THE EFFECT OF SPEAKER ON SPEECH INTELLIGIBILITY

Richard Wright[1], Benjamin V. Tucker[2,3], Matthew C. Kelley[1]

[1]University of Washington, [2]Northern Arizona University, [3]University of Alberta
rawright@uw.edu, benjamin.tucker@nau.edu, mattck@uw.edu

ABSTRACT

Most perception and intelligibility studies assume that noise masks spoken-sentence stimuli equivalently regardless of the speaker. That is, any sentence is equally intelligible across speakers when controlled for style and rate. However, previous studies have shown significant intelligibility variation between speakers. In this study, we predict speaker-specific intelligibility differences in noise for 720 IEEE/Harvard sentences produced by 20 Pacific Northwest speakers (10 women and 10 men). In this corpus, style and rate are controlled in the productions to avoid clear speech. Each sentence recording ($n = 14,400$) was mixed with corpus-shaped noise at three levels (+2, 0, -2 dB) resulting in 43,200 total stimuli. Sentence intelligibility was calculated from orthographic transcriptions provided by 1,868 English-speaking listeners to 120 randomly selected stimuli presented online. We find that different speakers have different intrinsic intelligibility even when reading the same set of sentences. Moreover, some speakers' speech was more affected by noise than others.

Keywords: speech intelligibility, speaker intelligibility, IEEE Corpus, speech in noise.

1. INTRODUCTION

The use of natural speech, as opposed to synthetic speech or pure tones, is well motivated in perceptual testing and hearing-related testing [1]. One might even say that *a word is worth a thousand pure tones*. However, the greater ecological validity of natural speech brings greater variation to the test or experiment. Previous studies have found individual differences in speaker intelligibility related to varying tasks [2, 3]. Even when controlling for task, individual speaker differences in intelligibility are strong enough to overwhelm other factors such as dialect differences [4]. Despite these previous findings, most perception and intelligibility studies implicitly assume that noise masks spoken-sentence stimuli equivalently regardless of the speaker. That is, any one sentence is equally intelligible across speakers when controlled for style and rate. This is particularly problematic if a small number of speakers are used in the stimuli and especially when different listeners are presented with different speaker-stimuli. While there has been significant work on individual speaker differences in clear speech [5, 6], with a few exceptions [7] less work has been done on individual differences in normal speaking styles with controlled materials.

In this study, we examine whether sentences are equivalently intelligible across speakers using the same speaking style. Our hypothesis is that there will be a significant effect of individual speakers on speech intelligibility in noise. Furthermore, we predict that there will be speaker-by-noise interactions resulting in non-linear intelligibility differences at different noise levels. To test this, we used an online perceptual-transcription procedure presenting stimuli from the Pacific Northwest subset of the UWNU IEEE Corpus [8] in three levels of noise. We chose this corpus because it has a large number of sentence stimuli read by a relatively large number of speakers ($n = 20$) and because the recordings were carefully controlled for regional accent and curated to select for equivalent, non-hyperarticulated, speech styles. In examining the results we used individual speaker, English as a first language, and noise as independent variables.

2. METHODS

To investigate the intelligibility across speakers, we conducted a massive online perception study where we asked listeners to orthographically transcribe what they heard.

2.1. Stimuli

The stimuli for this experiment came from recordings of 720 IEEE ('Harvard') sentences [8]. All 720 IEEE sentences were read by 20 native English Speakers (10 women, 10 men) originating from the Pacific Northwest (Washington, Oregon, & Idaho). The recording of the IEEE sentences was controlled for speaking style and rate to avoid the production of hyperarticulated or clear speech. This resulted in a total of 14,400 items. Each recorded

	Sum Sq	Mean Sq	NumDF	DenDF	F value	Pr(>F)
NoiseCond	7745201.56	3872600.78	2.00	178440.16	78403.12	0.0001
English FL	86248.14	86248.14	1.00	1819.56	1746.15	0.0001
Speaker ID	622507.55	32763.56	19.00	179752.55	663.32	0.0001
NoiseCond×English FL	40903.00	20451.50	2.00	178468.19	414.05	0.0001
English FL×Speaker ID	12636.05	665.06	19.00	179778.33	13.46	0.0001
NoiseCond×Speaker ID	266782.01	7020.58	38.00	178257.03	142.14	0.0001

Table 1: ANOVA table of the final linear mixed-effects model. NoiseCond = Noise Condition, English FL = English is first language.

item was subsequently masked using steady corpus-shaped noise at three different signal-to-noise ratios: +2 dB (N1), 0 dB (N2), -2 dB (N3). After the masking manipulation was applied, the total number of stimuli was 43,200.

