Across and within consonant errors for isolated syllables in noise

Joseph C. Toscano*
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign

Jont B. Allen
Department of Electrical and Computer Engineering
Beckman Institute for Advanced Science and Technology
University of Illinois at Urbana-Champaign

To appear in *Journal of Speech, Language, and Hearing Research*
Available online at: http://dx.doi.org/10.1044/2014_JSLHR-H-13-0244/

Purpose. A critical issue in assessing speech recognition involves understanding the factors that cause listeners to make errors. Models like the articulation index show that average error decreases logarithmically with increases in *signal-to-noise ratio* (SNR). We investigated whether (1) this log-linear relationship holds across consonants and for individual tokens, and (2) what accounts for differences in error rates at the across- and within-consonant levels.

Method. Normal-hearing listeners heard CV syllables (16 consonants and four vowels) spoken by 14 talkers, presented at six SNRs. Stimuli were presented randomly, and listeners indicated which syllable they heard.

Results. The log-linear relationship between error and SNR holds across consonants, but breaks down at the token level. These two sources of variability (across- and within-consonant factors) explain the majority of listeners' errors. Moreover, by simply adjusting for differences in token-level error thresholds, we can explain 62% of the variability in listeners' responses.

Conclusions. These results demonstrate that speech tests must control for the large variability between tokens, not average across them, as is commonly done in clinical practice. Accounting for token-level differences in error thresholds with normal-hearing listeners provides a basis for tests designed to diagnostically evaluate individual differences with hearing-impaired listeners.

A critical issue in speech perception involves understanding the factors that cause listeners to make errors. Normal-hearing (NH) listeners are remarkably good at speech recognition, making very few errors in quiet or low levels of background noise (Singh & Allen, 2012). At higher noise levels, however, they eventually make errors when identifying consonants and vowels. Are these errors random? Or, are there systematic differences between speech sounds that cause listeners to make errors (e.g., are certain sounds more robust to noise than others)?

These questions are closely related to the assessment of speech recognition with hearing-impaired (HI) listeners, who often have difficulty recognizing speech even in quiet. Clinicians would like to be able to assess a listener's ability to perceive speech in noise, which we focus on here and which can provide information for evaluating hearing aid performance (along with other measures). However, existing speech tests

* Corresponding author: joseph.toscano@villanova.edu

have had limited success (Walden, Schwartz, Williams, Holum-Hardegen, & Crowley, 1983; Dobie & Sakai, 2001; Haskell, Noffsinger, Larson, Williams, Dobie, & Rogers, 2002), though certain tests can be used for predicting hearing aid acceptance (Nabelek, Freyaldenhoven, Tampas, Burchfield, & Muenchen, 2006). Nonetheless, the lack of an effective test for evaluating speech perception in noise is problematic, since the main purpose of wearing a hearing aid is to improve speech understanding (Ward, 1983).

Minimally, we would like a test that allows us to quantify a listener's sensitivity (in one ear or both) to a degraded speech signal. Ideally, such a test would allow us to fit a hearing aid or adjust a communication system. This was the goal of Fletcher's *articulation index* (AI) measure, for example, which was used in World War II to optimize pilot-navigator communications (Allen, 1996). However, despite nearly 100 years of work on this problem, existing speech tests are still insufficient (Walden et al., 1983; Dobie, 2011; Taylor, 2006).

How can we resolve this problem? The approach we take here is to examine the nature of speech recognition with NH listeners at the level of individual consonants and tokens. This, in turn, can provide a baseline for developing speech tests that could be used for individual assessment with HI listeners.

At a very basic level, speech recognition can be characterized first as an acoustic signal processing stage (Allen, 2005a; b), followed by an information processing stage (Brokhorst, Bosman, & Smoorenburg, 1993; Bronkhorst, Brand, & Wagener, 2002): Given a speech signal in noise, listeners must extract the relevant acoustic cues that provide information about the underlying linguistic message.[1] The problem is made more difficult by the nearly open-set nature of the task (e.g., the large number of words that a listener might need to identify), and the fact that speech is highly affected by context (coarticulatory context, linguistic context, etc.).

Fletcher's early work on speech recognition (Fletcher, 1929) yielded many key insights about the relationship between errors that listeners make and the noise level (described by the *signal-to-noise ratio* [SNR]). Specifically, he observed that listeners' *average log error* decreases linearly as a function of the SNR measured in *auditory critical bands* (CBs; Allen, 1994). This observation (stemming from the *average* error) forms the basis of the AI, which is defined as the average SNR over 20 CBs covering the speech range (0.3 to 7.5 [kHz]).[2] For stimuli presented in speech-shaped noise, the AI is simply the wide-band SNR. The AI, and more recently, another measure, the *speech recognition threshold* (SRT), have been used for a number of years for research in audiometric speech assessment (Plomp & Mimpen, 1979; Plomp, 1986; Festen & Plomp, 1990; Humes, Dirks, Bell, & Ahlstbom, 1986; Pavlovic, Studebaker, & Sherbecoe, 1986; Rankovic, 1991).

However, both the AI and SRT have severe limitations, and researchers have noted the shortcomings when using these measures to assess recognition with HI listeners (Kamm, Dirks, & Bell, 1985; Humes et al., 1986). Why do speech tests based on these measures perform so poorly in predicting listeners' success? We will

[1] A great deal of research on speech perception has focused on how listeners distinguish between two phonemes (e.g., along a continuum varying from /b/ vs. /p/), or between sounds varying in distinctive features (e.g., voicing, manner, and place). Here, we focus specifically on how listeners recognize natural speech in noise.

[2] The current ANSI standard also includes band importance in its definition of the AI. Here, we refer to the version based on auditory CBs (as this is the aspect of the AI that is relevant to the current study).

show that these tests obscure listeners' errors by averaging across tokens. That is, they provide no information about listeners' responses to individual speech sounds. This is a major source of the problem with existing tests, since it ignores the considerable variability in natural speech attributable to talker identity, speaking rate, and coarticulation (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Repp, 1982; Fowler, 1984; Massaro & Cohen, 1983; Toscano & McMurray, 2012; Bronkhorst et al., 1993; 2002), as well as other factors not directly related to the signal, such as task demands, linguistic context, and individual differences between listeners (which are, of course, critical for assessing hearing loss). We will show that the problem is not with speech testing itself, but rather with the way the tests are scored (i.e., averaging across different sounds).

