frontiers in Psychology

TYPE Original Research
PUBLISHED 02 November 2026
DOI 10.3389/fpsyg.2026.1925812

Check for updates

OPEN ACCESS

EDITED BY
Luca Simione,
UNINT - Università degli studi Internazionali di Roma, Italy

REVIEWED BY
Zahra Jeddi,
Shiraz University of Medical Sciences, Iran
Mohammad Ali Salah Bani Younes,
Al al-Bayt University, Jordan

*CORRESPONDENCE
Yahya Aldholmi
✉ yaldholmi@ksu.edu.sa

RECEIVED 01 July 2026
REVISED 25 September 2026
ACCEPTED 19 October 2026
PUBLISHED 02 November 2026

CITATION
Aldholmi Y (2026) Mapping ASR systems onto the human auditory spectrum: speech-in-noise recognition of Arabic in normal-sighted, blind, and hearing-impaired listeners. *Front. Psychol.* 17:1925812. doi: 10.3389/fpsyg.2026.1925812

COPYRIGHT
© 2026 Aldholmi. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.

Mapping ASR systems onto the human auditory spectrum: speech-in-noise recognition of Arabic in normal-sighted, blind, and hearing-impaired listeners

Yahya Aldholmi*

Department of Linguistics, College of Language Sciences, King Saud University, Riyadh, Saudi Arabia

Background/objectives: This study examines how human listeners and automatic speech recognition (ASR) systems process speech under varying background noise levels.
Methods: Three groups of human participants—normal-sighted (NSLs), blind (BLs), and hearing-impaired listeners (HILs)—were compared with two state-of-the-art ASR systems, Whisper and TurboScribe. Forty Arabic sentences (20 imperative, 20 interrogative) were presented at four signal-to-noise ratios (SNRs: +10, +5, −5, −10 dB) using passenger boarding noise to simulate realistic acoustic environments.
Results: Accuracy declined with increasing noise for all groups, although rates varied substantially. BLs performed best, maintaining 85% accuracy even at −10 dB SNR, whereas HILs showed the steepest decline. The ASR systems achieved near-human performance in moderate noise but degraded sharply under severe masking, mirroring HIL trends. Statistical analysis confirmed significant group differences across SNRs ($p < 0.05$). A linear mixed-effects analysis further revealed a significant advantage for interrogative over imperative sentences in human listeners ($p < 0.05$), plausibly attributable to the rich prosodic cues of question intonation, but no comparable benefit was observed for ASR systems ($p = 0.30$), which appear to weight segmental over prosodic information.
Conclusion: These findings highlight that, although contemporary ASR models approach human performance in favorable conditions, they lack the adaptive mechanisms that allow humans—especially BLs—to sustain intelligibility in complex acoustic scenes. By situating ASR performance within the spectrum of human auditory variability, this research underscores both the progress and the limitations of machine hearing and points toward neurobiologically inspired strategies for achieving noise-robust speech recognition. Clinically, this framework suggests that benchmarking ASR systems against the full spectrum of human auditory variability—including hearing-impaired and blind listeners—can inform the design of accessibility tools, hearing-aid signal processing, and supplementary diagnostic measures of speech-in-noise perception in audiology practice.

KEYWORDS
auditory variability, automatic speech recognition, machine hearing, signal-to-noise ratio, speech-in-noise

1 Introduction

Automatic speech recognition (ASR) has become a routine component of virtual assistants, live captioning, and accessibility tools, and its accuracy under realistic acoustic conditions, rather than in the clean recordings on which most benchmarks are still built, increasingly determines whether these systems serve users equitably (Hinton et al., 2012; Ozarkar et al., 2020; Gong et al., 2023). Natural listening environments rarely offer clean speech: multiple talkers, environmental noise, and reverberation combine and fluctuate over time (Bregman, 1990; Dua and Dua, 2023), degrading machine transcription accuracy as measured by word error rate (WER; Kumalija and Nakamoto, 2022; Won, 2025). Recent models such as Whisper have narrowed the historical gap with human listeners for clean speech, but a gap persists in noisy or unpredictable conditions (Gong et al., 2023; Patman and Chodroff, 2024), raising the question of whether artificial systems process degraded speech by strategies comparable to those used by human listeners under the same constraints.

Direct human–ASR comparisons help specify where this gap remains. Kim et al. (2024) found that among four end-to-end systems tested against human listeners on second-language English in speech-shaped noise, only Whisper-large matched or exceeded human accuracy, and only at favorable SNRs; Patman and Chodroff (2024) reported that Whisper-large-v3 outperformed native listeners in most noise conditions but merely matched them in naturalistic pub noise; and Goswami (2026) showed that reverberation alone degrades Whisper's WER by 2–16 percentage points depending on model size. Together with industry benchmarks placing Whisper among the more noise-resilient systems currently deployed (Achorn et al., 2025), these results suggest that the human advantage has narrowed in stationary noise but remains substantial in naturalistic and reverberant conditions. Most of this evidence, however, comes from English. Arabic is comparatively underresourced: its rich morphology, nonstandardized orthography, and dialectal variation degrade ASR accuracy even before noise is introduced (Besdouri et al., 2024; Salhab and Abusheikh, 2025), Modern Standard Arabic accounts for only a small share of training data for models such as Whisper, and regional dialects such as Najdi Arabic are represented more sparsely still (Talafha et al., 2025). Reported WERs for Arabic already exceed 40% in clean conditions for some transformer models (Gerazov et al., 2025) and decline further once realistic noise is added (Ezzine et al., 2025). Any human–ASR comparison conducted on Arabic stimuli should therefore be expected to expose larger machine deficits than an equivalent English comparison, a consideration that motivates the present design and constrains the interpretation of our findings.

Human listeners are not uniform in their robustness to noise, and this variability is itself informative. Hearing-impaired listeners (HILs) show well-documented difficulty separating speech from competing sound, reflecting reduced audibility together with deficits in temporal and spectral resolution that amplification alone does not resolve (Lorenzi et al., 2006; Healy et al., 2013; Nabelek et al., 2004); accuracy under noise, moreover, need not reflect equal effort, since the Framework for Understanding Effortful Listening (FUEL; Pichora-Fuller et al., 2016) predicts that two listeners can reach the same accuracy at different cognitive cost, a distinction ASR output, which returns only a transcript, cannot capture (Peelle, 2018). Blind listeners (BLs), by contrast, often show enhanced speech-in-noise performance (Bae et al., 2022; Pang et al., 2020; Hertrich et al., 2009), consistent with evidence that early blindness recruits occipital cortex for auditory processing (Lewis et al., 2010; Guerreiro et al., 2015; Sabourin et al., 2022). This asymmetry across listener groups suggests that placing ASR systems along a spectrum of human auditory variability, rather than against a single idealized listener, may be more informative than accuracy alone.

The present study asks whether contemporary ASR systems process Arabic speech in noise in ways that more closely resemble typical, impaired, or enhanced human listening. We compare normal-sighted, hearing-impaired, and blind listeners with two ASR systems (Whisper and TurboScribe) on Arabic sentences presented at four SNRs, holding linguistic content constant across groups so that differences in recognition accuracy can be attributed to how each listener, human or machine, handles acoustic degradation. Beyond its bearing on ASR evaluation, this comparison speaks to an applied concern: as ASR is increasingly embedded in captioning, hearing-aid processing, and telehealth transcription, its behavior at adverse SNRs has direct consequences for hearing-impaired and visually impaired users (Healy et al., 2013; Aguiar et al., 2025), and speech-in-noise performance is itself a more ecologically valid index of communication ability than pure-tone audiometry alone (Wilson, 2011; Jain et al., 2025).

2 Materials and methods

2.1 Stimuli

The speech stimuli consisted of 20 Modern Standard Arabic (MSA) imperative sentences, as previously described by Aldhafeeri et al. (2022), and 20 Arabic interrogative sentences. Those sentences were scripted, had no nonlexical tokens (filled pauses), and were articulated carefully. Each sentence was recorded by both a male and a female native Arabic speaker, ensuring representation of gender-related acoustic variation. All recordings were amplitude-normalized to an average level of 70 dB SPL to maintain consistent presentation levels across conditions. The clean speech signals were then mixed with aircraft passenger-boarding noise at four signal-to-noise ratios (SNRs): +10 dB, +5 dB, −5 dB, and −10 dB. These SNRs were selected to represent a perceptual range from relatively easy listening (+10 dB) to highly challenging listening (−10 dB) conditions. This approach follows prior work in speech-in-noise perception research, which often employs a multilevel SNR design to assess performance robustness (Mattys et al., 2012; Patman and Chodroff, 2024). All audio files were resampled to 44.1 kHz, ensuring compatibility with the ASR systems used in this study and providing a uniform spectral resolution across machine and human analyses.

