NOVEMBER 12 2024

Speech recognition in adverse conditions by humans and machines

Chloe Patman ; Eleanor Chodroff

JASA Express Lett. 4, 115204 (2024)
https://doi.org/10.1121/10.0032473

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Speech recognition in adverse conditions by humans and machines

Chloe Patman[1,a] and Eleanor Chodroff[2]
[1]Theoretical and Applied Linguistics Section, Faculty of Modern and Medieval Languages and Linguistics, University of Cambridge, Sidgwick Avenue, Cambridge CB3 9DA, United Kingdom
[2]Department of Computational Linguistics, University of Zurich, Andreasstrasse 15, Zurich 8050, Switzerland
cep72@cam.ac.uk, eleanor.chodroff@uzh.ch

Abstract: In the development of automatic speech recognition systems, achieving human-like performance has been a long-held goal. Recent releases of large spoken language models have claimed to achieve such performance, although direct comparison to humans has been severely limited. The present study tested L1 British English listeners against two automatic speech recognition systems (wav2vec 2.0 and Whisper, base and large sizes) in adverse listening conditions: speech-shaped noise and pub noise, at different signal-to-noise ratios, and recordings produced with or without face masks. Humans maintained the advantage against all systems, except for Whisper large, which outperformed humans in every condition but pub noise.

[Editor: Douglas D O'Shaughnessy] https://doi.org/10.1121/10.0032473

Received: 17 July 2024 **Accepted:** 7 October 2024 **Published Online:** 12 November 2024

1. Introduction

Considerable evidence has demonstrated the deleterious impact of noise on speech recognition not only for humans (e.g., Brungart, 2001; Brungart *et al.*, 2020; Miller and Nicely, 1955), but also for automatic speech recognition (ASR) systems (e.g., Carey and Quang, 2005; Kim *et al.*, 2024). Throughout the development of ASR systems, a clear goal has been to achieve or even surpass human performance (the current gold standard) in speech recognition tasks, both with clean speech and in noisy environments (Baevski *et al.*, 2020; Radford *et al.*, 2023). Determining whether this goal has been achieved requires a direct comparison between humans and ASR systems. Early work consistently found that humans outperformed machines on most tasks and across a range of conditions (Cutler and Robinson, 1992; Moore and Cutler, 2001; see also Scharenborg, 2007 for an overview). In recent years, however, significant improvements have been made in ASR performance, largely attributable to advances in deep learning. While these improvements have significantly narrowed the performance gap between humans and ASR systems, this has yet to be tested across a wide range of conditions, particularly those involving adverse listening conditions, such as background noise or speech produced with a face mask.

Our study compares the current performance of humans and state-of-the-art ASR systems in adverse listening conditions to provide a more nuanced understanding of the performance gap between these two groups. Specifically, we test two state-of-the-art ASR systems for English (wav2vec 2.0 and Whisper) against L1 British English listeners across several listening conditions. These conditions are speech-shaped white noise (SSN) and pub noise at two signal-to-noise ratios (SNRs), 0 and 8 dB, with and without a face mask as produced by Southern Standard British English female speakers. The face mask condition was motivated by the COVID-19 pandemic and its relevance to forensic speech science (Geng *et al.*, 2023). Based on previous research on humans, we expect poorer speech recognition in pub noise compared to SSN (Mi *et al.*, 2013, where babble noise serves as a proxy for pub noise), at higher noise levels (Brungart *et al.*, 2020), and when a face mask is worn (Bottalico *et al.*, 2020).

Prior to 2020, it was widely accepted that humans outperformed ASR systems in most tasks, including phoneme identification (Sroka and Braida, 2005) and word identification (Carey and Quang, 2005; Lippmann, 1997), and with clean speech (Cutler and Robinson, 1992; Leeuwen *et al.*, 1995). As expected, the introduction of noise or a degraded speech quality substantially reduced ASR performance (Carey and Quang, 2005; Sroka and Braida, 2005). Since 2020, however, the rise of end-to-end deep neural architectures for ASR has led to dramatic improvements in word error rate (WER) in both clean and noisy conditions (e.g., Baevski *et al.*, 2020; Radford *et al.*, 2023). Nevertheless, direct comparisons to human performance have remained limited.

