AI vs. Human (Automatic) Speech Recognition: Silence-Replacement Paradigm as a Diagnostic

This study tests how vowels and consonants contribute to sentence-level word recognition in automatic speech recognition (ASR), using a silence-replacement paradigm modeled on classic human-perception research. We recorded 48 English sentences (adopted from Aldholmi, 2018) divided into two sets: 24 with a symmetrical (balanced vowel-to-consonant) ratio and 24 with an asymmetrical (consonant-heavy≈10 more consonants per sentence) ratio. For each sentence we created two processed versions: *CO* (consonant-only; vowels replaced by silence of equal duration) and *VO* (vowel-only; consonants replaced by silence). We then submitted all stimuli to two state-of-the-art ASR systems, TurboScribe (ChatGPT-integrated) and Whisper, and quantified word recognition as the percentage of original words correctly transcribed.

We modeled item-level recognized/total words with binomial generalized linear models and cluster-robust standard errors by sentence, estimating the effect of *Segment* (VO vs. CO) within *System* (TurboScribe, Whisper) and *Ratio* (Symmetrical vs. Asymmetrical). In the symmetroical set**,** VO outperformed CO: mean recognition was ~51% (VO) vs. ~45–46% (CO), consistent with sentence-level vowel advantages reported in human English listeners. In the asymmetrical set, the pattern reversed: CO averaged ~55% recognition, whereas VO collapsed to ~6–7%. Within-cell contrasts showed a nonsignificant VO advantage for symmetrical items in both TurboScribe and Whisper (ORs ≈1.8–1.9, p's > .14), but a large, reliable CO advantage for asymmetrical items in both systems (ORs ≈0.07 for VO vs. CO; p's < .001), indicating a strong *Segment-Ratio* interaction.

Methodologically, replacing one segment class with silence (equal in duration to the removed segments) isolates segment-type contributions while preserving temporal scaffolding. The symmetrical results suggest that, even for ASR, vowels tend to provide robust cues at the sentence level, echoing human data that attribute sentence-context benefits to vowel-borne envelope and suprasegmental information. However, when the vowel inventory is depleted by design (asymmetrical set), vowel-only input becomes too sparse to sustain recognition, and consonants carry the day. Thus, the intelligibility advantage is not an intrinsic property of vowels or consonants alone; it depends on segmental ratio and available temporal-contextual cues. *Theoretically*, these findings bridge psycholinguistic results and engineering practice: ASR systems mirror human-like reliance on vowel information in sentences when vowels are sufficiently available**,** but they pivot to consonant information when vowels are scarce. *Practically*, segment-aware preprocessing and training corpora that balance segmental distributions may improve ASR robustness under extreme degradations. More generally, the outcomes underscore that segment type and segment ratio jointly shape ASR performance, and that silence-replacement is a useful diagnostic for probing what cues modern systems actually use.