VARIABILITY IN NOISE-MASKED CONSONANT IDENTIFICATION

Noah H. Silbert & Lina Motlagh Zadeh

University of Cincinnati, Department of Communication Sciences & Disorders

ABSTRACT

Speech communication commonly occurs in the presence of noise. Perceptual errors in the perception of noise-masked speech vary as a function of noise type (e.g., white noise, speech-shaped noise, multi-talker babble), listener characteristics (e.g., listeners with hearing loss, non-native listeners), and target stimulus properties (e.g., native language of the talker, casual vs clear speech). There is evidence of talker-specific effects in multi-talker-babble-masked sentence intelligibility as well as token-specific effects in speech-shaped-noise-masked CV syllables. The present work analyzes talker- and token-level variation in the identification of a large number of tokens of four consonant categories - [t], [d], [s], [z] - produced by 20 talkers and masked by multi-talker babble. A fitted multilevel logistic regression model illustrates variation in intelligibility between talkers and with respect to within-talker (between-token) variation. The results are discussed in relation to landmark theory and the glimpsing model of speech-in-noise perception.

Keywords: Speech perception, multi-talker babble

1. PERCEPTION OF SPEECH IN NOISE

Speech is often produced and perceived in the presence of background noise. A number of factors influence the intelligibility of noise-masked speech, including characteristics of the noise, the listener, and the target speech.

1.1. Noise characteristics

Patterns of perceptual errors vary as a function of the properties of masker noise. White noise masks some phonological distinctions (e.g., place of articulation) more effectively than others (e.g., voicing) [1, 11, 22], as does signal-dependent noise [2]. On the other hand, speech-shaped noise and multi-talker babble produce very different patterns of perceptual confusions [7, 9, 14, 16, 17, 18].

When speech is masked by other speech, the number of talkers producing the noise has a large effect on perception of the target speech. When only a small number of talkers' speech constitutes masking noise, so-called 'informational masking' interferes with perception of a target speech signal, whereas larger numbers of talkers producing multi-talker babble seem to function more like speech-shaped, temporally modulated noise [6, 17]. There is evidence that spectra-temporal 'glimpses' play a key role in the perception of speech masked by multi-talker babble [7].

1.2. Listener characteristics

Noise masking also interacts with characteristics of the listener. It is particularly difficult for people with hearing loss and cochlear implant users to accurately perceive noise-masked speech [10, 12]. Perception of speech in noise can also be very difficult for non-native listeners and listeners with various language-related disorders [5, 9, 23].

1.3. Signal characteristics

Properties of the target speech signal also influence intelligibility. Speech produced by non-native talkers can be difficult to perceive accurately [4], and clear speech can increase the intelligibility of noise-masked speech significantly [10, 12, 15]. Talker-level variation influences sentence-level intelligibility [3], and some recent work has probed the relationship between token-level variability and intelligibility [18].

1.4. Talkers and tokens in phonetic identification

The present work is an exploratory analysis of simultaneous between- and within-talker variation in the intelligibility of noise-masked consonants in CV syllables. Observed and modeled response accuracies from a multi-talker-babble-masked consonant identification experiment indicate (a) that talkers vary with respect to the overall intelligibility of their speech and also with respect to the *degree of variation* in intelligibility of individual tokens, and (b) that between-talker variation also varies across consonant categories. These results will be discussed in the context of landmark theory [19] and the glimpsing model of noise-masked speech perception [7].

2. METHOD

2.1. Participants

Eleven normal-hearing listeners participated in the experiment. All 11 were female. Their ages ranged from 20 to 32. Two listeners were bilingual (English & Hindi; English & Vietnamese), and two listeners learned English as adults (L1 Korean & Persian).

2.2. Stimuli

The stimuli consisted of target CV syllables [ta], [da], [sa], and [za] embedded in 10-talker babble. The target speech consisted of 10 tokens of each syllable produced by 20 native English talkers. The multi-talker babble consisted of randomly selected 1.5s sections of randomly selected sentences. Each of 20 noise talkers (no overlap with the target talkers) produced 100 of the Harvard sentences.[1] Target syllables were embedded at -2 dB SNR, based on the peak RMS energy across five 100ms segments of the targets and the total RMS energy of the associated noise segment.

