Speech rate perception and interlocutor identification in human-directed vs. device-directed speech

Yahya Aldholmi, May Al-Sager, Arwa Alsahafi, Reema Alshiddi
King Saud University, Saudi Arabia

https://doi.org/10.36505/TheLinguisticProceedings/2025/16/01/001/000661

Abstract

This study investigates how listeners perceive differences between human-directed and device-directed speech, focusing on speech rate and interlocutor identification. Seventy-eight native Arabic speakers (aged 19–22; M = 20.46, SD = 1.11) participated in two tasks: rating the speed of 30 short recordings and determining whether each sample was directed toward a person or a device. The results showed that device-directed speech was consistently perceived as faster, while human-directed speech enabled more accurate interlocutor identification. Statistical analyses confirmed that these differences were significant, with moderate effect sizes. The findings suggest that devices produce speech efficiently but lack the natural variability that characterizes human communication. Incorporating more dynamic and expressive features into voice systems could improve user engagement. Future research should consider cultural differences and emotional tone in shaping speech perception.

Keywords: speech perception, human-directed speech, device-directed speech, speech rate, interlocutor identification

Introduction

Although voice assistants are direct and efficient, users still perceive key differences between device and human speech. Devices often sound faster and more precise, lacking the warmth and adaptability of human voices (Vonessen et al., 2024). Speech rate is a crucial characteristic in such perception; faster speech is efficient but less personal, whereas slower speech improves understanding and engagement (Huiyang & Min, 2022). While device speech rate has been studied in languages such as English (Jones et al., 2007) and Arabic (Aldholmi et al., 2021), the rate of speech directed to devices in Arabic has not. Interlocutor identification, which relies on subtle acoustic variations, is also more reliable with human speech (Zellou et al., 2023). Historically, since humans only spoke to other humans, listener identification was not a critical research topic; the emergence of AI-based systems, however, has opened new avenues for inquiry. Thus, this study explores how Arabic listeners evaluate speech rate in human- versus device-directed speech and their ability to identify the intended

© The International Linguistic Society

interlocutor (human or device). The findings will offer insights for creating more human-like and engaging voice systems.

Methodology

A within-subjects experimental design was employed. Seventy-eight native Arabic speakers aged 19 to 22 (M = 20.46, SD = 1.11) participated and were evenly divided into two counterbalanced groups, each exposed to a different order of stimulus presentation to control for order effects. Each participant completed a total of 60 trials. In the first task, listeners rated 30 recordings (15 human-directed and 15 device-directed) on a 7-point Likert scale ranging from "very fast" to "very slow." In the second task, the participants listened to another 30 recordings and indicated whether each sample was directed toward a human or a device. The stimuli were standardized to control for potential confounds. Each recording contained 6- to 10-word utterances (M = 7.8 words, SD = 1.2) spanning 13–36 syllables (M = 24.5 syllables, SD = 5.1) and lasting 2.2–6.3 seconds (M = 4.25 seconds, SD = 1.1) and was produced by the same speaker to ensure consistency in voice characteristics. For example, /wɒt ɪz ðə məʊst əˈtræktɪv ˈkʌlə tuː juː/ ("What is the most attractive color to you?"), a sentence in Modern Standard Arabic, was typical of the recordings used. Data were collected electronically under controlled listening conditions using noise-canceling headphones. The Wilcoxon signed-rank test was applied to assess differences between conditions, and effect sizes were computed to evaluate practical significance.

Results

The findings revealed apparent differences in how listeners perceived the two types of speech. Device-directed speech was generally judged to be faster, with an average rating of 4.09 (SD = 1.07) compared to 3.56 (SD = 1.04) for human-directed speech. This difference proved to be statistically significant (Z = –12.40, $p < .001$) and showed a medium effect size (r = .45). These results confirm a significant difference in perceived speed, supporting the first hypothesis—that device-directed speech would be perceived as faster than human-directed speech.

Regarding interlocutor identification, participants were more accurate when recognising human-directed speech. The average accuracy was 0.83 (SD = 0.12), whereas device-directed speech was identified less accurately at 0.78 (SD = 0.15). This difference was significant (Z = –9.23, $p < .001$) and had a medium effect size (r = .39).

These results support the study's hypotheses that listeners perceive device-directed speech as faster and achieve more accurate interlocutor identification with human-directed speech.

