AI vs. human (automatic) speech recognition: silence-replacement paradigm as a diagnostic

Yahya Aldholmi
King Saud University, Saudi Arabia

https://doi.org/10.36505/TheLinguisticProceedings/2025/17/02/002/000688

Abstract

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. I recorded 48 English sentences divided into two sets: 24 with a symmetrical ratio and 24 with an asymmetrical ratio. For each sentence I created two processed versions: CO (consonant-only; vowels replaced by silence) and VO (vowel-only; consonants replaced by silence). I then submitted all stimuli to two state-of-the-art ASR systems, TurboScribe and Whisper, and quantified word recognition as the percentage of original words correctly transcribed. When the material was symmetrical, VO speech outperformed CO speech, mirroring human patterns. However, with asymmetrical material, this advantage reversed dramatically, showing a strong interaction between segment type and stimulus structure.

Keywords: vowel importance, consonant importance, ASR, English, silence-replacement paradigm

Introduction

Consonants provide more information at the word level in many languages, such as English, Dutch, and Spanish (e.g., Van Ooijen, 1996; Cutler et al., 2000), while vowels provide more information at the sentence level in some languages such as English (e.g., Cole et al., 1996; Fogerty et al., 2012). In some tonal languages such as Chinese, vowels contribute to speech recognition and intelligibility at both word and sentence levels (e.g., Chen et al., 2013; 2015), while the opposite has been reported in Semitic languages such as Arabic (e.g., Aldholmi, 2018; Aldholmi & Pycha, 2023). The contribution of vowels and consonants is not a purely linguistic topic but has many implications in other relevant fields, such as human-machine interaction, and is important for understanding and developing automatic speech recognition (ASR) systems and hearing aids (e.g., Yan, Chen, and Li, 2025). In previous studies, silence- or noise-replacement was used in preparing the experimental stimuli, where in one version of the stimuli, vowels were replaced by silence or noise (consonant-only, CO) that was equal in duration to the original segment, while consonants were replaced in another version (vowel-only, VO). The human ability for speech recognition in such conditions varies due to both stimuli-related factors (e.g., word vs. sentence) and language-specific factors (e.g., concatenative vs. nonconcatenative), as well as the interaction thereof. Hence, this study attempts to examine how AI performs on

speech recognition when the silence-replacement paradigm is utilized in stimuli from a concatenative language, English. The silence-replacement paradigm could be a potential diagnostic for the future development of ASR systems.

Methods

I recorded 48 English sentences (adopted from Aldholmi, 2018) of an approximate length (*M* = 6 words per sentence), divided into two sets: 24 with a symmetrical (balanced vowel-to-consonant) ratio and 24 with an asymmetrical (consonant-heavy, with approximately 10 more consonants per sentence) ratio. For each sentence, I created two processed versions: CO and VO. Figure 1 below shows an example to illustrate the silence-replacement method.

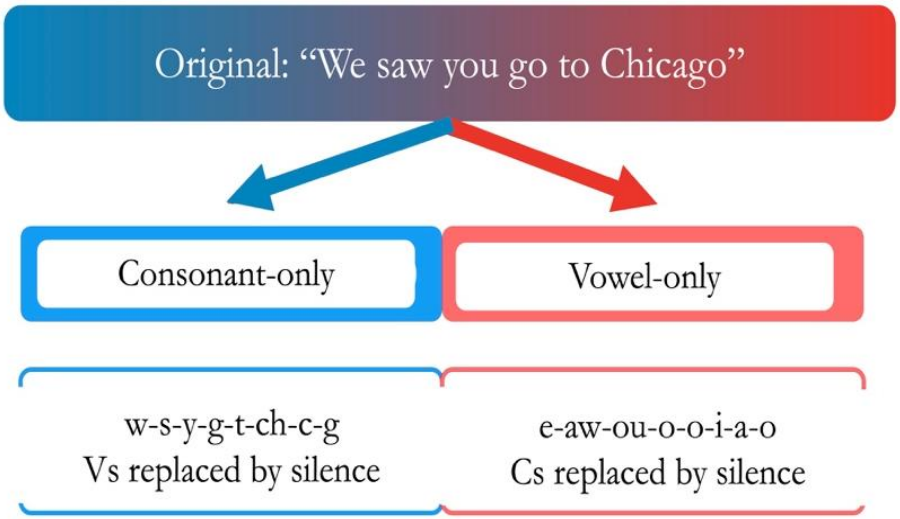

Figure 1. Illustration of the silence-replacement paradigm.

I then submitted all stimuli to two state-of-the-art ASR systems, TurboScribe (ChatGPT-integrated) and OpenAI/Whisper. The stimuli were sent to the two systems simultaneously to avoid any influence from time or system updates, and counterbalancing was used to prevent any impact from the order of conditions. I then quantified word recognition as the count and percentage of original words correctly transcribed by each system.