2.2 Participants

Three groups of human participants were recruited: NSLs, HILs, and BLs. The NSL group included 20 Saudi participants (10 men and 10 women) with self-reported normal hearing and vision. All were undergraduate students majoring in humanities and social sciences (e.g., Arabic language and literature, social work, and media). Their mean age was 21 years (SD = 0.6), and all spoke Najdi Arabic as their native dialect. The HIL group consisted of 10 Saudi participants (seven men and three women) with self-reported hearing loss of more than

1 year's duration; all wore bilateral hearing aids, but no audiometric records were available to verify the type or severity of the impairment. All HIL participants held a bachelor's degree, spoke Najdi Arabic (except one participant who came from Jeddah and had resided recently in Riyadh), and had a mean age of 31 years (SD = 3.5). The BL group comprised 10 male Saudi participants with early-onset blindness and normal hearing. Blindness onset was reported by participants and, where applicable, their relatives: nine participants were born blind, and one lost vision in childhood following a car accident; none acquired blindness in adulthood. The nine congenitally blind participants each held a university degree and had been accommodated by their university's disability-accessibility office on the basis of medical records documenting their blindness, although these records were not made available to the experimenters; residual vision was not formally assessed for any participant. The remaining participant held a secondary-school certificate. The mean age of this group was 26 years (SD = 2). All participants were native speakers of Arabic and provided informed consent prior to data collection. Some participants from the HIL and BL groups were assisted by their relatives in completing the consent process and in operating the experimental setup as needed, ensuring accessibility and full comprehension of all instructions. All participants (and their assistants, where involved) followed procedures approved by the institutional ethics committee and consistent with international standards for human-subjects research.

2.3 ASR systems

Two commercially available ASR systems were tested: Whisper (OpenAI) and TurboScribe (ChatGPT-integrated). Whisper (OpenAI) is a large-scale, multilingual speech recognition model trained on hundreds of thousands of hours of diverse audio data. Whisper has demonstrated strong robustness to background noise, varied accents, and spontaneous speech styles (Radford et al., 2023). TurboScribe (ChatGPT-integrated) is a proprietary ASR system optimized for real-time transcription and integrated with ChatGPT's language modeling framework. TurboScribe is designed for efficient decoding under variable acoustic conditions. These systems were selected to represent distinct architectural paradigms, one open-source and large-scale (Whisper) and one proprietary and tightly integrated with natural-language modeling (TurboScribe), thereby providing complementary baselines for current ASR performance in noise.

2.4 Experimental paradigm

A speech-in-noise recognition paradigm was implemented to investigate how human listeners and ASR systems process Arabic speech under varying degrees of acoustic interference. Each sentence was presented in passenger-boarding noise at one of four SNRs (+10, +5, −5, and −10 dB), with the speech signal preserved as a continuous waveform. Stimuli were delivered binaurally through headphones. Prior to the test trials, participants completed a short set of practice trials during which they adjusted playback volume to a comfortable level at their own pace; this self-selected level was then held constant throughout the test trials that followed. Hearing-impaired participants completed testing in the unaided condition, with hearing aids removed for the duration of the session. Participants listened to each sentence in a quiet, controlled environment and were instructed to repeat aloud what they heard. Repeated presentation before response is an established practice in speech-in-noise research, though the permitted number of repetitions varies across studies (e.g., up to five in Cole et al., 1996, and three in Yang and Xu, 2026); here, participants were allowed unlimited repetitions of each sentence before responding, but once a participant moved to the next stimulus, no further repetition of or correction to the prior response was permitted. This design choice was intended to offset a specific asymmetry between listeners and ASR systems: the ASR models received the complete, stable digital waveform as input rather than a single, fleeting exposure, and unlimited repetition gave human listeners comparable access to the acoustic signal itself before committing to a final response. The same stimuli were presented to both human groups and ASR systems to allow parallel analyses. For the ASR models, the same waveforms were input directly in digital form, and the resulting transcriptions were collected for quantitative comparison. Accuracy was measured using word error rate (WER), calculated for both human transcripts and ASR outputs, following standard metrics in speech recognition research (Li et al., 2014; Ali and Renals, 2018; Gong et al., 2023). WER was computed for each transcript relative to its reference using the standard formula adopted in recent ASR benchmarking work (Russell et al., 2024):

$$\mathrm{WER} = \left(\mathrm{S} + \mathrm{D} + \mathrm{I}\right) / \mathrm{N}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference. Lower values indicate higher transcription accuracy; WER = 0 indicates a perfect match. This experimental approach draws on established speech-in-noise paradigms used to compare listener groups (Mattys et al., 2012; Patman and Chodroff, 2024), adapted for use with ASR systems and Arabic speech stimuli (Aldhafeeri et al., 2022), as well as new stimuli constructed by the experimenter.

Human responses were scored in several stages to guard against transcription and coding errors. A research assistant, blind to each recording's participant identity and group membership, first transcribed all spoken responses; these transcriptions were then reviewed by the experimenter, who was not blind to participant identity, against the original audio. WER was subsequently computed twice for each transcript–reference pair: once using a Python implementation of the standard minimum-edit-distance algorithm, and once by the research assistant scoring by hand; both sets of scores were then checked again by the experimenter for discrepancies. Finally, all transcriptions and WER scores were sent to an independent phonetician, blind to participant identity, for review; this review resulted in only a small number of changes to the transcriptions and their associated WER scores. We did not compute a formal inter-rater reliability statistic (e.g., Cohen's kappa or an intraclass correlation coefficient) across these stages; instead, discrepancies were resolved through this sequential review-and-adjudication procedure rather than through independent parallel scoring.

In addition to the aggregate WER, each transcript–reference alignment was decomposed into its three constituent error types (substitutions, deletions, and insertions) using the standard minimum-edit-distance algorithm at the word level. The deleted words from the NSL and ASR transcripts were then classified into five closed-class linguistic categories by part of speech: articles and

determiners, prepositions, conjunctions, pronouns, and quantifiers. Coding was performed by a native speaker of MSA against a fixed inventory of function words for each category; ambiguous tokens were resolved with reference to the source sentence. Counts per category were aggregated across SNR levels and stimulus types and entered into the analyses reported in Figure 1.

3 Results

Table 1 summarizes the mean speech recognition error rates for the three groups of human listeners (BLs, NSLs, and HILs) alongside the two ASR systems, Whisper and TurboScribe, across four SNR levels: +10 dB, +5 dB, −5 dB, and −10 dB.

The declines in the mean speech recognition accuracy across SNR levels are illustrated in Figure 2. Recognition accuracy decreased consistently as noise increased, yet the slope of decline differed markedly among the listener groups. BLs maintained near-ceiling performance at mild and moderate SNR levels and showed only a gradual drop under more adverse conditions. Even at −10 dB SNR, BLs achieved an average accuracy of 85%, suggesting exceptional resilience to background interference. NSLs exhibited strong performance at higher SNRs (95% at +10 dB) but showed a clear inflection point around +5 dB SNR, below which recognition degraded more rapidly. By −10 dB SNR, accuracy for NSLs declined to 60%, marking a 35-percentage-point drop across the range—nearly quadruple the drop observed in BLs. HILs performed the poorest at all noise levels. Their accuracy decreased sharply from 80% at +10 dB SNR to only 25% at −10 dB SNR, reflecting a severe deficit in speech-in-noise perception even under moderate masking conditions. Among artificial systems, Whisper outperformed TurboScribe across all SNRs but with similar overall trends. Both systems showed robust recognition in favorable conditions (above 90% at +10 dB SNR), comparable to NSLs. However, their decline between +5 dB SNR and −10 dB SNR was notably sharper than that of BLs or even some NSLs, indicating reduced noise generalization capability. Whisper maintained 55% accuracy at −10 dB SNR, whereas TurboScribe dropped to 50%.

