Speech Rate and Interlocutor Identification in Human-Directed vs Device-Directed Arabic Speech

Received 2026-07-06
Accepted 2026-09-19
Published 2026-10-17

May Al-Sager[1], Yahya Aldholmi [2],
[1]Technical and Vocational Training Corporation,[2]Department of Linguistics, King Saud University, Saudi Arabia
malsager@tvtc, gov,sa [1], yaldolmi@ksu.edu.sa[2].

Abstract
This study investigates speech rate as both a perceptual cue and a production strategy in Arabic human-directed and device-directed speech. The study comprised two phases. The perception phase examined whether listeners perceive differences in speech rate between human-directed and device-directed speech and whether those differences aid interlocutor identification; 78 native Arabic listeners completed speech-rate rating and speaker-identification tasks. The production phase investigated speech-rate production across interlocutor conditions and sentence types in Najdi Arabic through acoustic analysis of recorded speech. Listeners consistently perceived device-directed speech as faster than human-directed speech and used these temporal differences to identify the intended interlocutor. Production results showed that speakers systematically adjusted speech rate across communicative contexts and sentence types, with shorter sentences eliciting faster rates and speech to a familiar human interlocutor eliciting faster rates than speech to an artificial device. Together, the findings reveal a coupling between speech production and perception, indicating that speech rate functions as a shared communicative cue linking speaker adaptation and listener interpretation. The research contributes to work on speech adaptation, Arabic temporal acoustics, and human–technology interaction by demonstrating that speech rate is not merely a phonetic property of speech but a meaningful temporal signal in communication.
Keywords: Device-Directed Speech; Human-Directed Speech; Speech Perception; Speech Production; Speech Rate

INTRODUCTION

Speech rate (SR) is a fundamental temporal property of spoken language, defined as the number of syllables produced per unit of time, including both articulated speech and pauses (Jacewicz et al., 2010). Because it incorporates pausing, SR reflects speech planning and communicative structuring rather than articulatory tempo alone, providing a comprehensive index of timing in natural communication. A foundational insight from the literature is that speakers do not produce speech at a fixed rate but continuously adapt their temporal behavior to their audience and communicative goals (Bell, 1984; Lindblom, 1990). Speech tends to become slower and more clearly articulated when intelligibility is at risk—for instance, when addressing non-native listeners or individuals with hearing difficulties—and faster and more reduced when communication is easy (Smiljanić & Bradlow, 2005).

The rapid expansion of voice-activated technologies has introduced communicative situations in which speakers address not only other humans but also artificial systems such as Siri, Alexa, and Google Assistant. A growing body of research shows that device-directed speech

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

DOI: 10.18860 /ijazarabi.V9i4.44864
ISSN(print): 2620-5912 | ISSN(online):2620-5947
ejournal.uin-malang.ac.id/index.php/ijazarabi/index

(DDS) differs systematically from human-directed speech (HDS) in pitch, articulation, and timing, suggesting that speakers treat technological systems as a distinct category of interlocutor (Cohn & Zellou, 2021b; Cohn et al., 2022). Most of this work, however, has concentrated on speech production in English, leaving open the question of whether listeners can use temporal cues such as SR to infer the type of interlocutor that a given utterance was directed toward. Arabic offers a particularly informative context for examining these questions, given its phonologically significant use of segment duration and its considerable dialectal diversity (Watson, 2007). Although the temporal acoustics of Arabic have been studied in some detail, SR in natural conversational Arabic—and its role in distinguishing communicative contexts or interlocutor types—remains underexplored. Whether SR functions as a perceptual cue to interlocutor identity, and whether speakers adjust SR when addressing human versus artificial interlocutors, has not been tested in Arabic.

The present study addresses this gap by investigating SR as both a production strategy and a perceptual cue in Arabic HDS and DDS. It comprises two complementary phases. A perception phase examines whether native Arabic listeners perceive differences in SR between HDS and DDS and whether those differences support above-chance identification of the intended interlocutor. A production phase examines SR production in Najdi Arabic across three interlocutor conditions—an AI chatbot (ChatGPT), a familiar human, and an unfamiliar human—and two sentence types, with both adolescent and adult speakers. The central hypothesis is that SR varies systematically across interlocutor conditions in both production and perception, and that these variations contribute to interlocutor identification. The remainder of the paper reviews the relevant literature, describes the study and its results, and discusses their implications for speech adaptation and human–technology interaction (Cohn et al., 2021).

