Do People Mirror Emotion Differently with a Human or Text-to-Speech Voice?

Comparing Listener Ratings and Word Embeddings

Michelle Cohn[1,*], Grisha Bandodkar[1], Raj Bharat Sangani[2], Kristin Predeck[1, 3,†], Georgia Zellou[1]
[1]University of California, Davis, [2]University of California, Irvine, [3]Amazon AGI - Information
*mdcohn@ucdavis.edu, †This work was done prior to joining Amazon

Phonetics Lab UCDAVIS

Background

People subconsciously mirror linguistic patterns during conversation with other humans

- Including emotional expressiveness [1, 2]

And people mirror patterns in text-to-speech (TTS) voices (e.g. Siri and Alexa) [2,3]

- But mirroring is often stronger for human than TTS voices [4]

The current study compares how people mirror emotion from a human and a TTS voice, with two methods:

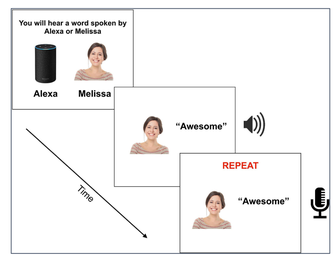
1. perceptual similarity by human listeners [6]
2. distance in latent space [5]

Dataset (from [2])

Audio recordings consisting of 36 L1 English speakers from the US who completed a controlled shadowing experiment.

Speakers produced baseline interjections (A) (e.g. "great", "awesome) and then repeated them (B) after hearing:

- a human and a US-English Alexa (female) TTS voice (X)
- producing the words expressively or neutrally

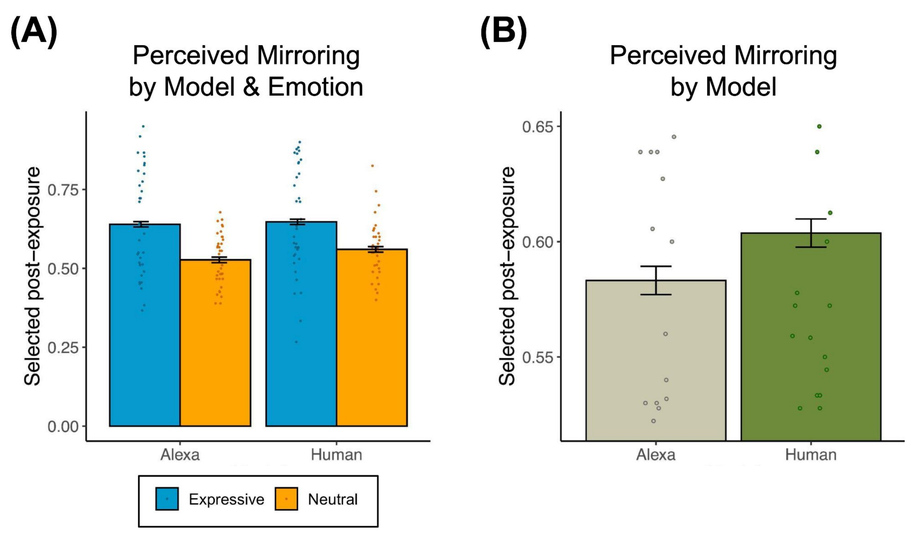


Experiment 1: Human raters

Raters (n=109, L1 English speakers from the US) were asked to select which speaker production (A or B) sounded more similar to the interlocutor's (X) production [6]. Analyzed with a mixed effects logistic regression (lme4) [7].

Results

- More perceived mirroring for expressive than neutral prosody ($p<0.001$)
- Participants mirrored the TTS voice less than the human voice ($p<0.05$)