A linear mixed-effects model was conducted to examine recognition accuracy across listener groups and noise levels. The analysis revealed a statistically significant main effect of noise level ($p < 0.05$), indicating that recognition accuracy varied systematically as the listening condition became more adverse. Pairwise comparisons further showed that BLs performed significantly better than HILs ($p < 0.05$). In addition, NSLs differed significantly from both BLs and HILs ($p < 0.05$), suggesting that the NSL group occupied an intermediate position between the two human listener groups. Because the three human groups differed in sex composition, sex was entered as an additional fixed-effect covariate in this model; it did not significantly predict recognition accuracy ($p > 0.05$), indicating that the observed group differences are not attributable to sex composition. No statistically significant difference was found between Whisper and TurboScribe ($p = 0.071$), nor did ASR systems differ significantly from NSLs ($p > 0.05$), indicating that the performance of Whisper and TurboScribe was statistically comparable to that of NSLs across the tested noise conditions. Figure 3 illustrates this differentiation between humans and ASR systems.

TABLE 1 SNR for all listener groups and ASR systems by error type (deletion, substitution, insertion).

Listener group/ ASR system	Deletion	Substitution	Insertion
BLs	11%	7%	3%
NSLs	19%	12%	6%
HILs	38%	25%	21%
Whisper	7%	18%	19%
TurboScribe	8%	16%	17%

Table 2 disaggregates the recognition data by sentence type to facilitate direct comparison between the imperative and interrogative conditions for each listener group and ASR system across the four SNRs. The results for each sentence type reflect similar overall ordering of groups to that observed in Table 1: BLs achieved the lowest WER at every SNR, followed by NSLs and the two ASR systems clustered closely together, with HILs demonstrating the highest error rates throughout. The monotonic increase in WER as the SNR diminished from +10 to −10 dB was also replicated within each subset, confirming that the noise-level effect is not driven by one stimulus type. More substantively, interrogative sentences were recognized more accurately than imperatives by every human group, and this advantage increased as listening conditions deteriorated. For BLs, WER at −10 dB was 11% for imperatives but 4% for interrogatives; for NSLs, 32% (for imperatives) vs. 18% (for interrogatives); and for HILs, 49% (for imperatives) vs. 26% (for interrogatives). The HIL group thus demonstrated the largest stimulus-type effect in absolute terms, with interrogatives reflecting nearly half the error rate of imperatives at −10 dB.

The ASR systems demonstrated a more uneven profile. At favorable and intermediate SNRs (+10, +5, −5 dB), Whisper and TurboScribe transcribed imperatives marginally more accurately than interrogatives, suggesting that short, fixed-syntax imperative templates are well represented in their training distribution. At −10 dB, however, this advantage reversed sharply: Whisper fell to 25% WER for imperatives vs. 20% for interrogatives, and TurboScribe to 30% vs. 20%, respectively. This shift indicates that under severe masking, the ASR models also benefited from the additional acoustic information present in interrogatives—most likely, the more salient pitch contour and longer voiced segments—although the human-side gain was consistently greater than that for machines, particularly for HILs. The results in Table 2 indicate that sentence type interacts with both noise level and listener type: Humans extract more information from the interrogative structure than ASR systems do, and this differential widens precisely in the conditions where ASR robustness is otherwise weakest.

Figure 4 visualizes the patterns reported in Table 2, displaying recognition accuracy (100 − WER) for the three human groups in the upper row and the two ASR systems in the lower row, with imperative and interrogative conditions plotted in adjacent columns. Two effects are immediately apparent. Among human listeners, every group performed better on interrogatives than on imperatives, and this gap widened systematically as the SNR worsened; this effect was most pronounced for HILs, where the difference in accuracy between imperative and interrogative sentences at −10 dB exceeded 20 percentage points. The two ASR systems, by contrast, showed much smaller and inconsistent stimulus-type



FIGURE 1
Word cloud of deletion patterns by linguistic category for NSL listeners (green) and ASR systems (orange). Each category label, and the Arabic exemplar words within it, are sized and positioned in proportion to the number of deletions of that category by the corresponding group, with more frequent categories placed closer to the center. NSL deletions concentrate in quantifiers and conjunctions; ASR deletions are distributed evenly across categories.

TABLE 2 Stimulus type (imperative vs. interrogative) for human and ASR listeners under four levels of SNR.

Listener group/ ASR system	+10 dB SNR	+5 dB SNR	−5 dB SNR	−10 dB SNR
BLs	2%	5%	10%	15%
NSLs	5%	12%	25%	40%
HILs	20%	35%	55%	75%
Whisper	6%	15%	30%	45%
TurboScribe	8%	18%	35%	50%

effects, with imperative and interrogative trajectories overlapping closely at every SNR.

A linear mixed-effects model with sentence type (imperative vs. interrogative), listener type, and SNR as fixed effects and participant or system as a random effect was fitted separately to the human and ASR data to formally test these patterns. For human listeners, the main effect of sentence type was statistically significant ($p < 0.05$), confirming that interrogatives were recognized more accurately than imperatives across BLs, NSLs, and HILs. By contrast, the same contrast tested on the two ASR systems was not statistically significant ($p = 0.30$), indicating that Whisper and TurboScribe transcribed the two sentence types with comparable accuracy. The perceptual benefit conferred by interrogative structure was reliably present in human listeners but absent in the ASR systems tested.

In Figure 5, the WER values are decomposed into the three constituent error types—deletions, substitutions, and insertions—averaged across all SNR levels and stimulus types and are presented separately for each listener group and ASR system.

Two patterns are immediately apparent. First, the data reflect a similar relative ordering of groups by total error rate to rankings established in Tables 1, 2: BLs reported the lowest error counts in every category, followed by NSLs, with HILs producing markedly more errors of each type. Second, and more substantively, the qualitative profile of errors was opposite for human listeners and ASR systems. Across all three human groups, deletions were the most frequent error type (BLs = 11%, NSLs = 19%, HILs = 38%), substitutions were intermediate, and insertions were the least frequent. The two ASR systems demonstrated an inverse pattern: deletions were the *least* frequent error type (Whisper = 7%, TurboScribe = 8%), with substitutions and insertions occurring at roughly comparable higher rates (16–19%). A linear mixed-effects model with error type, listener type (human vs. ASR), and their interaction as fixed effects and participant or system as a random effect confirmed a significant main effect of error type, a significant main effect of listener type, and a significant error-type × listener-type interaction (all $p < 0.05$). Pairwise comparisons within the human data showed significant differences between deletions and substitutions and between deletions and insertions (both $p < 0.05$) but not between substitutions and insertions. The same pattern of pairwise contrasts was observed for the ASR systems: Deletions differed significantly from both substitutions and insertions, whereas substitutions and insertions did not differ. However, the direction of the deletion contrast was reversed: Deletions were the most frequent error type for humans but the least frequent for ASR systems. The interaction therefore captures a true qualitative dissociation in error profile rather than a simple difference in overall accuracy.

The WERs for all listener groups and ASR systems are presented, broken out by error type, in Table 3.

Figure 1 unpacks the deletion column of Table 3 by linguistic category, indicating which classes of function words were most often omitted by NSL listeners and by the ASR systems. Each English

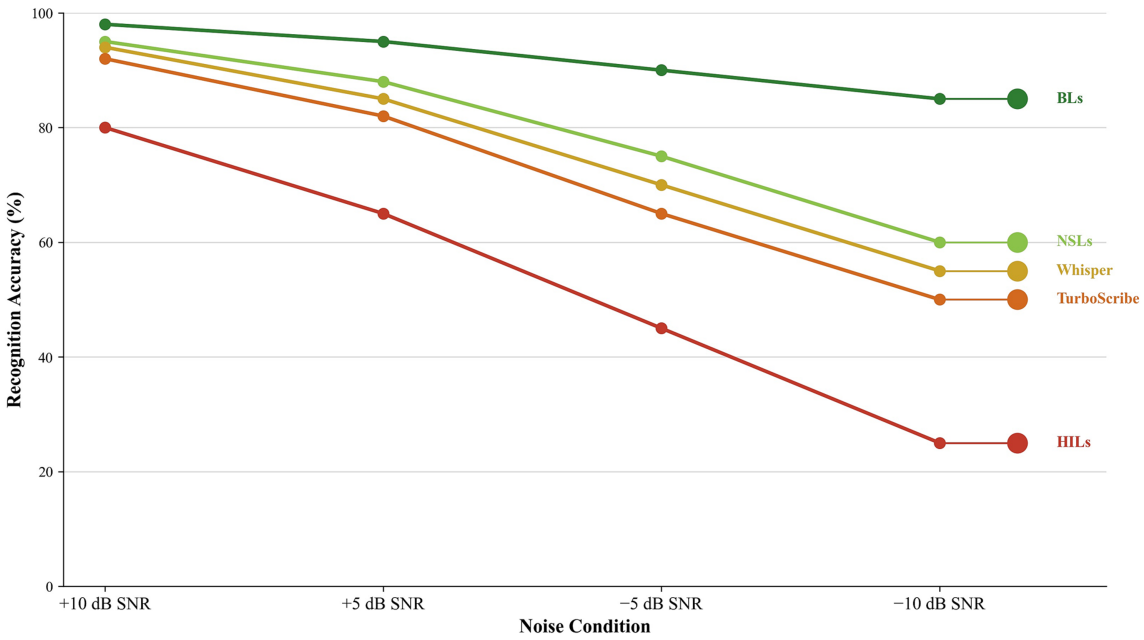


FIGURE 2
Mean speech recognition accuracy (percentage correct, i.e., 100 − WER) for human listeners and ASR systems under four levels of SNR.