SR is a crucial component of spoken communication that affects both intelligibility and listener perception. Slower rates frequently improve comprehension in difficult listening conditions: Picheny et al. (1986) demonstrated that clear speech, characterized by slower SR and greater articulatory precision, substantially enhances intelligibility for listeners with hearing impairment, and Smiljanić & Bradlow (2005) found that speakers reduce their tempo to sharpen acoustic contrasts and aid comprehension. SR interacts with vowel length, consonant articulation, and prosodic structure, and varies with linguistic and sociolinguistic factors (Jacewicz et al., 2010). Speakers may also adjust SR in response to their interlocutor: Giroud et al. (2024) reported stronger SR convergence during natural than artificially altered speech, suggesting that perceived naturalness shapes adaptation, while Buz et al. (2016) showed that communication failures prompt dynamic hyper-articulation. Collectively, these findings establish SR as a temporal signal that shapes both production and perception, though its role in signaling communicative context remains unclear.

The way SR and other temporal properties are realized depends in part on the language itself, and Arabic provides an especially rich case. Temporal structure plays an important role in Arabic, where segment duration and SR shape the timing and rhythm of speech (Jacewicz et al., 2010; Watson, 2007). Arabic is not a single uniform language but a family of related dialects whose differing phonological systems affect vowel quality, consonant realization, and syllable structure, and thus temporal patterns (Watson, 2007). Vowel length is a particularly salient feature: short and long vowels form a clear durational contrast that is phonologically rather than purely

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

physiologically determined (Mitleb, 1984; Alghamdi, 1998). Temporal effects also extend across segment boundaries; Ferrat & Guerti (2011) showed that geminate consonants are longer than singletons and influence neighboring vowels, with pre-geminate vowels shortening and post-geminate vowels lengthening.

Acoustic measurements confirm the strength of the vowel-length contrast. Aldholmi (2022) reported short vowels averaging about 62 ms and long vowels about 164 ms, a difference of roughly 102 ms, although realizations vary across Saudi, Sudanese, Jordanian, and Palestinian Arabic, with short-to-long ratios generally between 0.39 and 0.45 (Mitleb, 1984; Alghamdi, 1998). More recently, Aldholmi (2024) proposed two metrics for characterizing temporal structure—a duration metric (173.03 ms in Arabic) reflecting overall timing and a difference metric (64.08 ms) capturing variation between vowel types—providing a finer-grained description of temporal contrasts. Beyond segment duration, SR captures the global pace of speech, usually measured in syllables per second (Jacewicz et al., 2010). Research on Arabic SR in natural contexts remains limited, but work on synthetic speech indicates that listeners are sensitive to SR variation, treating rates around 4.5 syllables per second as slow and rates of 5.0–5.5 as more natural (Aldholmi et al., 2021a). Its role in identifying communicative context or interlocutor type, however, is still underexplored.

One communicative context in which such adaptation has become increasingly visible is interaction with technology. The spread of voice-activated systems has created new forms of human–computer interaction, and because communicating with systems that rely on automatic speech recognition (ASR) differs from ordinary human conversation, speakers often adjust their speech when addressing such devices. Cohn & Zellou (2021b) found that speakers modified both pitch and overall prosodic structure when addressing Alexa rather than a human, indicating that they treat technology as a distinct kind of communication partner. Cohn et al. (2021a) further showed that SR is adjusted dynamically during interaction with an Alexa socialbot, with speakers slowing down following recognition errors, suggesting that SR functions as an adaptive strategy for improving communication with voice assistants (Cohn & Zellou, 2021).

Subsequent work has broadened this line of research. Cohn et al. (2022) analyzed Siri-directed speech and found consistent differences in articulation and vowel space relative to HDS, with speakers producing clearer, more deliberately articulated output when addressing the device. Sociolinguistic and developmental variation has also been documented: Cohn et al. (2024a) found that African American English speakers adjusted pitch and rate according to the listener's perceived identity, and Cohn et al. (2024b) reported that both children and adults modify their speech for technological interlocutors, with children showing especially exaggerated patterns. Comparable effects have been observed beyond English, including in German (Siegert & Krüger, 2021). Although these studies offer valuable insight into production, most concentrate on production rather than perception, and relatively little work has examined whether listeners can use acoustic cues such as SR to determine the type of interlocutor in an interaction.

These patterns of adaptation—toward both human and technological interlocutors—are well captured by several theoretical frameworks that treat speech as flexible rather than fixed. Bell's (1984) Audience Design theory holds that speakers craft their speech with the audience in mind, modifying pronunciation, lexis, syntax, and prosodic features such as SR and intonation according to who they are addressing. Communication Accommodation Theory (Giles &

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Coupland, 1991) describes how speakers converge toward an interlocutor's speech patterns to signal solidarity or improve communication, or diverge from them to mark social distance or group identity; speakers commonly slow down and articulate more clearly when they anticipate comprehension difficulty, as with non-native listeners or those with hearing loss (Smiljanić & Bradlow, 2005).

