Unsupervised/semi-supervised prosodic unit discovery

Mahir Morshed

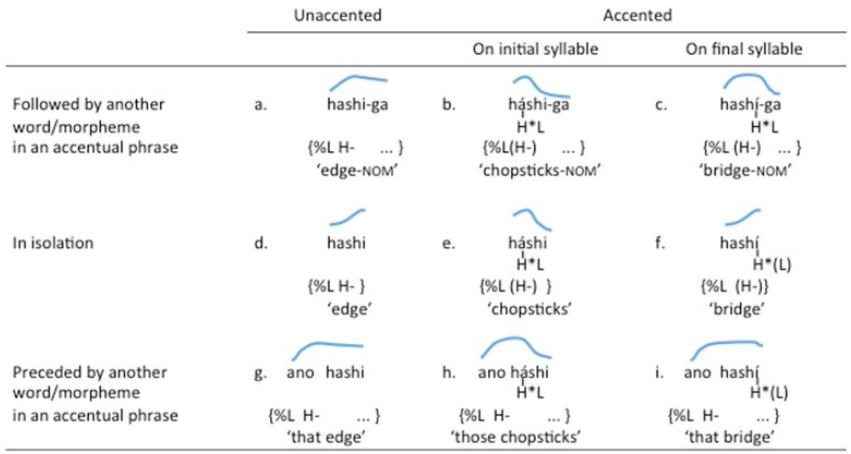
9 November 2021

Motivation

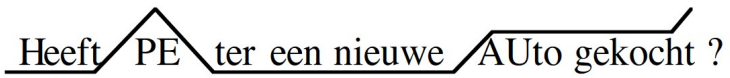
- Intonation patterns vary across languages, no matter the familial distance
  - Frequent connections of intonation to a language's semantic or pragmatic phenomena
  - For rule-based sentence production, selection/manipulation of base prosodic elements necessary
- What are these base prosodic elements for a new language?
- How is finding them affected by language tonality?
  - …these in addition to other identified challenges [1]: (increasing labeled data) and (developing pragmatic prosody theories) in particular

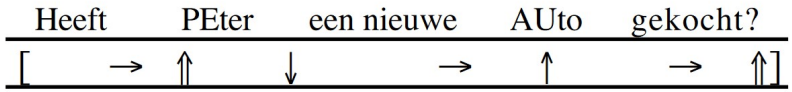
Background: lexical tone and intonation

- Suprasegmentals abound, even with pitch accent or lexical tone
- Intonation mismatch from phrasal, sememe pitch interaction [2]
  - Also present in non-tonal languages to some extent
- Often commensurate transcription mismatches, phonetic or otherwise
  - …these already complicated by inter-tone interactions (downstep/downdrift)



Background: tones/break indices



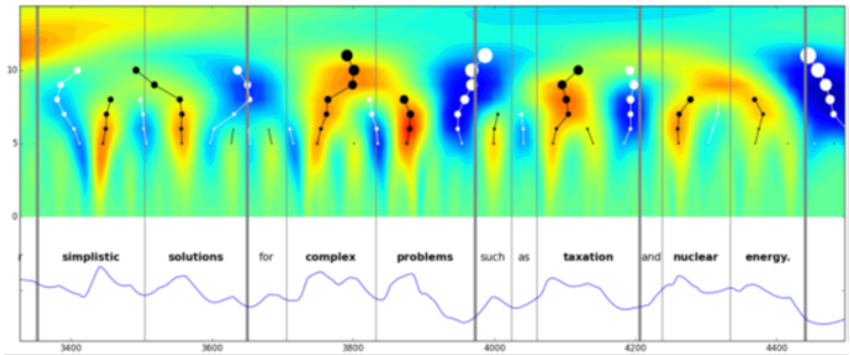


Language	Final sustained pitch	Final rise
English (Beckman et al., 2005)	H-L%	H-H%
German (Grice et al., 2005)	!H-%	^H%
Dutch (Gussenhoven, 2005)	Absence of boundary tone	H%
Greek (Arvaniti & Baltazani, 2005, p. 95)	!H!H%	H-H%
Spanish (Prieto & Roseano, 2010)	M%	H%
Portuguese (Frota, 2014)	!H%	H%
Catalan (Prieto, 2014)	!H%	H%
Serbian/Croatian (Godjevac, 2005, p. 152)	HL%	H%

