Quantifying perceptual similarity of connected speech

Seung-Eun Kim[1], Qingcheng Zeng[1], Bronya R. Chernyak[2], Joseph Keshet[2],
Matthew Goldrick[1], and Ann R. Bradlow[1]

[1]Northwestern University (USA), [2]Technion – Israel Institute of Technology (Israel)

Determining whether one speech signal is similar to another is central to many areas of phonetics research. For example, interlocutors' utterances are often perceived as becoming more similar over the course of a conversation ("phonetic convergence"; e.g., [1]). Speech similarity also plays a role in generalization of perceptual adaptation to second-language (L2) speech (e.g., [2]) with greater generalization associated with greater training-talker-to-test-talker similarity. It is, however, difficult to quantify phonetic similarity using objective phonetic measures (see [3]); this problem is aggravated for connected speech, where multiple interacting phonetic dimensions underly perceived similarity. To address this issue, many studies rely on perceptual measures of similarity (e.g., AXB tasks; [4]). However, such measures are inherently subjective (as shown by social biases on judgments; [5]) and require large numbers of participants.

We propose a novel way of quantifying phonetic similarity separate from social biases, utilizing a language-specific perceptual similarity space acquired via self-supervised learning. Specifically, waveforms of speech samples are passed through a self-supervised model that has been trained on a large speech sample from that language. Each speech sample is represented as a trajectory in this high dimensional space. Similarity of two utterances is then inversely related to distance between the trajectories in the space.

To validate this approach, we examined whether our method could capture perceived similarity among regional dialects of American English ([6]). The speech data consisted of two sentences (from TIMIT speech corpus; [7]), each read by 20 white male talkers from four dialect groups (five talkers per group) – i.e., New England, North, North Midland, South. The data were normalized in loudness and processed by HuBERT ([8]), a self-supervised model trained on a large sample of English. Distance between HuBERT trajectories was calculated for all possible talker pairs, and its additive inverse (which cues similarity) was analyzed. In [6], L1 and L2 (with various L1 backgrounds) English listeners were presented with 20 recordings of each sentence and asked to group similar talkers together (a free classification task). The resulting measure of perceived similarity – the number of participants who grouped the two talkers together – was compared to our trajectory distance measure.

As shown in Fig. 1, HuBERT captured overall dialect group similarity as humans do, but also exhibited differences. In sentence 1 (top row), both humans and HuBERT judged talkers of the same dialect to be more similar than talkers of different dialects (HuBERT diagonal mean (same dialect): 0.6, off-diagonal mean (different dialects): -0.2; Human L1 diagonal mean: 1.3, off-diagonal mean: -0.4; Human L2 diagonal mean: 0.6, off-diagonal mean: -0.2). Both humans and HuBERT had difficulty distinguishing North and North Midland talkers. While these broad patterns were similar, HuBERT groupings were not as categorical as L1 listeners and patterned more like L2 listeners. This suggests that both HuBERT and L2 listeners can use acoustic-phonetic cues to assess speech similarity, but they lack signal-independent, social knowledge that L1 listeners associate with each dialect ([6]). Similar though less stark patterns were found for sentence 2 (bottom row); here, HuBERT also showed differences from L2 listeners, exhibiting more categorical groupings (HuBERT diagonal mean: 0.6, off-diagonal mean: -0.2; Human L2 diagonal mean: 0.3, off-diagonal mean: -0.1).

Overall, the novel HuBERT-based method captured broad similarity relations amongst American English dialects like humans do, albeit with some differences from both L1 and L2 listeners. These results suggest that HuBERT-based objective and automatic measurement of perceptual similarity is a promising strategy for isolating the influence of purely signal-dependent information from social knowledge in similarity judgment.