Lindblom's (1990) Hyper- and Hypo-Speech (H&H) Theory offers a production-oriented account in which speakers balance clarity against efficiency along a continuum: hypo-speech involves reduced articulation and faster rates when communication is easy, whereas hyper-speech involves clearer articulation and slower rates when communication is difficult. Together, these frameworks extend naturally to human–technology interaction, since speakers may treat voice assistants as conversational partners that require slower, clearer, or prosodically adjusted speech to improve recognition (Cohn & Zellou, 2021b; Cohn et al., 2021a). They thus provide a conceptual grounding for examining how SR functions as a cue across different interlocutors. If speakers systematically adjust SR across interlocutors, a further question is whether listeners can recover that information. Speech perception is the process by which listeners interpret acoustic input to identify linguistic units and derive meaning. Although early accounts emphasized acoustic and phonetic cues, later work established that listeners also draw on cognitive, social, and contextual information, so that perception is now understood as combining bottom-up processing with top-down expectations (McGettigan, 2015). Listener expectations are central to this process. In Niedzielski's (1999) classic study, participants who were told a speaker was from Detroit or from Canada judged identical vowel sounds differently, showing that social assumptions shape perception. Hay & Drager (2010) similarly found that external information, such as social labels, influences how listeners perceive speech sounds (Steffman, 2021).

Acoustic properties themselves also shape perception. Temporal features such as SR and prosodic structure carry important information, and slower, more precise delivery tends to improve intelligibility in challenging conditions (Picheny et al., 1986; Smiljanić & Bradlow, 2005), implying that listeners exploit temporal cues when interpreting speech. Recent work has extended this question to human–technology interaction. Zellou et al. (2023b) showed that listeners' beliefs about whether speech came from a human or a device altered their perceptual learning even when the acoustic signal was held constant, demonstrating that beliefs about the interlocutor can shape processing itself. Vonessen et al. (2024) found that human listeners and ASR systems rely on different cues, with humans incorporating contextual and social information and machines depending largely on acoustic features. Despite this growing interest, few studies have tested whether listeners can use SR itself to identify the communicative context or the type of interlocutor, and fewer still have asked whether listeners can distinguish HDS from DDS on the basis of temporal features alone—an especially notable gap for Arabic that the present study addresses. A comparative overview of these and related studies is provided in Appendix A.

METHOD

To investigate SR as both a perceptual cue and a production strategy in Arabic HDS and DDS, a single study was conducted comprising two complementary phases. Phase 1 (perception) examined whether listeners perceive systematic differences in SR between HDS and DDS and whether those differences enable them to identify the intended interlocutor. Phase 2 (production)

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

examined whether speakers themselves adjust SR when addressing different interlocutors, and how such adjustments interact with sentence length and speaker age. Conducting both phases within a single study allows the temporal patterns that speakers produce to be evaluated against the temporal cues that listeners detect, providing an integrated view of SR in human–technology interaction.

Participants

Two groups of participants took part, one for each phase. In the perception phase, 78 female native Arabic speakers (aged 19–22 years; *M* = 20.46, *SD* = 1.11) served as listeners. They were recruited from a university setting, predominantly as undergraduate students, and reported no history of hearing or speech impairments. In the production phase, 20 native Najdi Arabic speakers served as talkers, divided into two age groups: adolescents (n = 10; ages 13–18) and adults (n = 10; ages 30–45). Production participants were recruited through social networks and local schools and reported normal hearing and no history of speech or language difficulties. All participants provided informed consent before taking part and were informed of their right to withdraw at any time without penalty; parental consent was additionally obtained for the adolescent talkers.

Both phases used interrogative sentences designed to reflect natural spoken Arabic. The perception stimuli consisted of interrogatives produced in matched HDS and DDS versions, each containing between 6 and 10 words (*M* = 7.8, *SD* = 1.2) and 13–36 syllables (*M* = 24.5, *SD* = 5.1); the two versions of each item were matched for lexical content, with only the interaction context varying. An example is given in (1):

(1) maː huːwa ˈakθaru alˈalwaːni ʒaːðibijah ladajka?
'What is the most attractive color to you?'

The production materials consisted of pre-written Najdi Arabic interrogatives divided into short (approximately 8 syllables; 3–5 words) and long (approximately 14–15 syllables; 6–10 words) types. The same set of sentences was used across all interaction conditions to control for lexical and structural variation. An example of a short interrogative is given in (2):

(2) wɛʃ ˈakθar ˈaklah tħibhaː?
'What food do you like the most?'

Procedure

In the perception phase, listeners completed two tasks. In the SR rating task, they heard each stimulus and rated its perceived speed on a seven-point Likert scale (1 = very slow, 7 = very fast). In the speaker-identification task, they indicated whether each stimulus was directed toward a human or a device. Stimuli were presented over headphones in a quiet room, and the order of presentation was randomized across listeners. In the production phase, talkers produced the same set of sentences under three interaction conditions: an AI chatbot (ChatGPT), a familiar human interlocutor, and an unfamiliar human interlocutor. In the ChatGPT condition, participants used a mobile interface to address the AI system; in the familiar-human condition, they interacted face-to-face with a known individual; and in the unfamiliar-human condition, they spoke to an unknown interlocutor by telephone. Each interaction lasted approximately five minutes, during which participants produced 10 sentences per condition.