2.3. Procedure

Participants completed 7 blocks. In each block, 10 target talkers and 10 noise talkers (each set of 10 split by sex: 5 male, 5 female) were randomly selected. Each block consisted of 400 trials, each trial corresponding to a unique token (10 talkers $\times$ 10 tokens $\times$ 4 consonant categories). Stimuli were presented binaurally at $\sim$60 DB SPL in a sound-attenuating booth. Responses were collected via button boxes. On-screen text indicated four response options (d, z, t, s). Feedback was given on every trial (i.e., 'correct' or 'incorrect'; the correct response was indicated by changing the correct option color to blue, and incorrect responses were changed to red). The experiment was implemented in PsychoPy [13].

2.4. Analysis

A multilevel logistic regression model was fit using PyStan [21]. The dependent variable for each trial i, $y_i \in \{0,1\}$, was modeled as a logistic function of a linear combination of listener (l) and token-level (k) terms, with the token-level terms specified for each combination of consonant (c) and talker (t):[2]

$$(1) \quad y_i \sim \text{Bernoulli}\left(\text{logit}^{-1}\left(\beta_l + \gamma_{c,t,k}\right)\right)$$

The listener parameters were governed by group-level mean and SD parameters:

$$(2) \quad \begin{aligned} \beta_l &\sim \text{Normal}(\mu_\beta, \sigma_\beta) \\ \mu_\beta &\sim \text{Normal}(0,1) \\ \sigma_\beta &\sim \text{Gamma}(2,4) \end{aligned}$$

The token-level parameters were governed by parameters at the talker and consonant level:

$$(3) \quad \begin{aligned} \gamma_{c,t,k} &\sim \text{Normal}(\alpha_{c,t}, \sigma_t) \\ \sigma_t &\sim \text{Gamma}(2,4) \end{aligned}$$

The α parameters were governed by group-level mean and SD parameters:

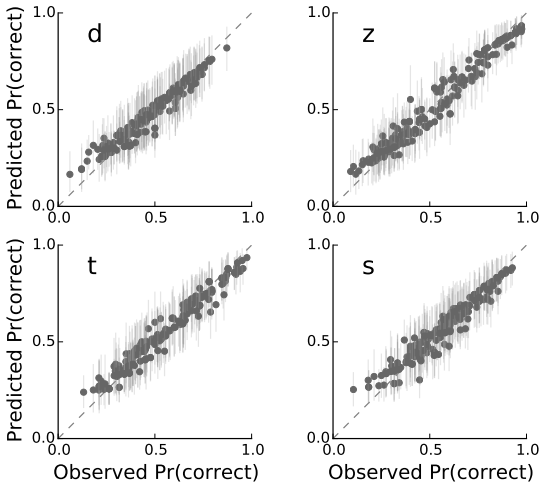
$$(4) \quad \begin{aligned} \alpha_{c,t} &\sim \text{Normal}(\mu_c, \sigma_c) \\ \mu_c &\sim \text{Normal}(0,1) \\ \sigma_c &\sim \text{Gamma}(2,4) \end{aligned}$$

3. RESULTS

3.1. Model fit

Figure 1 shows observed (x-axis) and estimated (y-axis) response accuracy (intelligibility) at the token level for each consonant category (as labeled in each panel).

Figure 1: Token-level observed (x-axis) and predicted (y-axis) response accuracy separately for each consonant category



The vertical lines indicate the 95% highest density intervals (HDIs; roughly, Bayesian CIs [8]), and the points indicate the observed and mean posterior response accuracies. The dashed diagonal lines indicate equal observed and estimated values.

The model provides a good overall fit to the data, as indicated by the close correspondence between

the observed and estimated accuracies, as well as the fact that the diagonal is within the HDIs for nearly all estimated accuracies (the multilevel structure of the model pulls estimates corresponding to extreme observations - very low or very high accuracies - toward more typical values).

