Title

Accented sentence and word recognition: Humans versus whisper automatic speech recognition

Journal

The Journal of the Acoustical Society of America, 156(4_Supplement)

ISSN

0001-4966

Authors

Chen, Junrong

Kwong, Jan

Creel, Sarah C

Publication Date

2024-10-01

DOI

10.1121/10.0035071

Chen, Kwong, & Creel
Conference abstract
ASA 2024

Despite advancements in speech recognition technology, questions remain about model generalizability and how much models mirror human perception. These questions are addressed by comparing OpenAI's Whisper model and 75 human transcribers on 300 English sentences (20 speakers, half F, half M, half US-accented, half [Mexican-]Spanish-accented). Sentences ended in 100 target words, with ⅓ high-predictability sentences (*The farmer milked the cows*) and ⅔ varying degrees of low-predictability (*The farmer/barmer milked the nose*). Target-word error rate (WER) was examined for final words in sentences and for isolated final words (recordings excised from same sentences).

WER decreased with increasing model size, but was higher for Spanish-accented than US-accented speech, suggesting imperfect generalizability. Both models and humans benefited from more-predictable vs. less-predictable sentences. However, using isolated-word WERs as a baseline revealed that sentence context affected models and humans differently: humans benefited only from high-predictability sentences, while models benefited somewhat from *any* sentence context. Humans outperformed models on isolated words, suggesting that Whisper may have a restricted distribution of single-word utterances or may need lengthier acoustic context than humans.

Findings suggest that more inclusive, varied training data may yield more generalizable ASR. Potential for using ASR to model human speech adaptability is discussed.