2.2. Listeners

In the present analysis, 1,868 English-speaking adults were recruited as listeners from the University of Alberta and the University of Washington. Participants received course credit for their participation in the experiment. Participants' age ranged from 16-70 years-old with a mean age of 20.93 (standard deviation: 4.11). A total of 1,143 listeners reported English as their first language and 752 reported another first language. In addition, 1,242 participants reported their gender as female, 589 as male, and 37 as other.

2.3. Procedure

Data was collected using a custom online experiment, presented using a browser, over a two-year period (Nov. 2020 – Nov. 2022). At the beginning of each experimental session, participants were asked to wear headphones and then completed a demographic questionnaire. Each participant was presented with 120 randomly selected test sentences from the full list of stimuli and was asked to type each sentence to the best of their ability. Sentence lists were controlled so that no sentence was repeated in any given experimental session.

There are many potential measures of intelligibility that can be used to investigate speech intelligibility [9, 10, 11]; we chose Levenshtein Distance (LD) to provide a general measure of accuracy. This is a string-edit metric for calculating the number of edits needed to change one string into another. For each response, we calculated the Levenshtein Distance from the participants' typed response to the correct response across the entire sentence [12, 11]. All responses that were blank or contained some variant of "I don't know" were removed from the data leaving just over 180,000 total responses.

2.4. Statistical Analysis

A linear mixed effects analysis was performed using the lme4 package [13] in R [14]. With listener and item as random intercepts, we used the calculated Levenshtein distance as the dependent variable with predictors of `Noise Condition` (NoiseCond: *N1*, *N2*, *N3*) `Speaker` (20 speakers), and whether English was the listener's first language, `English is First Language` (English FL: *yes* or *no*). We also included all possible two 2-way interactions of `English as a First Language`, `Noise Condition`, and `Speaker`. While `Speaker` as a predictor has 20 levels and is a bit difficult to interpret, we chose to include it as a predictor in the models as opposed to a random effect to allow for an investigation of speaker differences in terms of listener accuracy.

3. RESULTS

The statistical model is summarized in the ANOVA table in Table 1 of the linear-mixed effects model. An ANOVA table was selected to summarize the model effects without printing all the interactions in a coefficient table of 20 levels of `Speaker`. The model indicates that all interactions were significant, as were the main effects.

The results indicate that listeners' accuracy decreases as the SNR becomes worse (0dB: $\hat{\beta} = 9.28$, SE=0.18, $t = 51.41$; -2dB: $\hat{\beta} = 13.93$, SE=0.18, $t = 76.09$). When `English is First Language` listeners are more accurate ($\hat{\beta} = -7.63$, SE=0.26, $t = -29.12$) and the effect of `Speaker` is also significant. However, there are also two two-way interactions in the model. The results of the interaction between `English is First Language` and `Noise Condition` are illustrated in