With respect to human comparisons, one of the original end-to-end neural ASR architectures, wav2vec 2.0 (Baevski *et al.*, 2020), achieved impressive WERs, but was not tested against human speech recognition abilities. In

[a]Author to whom correspondence should be addressed.

contrast, the original paper introducing OpenAI's Whisper (Radford *et al.*, 2023), included a preliminary comparison with human performance and found roughly comparable performance. The primary analysis compared Whisper against just one transcriber in ideal studio conditions, while a secondary analysis compared Whisper against four professional transcribers in 25 mixed variety speech recordings. Radford *et al.* (2023) further tested Whisper in various noise conditions, including white noise and pub noise, with the latter simulating a more realistic noisy environment. As expected, increased noise levels corresponded to reduced performance. However, similar WERs were observed across both noise types, highlighting Whisper's robustness to naturalistic background noise like pub noise. These results, however, were not directly compared with human performance, leaving it unclear whether Whisper's high performance in optimal studio conditions would remain comparable in suboptimal, noisy conditions.

To our knowledge, the only other study to compare the effect of noise on human and machine performance comes from Kim *et al.* (2024). Unlike Radford *et al.* (2023), Kim *et al.* (2024) evaluated multiple human listeners and several end-to-end ASR systems (Whisper, Google Speech-to-Text, wav2vec 2.0, and HuBERT). Their study focused on transcribing second language (L2) English speech mixed with speech-shaped noise (SSN) at SNRs ranging from −4 to 8 dB. Of the four ASR systems tested, only Whisper large matched or exceeded human accuracy.

Despite advances made in Radford *et al.* (2023) and Kim *et al.* (2024), critical gaps remain in our understanding of human-like ASR performance. Namely, it is unclear how ASR systems compare to humans in naturalistic noise settings and with straightforward L1 speech transcription where the dialect matches that of the transcriber. While Radford *et al.* (2023) provided insight into Whisper's performance under noisy conditions, their comparison to human performance was limited. Kim *et al.* (2024) compared many human listeners against modern ASR systems, but only tested L2 English speech in speech-shaped noise. Humans and ASR systems are likely to perform better on L1 English in noise, possibly reducing the performance gap between them. Additionally, no study has yet compared the performance degradation between humans and ASR systems in more naturalistic noise conditions, such as pub noise or face mask speech.

The current study aims to evaluate modern ASR systems against human performance in adverse listening conditions to identify current limitations of ASR systems and areas for improvement. Specifically, we compare L1 British English speakers with wav2vec 2.0 and Whisper (base and large models) in transcribing L1 Southern Standard British English speech. The test conditions include two noise types (speech-shaped noise and pub noise), two noise levels (0 and 8 dB), and two mask conditions (cotton mask and no mask).

2. Methods

2.1 Participants and models

2.1.1 Humans

Sixty native speakers of British English participated in the online experiment distributed via Prolific (Prolific, 2023). After completing a consent form, participants underwent a headphone check (Milne *et al.*, 2021) and filled out a demographic questionnaire. To qualify, participants needed to be native British English speakers, over 18 years old, and with no hearing, language, or learning disabilities. Participants were excluded if they did not meet the eligibility criteria, failed the headphone check, or did not proceed to the test trials. Eleven additional participants were excluded for these reasons. Compensation was provided upon experiment completion.

2.1.2 Machines

Two openly available ASR systems, wav2vec 2.0 and Whisper, were used to transcribe audio files. Both the base size models (wav2vec2-base-960h[1] and Whisper base[2]) and the large models (wav2vec2-large-960h[3] and Whisper large-v3[4]) were set to transcribe in English. wav2vec 2.0 is a self-supervised encoder, trained and fine-tuned on all 960 h of the LibriSpeech dataset, which contains North American English clean read speech (Panayotov *et al.*, 2015). The base and large models differ in the number of parameters (Baevski *et al.*, 2020). Whisper, an encoder-decoder architecture trained via weak supervision, was trained on 680 000 h of multilingual labeled data for the base model, with approximately 65% of the data in English. No details were provided about noise in the training data. Whisper large-v3, used for the large model, was trained on 1 million hours of labeled multilingual speech and 4 million hours of speech generated by Whisper large-v2 (Hugging Face, 2024).