In both phases, recordings were analyzed in Praat (Boersma & Weenink, 2023), with each interrogative utterance manually segmented using TextGrid annotation. SR was calculated as the

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

number of syllables divided by total utterance duration (syllables per second). Silent pauses were excluded from duration calculations, while filled pauses were retained; when an utterance was repeated, only the final production was included. For the production data, SR was analyzed using a mixed-effects linear model implemented in R, with interlocutor condition (ChatGPT vs. familiar human vs. unfamiliar human), sentence type (short vs. long), and age group (adolescents vs. adults) as fixed effects and participant as a random effect.

RESULTS AND DISCUSSION

Results show that the SR ratings, presented in Table 1, were consistently higher for DDS than HDS, indicating that listeners perceived DDS as faster.

Table 1. SR Mean Ratings for HDS vs. DDS

Speech Direction	Mean	SD	Median
HDS	3.56	1.04	3.5
DDS	4.09	1.07	4.0

Note. Ratings were measured using a 7-point Likert scale ranging from very slow to very fast.

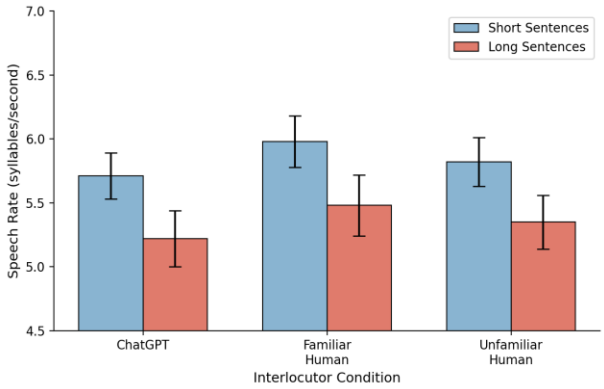
Table 2 presents identification accuracy for HDS and DDS stimuli. HDS was identified with higher accuracy than DDS, indicating that listeners found it easier to identify HDS than DDS. Both values exceeded the 50% chance level.

Table 2. Speaker Target Identification Accuracy

Speech Direction	Mean Accuracy	SD
HDS	0.83	0.12
DDS	0.78	0.15

Results show an overall mean production SR of 5.41 syllables per second (*SD* = 0.88), ranging from 1.97 to 8.40, with adults producing a slightly higher mean SR (*M* = 5.51, *SD* = 1.00) than adolescents (*M* = 5.31, *SD* = 0.73). The mixed-effects model revealed a significant main effect of sentence type: short sentences were produced at a higher SR than long sentences (β = 0.484, *p* = .001). A significant main effect of interlocutor condition was also found, with speech to a familiar human interlocutor produced at a higher SR than speech directed to ChatGPT (β = 0.299, *p* = .049). Figure 1 illustrates SR across interlocutor conditions and sentence types.

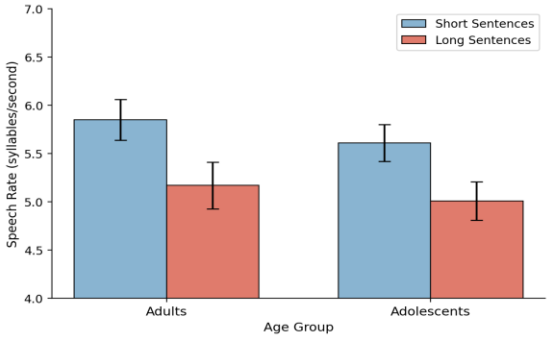
Figure 1. Mean SR Across Addressee Types And Sentence Length. Error Bars Represent ± 1 SE



The main effect of age group was not statistically significant (β = −0.248, *p* = .269), indicating that adolescents and adults did not differ reliably in overall SR. Figure 2 shows SR across age groups and sentence types.

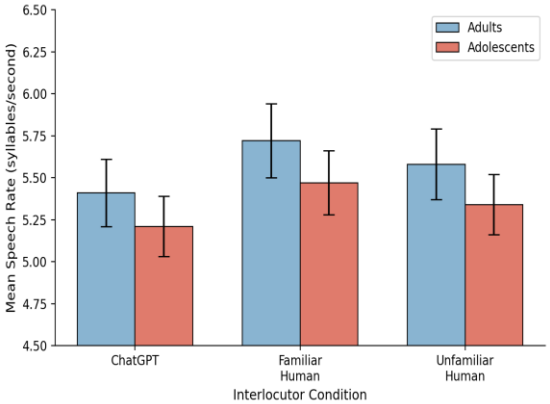
Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Figure 2. Mean Syllables Per Second By Age Group And Sentence Type



No significant interaction effects were observed (all $p > .10$), indicating that the effects of sentence type and interlocutor condition were comparable across age groups. Figure 3 presents mean SR across addressee types for each age group.