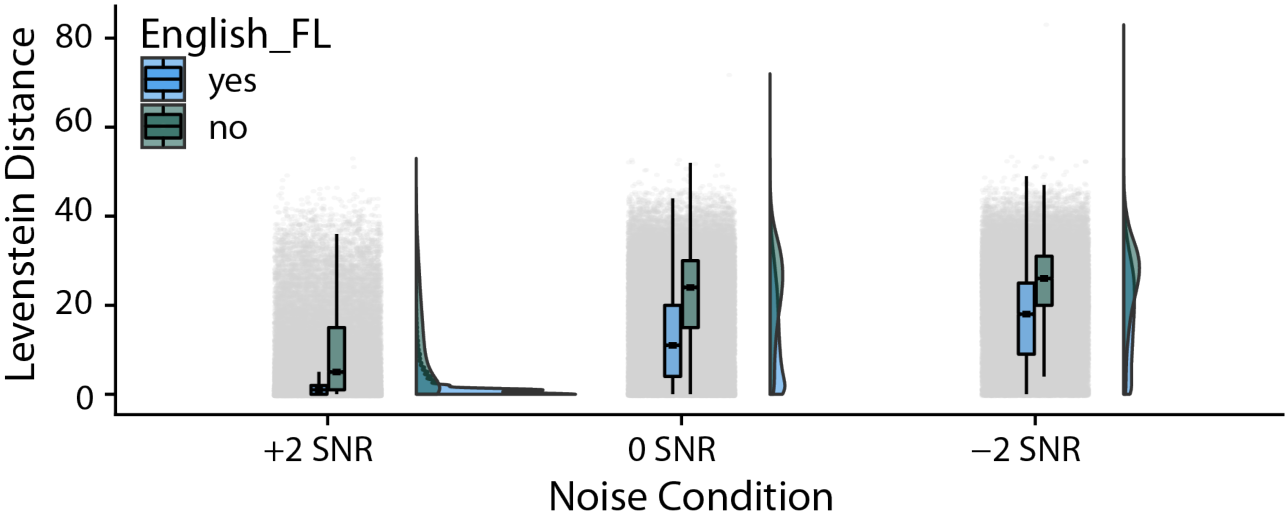

Figure 1: Raincloud plot of interaction between `Noise` and `English is First Language`.

Figure 1. We see that listeners with English as their first language are more accurate than listeners with other first languages and that the difference between these two groups increases as the SNR is degraded. The interaction between `English is First Language` and `Speaker` is also significant.

Figure 2 illustrates the main effects of `Speaker` and the three different noise conditions. It is clear that the individual speakers differ in terms of general intelligibility with some speakers being significantly more intelligible than other speakers. For example, PNM055 and PNF143 are the two most intelligible speakers in all noise conditions while PNM085 and PNF142 are the two least intelligible. The statistical model does not show a speaker gender effect; we simply find that some speakers are more intelligible than others.

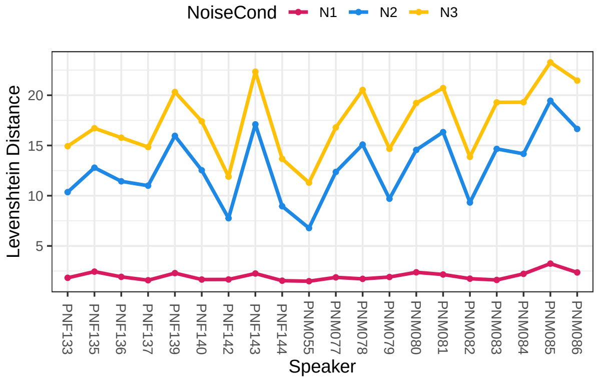

Figure 2: Levenshtein Distance split by talker and noise condition. The 10 speakers on the left are female and the 10 speakers on the right are male.

There is also a significant interaction between `English is First Language` and `Speaker`. Since there are 20 speakers and thus 20 levels in the model it is difficult to visualize the interaction between `Noise Condition` and `Speaker`. Figure 3 illustrates the least intelligible and the most intelligible speakers and their Levenshtein distance for individual items. Items have been sorted in this figure based on their average accuracy. It is interesting to observe the massive impact of increased noise in the signal for the least intelligible speaker while that impact is greatly reduced for the most intelligible speaker.

4. DISCUSSION

In the present study, we used a massive data approach (43,200 total stimuli, 1,868 listeners, and over 180,000 total data points after cleaning) to probe the effect of an individual speaker on sentence intelligibility. The advantage of a large data set like this is that we have increased power which allows us to truly investigate individual speaker effects. We replicated previous studies' findings that there is no reliable effect for speaker-gender on intelligibility [15, 16]. We saw a significant effect of English L1 on intelligibility scores, as expected with lower scores for L2 English listeners [17, 18] and interactions with both noise and speaker (see Figure 1. The main hypothesis, that we would see a significant effect of speaker on sentence intelligibility was confirmed (see Figures 2 & 3).