2.2 Stimuli

The stimuli were produced in a sound-treated room by three female speakers of Southern Standard British English. A multi-speaker design was employed to ensure that any observed effects were not specific to a single speaker. Across the three speakers, 40 sentences were recorded: 20 without a face mask and 20 with a two-layer cotton mask. The sentences were drawn from low-predictability carrier phrases sourced from Kalikow *et al.* (1977) and contained, at minimum, a subject noun or pronoun, a verb, and a final noun that formed a minimal pair known to cause confusion in everyday speech: e.g., /f/ and /s/, /f/ and /h/, /p/ and /k/, /p/ and /h/ and /s/ and /ʃ/. For instance, "The girl spoke about the fun/sun." The sentences otherwise varied in syntactic form. Read speech was chosen to maintain control over the material and to

minimize the semantic and syntactic predictability of the sentences. The stimuli did contain a high proportion of proper names and homophones (addressed in Sec. 2.4.2).

To assess the effects of different background noise types and cotton mask speech, studio-quality recordings were mixed with SSN and pub noise. The SSN was derived from the 40 sentences, while the pub noise, sourced from Islabonita (2013), featured multiple speakers talking in a pub along with sounds typical of a restaurant, such as plates, glasses, and cutlery. Despite variation in the noise sources, the sound pressure level varied only marginally around 70 dB. We acknowledge that realistic background noise can be more variable, involving multiple sound sources at different levels, which can affect word recognition to varying degrees (Barker *et al.*, 2015).

The background noise was mixed with the clean recordings at two SNRs: 0 dB, where the noise and speech are at the same loudness; and 8 dB, where the speech is 8 dB louder than the noise. Although mixing clean speech with noise is somewhat artificial, this method was implemented to ensure that each sentence was consistently affected by noise. Only two SNR levels were used due to experimental constraints, but these levels still represent distinct degrees of challenging listening conditions. The background noise was added using a custom Praat script (Harrison, 2022; Boersma and Weenink, 2023).

To mitigate differences in vocal intensity between speakers, all stimuli were normalized to 70 dB before mixing with noise. The recordings were resampled to 16 kHz and converted to MP3 format, meeting the requirements of the ASR systems and the experiment builder.

2.3 Procedure

The human experiment was designed using the online experiment builder Gorilla (Anwyl-Irvine *et al.*, 2020) and distributed via Prolific. Participants heard a total of 40 trials in only one SNR level (either 0 or 8 dB). Within their assigned SNR, participants were presented, in a random order, with an equal number of (i) SSN and pub noise sentences and (ii) no mask and cotton mask speech. All three speakers were represented at roughly equal rates, and participants heard each sentence only once. Before beginning the test trials, participants completed a demographic questionnaire and headphone check. Those who passed the check received instructions and completed three practice trials. Participants were instructed to transcribe the sentences as accurately as possible, paying attention to spelling. They were also asked to adjust the volume to a comfortable level during the practice trials, after which they were to keep the same volume throughout the experiment. Each sentence could be played only once, and transcriptions could not be edited after submission.

For machines, each sentence was transcribed by both base models (wav2vec2-base-960h and Whisper base) and large models (wav2vec2-large-960h and Whisper large-v3).

