A deep neural network approach to investigate tone space in Mandarin Chinese

Bing’er Jiang1, Meghan Clayards1,2, Mirco Ravanelli3, Timothy J. O’Donnell1,3

1Department of Linguistics, McGill University; 2School of Communication Sciences and Disorders, McGill University; 3Mila, Université de Montréal, Canada

binger.jiang@mail.mcgill.ca, meghan.clayards@mcgill.ca, mirco.ravanelli@gmail.com, tim.odonnell@mcgill.ca

177th ASA, Louisville, KY
May 13 – 17, 2019

BACKGROUND

- Tonal contrast in Mandarin Chinese is signaled by various acoustic cues, including but not limited to:
  - Pitch (Howie, 1976; Gaudour, 1984)
  - Intensity/amplitude (Chung et al., 1972; Lin, 1988)
  - Duration (Dreherr & Lee, 1996; Chuang et al., 1972)
  - Spectral information (Kong & Zeng, 2006)
  - Voice quality (Cao, 2012)

- We do not know how much information is available from the low-level acoustic signals, prior to forming any linguistic units
- Low-dimensional representation of speech (e.g. Weber et al., 2015 on phonemes) extracted from the Bottleneck Layer trained in Deep Neural Networks (DNNS) show similar properties to linguistic features (e.g. F1/F2 for vowels)
- Representations learnt in DNNS can be used to understand various phonological contrasts

Q1: How much information is available from the acoustic signals carried by each cue?

- Tone recognition accuracy when each cue is removed/neutralized
- The time course of tone recognition

Q2: What can we learn about tonal contrast from the low-dimensional representation derived from DNNS?

DATA

- Corpus: Mandarin Chinese Phonetic Segmentation (Yuan et al., 2015)
  - Test: 300 utterances, 6 speakers; Train: 7540 utterances (train/validation: 90%/10%)
- Input: 39 MFCCs (the first 13 cepstral coefficients with Δ and ΔΔ) + F0 estimation (z-scored)
  - Extracted from the rhyme (excluding onset)
  - Computed every 10ms, with window of length 25ms
- Manipulation: neutralize one or more cues from the natural data, up to all three cues
  - No Pitch: resynthesize all tones to have F0 = 200Hz, using PSOLA method in Praat (Boersma & Weenink, 2019)
  - No Intensity: flatten intensity to 70db (using Praat)
  - No Duration: normalize all tones to be 12 frames (= mean length of training data)

OBJECTIVES

- Pitch
- Intensity
- Duration

- Task: tone classification (one of the four tones)
- Model: Long short-term memory (LSTM) network (Hochreiter & Schmidhuber, 1997), a variant of Recurrent NN (Rumelhart, 1988)

- Advantages:
  - Compare to traditional GMM-HMM/ other deep learning models: allows input to have different lengths, representing duration
  - The prediction of the current state is dependent on previous states
  - Allows for high-dimensional acoustic input from raw speech before forming any linguistic abstraction, more similar to input humans receive

THREE CUES

Trained on bi-directional LSTMs without BN layer:

<table>
<thead>
<tr>
<th>Model (data condition)</th>
<th>Accuracy*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1) Natural speech</td>
<td>77.3%</td>
</tr>
<tr>
<td>2) no Pitch</td>
<td>67.2%</td>
</tr>
<tr>
<td>3) no Intensity</td>
<td>75.0%</td>
</tr>
<tr>
<td>4) no Duration</td>
<td>74.1%</td>
</tr>
<tr>
<td>5) no Pitch &amp; Intensity</td>
<td>62.8%</td>
</tr>
<tr>
<td>6) no Pitch &amp; Duration</td>
<td>59.3%</td>
</tr>
<tr>
<td>7) no Intensity &amp; Duration</td>
<td>71.9%</td>
</tr>
<tr>
<td>8) All three cues removed</td>
<td>55.8%</td>
</tr>
</tbody>
</table>

*All are significant (p < 0.001) compared against a random baseline proportional to the frequency of each tone, under the Wilcoxon test.

Effects of removing each cue for different tones:

- **Tone 1**
  - Natural speech: 76.7%
  - No Pitch: -15.4%
  - No Intensity: +2.0%
  - No Duration: -0.7%
- **Tone 2**
  - Natural speech: 75.5%
  - No Pitch: -12.0%
  - No Intensity: +7.3%
  - No Duration: -5.0%
- **Tone 3**
  - Natural speech: 63.7%
  - No Pitch: +4.4%
  - No Intensity: -7.9%
  - No Duration: +0.4%
- **Tone 4**
  - Natural speech: 84.9%
  - No Pitch: -7.9%
  - No Intensity: +0.5%
  - No Duration: -5.0%

TIME COURSE

Trained on uni-directional LSTMs without BN layer:

Averaging over four tones

- Presence/ absence of F0 facilitates recognition earlier in the tone
  - Four tones independently
  - Presence/ absence of F0 creates different patterns except for Tone 3
  - Tone 3 is more sensitive to intensity

DISCUSSION

- Pitch is the most important cue — evident from all three tasks
- Intensity is important for Tone 3; Duration is important for Tone 2 and 4
- 2D BN representations separate four tones in four quadrants
- When a cue is neutralized, four groups are pulled together in the 2D space
- Future work: map the BN dimensions with acoustic dimensions