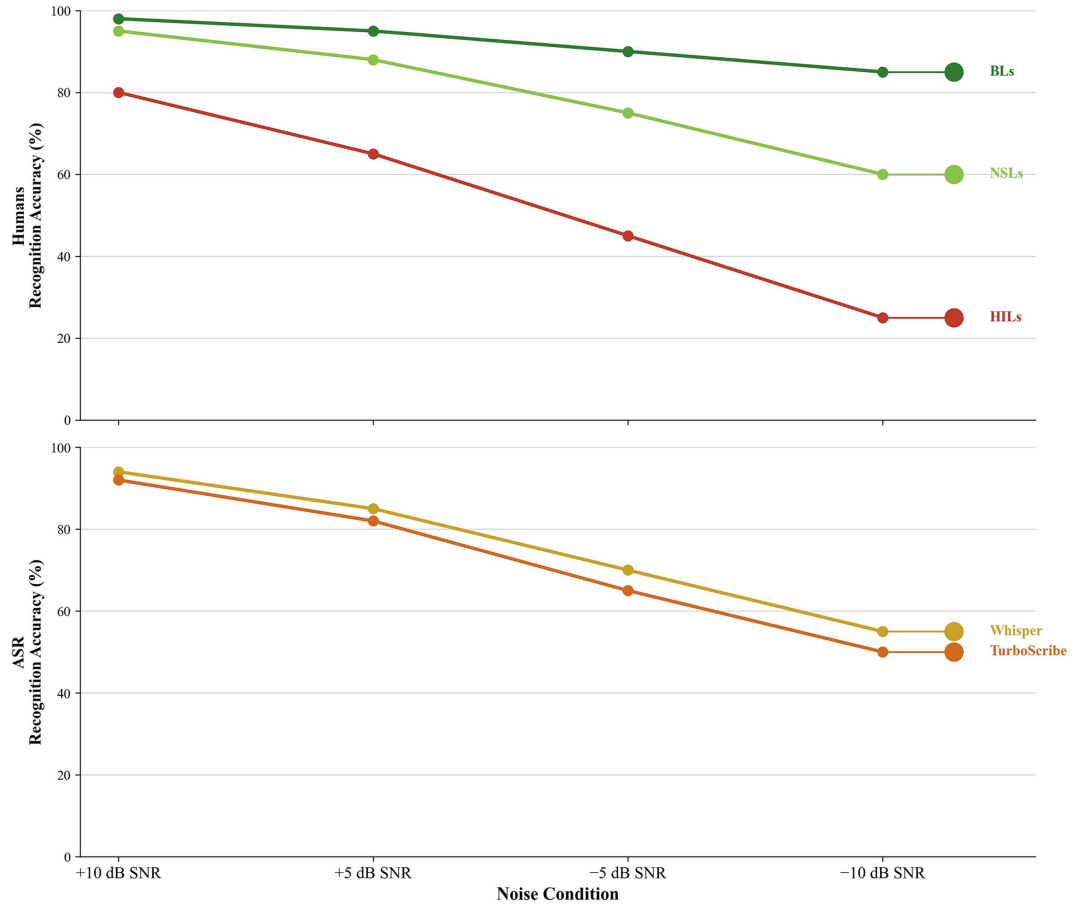


FIGURE 3
Mean speech recognition accuracy (percentage correct, i.e., 100 − WER) under four levels of SNR, shown separately for human listeners (top panel) and ASR systems (bottom panel).

TABLE 3 Error type (deletion, substitution, insertion) for human and ASR listeners under four levels of SNR by stimulus type (imperative vs. interrogative).

Listener group/ASR system	Imperative sentences				Interrogative sentences			
	+10 dB	+5 dB	−5 dB	−10 dB	+10 dB	+5 dB	−5 dB	−10 dB
BLs	1%	3%	7%	11%	1%	2%	3%	4%
NSLs	3%	6%	14%	32%	2%	6%	11%	18%
HILs	13%	25%	33%	49%	7%	10%	22%	26%
Whisper	4%	7%	13%	25%	2%	8%	17%	20%
TurboScribe	4%	8%	15%	30%	4%	10%	20%	20%

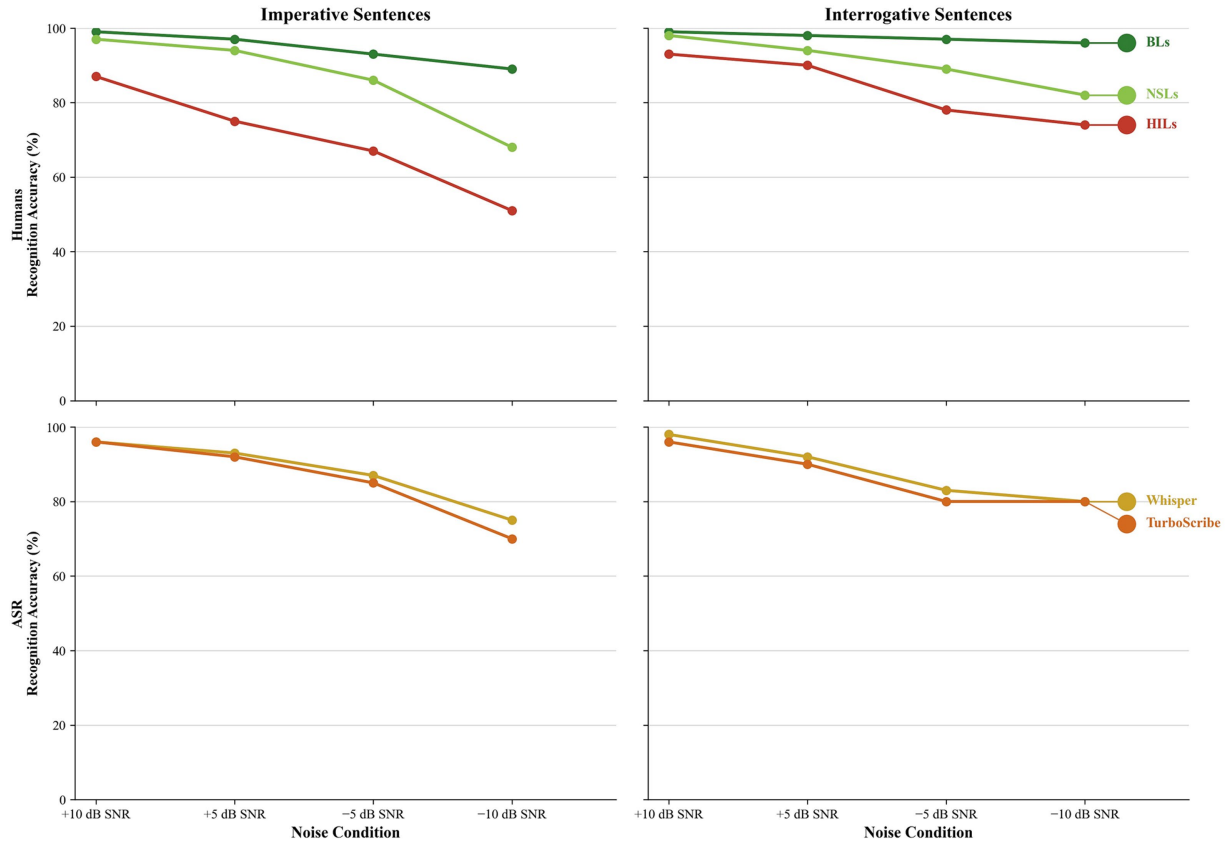


FIGURE 4
Mean speech recognition accuracy (proportion correct) for human listeners and ASR systems under four levels of SNR, broken down by sentence type. 'Imperative Sentences' and 'Interrogative Sentences.

category label is sized in proportion to how often that category was deleted by the corresponding group, and the Arabic exemplars beneath each label illustrate the kinds of words involved (e.g., quantifiers and conjunctions). Two patterns stand out. NSL deletions were heavily concentrated in two categories—quantifiers and conjunctions—accounting collectively for approximately three-quarters of all NSL omissions; the remaining three categories contributed only marginally. ASR deletions were distributed almost evenly across all five categories, with no category dominating. The figure therefore makes visible a property the aggregated counts in Table 3 only imply: Human deletions are selective and follow linguistic structure, whereas ASR deletions are indiscriminate and appear to be driven primarily by acoustic rather than syntactic factors.