As illustrated in Figure 1, the ratings distribution further emphasizes these differences. Participants' judgments of device speech speed clustered tightly

around higher values, with a median near 5.0, indicating strong consensus that it was faster. By contrast, ratings for human speech centered around 4.0 and displayed slightly greater variability, suggesting less uniformity in perception. The boxplot comparison clearly shows that device-directed speech was consistently perceived as faster across participants, strengthening the statistical findings.

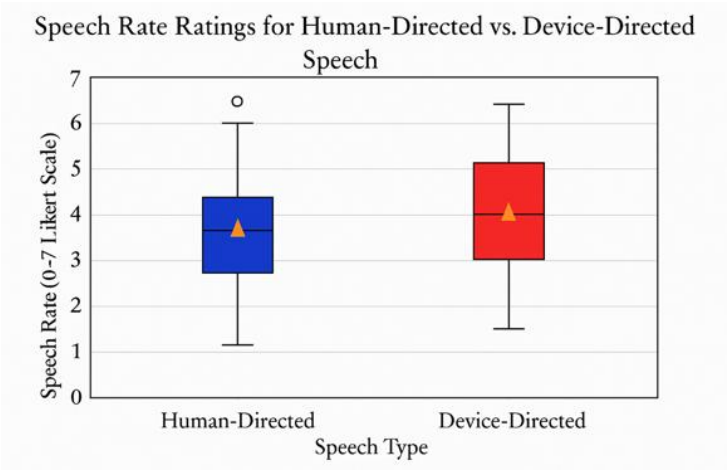

Figure 1. Boxplot of speech rate ratings for Human-Directed vs. Device-Directed Speech.

Discussion

Even though the same interlocutor produced all the recordings, listeners consistently rated device-directed speech as faster than human-directed speech. This finding suggests that people instinctively speak differently depending on the listener. When talking to a device, they tend to speed up and flatten their tone, perhaps aiming for clarity. While this delivery style resembles synthetic voices such as Google TTS (Aldholmi et al., 2021), human speech still carries minor variations that make it sound more natural to human listeners. These small cues matter. In the device-directed samples, participants had greater difficulty determining the intended recipient of the speech; however, they better understood the interlocutor's message since human-directed speech featured more tonal and rhythmic variations. These results remind us that communication requires connection as much as clarity. Voice technology must sound like humans to feel more human, not just to convey information rapidly.

Conclusion

This study illustrated that people's speech patterns naturally vary according to who or what they speak to. The difference was evident even when the same

interlocutor was used. Speech directed at devices was perceived as faster, while speech meant for humans felt more expressive and easier to connect with. That difference had real effects. Listeners could more accurately ascertain whom human-directed speech was meant for, demonstrating that tone and rhythm matter as much as clarity. These results highlight that communication is not just about being understood. Achieving the ideal mix between natural expression and efficiency will be crucial as speech devices advance. In addition, systems should mimic human speech patterns. Future studies can build on these findings by examining how these speech patterns appear in real conversations, across languages, and in emotionally rich situations.

References

Aldholmi, Y., Aldhafyan, R., Alqahtani, A. 2021. Perception of Standard Arabic synthetic speech rate. Interspeech 2021, 1704–1707. https://doi.org/10.21437/Interspeech.2021-39

Huiyang, S., Min, W. 2022. Improving interaction experience through lexical convergence: The prosocial effect of lexical alignment in human-human and human-computer interactions. International Journal of Human-Computer Interaction, 38(1), 28–41. https://doi.org/10.1080/10447318.2021.1921367

Jones, C., Berry, L., Stevens, C. 2007. Synthesized speech intelligibility and persuasion: Speech rate and non-native listeners. Computer Speech & Language, 21(4), 641–651. https://doi.org/10.1016/j.csl.2007.03.001

Vonessen, J., Aoki, N. B., Cohn, M., Zellou, G. 2024. Comparing perception of L1 and L2 English by human listeners and machines: Effect of interlocutor adaptations. Journal of the Acoustical Society of America, 155(5), 3060–3070. https://doi.org/10.1121/10.0025930

Zellou, G., Cohn, M., Pycha, A. 2023. Listener beliefs and perceptual learning: Differences between device and human guises. Language, 99(4), 692–725. https://dx.doi.org/10.1353/lan.2023.a914191