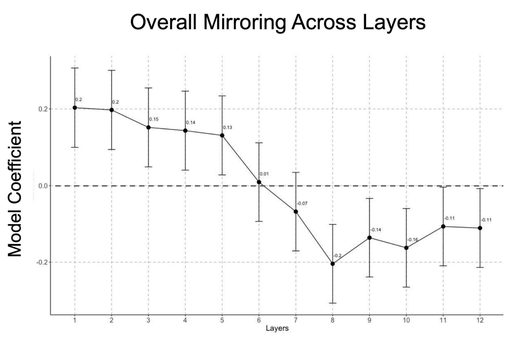
Experiment 2: Wav2Vec 2.0

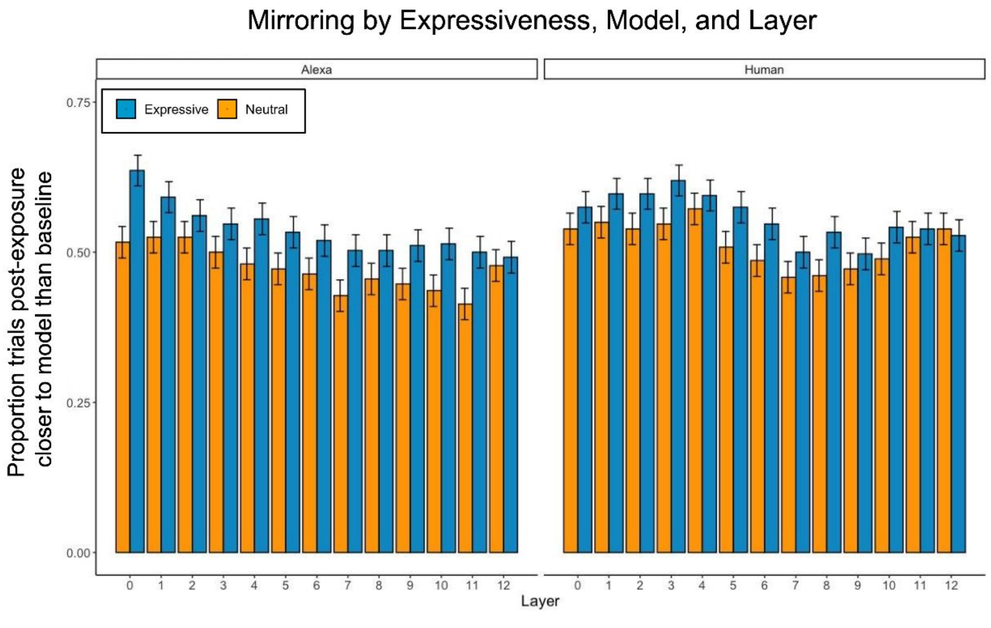
Extracted the word embedding in latent space using wav2vec 2.0 [5] based on 768 vectors

Calculated the cosine distance between **A ~ X** and **X ~ B** for each embedding layer. Perceived mirroring = distance (A ~ X) > (X ~ B).

Results

- More mirroring toward expressive than neutral speech ($p<0.01$)
- Interlocutor (n.s.)
- Overall mirroring in early layers
  - Layers 1-5 positive (all $p<0.05$)
  - Layers 6-7 (n.s.)
  - Layers 8-12 negative (all $p<0.05$)





Discussion

1. People mirror emotional prosody for both human and TTS voices — confirmed by human raters and wav2vec 2.0
2. Embeddings can be used to assess mirroring in speech
3. Unlike AXB, wav2vec 2.0 did not reveal differences in mirroring toward the human vs. TTS voices

Future Directions

1. Include additional emotions + types of voices
2. Compare emotion produced by different participant groups (language, gender, age, race/ethnicity)
3. Test which embedding layers capture mirroring
   - What feature does each layer focus on?
4. Examine mirroring in different interaction types beyond single word shadowing

[1] Pablo Arias, Pascal Belin, and Jean-Julien Aucouturier. 2018. Auditory smiles trigger unconscious facial imitation. Current Biology
[2] Michelle Cohn, Kristin Predeck, Melina Sarian, and Georgia Zellou. 2021. Prosodic alignment toward emotionally expressive speech: comparing human and alexa model talkers. Speech Communication
[3] Georgia Zellou, Michelle Cohn, and Tyler Kline. 2021. The influence of conversational role on phonetic alignment toward voice-ai and human interlocutors. Language, Cognition and Neuroscience
[4] Michelle Cohn, Bruno Ferenc Segedin, and Georgia Zellou. 2019. Imitating siri: socially-mediated alignment to device and human voices. In Proceedings of International Congress of Phonetic Sciences. Melbourne, Australia.
[5] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2vec 2.0: a framework for self-supervised learning of speech representations. Advances in neural information processing systems
[6] Jennifer Pardo. 2013. Measuring phonetic convergence in speech production. English. Frontiers in Psych
[7] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. en. J. Statistical Software