4 Discussion

The results of this study clarify how current ASR systems approximate—but do not yet replicate—human auditory capabilities, particularly under adverse acoustic conditions. The comparative analysis of BLs, NSLs, HILs, and two ASR systems (Whisper and TurboScribe) across four SNRs provides empirical evidence of both human superiority in noise robustness and the distinct computational strategies employed by artificial versus biological auditory systems. Across all SNR levels, performance decreased consistently as noise increased, yet human listeners—especially blind participants—demonstrated significantly greater resilience than the ASR systems. These findings mirror observations from previous research showing that human listeners

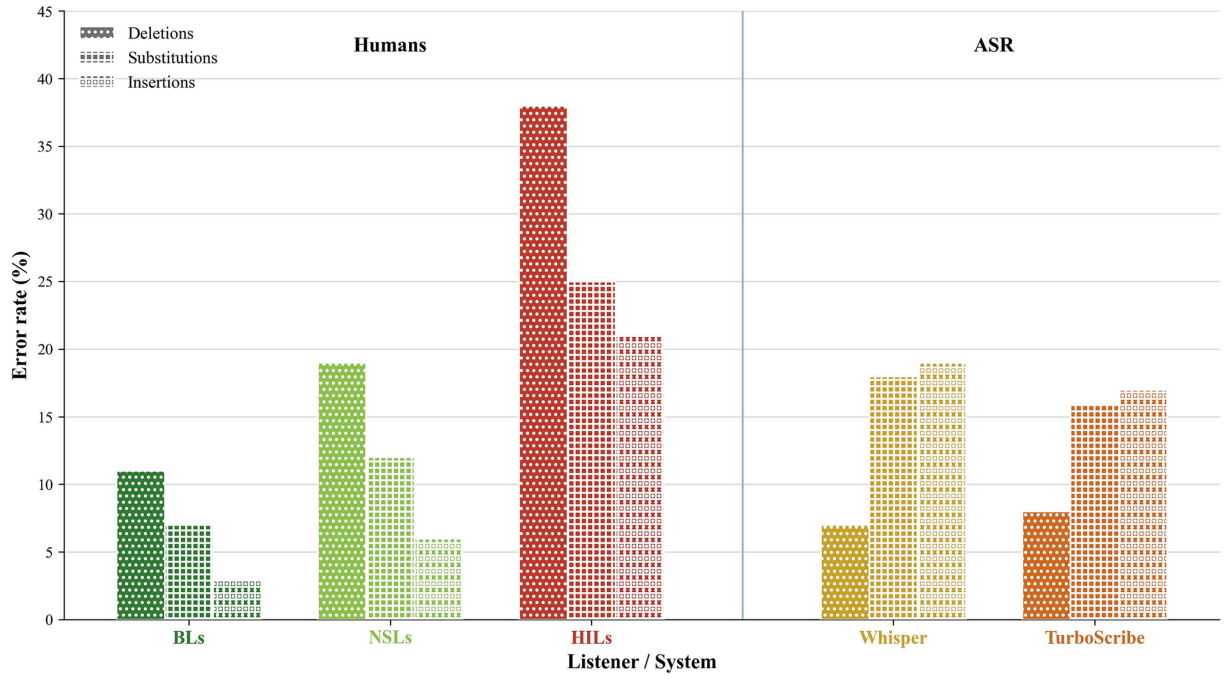


FIGURE 5
Decomposition of word error rate (WER) into deletions, substitutions, and insertions for human listeners and ASR systems averaged across all SNR levels and stimulus types. Each group is shown in its own color from Figures 2–4; error type is marked by fill pattern (dots = deletions, squares = substitutions, combined dots and squares = insertions) rather than by color, so each group's own color stays consistent across figures. Orange denotes the two ASR systems combined.

rely on contextual, linguistic, and cognitive redundancy to recover meaning from degraded acoustic signals (Mattys et al., 2012; Clonan et al., 2025). Whisper's accuracy, which matched NSLs in mild and moderate noise conditions, underscores the strength of large-scale models trained via weak supervision (OpenAI, 2022; Radford et al., 2023). However, its rapid decline at −5 dB and −10 dB SNR suggests that noise robustness in machine speech recognition remains far from human adaptability. Even the most sophisticated ASR systems display considerable degradation in real-world noisy Arabic speech, confirming that generalization under adverse acoustic contexts remains a key challenge.

The superior performance of BLs under all noise conditions highlights the remarkable adaptability of the human brain. Consistent with neuroimaging and behavioral findings, early blindness leads to cross-modal reorganization in which visual cortical regions are recruited for auditory processing (Lewis et al., 2010; Guerreiro et al., 2015; Sabourin et al., 2022). We did not collect neuroimaging or electrophysiological data in the present study, so this mechanism is inferred from convergence with prior literature rather than demonstrated directly in our sample. Our BL sample is consistent with this developmental window: nine participants were congenitally blind and the remaining participant lost vision in early childhood, and none acquired blindness in adulthood; blindness history was based on participant and relative report, corroborated for the university-educated participants by disability-accommodation records that were not directly examined by the experimenters, and residual vision was not formally assessed, so a contribution of unassessed partial vision cannot be entirely excluded. This neural plasticity supports enhanced temporal precision and spectral resolution, allowing for more efficient segregation of speech from background noise (Hertrich et al., 2009; Bae et al., 2022). Recent electrophysiological evidence aligns with this account: Congenitally blind adults show shorter MMN and N2b latencies and larger P3a amplitudes during speech-sound discrimination, indicating faster auditory change detection and greater attentional engagement than sighted controls (Turgeon et al., 2025). Crucially, King (2015) emphasized that cross-modal plasticity following blindness reflects a balance of compensatory gains and developmental trade-offs rather than uniform enhancement, helping explain why BL robustness in this study was substantial yet not absolute. The sustained high performance of blind participants, even under extreme masking, is consistent with the possibility that cross-modal adaptation endows unique resilience, and may offer a useful source of inspiration for designing more adaptive acoustic models, though this behavioral parallel does not itself establish the underlying neural mechanism or demonstrate that it can be transferred to machine architectures.

By contrast, HILs exhibited pronounced difficulty across all SNR levels. Their reduced intelligibility, even at favorable listening conditions, is consistent with longstanding evidence that audibility restoration alone often does not fully compensate for compromised temporal fine structure or spectrotemporal resolution in sensorineural hearing loss (Lorenzi et al., 2006; Healy et al., 2013); because hearing status in our sample was self-reported rather than audiometrically confirmed, we cannot establish that this specific mechanism accounts for the HIL group's performance here. A further, more parsimonious contributor is that HIL participants were tested without their hearing aids: some of the pronounced difficulty observed here may therefore reflect reduced audibility in the unaided condition rather than, or in addition to, a central deficit in processing degraded speech. Recent reviews of cross-modal plasticity in mild-to-moderate partial hearing loss have further indicated that adaptive neural changes do not consistently translate into perceptual gains in this population, in contrast to the

clearer enhancements observed after total sensory loss (Aguiar et al., 2025). If our HIL group's impairment falls within this range—which our self-reported classification cannot confirm—this distinction may help explain why these participants failed to recover in noise despite presumably preserved cognitive resources. Convergent evidence comes from a recent large-scale audiology study by Jain et al. (2025), who graded 100 listeners across five hearing-loss severities and showed that word-based recognition scores fell to floor at adverse SNRs precisely in the moderately severe and severe groups, whereas finer-grained phoneme-based scoring still discriminated severity at 98% accuracy under those same conditions. The collapse pattern they documented is broadly analogous to the floor we observed at −10 dB SNR in our HIL group, though without audiometric grading we cannot place our sample precisely within their severity bands; the parallel nonetheless suggests that the apparent saturation of HIL deficit at the most adverse SNR may partly reflect the granularity of WER as a measure rather than the absence of further auditory differentiation. The performance of HILs parallelled the steepest degradation observed in the ASR models, emphasizing both the technological proximity of ASR systems to impaired human processing and the extent of improvement still needed to achieve typical human robustness. This comparison reinforces the value of model–human benchmarking for identifying limitations in both machine-learning algorithms and human-assistive devices.