2.4 Performance evaluation

2.4.1 WER analysis

WER was calculated based on the number of substitutions, insertions, and deletions between the transcriber-provided transcript and the reference transcript. Before analysis, all punctuation and extra spaces were removed from the transcriptions. A statistical assessment of WER was then conducted using a Bayesian gamma mixed-effects regression model using the *brms* package in R (Bürkner, 2017; R Core Team, 2023). A separate model was run for each model size: one comparing humans to wav2vec 2.0 base and Whisper base and one comparing humans to wav2vec 2.0 large and Whisper large. Each model included fixed effects for noise type, noise level, mask, and transcriber with all interactions, along with a random intercept for file and speaker. Noise type, noise level, and mask were sum-coded, while transcriber was treatment-coded with the human transcriber level as the baseline. Effects were considered reliable in their direction if the 95% credible interval (CI) of the posterior distribution excluded 0. Further details about the model specifications can be found in the supplementary material.

2.4.2 Correction for proper names and homophones

Although WER is a useful metric for assessing transcription performance, it does come with limitations, including equal treatment of errors in content and function words and potential overestimation of inaccuracy due to variations in spelling for homonyms or proper names. In response to the latter issue, a modified set of transcriptions (subsequently referred to as the "corrected analysis") was produced by two native English speakers. Spellings of homonyms were standardized, as were proper names, provided the spelling suggested a phonological relationship to the original name (e.g., Elisa or Alyssia for Alicia, pronounced as [ɛlɪsia]). The primary analysis used the raw, uncorrected WER, but a secondary analysis was implemented using the corrected WER. These results are discussed briefly and can be found in the supplementary material.

3. Results

The data, analyses, and model outputs for the raw and corrected transcripts can be found on OSF.[5]

3.1 Base models

The noise type, noise level, their interaction, and the presence of a face mask reliably impacted WER for human transcribers (see Fig. 1). The pub noise, a 0 dB SNR, and the presence of a face mask increased WER relative to average and the

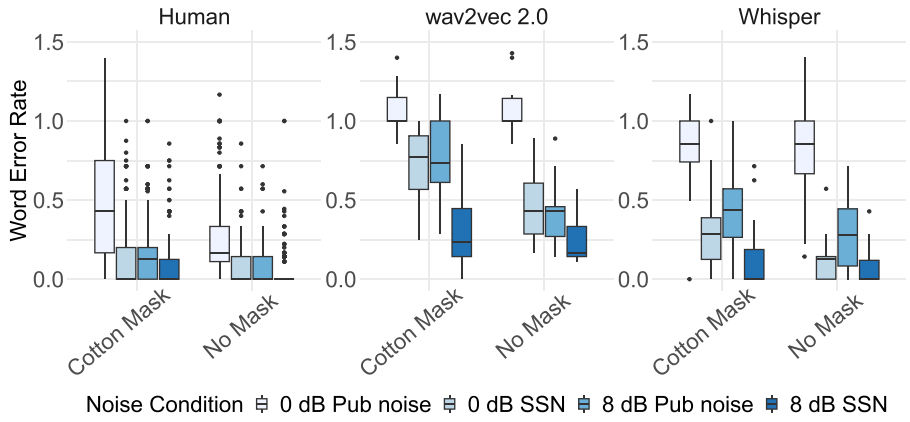

Fig. 1. The distribution of WER for the base raw results. Results are presented according to the transcriber (from left to right: Human, wav2-vec 2.0, Whisper). Mask condition is specified along the *x* axis. Noise conditions are grouped by color and shade.

respective opposing level (pub noise vs SSN: $\beta = 0.47$, 95% CI: [0.35, 0.59]; 0 vs 8 dB: $\beta = 0.51$, 95% CI: [0.38, 0.63]; face mask vs no mask: $\beta = 0.30$, 95% CI: [0.18, 0.42]). In addition, WER in pub noise at the 0 dB level was reliably worse than average (0 dB × pub noise: $\beta = 0.22$, 95% CI: [0.10, 0.34]). Human transcribers reliably outperformed wav2vec 2.0 and Whisper base models (wav2vec2: $\beta = 1.84$, 95% CI: [1.45, 2.26]; Whisper: $\beta = 1.09$, 95% CI: [0.70, 1.51]). No other credible interactions were observed, suggesting that the individual effects were consistent in their influence on WER between humans and machines.

