Large-scale assessment of speech intelligibility

Seung-Eun Kim, Matthew Goldrick, and Ann R. Bradlow
Northwestern University

Word recognition measures from human listeners – i.e., measures of how accurately listeners transcribe speech they heard – have for decades stood as the gold standard for assessing speech intelligibility in a wide range of research and practical applications (see Baese-Berk et al., 2023 for review). In addition to providing insight into the communicative impact of speech variation, the availability of standardized intelligibility scores can serve as a critical study-external basis for stimulus selection and procedure specification (e.g., selecting talkers or sentences to avoid floor or ceiling effects or determining experimental parameters such as signal-to-noise ratio (SNR)). Crowd-sourcing platforms have made intelligibility data collection easier by speeding up participant recruitment, but it is expensive and cumbersome to gather such data, because a substantial number of listeners are required to assess the intelligibility of any single talker. In this context, we introduce a newly-developed database that contains intelligibility measures for publicly-accessible recordings from 139 talkers with 120 English sentences for each talker.

The database has 114 second language (L2) English talkers from 22 L1 backgrounds as well as 25 L1 talkers. For each talker, 120 Hearing in Noise Test (HINT) sentences (Soli & Wong, 2008) were drawn from theArchive of L1 and L2 Scripted and Spontaneous Transcripts and Recordings corpus (ALLSSTAR; Bradlow, n.d.). The 120 sentences from a given talker were equally divided into 8 blocks that differed in levels of SNR. In the first block, sentences were mixed with speech-shaped noise at -4dB SNR, and subsequent blocks at -2, 0, +2, +4, +6, and +8dB SNR; sentences were presented without noise in the final block (Quiet; Q). The order of low-to-high SNR was adopted to elicit the widest possible range of intelligibility scores for each talker. In the first block (-4dB), we would obtain floor intelligibility as listeners do not have experience with the talker and the task, and the SNR is least favorable; in the final block (Q), ceiling intelligibility could be obtained due to the most favorable SNR and maximum exposureto thetalker/task. Sentenceorder within and acrossblockswasfixed acrossalltalkers. Experiments were implemented talker-specifically; for each talker, intelligibility data were collected from 10 L1 English listeners via Prolific. They were US residents, over the age of 18, and had no history of hearing, speech, or language impairments. Participants heard each sentence once and transcribed what they heard. They did 8 practice trials, one for each SNR, before the experiment; these were sentences in the story *The Little Prince* produced by an L1 English speaker who was not included in the test. Our database provides participant-generated transcripts for each sentence as well as word accuracy measures obtained viaAutoscore (Borrie et al., 2019).

There are several noteworthy features of our database. First, it contains talkers with a wide range of intelligibility. For instance, the average recognition accuracy varies from 43.7% to 92.7% for L2 talkers, whileitvariesfrom69.9% to 94.8% for L1 talkers. Second, by examining intelligibility at multiple SNRs, the database shows that this variability across talkers differs by SNR levels, as shown in Fig. 1. Together with audio recordings, this makes our data suitable to identify, for instance, phonetic correlates of speech intelligibility. Talkers in the database are also from diverse language backgrounds, which facilitates analysis of L1 vs. L2 differences or of specific language groups. Lastly, the raw transcripts allow researchers to apply their own scoring methods and/or analyze listener errors at the level of individual words and sounds.

Our database can also be used to assess the performance of speech intelligibility prediction algorithms (as for example in Spille et al., 2018) or automatic speech recognition systems (as in Kim et al., 2024). Comparison between the empirical and predicted/automated data will help identify areas where the algorithms need to be improved (see Scharenborg, 2007). Overall, we believe that our novel large-scale intelligibility dataset has great potential in human speech perception research and has practical applications both in clinical and technological settings.

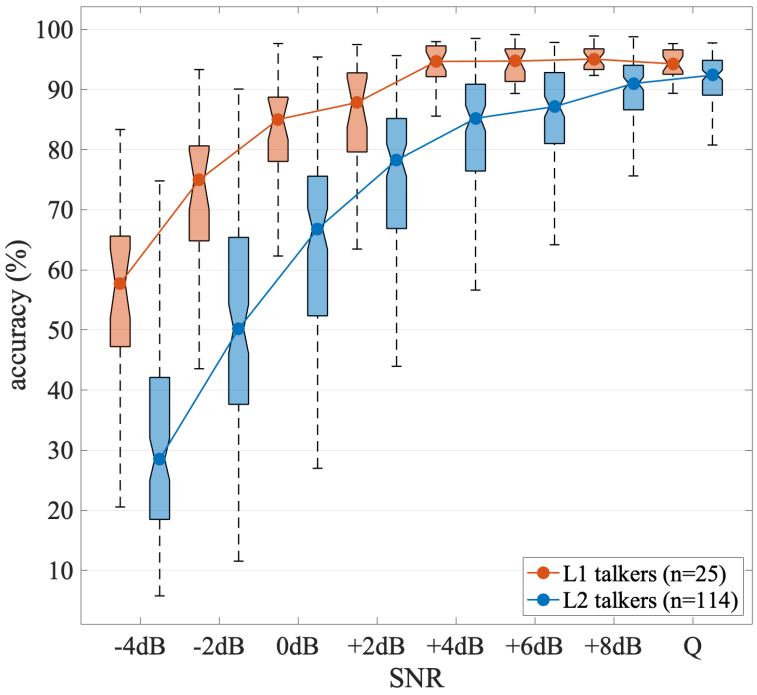

Fig. 1. Distributions of word recognition accuracy (measures of speech intelligibility) at each level of SNR in L1 (orange) and L2 (blue) English talkers. Boxes show 25th and 75th percentiles, and whiskers show minimum and maximum values. Dots in the middle indicate median values.

References

Baese-Berk, M. M., Levi, S. V., & van Engen, K. (2023). Intelligibility as a measure of speech perception: Current approaches, challenges, and recommendations. *The Journal of the Acoustical Society of America, 153*(1), 68-76.

Borrie, S. A., Barrett, T. S., and Yoho, S. E. (2019). Autoscore: An open-source automated tool for scoring listener perception of speech. *The Journal of the Acoustical Society of America, 145*(1), 392–399.

Bradlow, A. R. (n.d.). ALLSSTAR: Archive of L1 and L2 Scripted and Spontaneous Transcripts And Recordings. Retrieved from https://speechbox.linguistics.northwestern.edu/#!/?goto=allsstar

Kim, S.-E., Chernyak, B. R., Seleznova, O., Keshet, J., Goldrick M., & Bradlow, A. R. (2024). Automatic recognition of second language speech-in-noise. *JASA Express Letters, 4*(2), 025204.

Scharenborg, O. (2007). Reaching over the gap: A review of efforts to link human and automatic speech recognition research. *Speech Communication, 49*, 336-347.

Soli, S. D., & Wong, L. L. N. (2009). Assessment of speech intelligibility in noise with the Hearing in Noise Test. *International Journal of Audiology, 47,* 356-361.

Spille, C., Ewert, S. D., Kollmeier, B., & Meyer, B. T. (2018). Predicting speech intelligibility with deep neural networks. *Computer Speech & Language, 48*, 51-66.