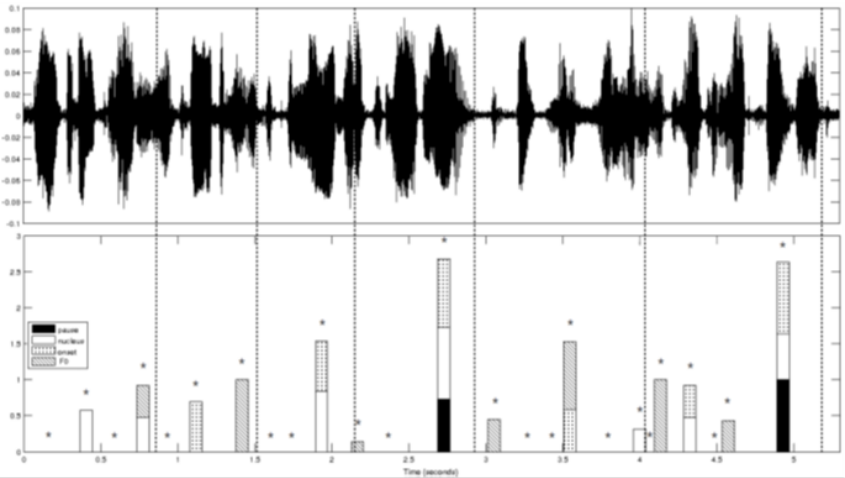
- Prosodic transcription method standardization difficult [3]
  - (top-left) the convention of one author in a volume at odds with its other authors
- Variation for similar phenomena, even among ToBI-like transcriptions [4]
  - Depends on distinctiveness in a given language
- What, then, of languages without any transcription system?

Background: prosodic event detection

- Extracting tone contours using zero-frequency filters to identify stable voice frames [5]
- Identifying boundaries and prominences with multi-resolution analyses [6]
- Combining purely phonetic cues as heuristics (20hr dataset) [7]
- Minimizing quantized F0/spectral tilt/energy n-gram statistics' predictability [8]



Background: low-resource prosodic event detection



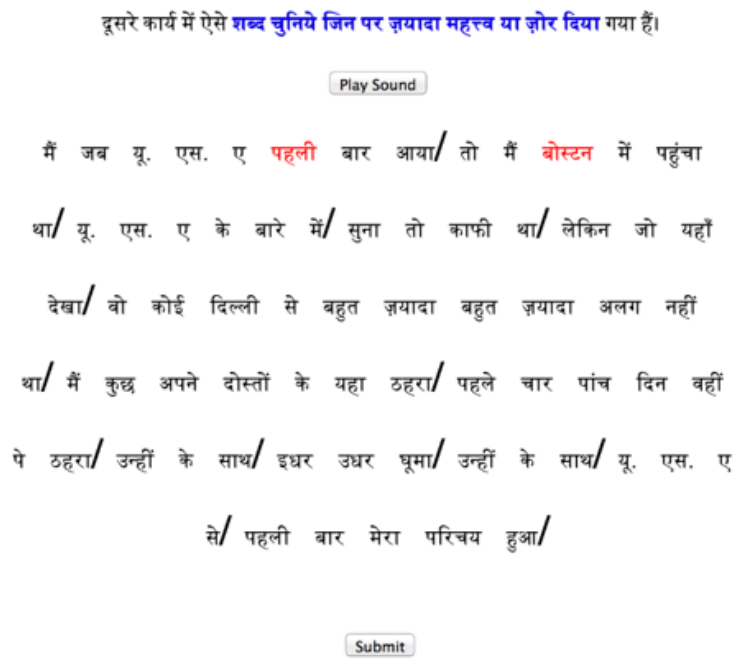
- Due to typical sizes of prosody-annotated corpora, largely similar efforts to those mentioned previously:
  - Combining purely phonetic cues as heuristics (3hr dataset) [9]
  - Attempting the same cross-linguistically (7.5hr across four languages) [10]
- Classifying tone contours into discrete classes with GRUs (233 utterances) [11]

Proposed datasets

- Among non-tonal languages (specifically English for the moment):
  - Boston University [12], NXT Switchboard [13] (as English baselines for such a system)
  - Helsinki Prosody Corpus [14] (as a much larger corpus, to assess scalability)
- Among tonal languages (to examine disentangling lexical tone):
  - Setswana/isiXhosa speech corpora—toneless transcripts [15]
  - Burmese speech corpus—some tones inferable, but not consistently [16]
  - Yoruba speech corpus—transcripts with lexical tone [17]