A complementary pattern, equally instructive for the human–machine comparison, emerges from the sentence-type contrast. The mixed-effects analysis confirmed a significant advantage for interrogative over imperative sentences in human listeners ($p < 0.05$) but no comparable effect for the ASR systems ($p = 0.30$). This dissociation aligns with a well-established finding that the interrogative–declarative contrast is acoustically marked in Arabic, as in many languages, by a prominent rising pitch contour and lengthening over the final portion of the utterance (Peng et al., 2009; Shen, 2021) and that humans rely heavily on these prosodic cues to support intelligibility when the segmental signal is degraded by noise (Shen, 2021). The fact that BLs, NSLs, and HILs all benefited from interrogative prosody—and that HILs benefited most in absolute terms—suggests that prosodic top-down support is preserved even when peripheral auditory resolution is compromised and may serve as a partial compensatory route. End-to-end ASR systems, by contrast, are optimized primarily on segmental and lexical accuracy and weight prosodic structure only weakly during decoding (Tran et al., 2018; Wagner et al., 2015), which plausibly explains the absence of an interrogative advantage for Whisper and TurboScribe. The applied implication is that the human–ASR gap observed at low SNR will be larger for short, prosodically flat utterances such as imperatives than for utterances with rich intonational structure, an asymmetry worth factoring into the design of voice interfaces deployed in noisy real-world settings.

A third dissociation between humans and ASR systems is indicated by the decomposition of WER into its constituent error types (Figure 5). Human listeners, including BLs, produced predominantly deletions when the signal was degraded—omitting words they could not confidently identify rather than guessing. Substitutions and insertions were comparatively rare. The two ASR systems showed an inverse pattern: Deletions were the least frequent error type, with most errors taking the form of substitutions or insertions of phonetically or lexically plausible material that was not actually present in the signal. This contrast is consistent with documented properties of end-to-end ASR systems trained with weak supervision: Such models are driven to produce fluent, locally plausible output and exhibit a known tendency to hallucinate or paraphrase rather than to report uncertainty (Russell et al., 2024; Koenecke et al., 2024). Whisper in particular has been shown to insert content not present in the audio under low-SNR or silent conditions (Koenecke et al., 2024). The human strategy—omit rather than fabricate—reflects metacognitive awareness of perceptual uncertainty that current end-to-end ASR architectures lack. From an applied standpoint, the two error profiles have very different downstream costs: A deletion in an air-cabin announcement is recoverable through repetition or context, whereas a confident substitution may produce a fluent but incorrect transcription that the listener has no way to flag. Reducing ASR insertion and substitution errors—ideally by introducing calibrated uncertainty estimates that allow systems to abstain when the signal is too degraded—would therefore be a higher-priority target for safety-critical deployments than further reductions in overall WER alone.

The linguistic-category breakdown shown in Figure 1 sharpens this interpretation. NSL deletions were not distributed uniformly across function-word classes but were concentrated in two categories—quantifiers and conjunctions—which together accounted for approximately three-quarters of all NSL omissions, whereas articles, determiners, prepositions, and pronouns were retained at much higher rates. This selective omission profile is consistent with information-theoretic accounts of speech production and comprehension, which predict that listeners under processing load preferentially drop high-probability, low-information items that can be reconstructed from context (Levy, 2008; Mahowald et al., 2013). The pattern is also consistent with classic findings on shadowing and cued recall under noise, in which content words are preserved at the expense of grammatical morphemes (Marslen-Wilson and Welsh, 1978). The two ASR systems showed a markedly different profile: Deletions were spread evenly across all five categories at counts between 3 and 6, with no category dominating. This flat distribution suggests that ASR omissions are governed primarily by acoustic salience—weakly stressed or briefly articulated tokens are dropped regardless of their syntactic role—rather than by any learned sensitivity to linguistic predictability. The contrast in error structure therefore points to a deeper architectural divergence: Human listeners deploy linguistic prior knowledge to allocate effort selectively, whereas end-to-end ASR systems treat all tokens as equally informative until the acoustic signal forces a decision. Closing this gap may require ASR architectures that explicitly model word-level uncertainty and predictability, perhaps by integrating language-model surprisal signals into the decoding process so that confidence can vary with linguistic context rather than acoustic energy alone.

An accuracy-only reading of the HIL pattern, however, risks missing a layer of the phenomenon. Within the FUEL framework, the same percent-correct score can be achieved at very different cognitive costs, and HILs typically pay a higher cost even when their intelligibility scores match those of normal-hearing peers (Pichora-Fuller et al., 2016; Peelle, 2018). Recent work using EEG alpha-band oscillations and pupillometry has shown that elevated effort persists even at SNRs where accuracy appears intact and that effort increases sharply once the SNR drops below the individual speech-reception threshold (Wendt et al., 2018; Simeone et al., 2025). The pronounced HIL deficit observed here at −5 and −10 dB SNR is therefore plausibly an underestimation of the perceptual burden for these listeners, since accuracy

had already collapsed before effort could be expended productively. In contrast, ASR systems have no analog of cognitive effort but also no graceful-degradation strategy when the signal becomes uninformative. The convergence between HIL and ASR error trajectories at low SNR may thus reflect a shared inability to recruit additional resources rather than a shared underlying mechanism, an interpretation consistent with the divergence between human top-down inference and the largely bottom-up statistical regularities exploited by current end-to-end models (Clonan et al., 2025; Hicks and McDermott, 2024).

A further dimension of these results concerns the language tested. The Whisper and TurboScribe declines observed at −5 and −10 dB SNR are steeper than several published English-stimulus benchmarks would predict (Patman and Chodroff, 2024; Kim et al., 2024), and this gap is consistent with the broader picture of Arabic ASR. Arabic is severely underrepresented in the training corpora of major end-to-end systems, and even on clean Arabic data, state-of-the-art models report markedly higher WERs than on comparable English material (Besdouri et al., 2024; Salhab and Abusheikh, 2025). Accuracy degrades further with ambient noise; for example, dialect-specific work in Moroccan Arabic under traffic noise has reported sharp declines tracking the SNR (Ezzine et al., 2025). Two implications follow. First, the apparent NSL–ASR equivalence at +10 and +5 dB in the present data should not be overgeneralized; it most likely reflects a ceiling on the relatively short, predictable imperative and interrogative sentences used here and would probably not extend to spontaneous or dialectal Arabic. Convergent cross-language evidence comes from Russell et al. (2024), who reported increases in Whisper's WER from 2–3% on read audiobook English to approximately 19% on Switchboard telephone conversations and 27% on the RoomReader videoconferencing corpus, confirming that any benchmark obtained on short, well-formed read sentences sets an upper bound on what current ASR systems can achieve in conversational deployment. Second, the pronounced ASR collapse at low SNR cannot be attributed solely to acoustic masking; it is compounded by data-scarcity effects specific to Arabic. The implication for applied settings such as in-flight communication is that ASR accuracy benchmarks generated on English speech do not translate directly to Arabic-speaking passengers, and dedicated dialectal training data, rather than scaling alone, is likely to be the rate-limiting factor for closing the human–machine gap.

Overall, the pattern of results is consistent with current ASR models approaching human-level accuracy under moderate noise while declining more steeply than human listeners under more adverse conditions; we interpret this as suggestive of a difference in adaptive perceptual mechanisms and higher-order prediction between artificial and biological systems, though the present study measured recognition accuracy alone and did not test these mechanisms directly. The contrast between the performance of blind listeners and that of ASR systems is one motivation, among others, for exploring neuroplasticity-inspired approaches in future ASR development, such as integrating multimodal cues and dynamic reweighting strategies; whether such biologically inspired designs would in fact narrow the gap observed here remains to be tested. These findings reaffirm the practical limitations of current machine hearing under adverse conditions and the noise robustness of the human listeners tested here, and they suggest, rather than establish, directions for future noise-robust ASR development.

5 Concluding remarks, limitations, and future directions

This study compared normal-sighted, hearing-impaired, and blind listeners with two ASR systems (Whisper and TurboScribe) on Arabic speech-in-noise recognition across four signal-to-noise ratios. Recognition accuracy declined for all groups and both ASR systems as noise increased, but the rate of decline differed: blind listeners were the most resilient across all SNRs, ASR systems approximated normal-sighted listeners at favorable and moderate SNRs but declined more steeply at the two lowest SNRs, and hearing-impaired listeners, tested unaided, showed the most pronounced difficulty throughout.