Recently, Singh and Allen (2012) demonstrated that, for stop consonants, there are indeed large differences in error rates both across consonants and for individual tokens of a given consonant. Two productions of the same stop consonant spoken by different talkers can be very different in their robustness to noise (i.e., listeners may be able to accurately recognize one token in noise but not another due to subtle differences in how the tokens are produced). Singh and Allen found that (1) for most stop consonant tokens, error rates increased abruptly beyond a critical SNR (i.e., the nature of the error is binary; NH listeners uniformly make errors below a particular SNR and almost never make errors above that SNR); (2) the NH listeners in the experiment performed nearly identically to each other; and (3) consonant error rates at -2 [dB SNR] and quiet (no noise) were extremely low (<5% on average). The majority of the error above -2 [dB SNR] is driven by a small number of tokens (e.g., 1 in 20) that have significant error in quiet. That is, most sounds have no error and a few sounds have large error (Phatak & Allen, 2007). This story here is clear: Responses at and above -2 [dB SNR] consist of a bimodal distribution with a large number of zero error sounds, plus a small number of high error sounds.

These results suggest that we need measures that capitalize on this bimodal error distribution. Such detailed information is necessarily lost in aggregate measures. By looking only at the mean error across tokens, the AI and SRT average over the large natural variability between speech sounds. The goal here is to address these issues by further examining listeners' error rates across a broad range of consonant classes, including voiced stops (/b, d, g/), voiceless stops (/p, t, k/), voiced fricatives (/v, ð, z, ʒ/), voiceless fricatives (/f, θ, s, ʃ/), and nasals (/m, n/), for a large set of individual tokens (from 14 different talkers, produced in four different vowel contexts). This will allow us to parse differences across and within consonant classes (i.e., token differences) and to quantify the total error attributable to each of these factors.

The remainder of the paper is organized as follows: First, we provide a brief review of previous approaches for explaining listeners' errors in speech recognition tasks. Next, we present the results of an experiment examining listeners' ability to identify CV syllables in noise, varying in the identity of the consonant, vowel, and talker. These data will be analyzed to determine the extent to which tokens differ in their error rates both across and within consonants.

The Articulation Index

One of the earliest attempts to quantify listeners' speech recognition performance in noise is Fletcher's articulation index (AI), defined as:

$$AI(SNR) \equiv \frac{1}{20}\sum_{k=1}^{20} snr_k \qquad (1)$$

where snr_k is the SNR (measured in [dB]) in the kth *critical band* (CB), normalized by 30 [dB] (Fletcher, 1929; French & Steinberg, 1947; Allen, 1994; see also Note 2). There are 20 CBs corresponding to 20 [mm] along the basilar membrane, covering the frequency range from 0.3 to 7.5 [kHz].

Due to the normalization, the AI ranges from 0 to 1, such that an AI of 1 indicates maximum performance, with an error of e_{min} (e.g., 2% error, 98% correct), while an AI of 0 indicates chance performance (e.g., 1/16 = 6.25% correct for a 16AFC task). Fletcher found that, for a CV listening test, the AI can be used to quantify the *across-consonant average error* as

$$e(SNR) = e_{\min}^{AI}, \qquad (2)$$

defined as the average probability of error in recognizing the CV, as a function of the AI, which, for speech-weighted noise, is the same as the *wide-band SNR*. Thus, the AI models the across-consonant average error (e) as the product of average band errors over 20 CBs (Allen, 1994; Allen, 2005a; Phatak & Allen, 2007; Li, Trevino, Menon, Allen, 2012). Note that if any of the 20 CBs contains a perceptual cue, the corresponding band error could theoretically go to zero, causing the total error (product of the band errors) to be zero (i.e., probability correct=1).

Across vs. within consonant errors

While Fletcher had a deep insight in creating this model (Allen, 1996), and while it nicely captures listeners' *average* error, it is unclear whether it works for individual consonants and tokens (Allen, 1994; 2005a). If token-level error functions are log-linear, as the average is, such an approach could potentially explain listeners' errors for both individual speech sounds and average scores. This is one of the questions addressed in the present study.

Miller and Nicely (1955) provided some early insights about errors for specific consonants, demonstrating that they vary in their intelligibility (Allen, 2005b). Sounds like /v/ have much higher errors at a given SNR than sounds like /p, t, k/. At a particular SNR, some consonants are easier to recognize than others. Thus, *across-consonant error rates* are highly variable (Phatak & Allen, 2007). Namely, the average error for each consonant (e_c) depends strongly on the consonant that was spoken (c).

Recently Singh and Allen (2012) examined this issue further to determine whether the predictions of the AI hold both across and within stop consonant tokens. They measured NH listeners' errors for 56 tokens of 24 CVs (six consonants, /p, t, k, b, d, g/, and four vowels, /ɑ, eɪ, ɪ, æ/), spoken by 14 talkers, at six SNRs (-22, -20, -16, -10, -2, and quiet), with 25 NH listeners. They found that for the -2 dB SNR and quiet conditions, listeners made almost no errors, recognizing 95% of the CV tokens with *zero error*.

Importantly, they also found across-consonant differences in error rates above -10 [dB SNR]: /g, k/ had low error (≈1%), /t, p, d/ had moderate error (≈3-5%), and /b/ had high error (≈18%). As with the AI (which averages different phonemes together), the log-error for each consonant was approximately linear as a function of SNR, increasing to chance performance below -22 dB SNR. Thus, the AI (Eq. 2) seems to hold across stop consonants, but has a consonant-dependent slope and intercept (Phatak & Allen, 2007).

In contrast, within-consonant errors (i.e., differences in error rates between tokens of the same consonant) above -10 [dB SNR] were caused by a small group of high error sounds. For example, 41 /p/ tokens had no errors above -10 [dB SNR], 11 showed a single error (no different from zero; $p>0.05$), and only 4 had high error ($p<0.001$). Above

the SNR at which the error function crosses 50% (defined as the SNR_{50}), within-consonant errors show an abrupt drop from chance to zero (i.e., above the SNR_{50}, NH listeners make virtually no errors).