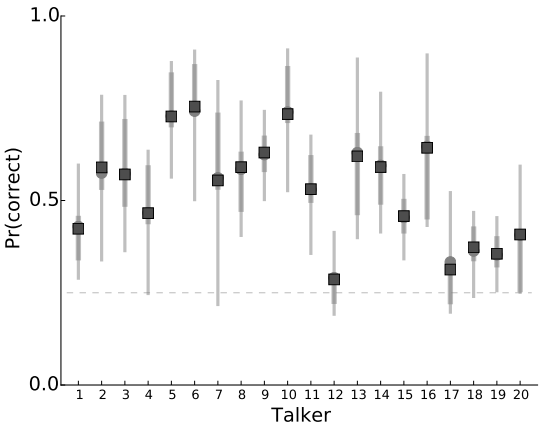
Figure 1 also shows that response accuracy varied somewhat across consonants. There are very few highly intelligible tokens of [d], whereas there are a number of highly intelligible [z], [t], and [s] tokens. Accuracy spanned a wider range for [z] than for any of the other categories, and the voiceless categories had relatively few very *unintelligible* tokens.

3.2. Variation in talker intelligibility

Figure 2 illustrates observed and estimated accuracies for each talker (female: 1-10, male: 11-20). The thin vertical lines indicate 95% HDIs, and the thick vertical lines indicate 50% HDIs (i.e., the interquartile range of the estimated accuracies). Squares indicate observed accuracies for each talker, and the horizontal dashed line at 0.25 indicates chance-level accuracy. There is substantial variation across talkers; observed accuracies ranged from $\sim$0.29 to $\sim$0.76.

Overall, the male talkers were somewhat less intelligible than the female talkers. Five of the male talkers had very low intelligibility ($\leq$0.40), whereas only one or two female talkers were similarly untintelligible ($\leq$0.50). The precision of the estimated talker-based accuracies (i.e., the extent of the HDIs) varied quite a bit across talkers, as well, with some fairly precise estimates (e.g., talkers 5, 9, 12, 15, 18, and 19) and some rather less precise (e.g., talkers 7, 13, and 16).

Figure 2: Observed (squares) and predicted (lines, circles) response accuracy by talker (blue: female; red: male).

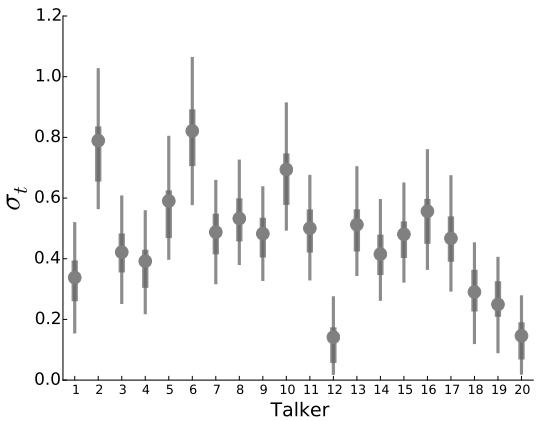


3.3. Variation in variation across tokens and talkers

The standard deviation (SD) parameters at the intermediate and group levels of the model allow us to also probe how *variation* itself varies at different levels of analysis. Figure 3 shows the estimated σ_t parameters, which govern variation across tokens within talkers. Across-token variation differs quite a bit across talkers, in much the same way that overall accuracy varies across talkers (Fig. 2).

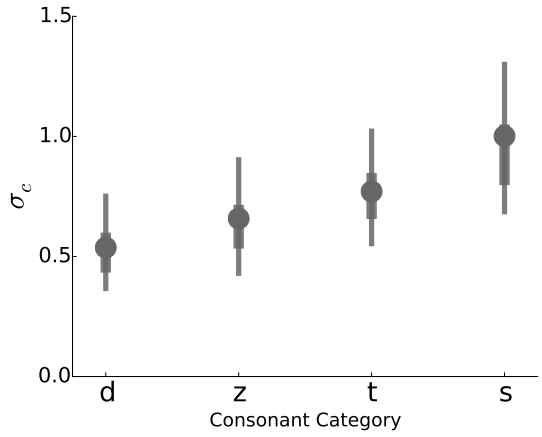
The mean posterior correlation between estimated talker accuracies and estimated σ_t parameters is 0.70 (95% HDI = $[0.54, 0.83]$), indicating that more intelligible talkers tend to have more variation in the intelligibility of their tokens. It is likely that a floor effect influences this pattern somewhat, since a very unintelligible talker's tokens are more free to vary in the direction of higher intelligibility (i.e., there is a hard limit to how *unintelligible* tokens can be). However, this can't be the whole story, since, e.g., talkers 12 and 17 had nearly equal (and equally low) intelligibility (Fig. 2), but 12's token-level variation is quite a bit lower than 17's (Fig. 3). There are similar dissociations between accuracy and token-variability among the more highly intelligible talkers (e.g., talkers 2 and 3).