3.2 Large models

As the data for the human transcribers remained the same for this comparison, the major changes in results involve the interactions with wav2vec 2.0 and Whisper (see Fig. 2). Human transcribers still reliably outperformed the large version of wav2vec 2.0 ($\beta = 1.83$, 95% CI: [1.42, 2.27]), and the lack of reliable interactions between wav2vec 2.0 and additional factors indicated that the overall influence of noise type, noise level, and face mask was not reliably different from humans. In contrast, Whisper outperformed human transcribers across almost all conditions ($\beta = -0.69$, 95% CI: [−1.10, −0.25]), except in the influence of noise type: Whisper took a particularly large hit in transcribing speech in pub noise, effectively putting its performance on par with humans ($\beta = 0.44$, 95% CI: [0.02, 0.87]). No other interactions were reliable in their direction.

3.3 Corrected WER analyses

Following correction for proper names and homonyms, a few differences emerged in the results for the base and large sizes, although the patterns were largely the same. For the base models, humans still outperformed wav2vec 2.0 and Whisper. For the large models, humans still outperformed wav2vec 2.0, and Whisper still outperformed humans, except in pub noise. Relative to humans, wav2vec 2.0 improved in the pub noise and 0 dB conditions, and Whisper improved

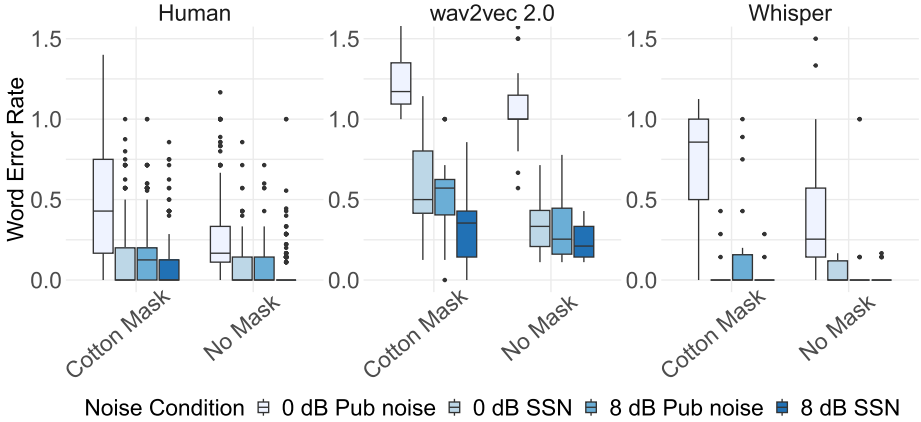

Fig. 2. The distribution of WER for the large raw results. Results are presented according to the transcriber (from left to right: Human, wav2-vec 2.0, Whisper). Mask condition is specified along the *x* axis). Noise conditions are grouped by color and shade.

considerably in the face mask condition, except in pub noise. These differences were, however, minor. The full analyses can be found in the supplementary material.

4. Discussion

For both humans and the tested ASR systems, WER increased in the presence of pub noise, a 0 dB SNR, and face mask speech. Humans outperformed both wav2vec 2.0 and Whisper base versions; and while humans outperformed wav2vec 2.0 large, Whisper large exceeded human performance. The only exception to this was in pub noise, where Whisper large was comparable to human performance. The present study evaluated performance on a mainstream dialect of English, a condition where both L1 listeners and the tested ASR systems were expected to perform well.[6] Compared to Kim *et al.* (2024), who examined the difference between humans and ASR systems on L2 English, overall WERs were considerably lower for L1 English. Contrary to our predictions, however, the magnitude of the difference between humans and ASR systems was similar for both L1 and L2 English.

These findings have important implications for ASR development and our understanding of the differences between human speech perception and ASR capabilities. The gap between human and machine speech recognition has been a long-standing topic of discussion, particularly for modular ASR systems (e.g., Scharenborg, 2007; Moore and Cutler, 2001). As demonstrated by the present study, the performance gap between humans and machines has been substantially narrowed, and in some cases, even bridged. Nonetheless, direct comparisons between human and machine speech recognition continue to provide valuable insights into areas for ASR enhancement, while also highlighting noteworthy similarities and differences among the speech recognition processes.