This observation fits a binary ("all or nothing") within-consonant error model, centered about the SNR_{50} threshold. This threshold provides an important measure of within-consonant noise robustness, and the standard deviation (SD) of these thresholds is large (>15-20 [dB SNR]). Thus, the AI model does not seem to be accurate at the level of individual tokens (at least, for stop consonants), due to the variability in SNR_{50} thresholds across tokens.

This result also means that NH listeners are highly consistent in their thresholds, as evidenced by the fact that the token-level error functions are highly non-linear and, therefore, correlated across listeners. If the error functions were nonlinear but thresholds were variable between listeners, the responses would have shown a more linear fit when averaged across them. Thus, there do not appear to be considerable individual differences among NH listeners (at least for this task), even though there are many individual differences for HI listeners (Trevino & Allen, 2013a; b). Our goal here is to examine data from NH listeners that may provide a baseline for tests that can be used to assess speech recognition with individual HI listeners.

Problem statement and approach

Overall, the results of Singh and Allen (2012) suggest that, while the average log error for NH listeners is close to linear as a function of SNR (as predicted by the AI), the variance in error rates across consonants is extremely large. In addition, this log-linear pattern breaks down for within-consonant errors, with individual tokens contributing further variance. Given this, we predict that there are two main components that make up most of the variability in listeners' errors: (1) an across-consonant component, and (2) a within-consonant component. The goal of the present study to investigate across and within consonant errors and variance as a function of SNR, consonant, and token, for a large dataset of speech sounds containing stop, nasal, and fricative consonants, and to determine how much variability in NH listeners' errors is attributable to each of these factors, and importantly, how much variability can be explained by accounting for differences in error thresholds (SNR_{50}).

We examined NH listeners' error rates for a dataset consisting of 896 tokens (56 CVs from 16 different consonants and four vowels, presented at six SNRs). If the AI formula (which holds for listeners' *average* error) accounts for listeners' responses, SNR would be the single factor predicting whether or not listeners' make an error. However, based on Singh and Allen (2012)'s results for stop consonants, we predict that the effect of SNR is relatively weak compared to across- and within-consonant differences between speech sounds.

We also ask whether the log-linear pattern predicted by Eq. 2 is valid across and within consonants. If it is, we would expect a similar relationship between listeners' errors and SNR at both the across- and within-consonant levels. If it is not, we would observe a different relationship, such as the highly non-linear binary responses seen by Singh and Allen (2012) for individual stop consonant tokens. This would, in turn, suggest that we should develop speech tests for HI listeners based on the error thresholds (SNR_{50}) for specific speech sounds. Accordingly, we will also examine how much of the variability in NH listeners' errors can be explained by accounting solely for the SNR_{50} error thresholds.

Method

Some of the data presented here were collected as part of the experiments previously reported by Phatak and Allen (2007) and Singh and Allen (2012). Additional details about the methods can be found in those papers. The data presented here include 14 additional subjects and the entire set of stimuli and responses (i.e., data not included in any of the previous studies).

Design

Participants performed a 64 AFC syllable identification task (16 consonants x 4 vowels). Each of the 64 consonant-vowel (CV) syllables was presented at six SNRs (-22, -20, -16, -10, -2, and quiet) with 14 talkers, for a total of 5,376 stimuli. Stimulus presentation was randomized across all stimulus variables, divided into 42 blocks of 128 trials each (5,376 trials total). The experiment was run over the course of several sessions, and on average each participant took 15 hours to complete the entire experiment. Each of the 5,376 conditions was presented once to each subject. However, by accident, a few of the subjects re-ran a portion of the experiment. These few additional data points were also included in the analysis.[3]

Participants

Twenty-eight NH listeners participated in the experiment. Phatak and Allen (2007) reported on 14 subjects, four of whom were removed based on their poor performance in quiet. The following year, the same experiment was re-run with a second cohort of 14 subjects. Thus, a total of 28 listeners participated in the joint experiment. Eighteen completed all the sessions. For participants that did not complete all sessions (i.e., those that did not hear each of the 5,376 stimuli), data from the sessions that were completed are included. After a careful evaluation of the responses, 25 subjects were included in the present analysis. The three removed were a subset of the four removed in Phatak and Allen (2007); one of the participants previously removed actually had excellent scores other than in quiet. All but one of the listeners were young native English speakers with self-reported normal hearing. Participants provided informed consent (approved by the University of Illinois IRB) and received monetary compensation.

Stimuli

CV stimuli were selected from the *Linguistic Data Consortium* LDC2005S22 "Fletcher" corpus (Fousek, Svojanovsky, Grezl, & Hermansky, 2004). Speech-weighted noise was generated on each trial and added to the CV at the appropriate SNR level.

Procedure

Participants were seated in front of a computer in a sound-attenuating booth. Stimuli were presented over Sennheiser headphones at the subject's most comfortable level. On each trial, a stimulus was presented, and the participant clicked on one of 64 buttons on the computer screen, corresponding to the 64 possible CV syllables they heard. The buttons were arranged in a grid by consonant and vowel to make the selection as easy as possible.

Participants were allowed to replay the stimulus if needed, and they were given the option of selecting a "Noise Only" response if they were unable to hear any speech sound at all. They were encouraged to use this only when needed, and to make their best guess about which sound they heard. Participants clicked on the "Noise Only" button on 12% of the trials, mostly on trials with very low

[3] The difference in the number of data points for these tokens was accounted for by running a weighted regression, as detailed in Results section below.