Figure 3. Mean Syllables Per Second Across Addressee Types For Adults And Adolescents



The present study examined SR as a perceptual cue and a production strategy in Arabic HDS and DDS. The findings from both experiments converge to indicate that SR varies systematically across communicative contexts and that these variations are both produced by speakers and detected by listeners. Taken together, the results suggest that SR functions as a meaningful temporal signal linking speaker adaptation and listener interpretation in human–technology interaction. The perception phase demonstrated that listeners perceived DDS as faster than HDS and used these perceived differences to identify the intended interlocutor at above-chance levels. This finding establishes SR as a functional perceptual cue in distinguishing communicative contexts, consistent with earlier work showing that listeners rely on temporal and rhythmic information when evaluating speech (Aldholmi & Park, 2016, 2019). The present results extend that literature in an important direction: SR differences arising from interlocutor type are not only perceptible but communicatively diagnostic. Listeners attended not to lexical content alone but to the speed and rhythm with which utterances were produced, using those temporal properties to draw inferences about whether speech was directed to a human or a device.

The asymmetry in identification accuracy between HDS (83%) and DDS (78%) is theoretically informative. Although both conditions substantially exceeded the 50% chance level, HDS was identified more reliably than DDS. One interpretation is that listeners have richer and more stable perceptual templates for HDS, given their extensive daily exposure to it, whereas the

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

acoustic characteristics of DDS may be less familiar and therefore less easily categorized. This explanation is consistent with the broader perception literature showing that listener experience and exposure shape the robustness of perceptual categories (Niedzielski, 1999; Hay & Drager, 2010). It also resonates with Zellou et al. (2023b), who showed that listener beliefs about whether speech originates from a human or a device modulate perceptual learning processes, even when the acoustic input is held constant. The present findings extend this line of work by demonstrating that temporal cues alone—without any manipulation of listener beliefs—are sufficient to support above-chance interlocutor identification in Arabic.

The perception results are also consistent with research on SR perception in Arabic. Aldholmi and Park (2016, 2019) showed that perceived SR depends not only on the number of syllables produced per second but also on broader acoustic patterns such as rhythm and indexical information. The present study adds to this picture by showing that interlocutor type constitutes a further dimension along which SR perception is organized. When listeners hear DDS, they rate it as faster and more device-like, suggesting that the temporal characteristics of DDS are sufficiently distinctive to support contextual categorization. The finding that listeners relied on SR to distinguish HDS from DDS is also consistent with Aldholmi et al. (2021a), who demonstrated that Arabic listeners are sensitive to SR variation in synthetic speech, and with Alshaikhmubarak and Aldholmi (2025, 2026), who found that rhythmic structure influences SR judgments across language backgrounds.

The production phase revealed that speakers systematically adjusted SR as a function of both sentence type and interlocutor condition. Shorter sentences were produced at a higher SR than longer sentences ($\beta = 0.484$, $p = .001$), a finding consistent with the well-established principle that longer utterances impose greater cognitive planning and articulatory demands, resulting in lower overall rates (Jacewicz et al., 2010). Of greater theoretical interest is the significant effect of interlocutor condition: speakers produced speech at a higher SR when addressing a familiar human interlocutor than when addressing ChatGPT ($\beta = 0.299$, $p = .049$). This suggests that speakers treat AI systems as a distinct communicative category, one that elicits a more deliberate and controlled speaking style relative to casual human interaction. Although unfamiliar human speech also showed a numerically lower SR than familiar human speech, this comparison did not reach statistical significance, indicating that the most pronounced adjustment was specifically between human interlocutors and the AI condition.

This pattern accords with the theoretical frameworks outlined in the literature review. From the perspective of Audience Design (Bell, 1984), speakers modify their speech in response to their assumptions about the audience's communicative needs and capacities. Addressing an AI system may trigger beliefs about its recognition limitations—for example, that it requires slower, more deliberate speech to process input accurately—leading to the SR reduction observed here. This interpretation is supported by Cohn and Zellou (2021b) and Cohn et al. (2021a), who found analogous SR adjustments in English speakers addressing Alexa, and by Cohn et al. (2022), who documented clearer, more hyper-articulated speech in Siri-directed versus human-directed speech. The present study demonstrates that comparable patterns extend to Arabic, specifically to Najdi Arabic, a variety that has received limited attention in the human–technology interaction literature. From the perspective of the H&H model (Lindblom, 1990), the slowing of SR in AI-directed speech can be understood as a shift toward the hyper-speech end of the continuum: when speakers

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

perceive potential difficulty in communication—whether due to the listener's linguistic background, hearing capacity, or, in this case, machine recognition constraints—they increase articulatory effort and reduce temporal compression.

