AutoRPT: Automatic Detection of Prosodic Prominence and Boundary

Seth Heiney and Jonathan Howell

Montclair State University

We present a method for automated detection of prosodic prominence and boundary. While attention to prosody continues to increase in speech sciences, the relative paucity of and lack of diversity of prosodically annotated corpora remains a challenge (cf. Rosenberg 2018).

Our approach takes inspiration from the AuToBI tool (Rosenberg 2010) for classification of prosodic events in Mainstream US English (MUSE) using the Tones and Breaks Indices (ToBI) standard (Silverman et al. 1992). Rather than committing to the ToBI standard and to a specific variety of English, however, we make use of the coarser, more theory- and language variety-agnostic Rapid Prosody Transcription (RPT) method (Cole et al. 2019; see also Ahn et al. 2019). Development of the tool is part of a larger project involving collection and annotation of African American English and Latine English conversational speech. RPT is a more appropriate standard for uncovering the prosodic inventory of these under-documented varieties and requires little training for native speaker annotators.

A Python-based program takes as input pairs of .wav and .TextGrids segmented at the word-and phone-level (using, for example, the Montreal Forced Aligner; McAuliffe et al. 2017). Textgrids in the training data include word-level intervals annotated for prominence and boundary. Acoustic measures–currently pitch and intensity, with plans to integrate duration, energy and spectral tilt–are extracted using Praat, via the Parselmouth Python library (Jadoul et al. 2018). F0 and intensity measures include maximum, mean, standard deviation from the mean, and speaker-normalized standard deviation. These values are fed into a Neural Network Binary Classifier model for prediction, and output a prosodic structure TextGrid tier. Future iterations will explore Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) models for enhanced sensitivity to syntagmatic context (e.g. Lin et al. 2020; Fernandez et al. 2017).

We test the tool on a prosodically-annotated subset of the Boston University Radio News Corpus (Ostendorf et al. 1996), featuring roughly 2 hours of annotated audio data from 6 speakers and divided into 80%/20% training/test sets. Annotations for pitch accent and phrase/boundary tone are collapsed into prominence and boundary, respectively. Results are summarized below. Preliminary models, based solely on pitch measures or on intensity measures, achieve lower accuracy compared with AuToBI (83% and 93% for prominence and boundary), although accuracy is already sufficient to facilitate manual annotation (cf. Escudero-Mancebo et al. 2014). Moreover, we anticipate improved performance with the inclusion of additional acoustic measures (e.g. duration, energy and spectral tilt) sensitive to the nuclei of stressed syllables (already identified in the phone-level transcription) and context-sensitive deep learning models (e.g. LSTMs).

Model	Metric	Boundary	Prominence
F0	F1	0.7652	0.6175
	Precision	0.7586	0.6821
	Recall	0.8321	0.6335
Intensity	F1	0.7692	0.6171
	Precision	0.7082	0.6568
	Recall	0.8416	0.6312

References

Ahn, Byron, Nanette Veilleux & Stefanie Shattuck-Hufnagel. 2019. Annotating prosody with PoLaR: Conventions for a decompositional annotation system. In Sasha Calhoun, Paola Escudero, Marija Tabain & Paul Warren (eds.), *Proceedings of the 19th international congress of phonetic sciences*, australia, 2019, 1302–1306. Canberra, Australia: Australasian Speech Science and Technology Association Inc.

Cole, Jennifer, José I Hualde, Caroline L Smith, Christopher Eager, Timothy Mahrt & Ricardo Napoleão de Souza. 2019. Sound, structure and meaning: The bases of prominence ratings in English, French and Spanish. *Journal of Phonetics* 75. 113–147.

Escudero, David and Aguilar-Cuevas, Lourdes and González-Ferreras, Cesar and Gutiérrez-González, Yurena and Cardeñoso-Payo, Valentín. 2014. On the use of a fuzzy classifier to speed up the Sp_ToBI labeling of the Glissando Spanish corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 1962-1969.

Fernandez, Raul, Asaf Rendel, Bhuvana Ramabhadran & Ron Hoory. 2014. Prosody contour prediction with long short-term memory, bi-directional, deep recurrent neural networks. In *Fifteenth annual conference of the international speech communication association*.

Jadoul, Yannick, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, *71*, 1–15. https://doi.org/10.1016/j.wocn.2018.07.001

Lin, Binghuai and Wang, Liyuan and Feng, Xiaoli and Zhang, Jinsong. 2020. Joint Detection of Sentence Stress and Phrase Boundary for Prosody. In *Proceedings of INTERSPEECH*, 4392-4396.

McAuliffe, Michael and Socolof, Michaela and Mihuc, Sarah and Wagner, Michael and Sonderegger, Morgan. 2017. Montreal forced aligner: Trainable text-speech alignment using Kaldi. In *Interspeech*, 498-502.

Ostendorf, Mari, Patti Price & Stefanie Shattuck-Hufnagel. 1996. *Boston University Radio Speech Corpus*. Linguistic Data Consortium.

Rosenberg, Andrew. 2010. AuToBI-a tool for automatic ToBI annotation. In Proceedings of INTERSPEECH, 146–149.

Rosenberg, Andrew. 2018. Speech, prosody, and machines: Nine challenges for prosody research. In *Proceedings of the international conference on speech prosody*, 784–793.

Silverman, K. E., Beckman, M. E., Pitrelli, J. F., Ostendorf, M., Wightman, C. W., Price, P., ... & Hirschberg, J. (1992, October). ToBI: A standard for labeling English prosody. In *Proceedings of the 1992 International Conference on Spoken Language Processing*, Vol. 2, 867-870.