As is illustrated in Figure 3, showing the most and least intelligible speakers, there are large differences in LD scores and an interaction of speaker with noise. The interaction of speaker with

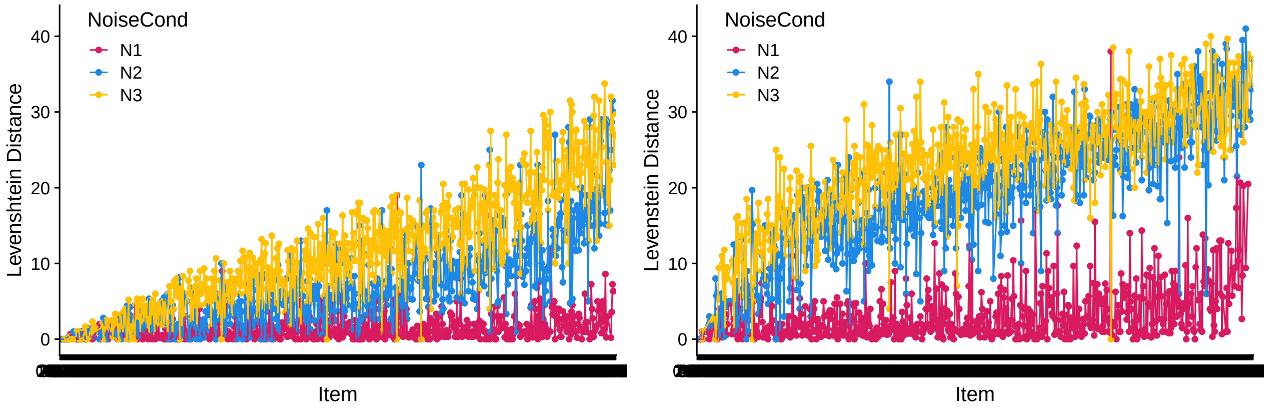

Figure 3: Levenshtein distance for each item (sorted by average Levenshtein distance) split across the three noise conditions for the most intelligible speaker (top, PNM085) and the least intelligible speaker (bottom, PNM055).

noise is important because it indicates a non-linear degradation of speech intelligibility with increased masking noise. That is, a low-intelligibility speaker is masked more effectively than a high-intelligibility speaker even when the linguistic content of the speech is controlled. We also observe with the least intelligible speaker that the difference between N2 and N3 in the noise condition is diminished while it seems to be maintained for the most intelligible speaker. There also appear to be speaker and noise interactions (see Figure 3) with individual sentences, which remains to be investigated in future work with this dataset.

Previous research calculating different measures of intelligibility, including phoneme and feature distance measures [9, 10], makes us confident that the LD distances are representative of listeners' sentence perception. Future work will include analysis of the acoustic characteristics of the individual talkers using vowel space, intensity, fundamental frequency, and dynamic measures [2, 19, 4]. It will also include an investigation of the effect of sentence material on talker intelligibility; as can be seen in Figure 3, there is a sharper rise for the least intelligible speaker in LD scores across sentences (items), hinting that there may be talker-by-sentence interactions.

One implication of this study is that when you recruit a speaker for an intelligibility experiment, or when you select a speaker from a corpus, you don't know where they land on the spectrum of speaker intelligibility and how stimuli based on their speech will interact with masking noise. It is also possible that when sampling a small number of speakers, for example only 2-4, effects of individual speaker intelligibility may inadvertently influence a generalization about gender or dialect. There is a speaker-by-second language interaction that suggests the need for more research on the impact of speaker and noise on second language speakers. It is also possible that a speaker's intelligibility may interact with L2 learners' proficiency or age of acquisition.

5. ACKNOWLEDGMENTS

This work was supported by NIH NIDCD R01 DC006014.

6. REFERENCES

[1] T. Beechey, "Ecological Validity, External Validity, and Mundane Realism in Hearing Science," *Ear and Hearing*, vol. 43, no. 5, pp. 1395–1401, Oct. 2022.

[2] A. R. Bradlow, G. M. Torretta, and D. B. Pisoni, "Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics," *Speech Communication*, vol. 20, no. 3–4, pp. 255–272, Dec. 1996.

