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

- 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]