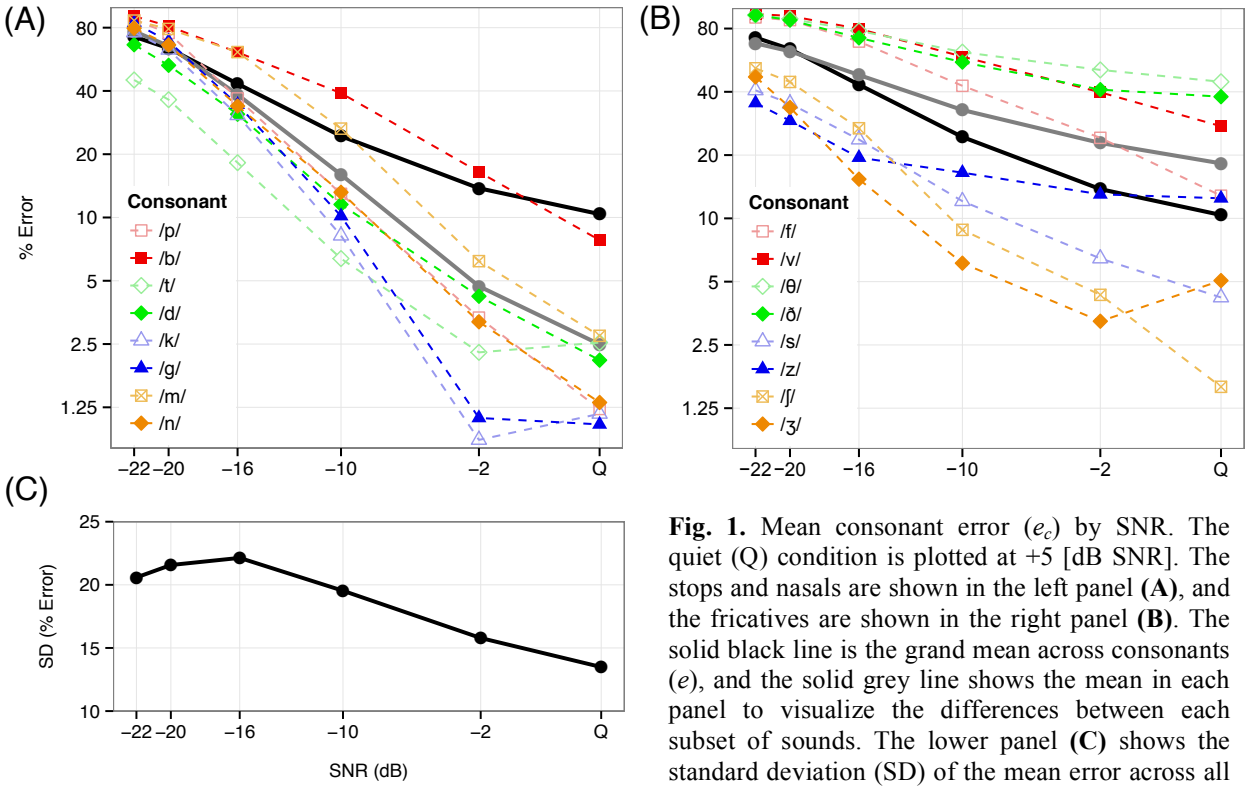

Fig. 1. Mean consonant error (e_c) by SNR. The quiet (Q) condition is plotted at +5 [dB SNR]. The stops and nasals are shown in the left panel **(A)**, and the fricatives are shown in the right panel **(B)**. The solid black line is the grand mean across consonants (e), and the solid grey line shows the mean in each panel to visualize the differences between each subset of sounds. The lower panel **(C)** shows the standard deviation (SD) of the mean error across all the consonants.

SNRs (correlation between SNR and probability of "Noise Only" response: –0.76). For the analyses, these trials were coded by distributing the error evenly across each of the 16 possible consonant responses. That is, each trial with a "Noise Only" response was coded as having 15/16 of an error.[4]

Results

Across-consonant errors

Our first goal is to characterize NH listeners' errors as a function of SNR and consonant. Figure 1 shows the proportion of trials on which participants made errors as a function of these two factors. The stops and nasals are shown in the left panel (Fig. 1A) and the fricatives are shown in the right (Fig. 1B). The grand mean (solid black line) and the means of the group of sounds in each panel (solid grey line) are also shown.

As the figure illustrates, several properties of the grand mean are consistent with the predictions from the AI model. First, the range of SNRs (-22 [dB] to quiet) covers nearly the entire range of error rates. This fits with the prediction from Eq. 1 that listeners' errors span approximately a 30 [dB] range. In addition, the log-error decreases linearly with increasing SNR (Eq. 2). Thus, the AI model holds for individual consonants (Phatak & Allen, 2007). A minor exception to this rule is the floor effect seen for /k, g, ʒ/ where the error function asymptotes near 0% at low noise levels (since these sounds are so robust to noise).

The consonants also vary considerably in their overall error rates. For stops, the

[4] Because this led to a non-integer number of errors in some conditions, the total number of errors was rounded when computing the condition weights for the empirical logit analyses described below.

voiced sounds (/b, d, g/) tend to have higher error rates than their voiceless counterparts (/p, t, k/). In addition, there are about twice as many errors for /m/ than for /n/. Fricatives have a larger error rate overall, and within the fricatives, alveolar and palato-alveolar sounds (/s, z, ʃ, ʒ/) have much lower error rates than the labio-dental and dental sounds (/f, v, θ, ð/).

Equation 3 quantifies the SD (σ_c) of the 16 consonant means (e_c) as a function of SNR

$$\sigma_c^2 = \frac{1}{16}\sum_{c=1}^{16}\left(e_c - e\right)^2 \qquad (3)$$

where e is the grand mean. This is shown in Fig. 1C. The SD (measured in units of percent error) is >22% below -10 [dB SNR], and decreases to 13% for the quiet condition. Thus, despite the fact that the 16 consonant means follow the log-linear pattern predicted by the AI model, there is considerable across-consonant variability in error rates (13-22%). This large across-consonant SD, at every SNR, is notable; SNR is not the only factor driving listeners' errors. This contrasts with the prediction from the AI, which considers only the SNR (since it is based on average responses, not on individual consonants; see also Note 2 regarding definitions of the AI that include band importance). The results shown in Fig. 1 demonstrate that both SNR and across-consonant differences must be considered when describing listeners' errors.

These results are consistent with earlier work showing differences in the audibility of specific acoustic cues that listeners use to identify consonants (Régnier & Allen, 2008; Li, Menon, & Allen, 2010; Li et al., 2012) and with overall differences between consonants in their robustness to noise (Miller & Nicely, 1955). These differences are the source of the variability across consonants seen here. In addition, as shown by Trevino and Allen (2013a; b), the across-consonant factor plays a key role for HI listeners, making consonant and SNR two of the main sources of errors for HI listeners.