Proposed methods: mismatched transcription

- Rapid Prosody Transcription, with alterations in italics:
  - *Non*-speakers of a language presented with text ($<$1 minute) in that language
  - Asked to annotate prosodic boundaries, prominent words therein
  - Typically done without playback control, *but speed options may be presented*
  - *Depending on the language, option for syllabication may be presented*

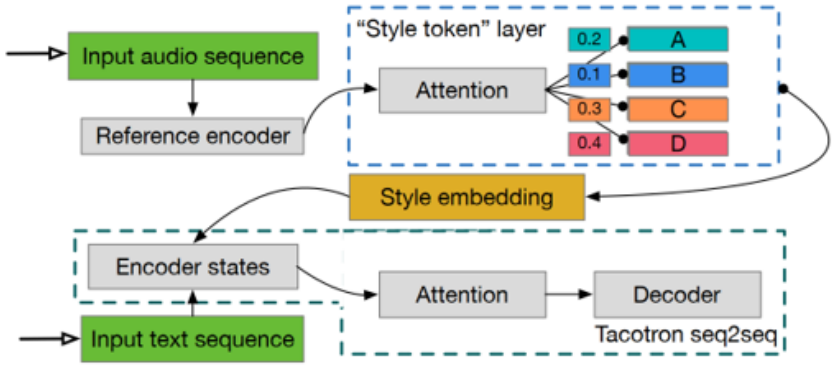


Proposed methods: uses and assessments

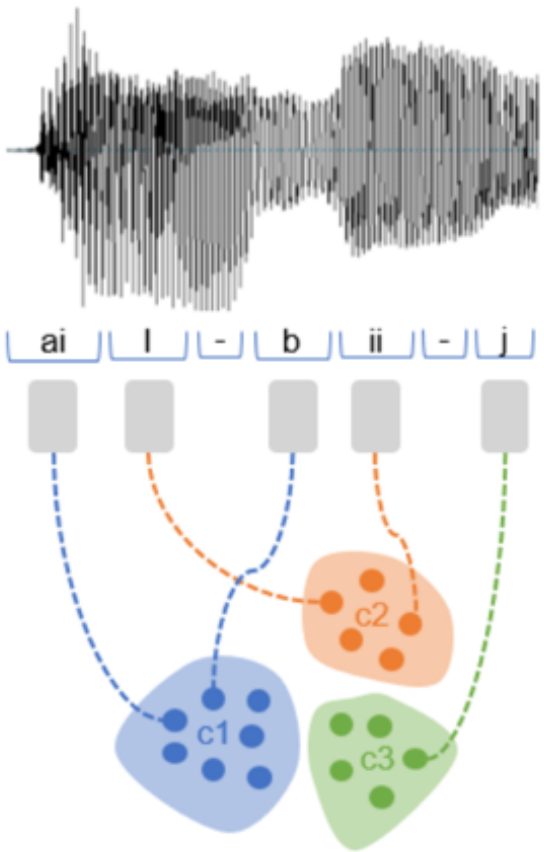
- Using the mismatched annotations *as a baseline*:
  - Reduce to ToBI-like outputs
  - Time-align words/phrases based on aggregate annotator judgments versus F0/other acoustics
- Two sets of model output assessments possible:
  - Accuracy (comparing model outputs from clustering to aforementioned ToBI-like outputs)
  - MSE (computing time differences between positions of the above)

Proposed methods: end-to-end generation

- Similar to ‘style token’ derivation with Tacotron [18]:
  - Generate (higher-dimension) continuous representations of intonation, tone type from input at each timestep
  - Need not correspond to F0 contours, silence points, other specific acoustic cues (model adjustments to encourage such correspondences to be considered later)
  - Defer discretization to later



Proposed methods: clustering



- Once intonation, tone type representations produced, cluster them all
  - Discretize intonations, tone-types separately first
  - Join time-adjacent intonation-tone type combinations
  - Combine frequently occurring time-adjacent intonation combinations
- K-means or GMM-fitting possible here [19]

References I


A. Rosenberg, "Speech, prosody, and machines: Nine challenges for prosody research," in *Proc. Speech Prosody*, 2018, pp. 784–793.