4.1 The effect of noise

In line with expectations, both the degree and type of noise were challenging for humans and machines. The impact of SNR (an 8 dB difference) on speech recognition was comparable to the effect of noise type for both groups.

Transcribing speech in pub noise was substantially more difficult compared to speech in speech-shaped noise. While it might seem intuitive for ASR systems to respond similarly to humans, this is not guaranteed. Pub noise has high temporal variation in the spectrum, which can lead to increased masking of the speech content; in turn, speech-shaped noise is steady-state with regular masking of energy, but less masking of information content (Zhang *et al.*, 2021). Despite wav2vec 2.0's overall lower transcription performance, the ASR systems mirrored human behavior in their response to speech in noise, even though their architectures and training data differed significantly, particularly given that wav2vec 2.0 was not trained on noisy speech.

Moreover, the difference between noise types was unexpected for the ASR systems, particularly given that Radford *et al.* (2023) found similar results for Whisper in both SSN and pub noise conditions. Their study tested Whisper on the clean test set of the LibriSpeech ASR Corpus (Panayotov *et al.*, 2015) mixed with static white noise and pub noise from the Audio Degradation Toolbox in MATLAB (Mauch and Ewert, 2013) at 0 dB SNR. Static white noise yielded a WER of approximately 17%, whereas pub noise yielded only a marginally higher WER of around 18% (estimated from Radford *et al.*, 2023, p. 8, their Fig. 5). In contrast, the present study observed a higher WER for stimuli mixed with 0 dB pub noise (mean WER = 44%), even when using the largest Whisper model. This discrepancy could be due to differences in the exact pub noise recordings or the speech recordings used in the two studies.

To explore this discrepancy, we mixed our speech recordings with the same pub noise recording used in Radford *et al.* (2023) and re-tested the Whisper large model. As shown in Fig. 3, this pub noise yielded WERs comparable to those in the current study, but higher than those originally reported in Radford *et al.* (2023). The likely explanation for the increased WER thus appears to be the difference in the speech stimuli used across the two studies.

The similar performance patterns in response to noise suggest key commonalities between humans and ASR systems. Even Whisper large, which generally outperformed human transcription abilities, struggled with the 0 dB pub noise

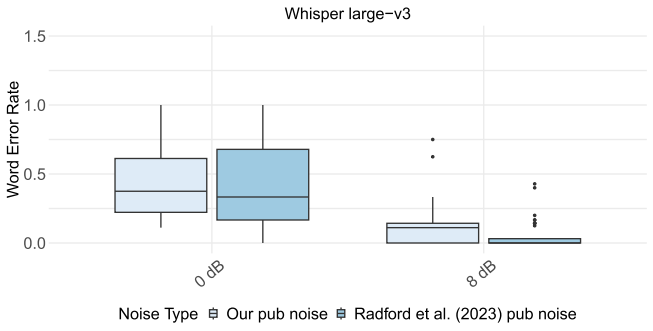

Fig. 3. The distribution of WER for Whisper large-v3 when transcribing our speech data mixed with our pub noise and the pub noise used in Radford *et al.* (2023). Noise level is specified along the *x* axis. Noise type is grouped by color and shade.

condition. By directly comparing humans and ASR systems using the same stimuli and controlled conditions, we can more precisely quantify both the strengths and limitations of ASR systems in recognizing speech under adverse listening conditions. Overall, these findings highlight the need for further testing ASR systems across a range of noise types and levels (e.g., fluctuating noise, such as pub noise, multi-talker babble, or street noise).

