Do speech models develop human-like perception?

A comparison between English stop voicing classification by humans and wav2vec2

Suyuan Liu[1] (suyuan.liu@ubc.ca)

To view references →

← To try the speech models

1. Introduction

- **Wav2vec2** [1,2]: self-supervised speech model; accurate performance in sound categorization; black-box manner
- **Humans** [3,4]: attend to Voice Onset Time (VOT) and post-stop f0 as primary and secondary cue for English stop voicing categorization
- **Do self-supervised speech models like wav2vec2 "attend" to both VOT and f0 when categorizing English stop voicing?**

2. Methods

2.1 Speech models (w/ wav2vec2 framework [1])

- Trained to categorize English monosyllabic words based on the voicing of the word-initial stop (voiced vs. voiceless) [5,6]
- Fine-tuned models: trained with pre-determined weights (n=5, $M_{accuracy}$ = 92.4%)
- Randomly-initialized models: trained with randomly initialized weights (n = 5, $M_{accuracy}$ = 86.62%)

2.2 Human participants

- Lexical categorization task of one group of following stimuli

2.3 Evaluation stimuli

- Audio continua: voiced ⟷ voiceless
- (3a) VOT-f0 [7]: VOT & f0
- (3b) VOT stimuli [7]: VOT
- (3c) TANDEM stimuli [8]: multiple acoustic dimensions

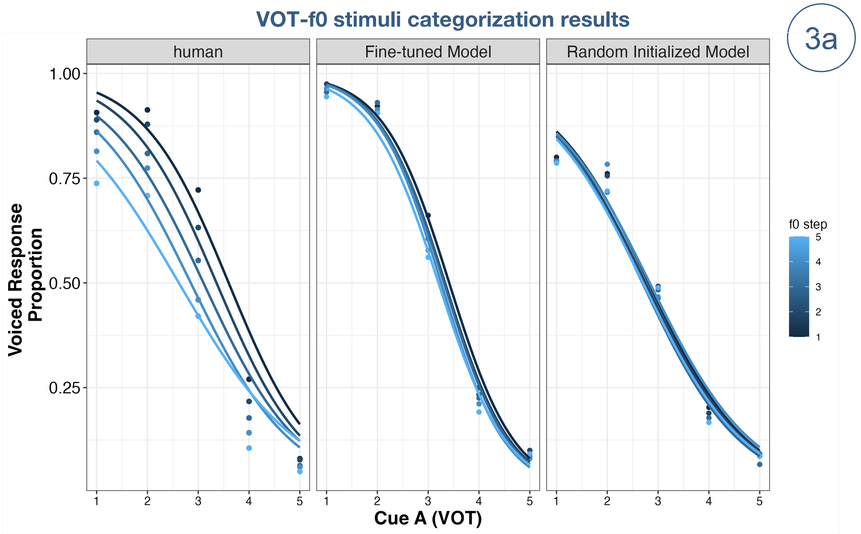
3. Results & Conclusion

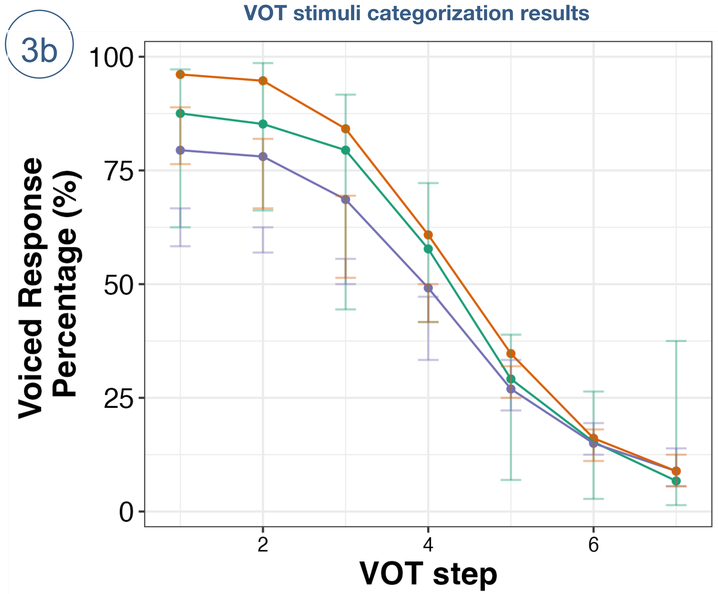
3.1 Speech models' lack of sensitivity to f0

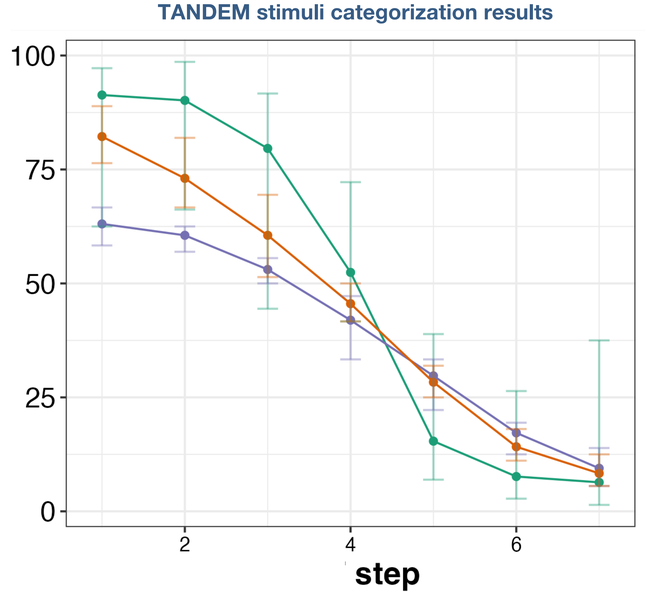
- Human and both speech models are sensitive to VOT
- Speech models are not sensitive to f0 changes; while humans are (3a)

3.2 Speech models' change in categoricalness

- Fine-tuned models are more categorical than randomly-initialized models (3a, 3b & 3c), and even humans (3b)
- Both speech models are more categorical when **single acoustic cue** is manipulated than morphed holistically (3b > 3c)







THE UNIVERSITY OF BRITISH COLUMBIA