M. Ota, N. Yamane, and R. Mazuka, "The effects of lexical pitch accent on infant word recognition in japanese," *Frontiers in Psychology*, vol. 8, p. 2354, 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2017.02354


C. A. De, D. L. Bolinger, D. Gibbon, E. Garding, J. T'Hart, N. Gronnum, S. Alcoba, J. Murillo *et al.*, *Intonation systems: a survey of twenty languages*. Cambridge University Press, 1998.


J. I. Hualde and P. Prieto, "Towards an international prosodic alphabet (ipra)," p. 5, Jun 2016. [Online]. Available: http://dx.doi.org/10.5334/labphon.11


J. Ni, Y. Shiga, and H. Kawai, "Using zero-frequency resonator to extract multilingual intonation structure." in *INTERSPEECH*, 2016, pp. 1522–1526.


A. Suni, J. Šimko, D. Aalto, and M. Vainio, "Hierarchical representation and estimation of prosody using continuous wavelet transform," p. 123–136, Sep 2017. [Online]. Available: http://dx.doi.org/10.1016/j.csl.2016.11.001


T. Biron, D. Baum, D. Freche, N. Matalon, N. Ehrmann, E. Weinreb, D. Biron, and E. Moses, "Automatic detection of prosodic boundaries in spontaneous speech," *PLOS ONE*, vol. 16, no. 5, pp. 1–21, 05 2021. [Online]. Available: https://doi.org/10.1371/journal.pone.0250969

References II


S. Kakouros and O. Räsänen, "3pro – an unsupervised method for the automatic detection of sentence prominence in speech," p. 67–84, Sep 2016. [Online]. Available: http://dx.doi.org/10.1016/j.specom.2016.06.004


B. Ludusan and E. Dupoux, "Towards low-resource prosodic boundary detection." in *SLTU*. Citeseer, 2014, pp. 231–237.


V. Soto, E. Cooper, A. Rosenberg, and J. Hirschberg, "Cross-language phrase boundary detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 8460–8464.


A. Saha, C. Yarra, and P. K. Ghosh, "Low resource automatic intonation classification using gated recurrent unit (gru) networks pre-trained with synthesized pitch patterns," Sep 2019. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2351


M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," *Linguistic Data Consortium*, pp. 1–19, 1995.


S. Calhoun, J. Carletta, D. Jurafsky, M. Nissim, M. Ostendorf, and A. Zaenen, "Nxt switchboard annotations," Nov 2009. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2009T26


A. Talman, A. Suni, H. Celikkanat, S. Kakouros, J. Tiedemann, and M. Vainio, "Predicting prosodic prominence from text with pre-trained contextualized word representations," in *Proceedings of the 22nd Nordic Conference on Computational Linguistics*. Turku, Finland: Linköping University Electronic Press, Sep.–Oct. 2019, pp. 281–290. [Online]. Available: https://aclanthology.org/W19-6129

References III

D. van Niekerk, C. van Heerden, M. Davel, N. Kleynhans, O. Kjartansson, M. Jansche, and L. Ha, "Rapid development of tts corpora for four south african languages," in *Proc. Interspeech 2017*, 2017, pp. 2178–2182. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-1139

Y. M. Oo, A. Theeraphol, C. F. Li, P. D. Silva, S. Sarin, K. Pipatsrisawat, M. Jansche, O. Kjartansson, and A. Gutkin, "Burmese speech corpus, finite-state text normalization and pronunciation grammars with an application to text-to-speech," in *Proc. 12th Language Resources and Evaluation Conference (LREC 2020)*, 11–16 May, Marseille, France, 2020, pp. 6328–6339. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2020/pdf/2020.lrec-1.777.pdf

A. Gutkin, I. Demirsahin, O. Kjartansson, C. E. Rivera, and K. Túbòsún, "Developing an open-source corpus of yoruba speech," in *Proc. of Interspeech 2020*, October 25–29, Shanghai, China, 2020., 2020, pp. 404–408. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1096

Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

A. Vioni, M. Christidou, N. Ellinas, G. Vamvoukakis, P. Kakoulidis, T. Kim, J. S. Sung, H. Park, A. Chalamandaris, and P. Tsiakoulis, "Prosodic clustering for phoneme-level prosody control in end-to-end speech synthesis," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5719–5723.