. The MIT Press, Cambridge, Massachussetts.

P. Stenetorp, S. Pyysalo, G. Topić, T. Ohta, S. Ananiadou, and J. Tsujii. 2012. Brat: A web-based Tool for NLP-assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 102–107, Stroudsburg, PA, USA. Association for Computational Linguistics.

G. Tsatsaronis, I. Varlamis, and K. Nørvåg. 2012. Semafor: Semantic document indexing using semantic forests. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, CIKM '12, pages 1692–1696, New York, NY, USA. ACM.

Y. Xu. 2013. Prosodypro a tool for large-scale systematic prosody analysis. In *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP)*, pages 7–10, Aix-en-Provence, France.