[3] M. A. Picheny, N. I. Durlach, and L. D. Braida, "Speaking Clearly for the Hard of Hearing I," *Journal of Speech, Language, and Hearing Research*, vol. 28, no. 1, pp. 96–103, Mar. 1985.

[4] D. R. McCloy, R. A. Wright, and P. E. Souza, "Talker Versus Dialect Effects on Speech Intelligibility: A Symmetrical Study," *Language and Speech*, vol. 58, no. 3, pp. 371–386, Sep. 2015.

[5] A. R. Bradlow and T. Bent, "The clear speech effect for non-native listeners," *The Journal of the Acoustical Society of America*, vol. 112, no. 1, pp. 272–284, Jul. 2002.

[6] S. H. Ferguson and D. Kewley-Port, "Talker Differences in Clear and Conversational Speech: Acoustic Characteristics of Vowels," *Journal of Speech, Language, and Hearing Research*, vol. 50, no. 5, pp. 1241–1255, Oct. 2007.

[7] J. C. Krause and L. D. Braida, “Acoustic properties of naturally produced clear speech at normal speaking rates,” *The Journal of the Acoustical Society of America*, vol. 115, no. 1, pp. 362–378, Jan. 2004.

[8] L. Panfili, J. Haywood, D. R. McCloy, P. Souza, and R. A. Wright, “The UW/NU Corpus 2.0,” Tech. Rep., 2017. [Online]. Available: https://depts.washington.edu/phonlab/projects/uwnu.php

[9] E. Felker, M. Ernestus, and M. Broersma, “Evaluating Dictation Task Measures for the Study of Speech Perception,” in *Proceedings of the 19th International Congress of Phonetic Sciences*, S. Calhoun, P. Escudero, M. Tabain, and P. Warren, Eds. Melbourne: Australasian Speech Science and Technology Association Inc., 2019, pp. 2690–2694.

[10] R. G. Podlubny, T. M. Nearey, G. Kondrak, and B. V. Tucker, “Assessing the importance of several acoustic properties to the perception of spontaneous speech,” *The Journal of the Acoustical Society of America*, vol. 143, no. 4, pp. 2255–2268, Apr. 2018.

[11] H. R. Bosker, “Using fuzzy string matching for automated assessment of listener transcripts in speech intelligibility studies,” *Behavior Research Methods*, Mar. 2021.

[12] E. Sohoglu and M. H. Davis, “Perceptual learning of degraded speech by minimizing prediction error,” *Proceedings of the National Academy of Sciences*, vol. 113, no. 12, pp. E1747–E1756, Mar. 2016.

[13] D. Bates, M. Mächler, B. Bolker, and S. Walker, “Fitting Linear Mixed-Effects Models Using **lme4**,” *Journal of Statistical Software*, vol. 67, no. 1, 2015.

[14] R. C. Team, “R: A language and environment for statistical computin,” Vienna, Austria, 2021. [Online]. Available: http://www.R-project.org/

[15] D. McCloy, L. Panfili, C. John, M. Winn, and R. Wright, “Gender, the individual, and intelligibility,” *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1722–1722, Sep. 2018.

[16] R. A. Wright, B. V. Tucker, M. C. Kelley, and M. Oganyan, “Effects of noise, native language, age, and speaker gender on intelligibility in a large corpus of read speech,” *The Journal of the Acoustical Society of America*, vol. 151, no. 4, pp. A264–A264, Apr. 2022.

[17] M. L. G. Lecumberri and M. Cooke, “Effect of masker type on native and non-native consonant perception in noise,” *The Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2445–2454, Apr. 2006.

[18] S. L. Mattys, M. H. Davis, A. R. Bradlow, and S. K. Scott, “Speech recognition in adverse conditions: A review,” *Language and Cognitive Processes*, vol. 27, no. 7-8, pp. 953–978, Sep. 2012.

[19] R. Smiljanić and A. R. Bradlow, “Production and perception of clear speech in Croatian and English,” *The Journal of the Acoustical Society of America*, vol. 118, no. 3, pp. 1677–1688, Sep. 2005.