The absence of a significant age effect (p = .269) is notable. Adolescents and adults produced comparable SRs across all conditions, suggesting that contextual factors—specifically interlocutor type and sentence length—exert a stronger influence on SR than developmental stage, at least within the age ranges sampled. This result aligns with Cohn et al. (2024b), who similarly found that children and adults both modify their speech when addressing technological systems, indicating that the capacity for speech adaptation to interlocutor type is not age-restricted. At the same time, the current sample was small and restricted in age range, so definitive conclusions about developmental trajectories would require broader participant groups. It is possible that more pronounced age-related differences would emerge with a wider developmental span or in more naturalistic interaction contexts.

A central contribution of the present study is its joint examination of production and perception within a single framework, which reveals a coherent alignment between the two phases. Listeners perceived DDS as faster than HDS (perception phase), while speakers produced speech at a lower SR when addressing an AI system than when addressing a familiar human (production phase). At first glance these results might appear contradictory—speakers slow down for AI, yet listeners perceive AI-directed speech as faster. However, these findings operate at different levels of analysis and likely reflect different mechanisms. In the perception phase, listeners judged prerecorded stimuli produced in naturalistic HDS and DDS contexts, and their ratings may reflect not only raw SR but also qualitative differences in prosodic structure, pause distribution, and articulatory precision that collectively create an impression of speed. In the production phase, the SR reduction for AI-directed speech was statistically modest and may be offset by reductions in pausing and hesitation behavior, which are more characteristic of casual human conversation. Together, the two phases suggest that SR functions as a shared communicative resource: speakers systematically adjust their temporal behavior across communicative contexts, and listeners extract and interpret those temporal patterns when making contextual judgments. This coupling is consistent with Smiljanıć and Bradlow's (2005) account of production–perception linkages in clear speech, and with the broader view that communicative adaptation involves coordinated speaker–listener dynamics rather than unidirectional production adjustments.

The present findings carry implications for the design and evaluation of voice-based AI systems. If speakers systematically modify their SR and temporal organization when addressing AI interlocutors, and if listeners can exploit these cues to identify communicative context, then ASR systems and voice interfaces should ideally be designed to accommodate the range of speaking styles that naturally emerge in human–technology interaction rather than optimizing exclusively for a single canonical speaking style. The results also suggest that the acoustic profile of AI-directed speech is sufficiently distinctive that it may require tailored processing strategies, particularly for languages like Arabic where temporal contrasts play a phonologically significant role. More broadly, as conversational AI systems become increasingly integrated into daily life, understanding how speakers adapt their temporal behavior to technological interlocutors becomes relevant not only for linguistic theory but also for the practical design of speech technologies that serve diverse linguistic populations.

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Finally, it is worth situating the current findings within the broader literature on speech adaptation. The adaptations documented here—slower SR for AI, faster SR for familiar humans, and sensitivity to sentence length—are not unique to human–technology interaction but reflect more general principles of speech planning and audience design. What is distinctive about the technological context is that the interlocutor's perceived communicative needs are based on assumptions about machine rather than human cognition. Speakers draw on their beliefs about how AI systems process speech, and those beliefs shape their temporal behavior in ways that are detectable by human listeners. This points to a reciprocal relationship between technological systems and the speakers who address them: as AI systems evolve and as users accumulate experience interacting with them, the acoustic norms of DDS may themselves shift. Longitudinal and cross-platform studies will be needed to track these developments and to determine how established the current patterns are across different Arabic-speaking communities and different AI interfaces.

CONCLUSION

This study examined SR as both a perceptual cue and a production strategy in Arabic HDS and DDS. Through two complementary phases, the findings demonstrated that SR varies systematically across communicative contexts and that listeners use these differences to identify interlocutor type. The perception phase showed that DDS was perceived as faster than HDS and that listeners could identify interlocutor type at above-chance levels using temporal information alone. The production phase showed that speakers adjusted SR as a function of interlocutor condition and sentence type, with AI-directed speech produced at a slower rate than speech to a familiar human, and shorter sentences produced at faster rates than longer ones. Together, the findings establish SR as a shared communicative cue linking speaker adaptation and listener interpretation in Arabic human–technology interaction. The study contributes to research on speech adaptation, Arabic temporal acoustics, and human–technology interaction by demonstrating that SR is not merely a phonetic property of speech but a meaningful temporal signal in communication.

Several limitations should be noted. The participant samples were relatively small and narrow in demographic scope, which may limit the generalizability of the findings. The interaction contexts examined represent a limited range of communicative situations, and only SR was analyzed as a potential cue, whereas pausing, pitch movement, and rhythm may also contribute to interlocutor identification. Future research should examine additional Arabic dialects and more diverse speaker populations, investigate how SR interacts with other prosodic cues, and explore speech adaptation across different AI systems and more naturalistic conversational tasks.