Within-consonant errors

Next, we characterize listeners' errors for individual tokens of a given consonant, that is, within-consonant errors. In particular, we are interested in whether the log-linear relationship seen in the grand average and the consonant averages is maintained at the token level, and whether there is additional variability in NH listeners' responses that is attributable to differences between individual tokens.

Figure 2A shows the log error for each token (e_t) in the dataset. As the figure clearly illustrates, there is a large amount of variability in the error rates of individual tokens. To examine the extent of the token-level variability, we estimated the SNR at which each token crosses 50% error (denoted as the SNR_{50}) by fitting a spline curve to the token error functions. We then interpolated between the SNR points with the help of the spline to estimate the SNR_{50} for each token. The token error responses were then shifted along the SNR dimension by subtracting off the SNR_{50} for each token

$$e_t\left(SNR - SNR_{50}\right), \qquad (4)$$

so that all the tokens are aligned with their 50% point at 0 [dB SNR].

For tokens where listeners' responses did not cross 50% in the range of SNRs used in the experiment, an estimate of the SNR_{50} was obtained by interpolating the spline at a broader range of SNRs. Table 1 contains a summary of the number of tokens with SNR_{50} values that were estimated based on (1) tokens with an observed 50% error point, (2) tokens with no errors *above* 50%, and thus, having SNR_{50} values below -22 [dB

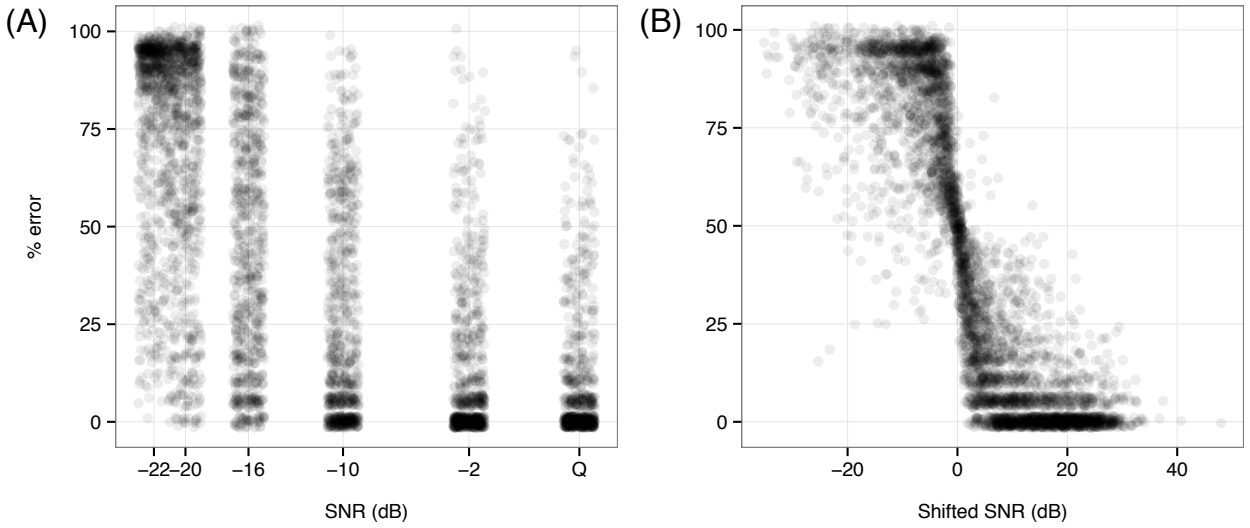

Fig. 2. (A) Percent log error for each token, as a function of SNR ($e_t(SNR)$). **(B)** Percent log-error shifted by the token's SNR_{50} ($e_t(SNR–SNR_{50})$). In the left panel **(A)**, the considerable variability between the error rates for each token is evident. After shifting by the token's SNR_{50} (panel **B**), this variability is greatly reduced, revealing the step-function (e.g., binary) nature of the error. Data points are jittered slightly to prevent overplotting.

SNR], (3) tokens with no errors *below* 50%, having SNR_{50} values above quiet (coded as +5 [dB SNR]), and (4) tokens for which no reliable SNR_{50} estimate could be obtained (i.e., the spline never crossed 50% error at any SNR). In cases where the mean error crossed 50% at multiple SNRs (50 of the tokens), the highest SNR with 50% error was used as the SNR_{50}.

Figure 2B shows the results of the token shifting procedure for each token that has a well-defined SNR_{50}. It is clear that shifting the token error function by its SNR_{50} greatly reduces the variability between individual tokens. Thus, by simply measuring the mean error for each token (indicated by the SNR_{50}), we can quantify most of the within-consonant error.

Figure 3 shows the token errors separated by each consonant with the curves shifted by each token's SNR_{50}. Tokens without a well-defined SNR_{50} are not shown, and the numbers after each consonant label indicate the number of tokens shown for that consonant (the maximum is 56). As the figure illustrates, listeners' error functions for individual tokens do not map well to the AI model's log-error dependence. Rather, shifting each token by its SNR_{50} reveals that individual token error curves (Eq. 4) are highly non-linear (i.e., nearly binary).

Figure 4 shows the distribution of SNR_{50} values for each consonant. As with Figure 3, responses that did not have a well-defined SNR_{50} are not shown. Due to the large range of SNR_{50} values, the variance of $e_{t|c}$ (across tokens, within a consonant class) is huge. Given this result, it follows that the AI's log-linear relationship is simply the result of averaging across nearly binary responses having the distribution of SNR_{50} values shown in Figure 4.