The two ASR systems and the human listeners also differed in kind rather than only in degree: humans, including blind listeners, predominantly omitted words they could not identify, whereas the ASR systems more often substituted or inserted plausible but incorrect material. Taken together, these results indicate that current ASR systems can approximate typical human accuracy under moderate noise but do not yet match the resilience of our most robust human group under severe noise, and that they fail differently than humans do when they fail. We interpret the contrast with blind listeners as consistent with, though not a direct test of, neuroplasticity-based accounts of auditory reorganization after early blindness (see Future Directions for the research and clinical questions this raises).

A few limitations should be considered when interpreting these findings. The relatively small sample (n = 10–20 per group) limits statistical power and generalizability. Hearing loss in the HIL group was self-reported rather than audiometrically verified, leaving the precise degree and configuration of impairment uncharacterized. Although sex was entered as a fixed-effect covariate in the group-level model and did not significantly predict recognition accuracy, the BL group was exclusively male, limiting how confidently the present findings generalize to female blind listeners. The three groups also differed in mean age (NSL: 21 years; BL: 26 years; HIL: 31 years); age was not entered as a covariate in the present analyses, and although the NSL group's younger age reflects its role as a normative baseline rather than a matched comparison group, age cannot be ruled out as a contributor to the observed group differences. Unlimited repetitions were permitted before response to give human listeners access to the acoustic signal comparable to the ASR systems' access to the complete waveform (see Materials and Methods); however, repeated listening also affords memory, lexical prediction, and learning processes unavailable to the ASR systems, so the human–ASR comparison reported here should be interpreted as bounded by this asymmetry. Similarly, WER scoring relied on sequential review and adjudication across a research assistant, the experimenter, and an independent phonetician (see Materials and Methods) rather than on independent parallel scoring, so no formal inter-rater reliability statistic could be computed. The acoustic environment was restricted to a single noise type (aircraft passenger-boarding noise), only two ASR systems (Whisper and TurboScribe) were evaluated, and the stimuli were limited to imperative and interrogative sentences in MSA, which together bound the generalizability of the reported human–machine asymmetries. Finally, WER, used here as the sole accuracy metric, may underestimate

residual speech-perception ability in the HIL group at the most adverse SNRs, where word recognition approaches floor.

Several directions follow from the limitations above. Future replications should recruit larger, gender-balanced samples with audiometrically confirmed hearing profiles rather than self-report, and should either match groups on age or enter it explicitly as a covariate. Broadening the range of background noise types and reverberation conditions, evaluating a wider set of ASR architectures, and extending the paradigm to additional languages, dialects, and speech genres would help establish how far the human–machine asymmetries reported here generalize. Complementing WER with finer-grained phoneme- or syllable-level scoring, which has been shown to discriminate hearing-loss severity even when word-level recognition is at floor (Jain et al., 2025), would also help characterize residual perceptual ability in listeners whose word-recognition scores saturate at the most adverse SNRs.

The BL–ASR contrast observed here raises a further, more speculative research direction: whether ASR architectures that incorporate adaptive, multimodal, or context-sensitive attention mechanisms—loosely inspired by the cross-modal reorganization documented in early blindness—can narrow the gap seen at low SNR. Testing this would require engineering work well beyond the present behavioral comparison and should not be read as already demonstrated by these data.

Finally, the human–ASR asymmetries documented here may be worth exploring for clinical and applied purposes, though we present these as hypotheses for future work rather than established applications. It remains to be tested whether ASR error profiles could serve as a low-cost adjunct to speech-in-noise testing in audiology, for example in screening or longitudinal monitoring; whether the deletion-versus-substitution asymmetry between humans and ASR systems has practical consequences for hearing-aid or telehealth transcription pipelines, where confidently incorrect ASR output could go unflagged at adverse SNRs; and whether benchmarking ASR systems against the full spectrum of human auditory variability, rather than against typical listeners alone, could usefully complement existing diagnostic tools (Wilson, 2011; Jain et al., 2025) for characterizing where a given ASR engine is likely to support, match, or fail particular patient populations.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of King Saud University (KSU-HE-25–1,494). The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

YA: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declared that financial support was received for this work and/or its publication. This research was funded by the Ongoing Research Funding Program (ORF-2026-580), King Saud University, Riyadh, Saudi Arabia.

Conflict of interest

The author(s) declared that this work was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declared that Generative AI was used in the creation of this manuscript. I used AI for language editing and document organization in the Response to Reviewer file. All the content is mine. I used AI to edit the manuscript and check for any grammatical errors or language ambiguity. I used AI to enhance the quality of the figures and to introduce some small changes in the colors and patterns of the bars.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Achorn, K., Lai, T.-Y., and Li, T. (2025) Whisper: An MLPerf Inference Benchmark for Automatic speech Recognition (ASR). Available online at: https://mlcommons.org/2025/09/whisper-inferencev5-1/ (Accessed April 12, 2026).

Aguiar, P. V., Preman, J., and Paul, B. T. (2025). Cross-modal neuroplasticity in partial hearing loss: a mini-review. *Front. Neurosci.* 19:1627888. doi: 10.3389/fnins.2025.1627888

Aldhafeeri, M., Alqahtani, G., and Aldholmi, Y. (2022). "Speech rate and intelligibility of Arab airlines safety instructions," in *ExLing 2022 Paris: Proceedings of 13th Annual Conference of Experimental Linguistics, Paris, France*, ed. A. Botinis (Athens, Greece: International Society of Experimental Linguistics), 5–8. doi: 10.36505/ExLing-2022/13/0002/000544

Ali, A., and Renals, S. (2018). "Word error rate estimation for speech recognition: e-WER," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 20–24.

Bae, E. B., Jang, H., and Shim, H. J. (2022). Enhanced dichotic listening and temporal sequencing ability in early-blind individuals. *Front. Psychol.* 13:840541. doi: 10.3389/fpsyg.2022.840541

Besdouri, F. Z., Zribi, I., and Belguith, L. H. (2024). Arabic automatic speech recognition: challenges and progress. *Speech Comm.* 163:103110. doi: 10.1016/j.specom.2024.103110

Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: The MIT Press.

Clonan, A. C., Zhai, X., Stevenson, I. H., and Escabí, M. A. (2025). Interference of mid-level speech and noise statistics underlies human speech recognition sensitivity in natural environmental noise. *J. Neurosci.* 45:e1751242025. doi: 10.1523/JNEUROSCI.1751-24.2025

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

Dua, M., and Dua, S. (2023). Noise robust automatic speech recognition: review and analysis. *Int. J. Speech Technol.* 26, 475–519. doi: 10.1007/s10772-023-10033-0

Ezzine, K., Laadi, N., and Satori, H. (2025). Automatic speech recognition for Moroccan dialect in noisy traffic environments. *Eng. Appl. Artif. Intell.* 141:109751. doi: 10.1016/j.engappai.2024.109751

Gerazov, B., Politi, M., and Bratières, S. (2025). "Arabic ASR on the SADA large-scale Arabic speech corpus with transformer-based models," in *International Conference on Speech and Computer (SPECOM 2025)* (Cham: Springer Nature Switzerland), 70–84.

Gong, Y., Yoon, J., and Glass, J. (2023). "WhisperAT: Noise-robust automatic speech recognizers," in *Proceedings of Interspeech 2023*, (Dublin: Antwerp International Speech Communication Association), 4488–4492. doi: 10.21437/Interspeech.2023-2193

Goswami, M. (2026). Whisper-RIR-mega: a paired clean-reverberant speech benchmark for ASR robustness to room acoustics. *ArXiv.* doi: 10.48550/arXiv.2603.02252

Guerreiro, M. J. S., Putzar, L., and Röder, B. (2015). The effect of early visual deprivation on the neural bases of multisensory processing. *Brain* 138, 1499–1504. doi: 10.1093/brain/awv076

Healy, E. W., Yoho, S. E., Wang, Y., and Wang, D. (2013). An algorithm to improve speech recognition in noise for hearing-impaired listeners. *J. Acoust. Soc. Am.* 134, 3029–3038. doi: 10.1121/1.4820893

Hertrich, I., Dietrich, S., Moos, A., Trouvain, J., and Ackermann, H. (2009). Enhanced speech perception capabilities in a blind listener are associated with activation of fusiform gyrus and primary visual cortex. *Neurocase* 15, 163–170. doi: 10.1162/jocn.2008.21019

Hicks, J. M., and McDermott, J. H. (2024). Noise schemas aid hearing in noise. *Proc. Natl. Acad. Sci. USA* 121:e2408995121. doi: 10.1073/pnas.2408995121

