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

4pSC28

Bing'er Jiang[1], Meghan Clayards[1,2], Mirco Ravanelli[3], Timothy J. O'Donnell[1,3]

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

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

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; Gandour, 1984)
  - **Intensity/ amplitude** (Chuang et al., 1972; Lin, 1988)
  - **Duration** (Dreher& Lee, 1996; Chuang et al., 1972)
  - Spectral information (Kong & Zeng, 2006)
  - Voice quality (Cao, 2012)

Mandarin Tones	T1	T2	T3	T4
Chao number	55	35	214	51
Description	High level	Low rising	Low dipping	High falling

- 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

OBJECTIVES

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

→
- **Pitch**
- **Intensity**
- **Duration**

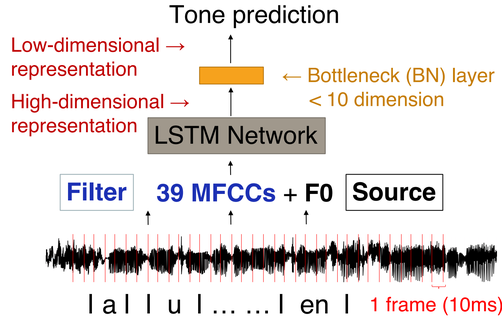
Q2: What can we learn about tonal contrast from the low-dimensional representation derived from DNNs?

DATA

- **Corpus**: Mandarin Chinese Phonetic Segmentation (Yuan et al., 2015)
  Test: 300 utterances, 6 speakers; Train: 7549 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)

MODEL

- **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



Model/ training detail:
- Bidirectional (uni- for time-course)
- Hidden state size =1024
- Loss function: cross-entropy loss
- Optimiser = Adam
- Dropout = 0.2
- Batch size = 32
- Trained the models until the validation loss failed to improve
- Bottleneck layer dimension was attempted from 1 to 128, only used for visualization task

THREE CUES

Trained on bi-directional LSTMs without BN layer:

	Model (data condition)	Accuracy*	
1)	Natural speech	77.3%	
2)	no Pitch	67.2%	← Most important
3)	no Intensity	75.0%	← Least important
4)	no Duration	74.1%	
5)	no Pitch & Intensity	62.8%	
6)	no Pitch & Duration	59.3%	
7)	no Duration & Intensity	71.9%	
8)	**All three cues removed**	**55.8%**	

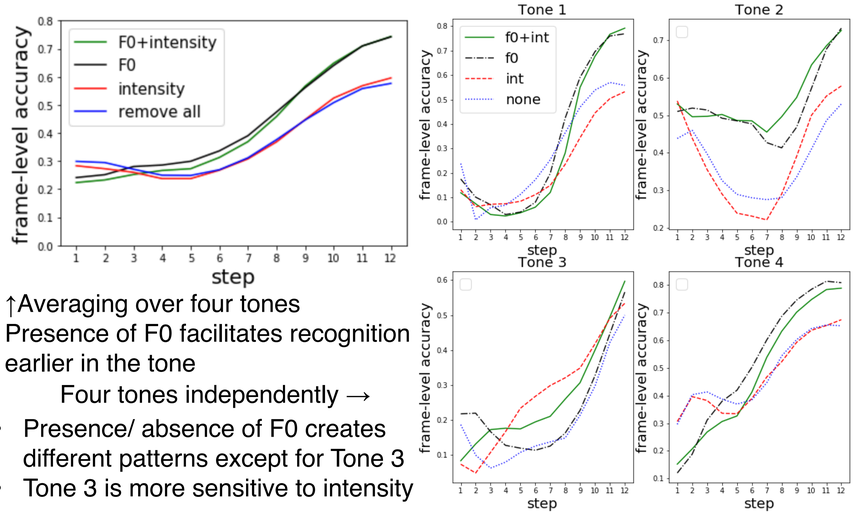
*All are significant ($p < 0.001$) compared against a random baseline proportional to the frequency of each tone, under the Wilcoxon test.