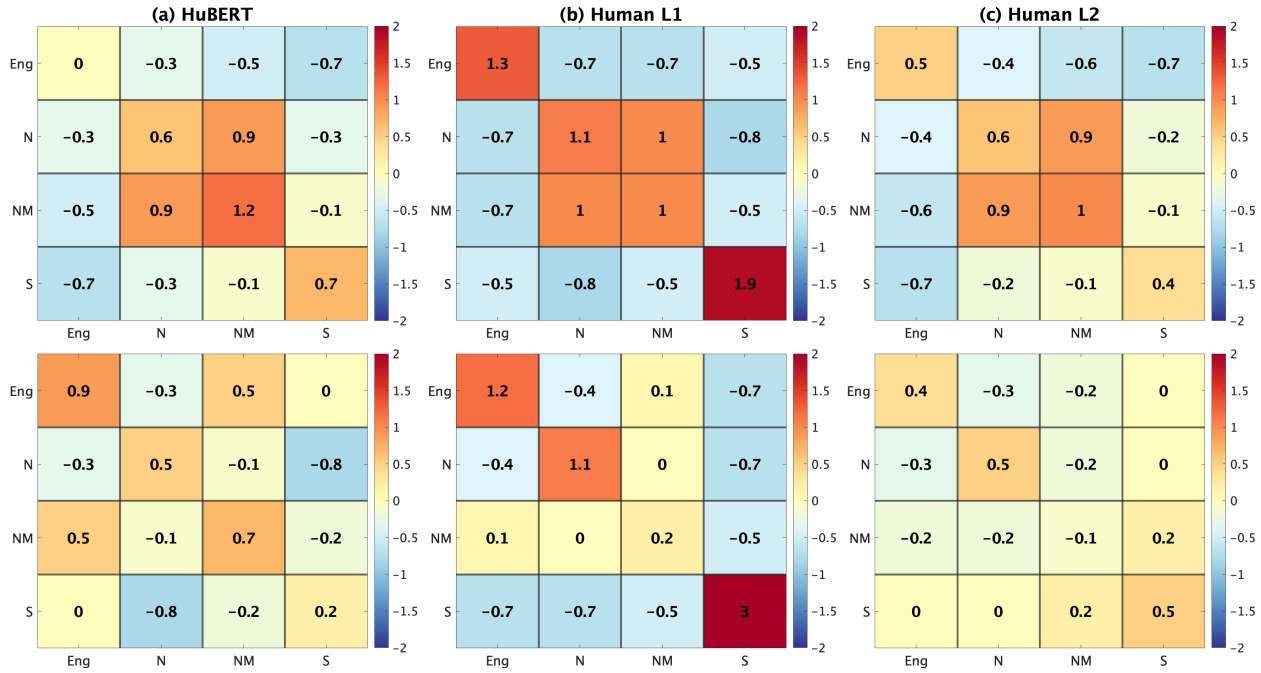

Fig. 1. Normalized average similarity measures obtained from (a) HuBERT, (b) L1 English listeners (sentence 1, n = 36; sentence 2, n = 27), and (c) L2 English listeners (sentence 1, n = 36; sentence 2, n = 32). Top: sentence 1 (*She had your dark suit in greasy wash water all year*), bottom: sentence 2 (*Don't ask me to carry an oily rag like that*). Eng: New England, N: North, NM: North Midland, S: South. In each panel, the number and color of the tile indicates average similarity between the dialect group in the column and the row (e.g., the first tile shows average similarity among New England talkers; the tile below shows average similarity between North and New England talkers); the larger the number is (more red), the more similar the two groups are. The similarity measures (i.e., HuBERT: additive inverse of distance between trajectories, Human: number of participants who grouped the talkers together) were z-score normalized within each sentence/listener type (i.e., within each panel).

References

[1] Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*(4), 2382-2393.

[2] Xie, X., & Myers, E. B. (2017). Learning a talker or learning an accent : Acoustic similarity constrains generalization of foreign accent adaptation to new talkers. *Journal of Memory and Language, 97,* 30-46.

[3] Pardo, J. S. (2013). Measuring phonetic convergence in speech production. *Frontiers in Psychology, 4,* 559.

[4] Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251-279.

[5] Bent, T., Atagi, E., Akbik, A., & Bonifield, E. (2016). Classification of regional dialects, international dialects, and nonnative accents. *Journal of Phonetics, 58,* 104-117.

[6] Clopper, C. G. & Bradlow, A. R. (2009). Free classification of American English dialects by native and non-native listeners. *Journal of Phonetics, 37,* 436-451.

[7] Garofolo, J. S. et al. (1993). TIMIT Acoustic-phonetic continuous speech corpus LDC 93S1. Web Download. Philadelphia : Linguistic Data Consortium.

[8] Hsu, W.-N., Bolte, B., Tsai, Y-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29,* 3451-3460.