4.2 Qualitative analysis of error types

A notable difference between the systems emerged in the types of errors produced. A *post hoc* qualitative analysis was conducted on the error types, with a focus on the 0 dB, pub noise, face mask condition (in principle, the most difficult listening condition). The analysis showed that humans generally maintained grammaticality and transcribed in English, i.e., *Aliasara was chatting about the story* (intended: Olivia was chatting about the cartridge). Whisper exhibited similar errors, producing grammatically correct, even if inaccurate phrases, i.e., *Hi, Elena. How's it going?* (intended: I hope Elena asked about the cell). In contrast, wav2vec 2.0 often produced gibberish outputs, i.e., *i a mas taking about so wer* (intended: Elena was talking about sailing), even with the large model size. This difference likely stems from wav2vec 2.0's character-level predictions, as opposed to Whisper's word-level predictions. Across conditions, wav2vec 2.0 was more prone to ungrammatical or incoherent responses, whereas humans and Whisper generally adhered to English lexical choices and grammatical structure. These distinctions have implications for detecting machine-generated responses and further understanding the nuances between human and machine speech recognition.

5. Conclusion

The primary goal of this study was to better understand the performance boundaries of human and ASR systems (wav2vec 2.0 and Whisper) in recognizing speech under adverse, yet naturalistic, listening conditions. For human participants, pub noise, a 0 dB SNR, and face mask conditions reliably increased WER compared to less challenging conditions. Both wav2vec 2.0 and Whisper base models performed worse than human participants across all scenarios. However, while humans outperformed wav2vec 2.0 large, Whisper large outperformed human participants in all conditions except pub noise, where the two performed comparably. These results have important implications for advancing ASR technology and enhancing our understanding of human vs machine speech recognition.

Supplementary Material

See the supplementary material at https://osf.io/vqwu5/?view_only=effeebc4f9dd44258730f96584d74576, which includes the data, code, analysis, stimuli, and experimental manipulations.

Acknowledgments

This work was supported by a COST Action, Language in the Human Machine Era, Short Term Scientific Mission Grant and the Harding Postgraduate Distinguished Scholarship. We thank Andrew Clark for help with MATLAB and the UZH Phonetics and Speech Sciences group, Cambridge Phonetics Laboratory, and Oxford Wave Research for helpful feedback.

Author Declarations

Conflict of Interest

The authors have no conflicts of interests to disclose.

Ethics Approval

Ethics approval was obtained from the University of Cambridge Faculty of Modern and Medieval Languages and Linguistics Research Committee.

Data Availability

The data that support the findings of this study are openly available on OSF at https://osf.io/vqwu5/?view_only=effeebc4f9dd44258730f96584d74576.

References

[1]https://huggingface.co/facebook/wav2vec2-base-960h
[2]https://huggingface.co/openai/whisper-base
[3]https://huggingface.co/facebook/wav2vec2-large-960h
[4]https://huggingface.co/openai/whisper-large-v3
[5]https://osf.io/vqwu5/?view_only=effeebc4f9dd44258730f96584d74576
[6]Although wav2vec 2.0 was technically trained on North American English speech, major performance differences were not expected between these two mainstream varieties of English, particularly given the consistent speech style (i.e., read speech).