Effects of removing each cue for different tones:

Model	T1	T2	T3	T4
Natural speech	76.7%	75.5%	63.7%	84.9%
no Pitch	-15.4%	-12.0%	-4.4%	-7.9%
no Intensity	+2.0%	**-7.3%**	**-7.9%**	+0.5%
no Duration	-0.7%	**-5.0%**	+0.4%	**-5.0%**

TIME COURSE

Trained on uni-directional LSTMs without BN layer:



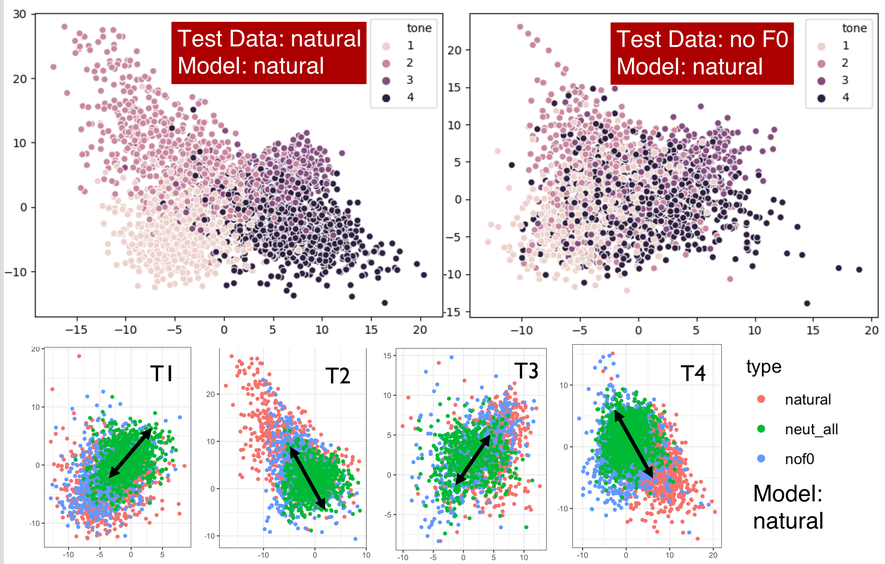
↑Averaging over four tones
Presence of F0 facilitates recognition earlier in the tone

Four tones independently →

- Presence/ absence of F0 creates different patterns except for Tone 3
- Tone 3 is more sensitive to intensity

LOW-DIMENSIONAL REPRESENTATION

2-dimension is sufficient for natural data; more if some cue missing.



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

The first author thanks Morgan Sonderegger, Jiangtian Li, and Michael McAuliffe for helpful discussion, and Lei Yu and Jingyi He for help on model training.

Selected References:S Chuang, C. K., Hiki, S., Sone, T., and Nimura, T. (1972). The acoustical features and perceptual cues of the four tones of Standard Colloquial Chinese. Proceedings of the Seventh International Congress on Acoustics (Adadémial Kiado, Budapest), 297-300.; F Dreher, J. and Lee, P.C. (1966). Instrumental investigation of single and paired Mandarin tonemes. Research Communication 13, Douglas Advanced Research Laboratories.; Gandour, J. (1984). Tone dissimilarity judgments by Chinese listeners. Journal of Chinese Linguistics 12, 235-261.; Hochreiter, Sepp, and Jürgen Schmidhuber. "LSTM can solve hard long time lag problems." Advances in neural information processing systems. 1997.; Howie, J. M. (1976). Acoustical studies of Mandarin vowels and tones (Cambridge University Press, Cambridge).; Lin, M. C. (1988). Putong hua sheng diao de sheng xue texing he zhi jue zhengzhao [Standard Mandarin tone characteristics and percepts]. Zhongguo Yuyan 3, 182-193.; Kong, Ying-Yee, and Fan-Gang Zeng. "Temporal and spectral cues in Mandarin tone recognition." The Journal of the Acoustical Society of America 120.5 (2006): 2830-2840.; Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors." Cognitive modeling 5.3 (1988): 1.; Yuan, Jiahong, Neville Ryant, and Mark Liberman. Mandarin Chinese Phonetic Segmentation and Tone LDC2015S05. Web Download. Philadelphia: Linguistic Data Consortium, 2015.