We can further quantify the differences between individual tokens by examining the slope of the error functions (as measured by the spline fits). Figure 5 shows the distribution of the steepest slope of the spine fit for each token. The slopes vary by consonant, but overall, they are around -10 [%/dB], consistent with Figure 3 which shows that the error typically goes from zero to chance over a 10 [dB] range. The slopes

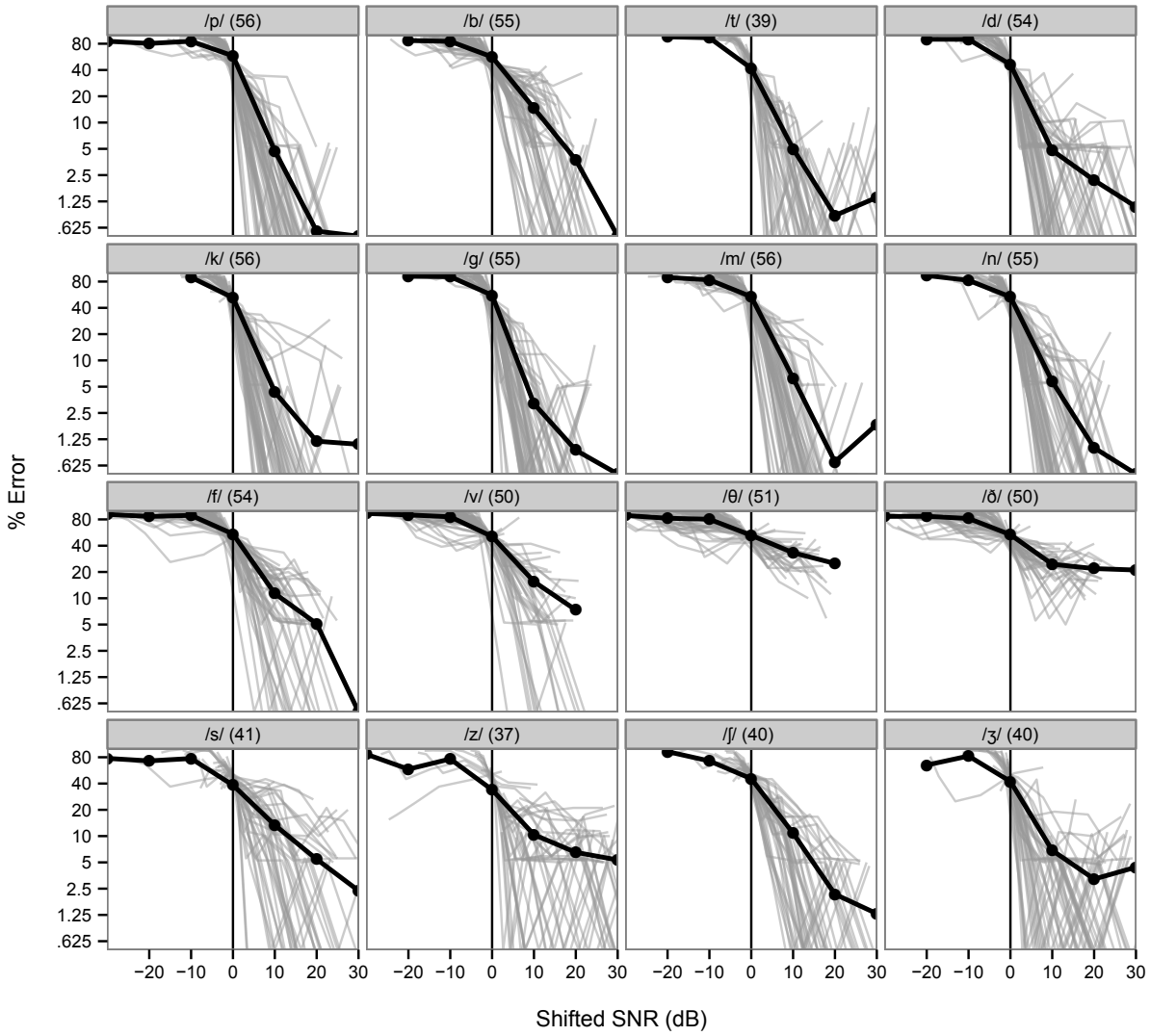

Fig. 3. Percent error by SNR for each token (lines) sorted by consonant (panels). Each response was shifted along the SNR axis to the SNR_{50} for that token (i.e., $e_{t|c}(SNR–SNR_{50})$ for each token). Above the SNR_{50}, $e_{t|c}$ abruptly drops to zero, and below the SNR_{50}, it rises to chance. Thus, the differences in SNR_{50} across individual tokens account for a major portion of the variance in $e_{t|c}$. Most of the natural variance has been removed in the shifted curves. The numbers after each consonant label indicate the number of tokens shown for that consonant (the maximum is 56).

are also tightly clustered. Together, the distributions of SNR_{50} values and slopes imply a robust threshold measure that can be used to quantify the differences in error rates between the tokens. As the figures illustrate, the within-consonant errors account for a major portion of the variability in the overall error. The residual variance that remains unaccounted for is small by comparison (shown as the differences between the individual tokens and the average in each panel of Figure 3).

In summary, the total variance in the average error measured by Eq. 1 is mainly due to (1) across-consonant and (2) within-consonant variability. The dependence on SNR is largely captured in the distribution of token SNR_{50} values.