Results and discussion

I modeled the proportion of recognized words per item using binomial generalized linear models with cluster-robust standard errors by sentence, estimating the effect of segment (VO vs. CO) within each system (TurboScribe,

Whisper) and ratio (Symmetrical vs. Asymmetrical). In the symmetrical set, VO outperformed CO: mean recognition was approximately 51% (VO) versus 45–46% (CO), consistent with sentence-level vowel advantages reported in human English listeners. In the asymmetrical set, the pattern reversed: CO averaged approximately 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. Figures 2 and 3 below summarize the results.

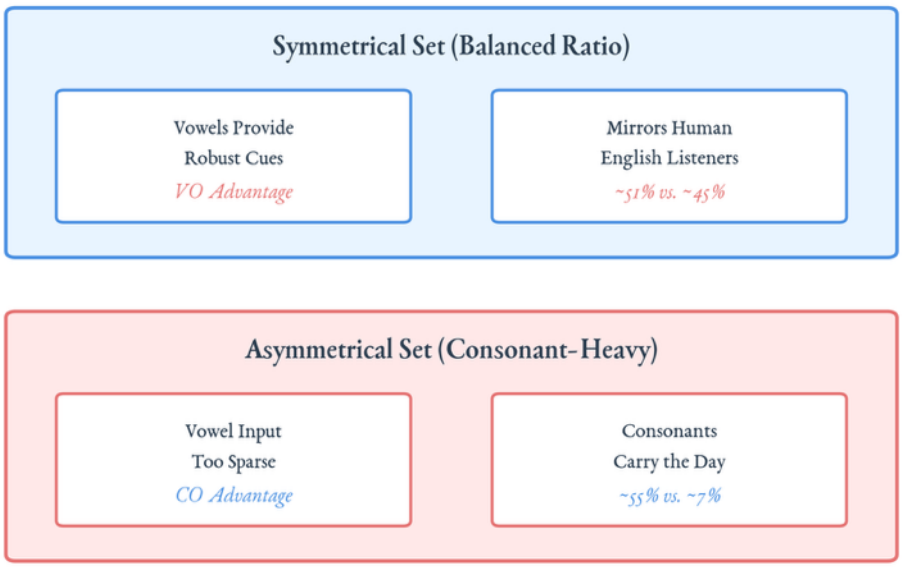

Figure 3. ASR performance by segment type and vowel-to-consonant ratio.

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), VO input becomes too sparse to sustain recognition, and consonants carry the load. Thus, the intelligibility advantage is not an intrinsic property of vowels or consonants alone; it depends on segmental ratio and available temporal-contextual cues.

Conclusion

This study tested how vowels and consonants contribute to sentence-level word recognition in ASR using a silence-replacement paradigm modeled on classic human-perception research. The 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. Segment-aware preprocessing and training corpora that balance segmental distributions may improve ASR robustness under extreme degradations. 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 use.

References

Aldholmi, Y. 2018. Segmental contributions to speech intelligibility in nonconcatenative vs. concatenative languages. The University of Wisconsin-Milwaukee.

Aldholmi, Y., Pycha, A. 2023. Segmental contributions to word recognition in Arabic sentences. Poznan Studies in Contemporary Linguistics, 59(2), 257-287.

Chen, F., Wong, L.L., Wong, E.Y. 2013. Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility. The Journal of the Acoustical Society of America, 134(2), EL178-EL184.

Chen, F., Wong, M.L., Zhu, S., Wong, L. L. 2015. Relative contributions of vowels and consonants in recognizing isolated Mandarin words. Journal of Phonetics, 52, 26-34.

Cole, R.A., Yan, Y., Mak, B., Fanty, M., Bailey, T. 1996. The contribution of consonants versus vowels to word recognition in fluent speech. In 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings (Vol. 2, pp. 853-856). IEEE.

Cutler, A., Sebastián-Gallés, N., Soler-Vilageliu, O., Van Ooijen, B. 2000. Constraints of vowels and consonants on lexical selection: Cross-linguistic comparisons. Memory & cognition, 28(5), 746-755.

Fogerty, D., Kewley-Port, D., Humes, L. E. 2012. The relative importance of consonant and vowel segments to the recognition of words and sentences: Effects of age and hearing loss. The Journal of the Acoustical Society of America, 132(3), 1667-1678.

Van Ooijen, B. 1996. Vowel mutability and lexical selection in English: Evidence from a word reconstruction task. Memory & Cognition, 24(5), 573-583.

Yan, Y., Chen, F., Li, J. 2025. An overview of the impacts of vowels and consonants in speech understanding and their applications. npj Acoustics, 1, 1-8.