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., and Jaitly, N. (2012). Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signal Process. Mag.* 29, 82–97. doi: 10.1109/MSP.2012.2205597

Jain, S., Narne, V. K., Bharani, Valayutham, H., Madan, T., Ravi, S. K., et al. (2025). Speech recognition in noise: analyzing phoneme, syllable, and word-based scoring methods and their interaction with hearing loss. *Diagnostics* 15:1619. doi: 10.3390/diagnostics15131619

Kim, S.-E., Chernyak, B. R., Seleznova, O., Keshet, J., Goldrick, M., and Bradlow, A. R. (2024). Automatic recognition of second language speech-in-noise. *JASA Express Lett.* 4:025204. doi: 10.1121/10.0024877

King, A. J. (2015). Crossmodal plasticity and hearing capabilities following blindness. *Cell Tissue Res.* 361, 295–300. doi: 10.1007/s00441-015-2175-y

Koenecke, A., Choi, A. S. G., Mei, K. X., Schellmann, H., and Sloane, M. (2024) "Careless whisper: Speech-to-text hallucination harms," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), Rio de Janeiro, Brazil, 3–6 June 2024* (New York: Association for Computing Machinery), 1672–1681. doi: 10.1145/3630106.3658996.

Kumalija, E., and Nakamoto, Y. (2022). Performance evaluation of automatic speech recognition systems on integrated noise-network distorted speech. *Front. Signal Process.* 2:999457. doi: 10.3389/frsip.2022.999457

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition* 106, 1126–1177. doi: 10.1016/j.cognition.2007.05.006

Lewis, L. B., Saenz, M., and Fine, I. (2010). Mechanisms of crossmodal plasticity in the early blind. *J. Neurosci.* 104, 2995–3008. doi: 10.1152/jn.00983.2009

Li, J., Deng, L., Gong, Y., and Haeb-Umbach, R. (2014). An overview of noise-robust automatic speech recognition. *IEEE/ACM Trans. Audio Speech Lang. Process.* 22, 745–777. doi: 10.1109/TASLP.2014.2304637

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., and Moore, B. C. J. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proc. Natl. Acad. Sci. USA* 103, 18866–18869. doi: 10.1073/pnas.0607364103

Mahowald, K., Fedorenko, E., Piantadosi, S. T., and Gibson, E. (2013). Info/information theory: speakers choose shorter words in predictive contexts. *Cognition* 126, 313–318. doi: 10.1016/j.cognition.2012.09.010

Marslen-Wilson, W. D., and Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cogn. Psychol.* 10, 29–63.

Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (2012). Speech recognition in adverse conditions: a review. *Lang. Cogn. Process.* 27, 953–978. doi: 10.1080/01690965.2012.705006

Nabelek, A. K., Tampas, J. W., and Burchfield, S. B. (2004). Comparison of speech perception in background noise with acceptance of background noise in aided and unaided conditions. *J. Speech Lang. Hear. Res.* 47, 1001–1011. doi: 10.1044/1092-4388(2004/074)

OpenAI (2022) Introducing Whisper. Available online at: https://openai.com/index/whisper/ (Accessed May 8, 2026).

Ozarkar, S., Chetwani, R., Devare, S., Haryani, S., and Giri, N. F. (2020). "AI for accessibility: Virtual assistant for hearing impaired," in *Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) Kharagpur, India* (New York: IEEE), doi: 10.1109/ICCCNT49239.2020.9225392.

Pang, W., Xing, H., Zhang, L., Shu, H., and Zhang, Y. (2020). Superiority of blind over sighted listeners in voice recognition. *J. Acoust. Soc. Am.* 148, EL208–EL213. doi: 10.1121/10.0001804

Patman, C., and Chodroff, E. (2024). Speech recognition in adverse conditions by humans and machines. *JASA Express Lett.* 4:115204. doi: 10.1121/10.0032473

Peelle, J. (2018). Listening effort: how the cognitive consequences of acoustic challenge are reflected in brain and behavior. *Ear Hear.* 39, 204–214. doi: 10.1097/aud.0000000000000494

Peng, S.-C., Lu, N., and Chatterjee, M. (2009). Effects of cooperating and conflicting cues on speech intonation recognition by cochlear implant users and normal hearing listeners. *Audiol. Neurootol.* 14, 327–337. doi: 10.1159/000212112

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M. A., Edwards, B., Hornsby, B. W. Y., Humes, L. E., et al. (2016). Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). *Ear Hear.* 37, 5S–27S. doi: 10.1097/aud.0000000000000312

Radford, A., Kim, J. W., Xu, T., Brockman, G., and McLeavey, C. (2023). Robust Speech Recognition via large-scale Weak Supervision (Whisper). Available online at: https://cdn.openai.com/papers/whisper.pdf (Accessed March 4, 2026).

Russell, S. O., Gessinger, I., Krason, A., Vigliocco, G., and Harte, N. (2024). What automatic speech recognition can and cannot do for conversational speech transcription. *Res. Methods Appl. Linguist.* 3:100163. doi: 10.1016/j.rmal.2024.100163

Sabourin, C. J., Merrikhi, Y., and Lomber, S. G. (2022). Do blind people hear better? Trends Cogni. *For. Sci.* 26, 999–1012. doi: 10.1016/j.tics.2022.08.016

Salhab, M., and Abusheikh, M.. (2025). "Pushing the boundaries of multidialectal Arabic ASR with weakly supervised pretraining and continual supervised fine-tuning," in *Proceedings of the Third Arabic Natural Language Processing Conference (ArabicNLP 2025) Suzhou, China* (Kerrville, TX: Association for Computational Linguistics), 734–739.

Shen, J. (2021). Pupillary response to dynamic pitch alteration during speech perception in noise. *JASA Express Lett.* 1:115202. doi: 10.1121/10.0007056

Simeone, D., Polo, E. M., Barbieri, R., and Paglialonga, A. (2025). Dynamic assessment of listening effort by EEG alpha oscillations during an adaptive speech-in-noise test. *Sci. Rep.* 15:34421. doi: 10.1038/s41598-025-17451-x

Talafha, B., Waheed, A., and Abdul-Mageed, M. (2025). "Dialectal coverage and generalization in Arabic speech recognition," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, eds. W. Che, J. Nabende, E. Shutova and M. T. Pilehvar (Kerrville, TX: Association for Computational Linguistics), 29490–29502.

Tran, T., Toshniwal, S., Bansal, M., Gimpel, K., and Livescu, K., (2018). "Parsing speech: A neural approach to integrating lexical and acoustic-prosodic information," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT) New Orleans, LA, USA*, eds. Walker, H., and Ji, A (Stent Stroudsberg, PA: Association for Computational Linguistics), 69–81, doi: 10.18653/v1/N18-1007

Turgeon, C., Hadid, V., Trudeau-Fisette, P., Knoth, I., Lepore, F., Lippé, S., et al. (2025). Impact of visual and auditory deprivation on speech perception: an EEG study. *Cereb. Cortex* 35:bhaf086. doi: 10.1093/cercor/bhaf086

Wagner, P., Origlia, A., Avesani, C., Christodoulides, G., Cutugno, F., D'Imperio, M., et al (2015). "Different parts of the same elephant: A roadmap to disentangle and connect different perspectives on prosodic prominence," in *Proceedings of the 18th International Congress of Phonetic Sciences Glasgow, United Kingdom*, ed. The Scottish Consortium for ICPhS 2015 (Glasgow, UK: The University of Glasgow).

Wendt, D., Koelewijn, T., Książek, P., Kramer, S. E., and Lunner, T. (2018). Toward a more comprehensive understanding of the impact of masker type and signal-to-noise ratio on the pupillary response while performing a speech-in-noise test. *Hear. Res.* 369, 67–78. doi: 10.1016/j.heares.2018.05.006

Wilson, R. H. (2011). Clinical experience with the words-in-noise test on 3430 veterans: comparisons with pure-tone thresholds and word recognition in quiet. *J. Am. Acad. Audiol.* 22, 405–423. doi: 10.3766/jaaa.22.7.3

Won, Y. (2025). Assessing the efficacy of word error rate as a proxy for pronunciation quality: a comparative study of ASR systems and human evaluations among young EFL learners. *J. Second Lang. Pronunc.* 11, 394–422. doi: 10.1075/jslp.25012.won

Yang, J., and Xu, L. (2026). Recognition of noise-vocoded accented speech by non-native listeners. *Phonetica*. 83, 505–531.