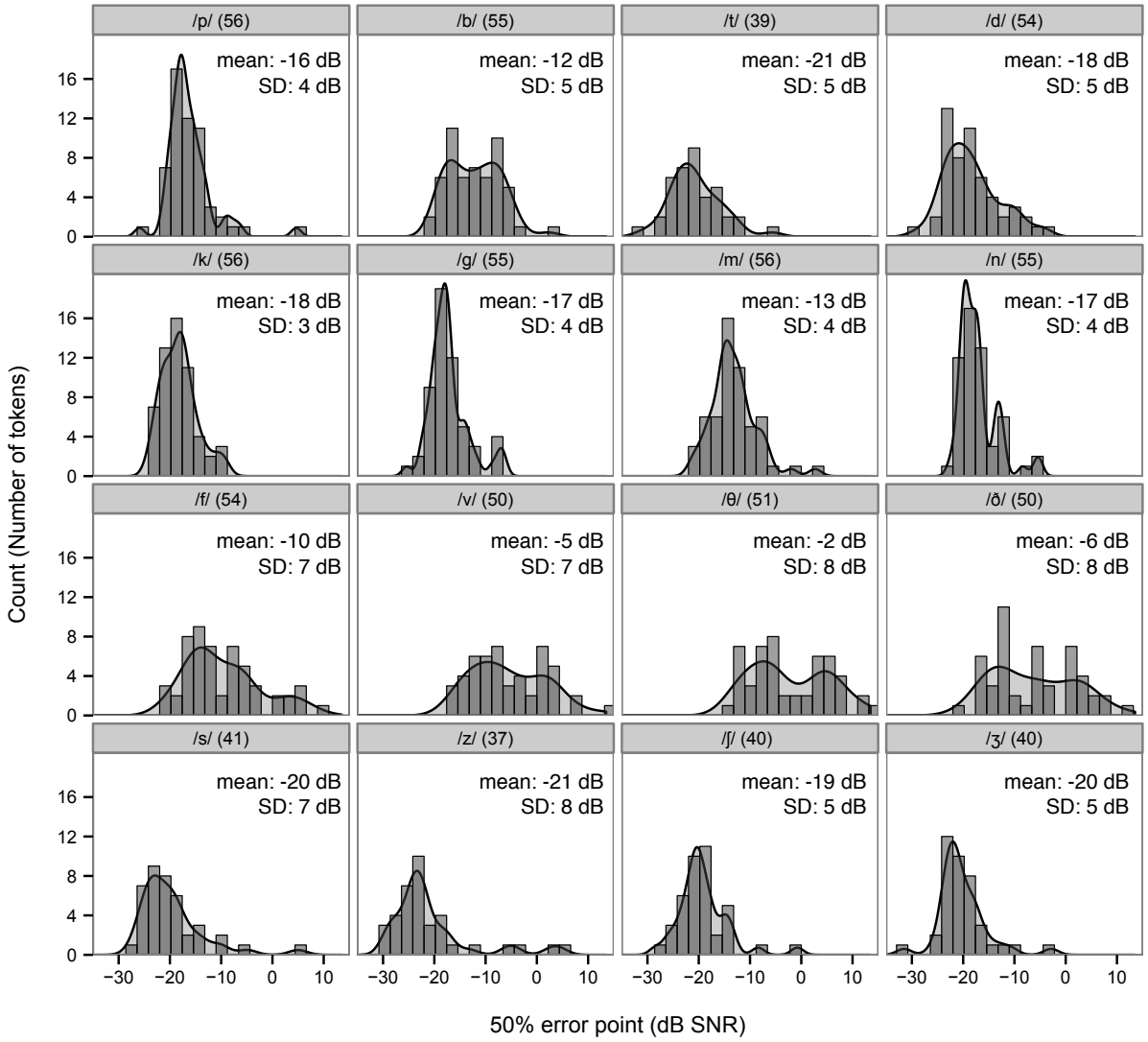

Fig. 4. Histogram of SNR_{50}, defined by $e_{t|c}$ = 0.5, for each token (t), sorted by consonant (c). The SNR_{50} may be thought of as a detection threshold measure for the token. The mean and standard deviation (SD) are provided of each distribution, in the upper right corner of each panel. Some consonants have a very tight distribution with just a few sounds outside the 1 SD range. Others are nearly uniform about the mean. The further to the right from the mean, the less robust the token. Not surprisingly high error sounds (mostly /θ, ð/) have a significant number of thresholds above 0 [dB SNR]. Note that one /s/ token is not plotted because its SNR_{50} is very low (less than –35 [dB SNR]).

Statistical analyses

To validate these observations statistically, we ran several regression analyses. Two sets of analyses were conducted: (1) an analysis examining how much variability is explained by shifting response curves by their SNR_{50}, and (2) an analysis examining the extent to which specific factors (SNR, across-consonant, and within-consonant factors) contribute to listeners' errors. Both analyses consist of fitting a series of regression lines to the average error for the 25 NH listeners, with each condition weighted by the number of