FUNDING STATEMENT

This research received no external funding.

CONFLICT OF INTEREST

The author(s) declare that there is no conflict of interest related to the research, authorship, or publication of this article.

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

REFERENCES

Aldhafeeri, M., Alqahtani, G., & Aldholmi, Y. (2022). *Speech rate and intelligibility of Arab Airlines' safety instructions*. In Proceedings of the 13th International Conference of Experimental Linguistics (ExLing 2022) (pp. 5–8). https://doi.org/10.36505/ExLing-2022/13/0002/000544

Aldholmi, Y. (2020). *Syllable rate vs. segment rate in perceived speech rate*. In Proceedings of the 11th International Conference of Experimental Linguistics. ExLing Society. https://doi.org/10.36505/ExLing-2020/11/0005/000420

Aldholmi, Y. (2022). *Medial vowel temporal acoustics in Arabic and Japanese polysyllabic words*. Archives of Acoustics, 47(3), 319–329. https://doi.org/10.24425/aoa.2022.142006

Aldholmi, Y. (2024). Statistical metrics for the temporal acoustics of durationally contrastive vocalics: A proposal tested with data from Arabic and Japanese. Archives of Acoustics. https://doi.org/10.24425/aoa.2024.148779

Aldholmi, Y. (2025). AI vs. human (automatic) speech recognition: Silence-replacement paradigm as a diagnostic. *Linguistic Proceedings Series*, 16th International Conference on Linguistic Research and Applications, Paris, France, *17*(2), 5-8. https://doi.org/10.36505/TheLinguisticProceedings/2025/17/02/002/0006

Aldholmi, Y., Aldhafyan, R., & Alqahtani, A. (2021a). Perception of Standard Arabic synthetic speech rate. *In Proceedings of Interspeech 2021* (pp. 1704–1707). https://doi.org/10.21437/Interspeech.2021-39

Aldholmi, Y., Alqahtani, A., & Aldhafyan, R. (2021b). Teen production and perception of dialectal Arabic speech rate. In *Proceedings of the 12th International Conference of Experimental Linguistics* (pp. 13–16). ExLing Society. https://doi.org/10.36505/ExLing-2021/12/0004/000477

Aldholmi, Y., & Park, H. (2016). Perception of speech rate in speech rate perception. *The Journal of the Acoustical Society of America, 140*(4), 3333. https://doi.org/10.1121/1.4969160

Aldholmi, Y., & Park, H. (2019). Interactions between temporal acoustics and indexical information in speech rate perception. In M. Alhawary (Ed.), *Studies in Arabic Linguistics* (Vol. 8). John Benjamins. https://doi.org/10.1075/sal.8.10ald

Alghamdi, M. M. (1998). A spectrographic analysis of Arabic vowels: A cross-dialect study. *Journal of King Saud University, 10*, 3–24. https://chss.ksu.edu.sa/sites/arts.ksu.edu.sa/files/imce_images/v10m190r1962.pdf

Alshaikhmubarak, L., & Aldholmi, Y. (2025). The influence of native language rhythm on speech rate perception in unfamiliar languages. *Journal of Arabic Sciences & Humanities, 18*(4), 9–49. https://doi.org/10.71422/0198-018-004-001

Alshaikhmubarak, L., & Aldholmi, Y. (2026). The effect of rhythmic system on the perception of L1, L2 and unfamiliar languages. Ijaz Arabi: *Journal of Arabic Learning, 9*(1), 64–82. https://doi.org/10.18860/ijazarabi.V9i1.32966

Bell, A. (1984). Language style as audience design. *Language in Society, 13*(2), 145–204. https://doi.org/10.1017/S004740450001037X

Bradlow, A. R., & Bent, T. (2002). The clear speech effect for non-native listeners. *The Journal of the Acoustical Society of America, 112*(1), 272–284. https://doi.org/10.1121/1.1487837

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

Buz, E., Tanenhaus, M. K., & Jaeger, T. F. (2016). Dynamically adapted context-specific hyper-articulation: Feedback from interlocutors affects speakers' subsequent pronunciations. *Journal of Memory and Language, 89*, 68–86. https://doi.org/10.1016/j.jml.2015.12.009

Cohn, M., Barreda, S., Graf Estes, K., Yu, Z., & Zellou, G. (2024a). Children and adults produce distinct technology- and human-directed speech. *Scientific Reports, 14*, 15611. https://doi.org/10.1038/s41598-024-66313-5

Cohn, M., Liang, K.-H., Sarian, M., Zellou, G., & Yu, Z. (2021a). Speech rate adjustments in conversations with an Amazon Alexa socialbot. *Frontiers in Communication, 6*, 671429. https://doi.org/10.3389/fcomm.2021.671429

Cohn, M., Mengesha, Z., Lahav, M., & Heldreth, C. (2024b). African American English speakers' pitch variation and rate adjustments for imagined technological and human addressees. *JASA Express Letters, 4*(4). https://doi.org/10.1121/10.0025484

