INTERSPEECH 2017: Show & Tell Contribution
August 20–24, 2017, Stockholm, Sweden

Prosograph: A Tool for Prosody Visualisation of Large Speech Corpora

Alp Öktem[1], *Mireia Farrús*[1], *Leo Wanner*[1,2]

[1]Universitat Pompeu Fabra, Spain
[2]Catalan Institute for Research and Advanced Studies (ICREA), Spain
{alp.oktem, mireia.farrus, leo.wanner}@upf.edu

Abstract

This paper presents an open-source tool that has been developed to visualize a speech corpus with its transcript and prosodic features aligned at word level. In particular, the tool is aimed at providing a simple and clear way to visualize prosodic patterns along large segments of speech corpora, and can be applied in any research that involves prosody analysis.
Index Terms: prosody, visualization tool, speech corpora

1. Introduction

Prosody conveys several communication elements such as meaning, intention, and emotions, among others. Being able to clearly visualize the different elements involved in prosody —intonation, rhythm, and stress— may be helpful for computational prosody research. Several speech analysis tools (e.g. Praat [1]), together with derived scripts and tools [2, 3, 4] partially cover these needs by helping to visualize quantifiable speech features like fundamental frequency ($f0$) and intensity contours, word stress marking, or prosodic labeling. These tools work well when showing detailed analyses on data and visualizing one single utterance at a time, but fail in visualizing generalized word-averaged speech features of many utterances, i.e., a discourse, at once.

Inspired by music scores and piano rolls, we developed *Prosograph*, a software that resembles a digital musical analysis tool [5]. Prosograph takes the whole speech corpus and shows all the utterances aligned with their corresponding prosodic features, making it easy to navigate through the corpus. Since that prosodic patterns can be easily observed and compared, this application can be used in many areas of research involving prosody, such as language learning and acquisition, comparative studies in different languages, tone languages, audiovisual prosody, etc. Also, it can aid the feature selection process in various speech related applications such as punctuation detection, speech synthesis, recognition and understanding, prosody transfer in speech-to-speech translation.

In what follows, we outline how the tool works and give examples of visualizations of different prosodic characteristics. The tool is publicly available in github[1] under the GNU General Public License[2].

2. Methodology

Prosograph is written in Python mode of Processing[3] because of its simplified access to graphical elements. In order to simulate music scores, the speech prosodic features are plotted in the vertical axis over a temporal horizontal axis. Words are put in order together with pauses and punctuation, and the prosodic features are drawn under each corresponding word. An overview of the tool can be seen in Figure 1.

2.1. Prosodic data

As input, Prosograph takes a set of prosodic features, which can be extracted by using any prosody extraction toolkit (e.g. [2, 6, 7]). For the examples in the current paper, we used [7] to extract the following prosodic features from TED Talks[4]: pauses and word durations, f0 and intensity contours —normalized over talks in semitones distance so that zero values correspond to mean values—, and their corresponding aggregate statistics: mean, standard deviation, maximum, minimum, median, and slope.

Prosograph reads prosodically annotated speech data from a list stored in a pickle file, where every list item holds the multidimensional data of one utterance as a dictionary. Dictionary keys define the name of the data (e.g. pause, duration, stress etc.) and dictionary values hold the sequences of these data. A sample dataset is included in the provided source code.

2.2. Data types in Prosograph

Prosodic features differ in the way they encode words or sentences. For instance, word stress is a feature that represents saliency among a group of words, intonation and intensity are continuous encodings throughout successive voiced phonemes, accent is a peak that occurs at a certain point in a word, etc. In order to visualize different types of prosodic encodings, we have compiled the following data sequence types, which are aligned with the word sequence in the utterance.

pause-value holds the silence duration in milliseconds coming after the corresponding word. Pauses are visualized as empty intervals with a width proportional to their length.

duration-value holds the duration of the corresponding word. If set to be visualized, bounding boxes of words are drawn with a width proportional to their duration.

punctuation-value holds the id's of punctuation marks coming after the corresponding word. Punctuation marks are placed in the same axis with words. If a punctuation mark coincides with a pause, then it is placed inside the pause interval.

binary-value holds a binary value determining if the corresponding word carries a certain feature (1) or not (0). This feature type can be used e.g. for word-stress. Bounding boxes of these words are drawn with a salient color.

point-features holds a real numbered value that belongs to the corresponding word (e.g. standard f0 deviation, mean f0,

[1]https://github.com/TalnUPF/Prosograph
[2]http://www.gnu.org/licenses/
[3]http://py.processing.org/
[4]https://www.ted.com/