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., and Evershed, J. K. (**2020**). "Gorilla in our midst: An online behavioral experiment builder," Behav. Res. **52**(1), 388–407.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (**2020**). "wav2vec 2.0: A framework for self-supervised learning of speech representations," Neural Inf. Process. Syst. **33**, 12449–12460.
Barker, J., Vincent, E., Ma, N., and Watanabe, S. (**2015**). "The CHiME-3 Challenge: Dataset, task and baselines," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, pp. 1–8.
Bürkner, P. C. (**2017**). "brms: An R package for Bayesian multilevel models using Stan," J. Stat. Softw. **80**(1), 1–28.
Boersma, P., and Weenink, D. (**2023**). "Praat: Doing Phonetics by Computer (version 4.3.14) [Computer Program]," http://www.praat.org (Last viewed June 2024).
Bottalico, P., Murgia, S., Puglisi, G. E., Astolfi, A., and Kirk, K. I. (**2020**). "Effect of masks on speech intelligibility in auralized classrooms," J. Acoust. Soc. Am. **148**(5), 2878–2884.
Brungart, D. S. (**2001**). "Informational and energetic masking effects in the perception of two simultaneous talkers," J. Acoust. Soc. Am. **109**(3), 1101–1109.
Brungart, D. S., Bowers, A. W., and McLaughlin, J. M. (**2020**). "Objective assessment of speech intelligibility in crowded public spaces," Ear Hear. **41**(5), 68S–78S.
Carey, M. J., and Quang, T. P. (**2005**). "A speech similarity distance weighting for robust recognition," in *Proceedings of the 9th Interspeech*, September 4–8, Lisbon, Portugal, pp. 1257–1260.
Cutler, A., and Robinson, T. (**1992**). "Response time as a metric for comparison of speech recognition by humans and machines," in *Proceedings of the International Conference on Spoken Language Processing*, October 13–16, Banff, Alberta, Canada, pp. 189–192.
Geng. P., Lu. Q., Guo. H., and Zeng. J. (**2023**). "The effects of face mask on speech production and its implication for forensic speaker identification: A cross–linguistic study," PLoS ONE **18**(3), e0283724.
Harrison, P. (**2022**). *batchCombineSpeechAndNoiseMatchedNoise.praat*, Praat Script Code.
Hugging Face. (**2024**). "OpenAI Whisper Collection" https://huggingface.co/collections/openai/whisper-release-6501bba2cf999715fd953013 (Last viewed July 16, 2024).
Islabonita. (**2013**). "Freesound: Pub.wav by Islabonita," https://freesound.org/people/Islabonita/sounds/178525/ (Last viewed September 2023).
Kalikow, D. N., Stevens, K. N., and Elliott, L. L. (**1977**). "Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability," J. Acoust. Soc. Am. **61**(5), 1337–1351.
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**(2), 025204.
Leeuwen, D. A., van den Berg, L. G., and Steeneken, H. J. M. (**1995**). "Human benchmarks for speaker independent large vocabulary recognition performance," in *Proceedings of the 4th Eurospeech*, September 18–21, Madrid, Spain, pp. 1461–1464.
Lippmann, R. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**(1), 1–15.
Mauch, M., and Ewert, S. D. (**2013**). "The Audio Degradation Toolbox and its application to robustness evaluation," *in Proceedings of the 14th International Society for Music Information Retrieval Conference*, November 4–8, Curitiba, Brazil, pp. 83–88.
Mi, L., Tao, S., Wang, W., Dong, Q., Jin, S.-H., and Liu, C. (**2013**). "English vowel identification in long-term speech-shaped noise and multi-talker babble for English and Chinese listeners," J. Acoust. Soc. Am. **135**(5), EL307–EL312.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**(2), 338–352.
Milne, A. E., Bianco, R., Poole, K. C., Zhao, S., Oxenham, A. J., Billig, A. J., and Chait, M. (**2021**). "An online headphone screening test based on dichotic pitch," Behav. Res. **53**(4), 1551–1562.
Moore, R. K., and Cutler, A. (**2001**). "Constraints on theories of human vs. machine recognition of speech," in *Proceedings of the Workshop on Speech Recognition as Pattern Classification*, April 25–27, 2001, Nijmegen, Netherlands, pp. 145–150.
Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (**2015**). "Librispeech: An ASR corpus based on public domain audio books," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, April 19–24, South Brisbane, Australia, pp. 5206–5210.
Prolific. (**2023**). "Prolific" https://www.prolific.com (Last viewed September 2023).
R Core Team (**2023**). "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria (Last viewed September 2024).
Radford, A., Kim, J. W., Xu, T., Brockman, G., Mcleavey, C., and Sutskever, I. (**2023**). "Robust speech recognition via large-scale weak supervision," in *Proceedings of the International Conference of Machine Learning*, Vol. 202, pp. 28492–28518.
Scharenborg, O. (**2007**). "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Commun. **49**(5), 336–347.
Sroka, J. J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Commun. **45**, 401–423.
Zhang, L., Schlaghecken, F., Harte, J., and Roberts, K. L. (**2021**). "The influence of the type of background noise on perceptual learning of speech in noise," Front. Neurosci. **15**, 646137.