Cohn, M., Segedin, B. F., & Zellou, G. (2022). Acoustic-phonetic properties of Siri- and human-directed speech. *Journal of Phonetics, 90*, 101123. https://doi.org/10.1016/j.wocn.2021.101123

Cohn, M., & Zellou, G. (2021b). Prosodic differences in human- and Alexa-directed speech, but similar local intelligibility adjustments. *Frontiers in Communication, 6*, 675704. https://doi.org/10.3389/fcomm.2021.675704

Cohn, M., & Zellou, G. (2021). Prosodic Differences in Human- And Alexa-Directed Speech, but Similar Local Intelligibility Adjustments. *Frontiers in Communication*, *6*(July), 1–13. https://doi.org/10.3389/fcomm.2021.675704

Ferrat, K., & Guerti, M. (2011). Acoustic and articulatory analysis of the gemination in Modern Standard Arabic. *Al-Lisaniyyat, 21*, 73–88. https://doi.org/10.61850/allj.v21i2.306

Field, A. (2013). Discovering statistics using *IBM SPSS statistics* (4th ed.). SAGE Publications.

Gessinger, I., Raveh, E., Steiner, I., & Möbius, B. (2021). Phonetic accommodation to natural and synthetic voices: Behavior of groups and individuals in speech shadowing. *Speech Communication, 127*, 43–63. https://doi.org/10.1016/j.specom.2020.12.004

Giles, H., & Coupland, N. (1991). Language: *Contexts and consequences.* Open University Press.

Giroud, J., Lei, J., Phillips, K., & Davis, M. (2024). Behavioral evidence for higher speech rate convergence following natural than artificial time altered speech. *Proceedings of Interspeech 2024.* https://doi.org/10.21437/Interspeech.2024-1001

Hay, J., & Drager, K. (2010). Stuffed toys and speech perception. *Linguistics, 48*(4), 865–892. https://doi.org/10.1515/ling.2010.027

Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America, 128*(2), 839–850. https://doi.org/10.1121/1.3459842

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In W. J. Hardcastle & A. Marchal (Eds.), *Speech production and speech modelling* (NATO ASI Series, Vol. 55, pp. 403–439). Springer. https://doi.org/10.1007/978-94-009-2037-8_16

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)

McGettigan, C. (2015). The social life of voices: Studying the neural bases for the expression and perception of the self and others during spoken communication. *Frontiers in Human Neuroscience, 9*, 129. https://doi.org/10.3389/fnhum.2015.00129

Mitleb, F. (1984). Vowel length contrast in Arabic and English: A spectrographic test. *Journal of Phonetics, 12*(3), 229–235. https://doi.org/10.1016/S0095-4470(19)30879-4

Niedzielski, N. (1999). The effect of social information on the perception of sociolinguistic variables. *Journal of Language and Social Psychology, 18*(1), 62–85. https://doi.org/10.1177/0261927X99018001005

Picheny, M. A., Durlach, N. I., & Braida, L. D. (1986). Speaking clearly for the hard of hearing I: Intelligibility differences between clear and conversational speech. *Journal of Speech and Hearing Research, 29*, 434–446. https://doi.org/10.1044/jshr.2904.434

Siegert, I., & Krüger, J. (2021). Speech melody and speech content didn't fit together: Differences in speech behavior for device-directed and human-directed interactions. *In Advances in Data Science* (pp. 65–95). https://doi.org/10.1007/978-3-030-51870-7_4

Smiljanić, R., & Bradlow, A. R. (2005). Production and perception of clear speech in Croatian and English. *The Journal of the Acoustical Society of America, 118*(3), 1677–1688. https://doi.org/10.1121/1.2000788

Steffman, J. (2021). Rhythmic and speech rate effects in the perception of durational cues. *Attention, Perception, and Psychophysics*, *83*(8), 3162–3182. https://doi.org/10.3758/s13414-021-02334-w

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. *The Journal of the Acoustical Society of America, 155*(5), 3060–3070. https://doi.org/10.1121/10.0025930

Watson, J. C. E. (2007). The phonology and morphology of Arabic. Oxford University Press.

Zellou, G., & Cohn, M. (2023a). Clear speech in the new digital era: Speaking and listening clearly to voice-AI systems. *The Journal of the Acoustical Society of America, 153*(3). https://doi.org/10.1121/10.0018378

Zellou, G., Cohn, M., & Pycha, A. (2023b). 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

Zellou, G., & Holliday, N. (2024). Linguistic analysis of human-computer interaction. *Frontiers in Computer Science, 6,* 1384252. https://doi.org/10.3389/fcomp.2024.1384252

Copyright © 2026, This is an open access article under the CC BY-SA 4.0 license
(https://creativecommons.org/licenses/by-sa/4.0/)