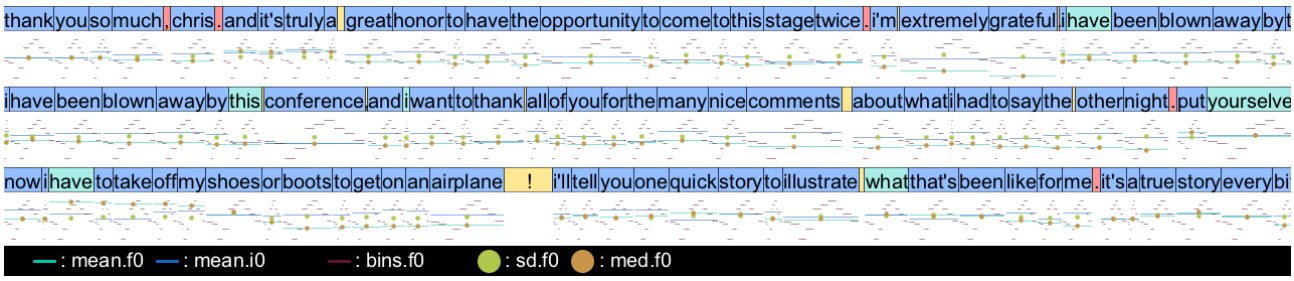

Figure 1: *An example of a visualization frame of a speech from the TED corpus with Prosograph*

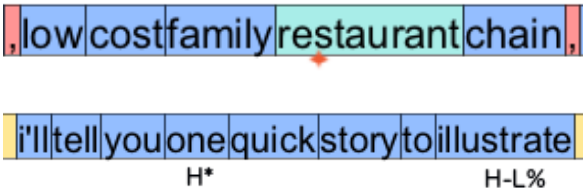

Figure 2: *Percentage-features showing word stress (top) and label-features (bottom) showing ToBI labels.*

median f0, etc.). It is placed at its value below the middle of the word's bounding box.

line-features holds a real numbered value as point-features. They are visualized as a line below and parallel to the word unlike point-features. This feature type could be useful e.g. for visualizing better the mean f0 movement across the utterance.

curve-features holds sequences of same length corresponding to each word. Each value is treated as curve bins and drawn as a line below the word in same length intervals. It is to be used e.g. for visualizing f0 curves or intensity curves or quantiles.

percentage-features holds sequences of varying lengths where each value in the sequence corresponds to a percentage of time with respect to the duration of the word. A mark is placed at the corresponding time position below the word's bounding box. This feature type can be used e.g. to mark the point where the accent occurs in a word, f0 or intensity peaks (see Figure 2).

label-features holds a string label for their respective words. The label is written just below the respective word's bounding box. This feature type can be used to visualize prosodic labels such as ToBI or part of speech labels of words (see Figure 2).

2.3. Usage

The path to the dataset and vocabulary file, names of dictionary keys and the type of data they refer to (as described in Section 2.2) should be set in the configuration file. Once the configurations are set, Prosograph can be run to navigate over the dataset. Utterances are shown in batches and user can navigate over the batches using keyboard shortcuts N(next) and B(previous). The current batch frame can be saved as an image by pressing S.

Colors of different prosodic features are set randomly at runtime. A legend showing which color belongs to which feature is shown at the bottom of the screen. If not easily distinguishable, the colors can be changed (again randomly) by pressing C.

3. Conclusions and Future Work

We have presented a tool, which can be used for the analysis of prosodic features and patterns in a speech corpus. It has been designed to be robust for handling different types of prosodic data aligned on word level. We have yet to design it to cover other types of prosodic data. To make it more usable, we plan to develop a configuration interface to fully isolate the non-programmer user from the source code. Also, we will make the interface more interactive by adding an information box that appears when mouse is hovering over a word.

Prosograph is the first attempt to develop a tool that visualizes speech at the discourse level. It is open for development and collaboration for the prosody research community. We hope that our simple start may help many sub-areas of prosodic research in the future.

4. Acknowledgements

This work is part of the KRISTINA project, which has received funding from the *European Union's Horizon 2020 Research and Innovation Programme* under the Grant Agreement number 645012. The second author is partially funded by the Spanish Ministry of Economy, Industry and Competitiveness through the *Ramón y Cajal* program.

5. References

[1] P. Boersma and D. Weenink, "Praat: Doing phonetics by computer [Computer software], retrieved from http://www.praat.org/," 2017.

[2] Y. Xu, "ProsodyPro — A tool for large-scale systematic prosody analysis," in *Proceedings of Tools and Resources for the Analysis of Speech Proso*, Aix-en-Provence, France, 2013, pp. 7–10.

[3] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Proceedings of the Speech Prosody*, Nara, Japan, 2004.

[4] M. Domínguez Bajo, I. Latorre, M. Farrús, J. Codina-Filbà, and L. Wanner, "Praat on the Web: an upgrade of Praat for semi-automatic speech annotation," in *Proceedings of COLING 2016*, Osaka, Japan, 2016, pp. 218–222.

[5] S. Şentürk, A. Ferraro, A. Porter, and X. Serra, "A Tool for the Analysis and Discovery of Ottoman-Turkish Makam Music," 2015.

[6] N. G. Ward, "Midlevel prosodic features toolkit," 2017. [Online]. Available: https://github.com/nigelgward/midlevel

[7] C. Lai, "Code for experiments with prosodic features," 2015. [Online]. Available: https://github.com/laic/prosody