Figure 3: Estimated within-talker standard deviation (σ_t) by talker (blue: female; red: male)



Estimates and HDIs of the σ_c parameters governing between-talker variation across consonants are shown in Figure 4. Although there is a fair amount of overlap in the HDIs across the four categories, there are a two suggestive patterns. First, there seems to be more variation across talkers in the voiceless categories than in the voiced. Second, there seems to be more variation across talkers in the fricatives than in the stops within each voicing category (i.e., more variability in [z] than [d], and in [s] than [t]).

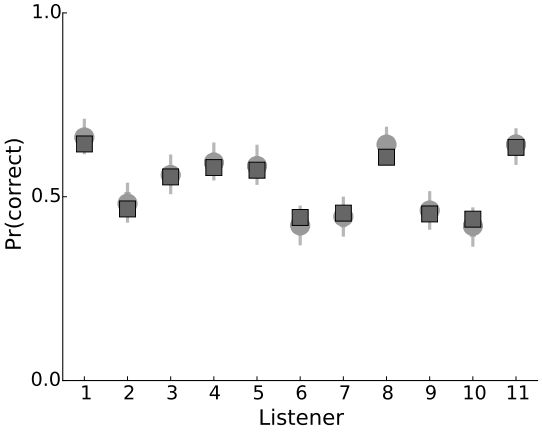
Figure 4: Estimated between-talker standard deviation (σ_c) by consonant category



3.4. Variation in listener accuracy

Figure 5 shows observed and estimated response accuracies for each listener. The vertical lines indicate 95% HDIs, the circles indicate mean posterior estimated accuracies, and the squares indicate observed accuracies. Estimated and observed accuracies ranged from ∼0.44 to ∼0.64. The two bilinguals were listeners 2 and 10, while listeners 5 and 7 were second-language speakers of English.

Figure 5: Observed (squares) and predicted (lines, circles) response accuracy by listener. Lines indicate 95% HDIs.



4. CONCLUSION AND DISCUSSION

Numerous factors influence the perception of speech in noise, including the properties of the noise, the target signal, and the listener. Previous work has shown that sentence intelligibility can vary significantly across talkers [3] and that consonant identification error rates can differ across individual tokens [18]. In the present work, we analyzed talker- and token-level variation in a large corpus of multi-talker-babble-masked consonant identifications.

Consistent with previous work, we observed sizable variation in both talker- and token-level intelligibility. Within each consonant category, variability in token-level intelligibility is large (Fig 1), as is variation across talkers (Fig. 2). Estimates of SD parameters in the multilevel model indicate that talkers also vary with respect to token-level *variability* (Fig. 3) and that between-talker variation differs across consonant categories (Fig. 4).

The properties of the tokens that drive these differences in intelligibility and levels of variation are currently unknown. However, the combination of landmark theory [19] and the glimpsing model of babble-masked speech perception [7] offer a promising approach to analyzing these kinds of variation.

Landmark theory concerns the perceptual impact spectro-temporal discontinuities in the speech signal associated with particular articulations, and there is strong evidence that the portions of speech with the greatest spectro-temporal change bear the most perceptual information [20]. In the glimpsing model of perception, noise-masked speech is perceived on the basis of the spectro-temporal portions of the signal that are more energetic than the masking noise. Putting these two ideas together, it seems reasonable to hypothesize that talker- and token-level variation in intelligibility are driven by variation in the presence of landmarks in the signal and the robustness of those landmarks with respect to noise.

We further hypothesize that differences in token-level variability across talkers are driven by variability in the production of noise-robust landmarks, and that between-talker variability with respect to phonological category is closely related to the nature of the landmarks corresponding to particular phonological features (e.g., voicing, manner of articulation). Interactions between the landmarks for different phonological distinctions and the spectra-temporal properties of different noise types may explain the differences in perceptual errors due to white noise, speech-shaped noise, and multi-talker babble, as well.

A deeper understanding of the properties of noise-robust speech and the mechanisms by which noise-distorted speech can be accurately perceived promises to provide useful tools for mitigating common communication difficulties caused by noise. Listeners with hearing loss, non-native listeners, people who work in noisy environments, and their interlocutors stand to benefit from such tools.

5. REFERENCES

[1] Allen, J. B. 2005. Consonant recognition and the articulation index. *The Journal of the Acoustical Society of America* 117(4), 2212–2223.

[2] Benkí, J. R. June 2003. Analysis of english nonsense syllable recognition in noise. *Phonetica* 60(2), 129–157.

[3] Bent, T., Buchwald, A., Pisoni, D. B. 2009. Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *The Journal of the Acoustical Society of America* 126(5), 2660–2669.

[4] Bradlow, A. R., Bent, T. 2008. Perceptual adaptation to non-native speech. *Cognition* 106(2), 707–729.

[5] Bradlow, A. R., Kraus, N., Hayes, E. 2003. Speaking clearly for children with learning disabilities: Sentence perception in noise. *Journal of Speech Language and Hearing Research* 46(1), 80.

[6] Brungart, D. S. 2001. Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America* 109(3), 1101–1109.

[7] Cooke, M. 2006. A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America* 119(3), 1562–1573.

[8] Kruschke, J. 2010. *Doing Bayesian Data Analysis: A Tutorial Introduction with R and BUGS*. Academic Press.

[9] Lecumberri, M. L. G., Cooke, M. 2006. Effect of masker type on native and non-native consonant perception in noise. *The Journal of the Acoustical Society of America* 119(4), 2445–2454.

[10] Liu, S., Rio, E. D., Bradlow, A. R., Zeng, F.-G. 2004. Clear speech perception in acoustic and electric hearing. *The Journal of the Acoustical Society of America* 116(4), 2374–2383.

[11] Miller, G. A., Nicely, P. E. 1955. An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America* 27, 338–352.

[12] Payton, K. L., Uchanski, R. M., Braida, L. D. 1994. Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. *The Journal of the Acoustical Society of America* 95(3), 1581–1592.

[13] Peirce, J. W. 2007. PsychoPy—psychophysics software in python. *Journal of Neuroscience Methods* 162(1–2), 8–13.

[14] Phatak, S. A., Allen, J. B. 2007. Consonant and vowel confusions in speech-weighted noise. *The Journal of the Acoustical Society of America* 121(4), 2312–2326.

[15] Picheny, M. A., Durlach, N. I., Braida, L. D. 1985. Speaking clearly for the hard of hearing i: Intelligibility differences between clear and conversational speech. *Journal of Speech Language and Hearing Research* 28(1), 96.

[16] Silbert, N. H. 2014. Perception of voicing and place of articulation in labial and alveolar english stop consonants. *Laboratory Phonology* 5(2), 289–335.

[17] Simpson, S. A., Cooke, M. 2005. Consonant identification in n-talker babble is a nonmonotonic function of n. *The Journal of the Acoustical Society of America* 118(5), 2775–2778.

[18] Singh, R., Allen, J. B. 2012. The influence of stop consonants' perceptual features on the articulation index model. *The Journal of the Acoustical Society of America* 131(4), 3051–3068.

[19] Stevens, K. N. 2002. Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America* 111, 1872.

[20] Stilp, C. E. 2014. Information-bearing acoustic change outperforms duration in predicting intelligibility of full-spectrum and noise-vocoded sentencesa). *The Journal of the Acoustical Society of America* 135(3), 1518–1529.

[21] Team, S. D. 2014. Pystan: The python interface to stan.

[22] Wang, M. D., Bilger, R. C. 1973. Consonant confusions in noise: a study of perceptual features. *The Journal of the Acoustical Society of America* 54(5), 1248–1266.

[23] Ziegler, J. C., Pech-Georgel, C., George, F., Lorenzi, C. 2009. Speech-perception-in-noise deficits in dyslexia. *Developmental Science* 12(5), 732–745.

[1] http://www.cs.columbia.edu/ hgs/audio/harvard.html

[2] Preliminary analyses provided no evidence of systematic changes in response accuracy across blocks, so no block-based terms are included in the model.