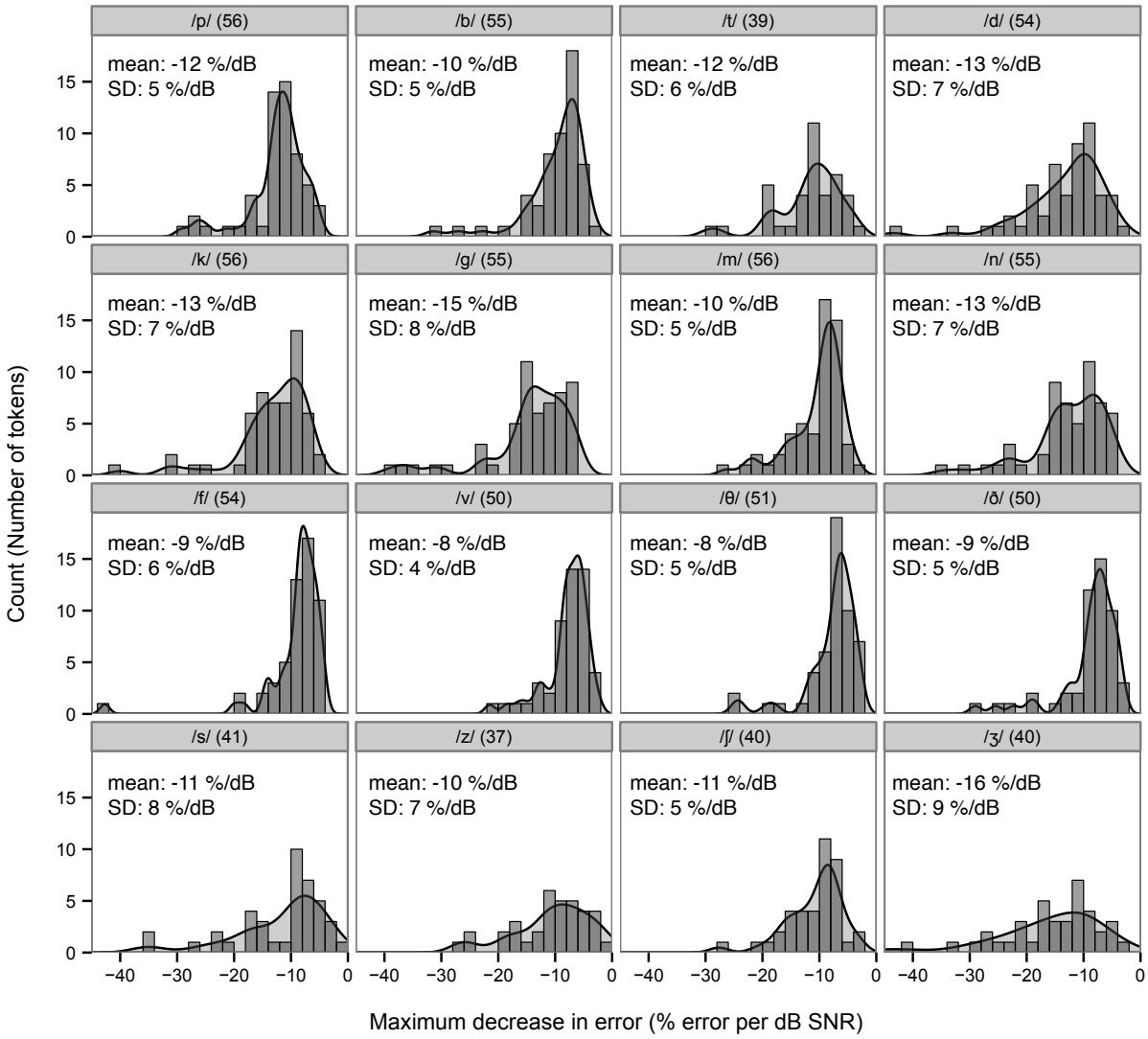

Fig. 5. Histogram of the maximum slope (in magnitude) of the error function for each token as characterized by splines fits to e_t. The units are in %/dB. Many of the distributions peak –10 %/dB, meaning that over a 10 dB range the score would vary from no error to chance performance. The means and SDs of the distributions for each consonant are shown in the upper left corner of each panel.

data points in that condition.

Note that these analyses do not examine responses for individual listeners. There are several reasons for this. First, the questions we are interested in here primarily concern the factors that lead listeners to make errors (SNR, across-, and within-consonant differences). Second, as noted above, NH listeners are highly consistent with each other in their errors for these sounds. That is, they all show approximately the same threshold for a given token. Thus, differences between NH listeners do not contribute much of the variance. Of course, for a speech test designed to assess performance with HI listeners, individual differences are paramount, and one would not want to average across HI listeners. This is, however, different from our goal here, which is to characterize errors at the level of

individual tokens for a group NH listeners (which can, in turn, provide data that will be useful for developing speech tests with HI listeners).

Given the highly non-linear responses observed for the token-level error functions, one would like to transform the data so that they lie along a linear scale. In the figures, this is achieved by plotting the data on a log-scale, which produces a linear relationship between listeners' errors and SNR for the overall mean (the effect predicted by the AI) and for individual consonant means. Thus, one reasonable transformation would be a log transform:

$$L(e) = \log(e) \tag{5}$$

One problem with this approach is that the log-linear relationship between listeners' errors and SNR breaks down at the level of individual tokens, as discussed above. Instead, the token error functions resemble step functions. For proportional data of this type, a common approach is to use the *empirical logit* transform (Barr, 2008):

$$L(e) = \log \frac{n_e + 0.5}{n_t - n_e + 0.5} \tag{6}$$

where n_e is the total number of observed errors, and n_t is the total number of data points in that condition. This is particularly useful for the individual token data, which approximate step functions (i.e., sigmoids with an infinite slope). Moreover, for data that are distributed log-linearly and bounded between 0 and 1 (as the consonant means and overall mean are), the transform will still produce values on an approximately linear scale. This allows us to compare models with coefficients for different factors at both the across- and within-consonant level. The 0.5 term ensures that *L(e)* is defined when n_e is 0 (i.e., no errors) or equal to n_t (100% error). For the analyses presented here, both transforms (Eq. 5 and 6) yielded similar results. The data from the empirical-logit analyses are given in the text below and summarized in Table 2.

Effect of SNR_{50}

First, we wanted to quantify the amount of variability in listeners' responses that can be explained by the error threshold (SNR_{50}) alone. Note that this analysis can only be run on the subset of sounds with well-defined SNR_{50} values (789 of 896 tokens); the analyses presented in the next section were run on the full dataset of 896 tokens (since SNR_{50} did not enter into those models directly). We compared two regression models. The first model,

$$L(e) = \beta_0 + \beta_1 SNR \tag{7}$$

examines the effect of SNR by itself, where β_1 is the regression coefficient for SNR (corresponding to the slope of the average error shown in Figure 1) and β_0 is the intercept (corresponding to the grand mean error). As expected, SNR accounted for a significant proportion of the total variance in listeners' errors (R^2=0.290, $p<0.001$). However, it explains only 29.0% of the total variability. Thus, SNR is only one factor predicting listeners' errors.

The second model,

$$L(e) = \beta_0 + \beta_1 \left(SNR - SNR_{50} \right) \tag{8}$$

examines the effect of shifting the error functions by SNR_{50}. This model provided a much better fit to listeners' responses (R^2=0.624, $p<0.001$). Thus, adjusting for the SNR_{50} allows us to explain more than twice the variability explained by SNR alone. This result fits with the observations above and suggests that by simply measuring error thresholds, we can account for the majority of listeners errors (62.4%) in speech recognition tasks. Figure 6 shows the

relationship between the observed empirical log-odds error and the predicted log-odds error from this model.

Effect of across- and within-consonant factors

Next, we examined the amount of variance explained by across- and within-consonant differences using the full dataset, including those sounds that did not have well-defined SNR_{50} values. This analysis can provide a more complete account, but it also relies on more complex models. We used multi-step regression to look at the contribution of (1) SNR, (2) consonant, (3) token means (corresponding approximately to SNR_{50}) and (4) token slopes. At each step, one factor was added, and the proportion of variance accounted for by the model (R^2) was calculated. The change in R^2 on each step quantifies how much each factor contributed to listeners' errors.

On the first step of the regression, we entered SNR as the only factor, yielding the same regression equation used above (Eq. 7). Here, less variance is accounted for than in the first analysis (R^2=0.228, p<0.001), though SNR still has a significant effect.

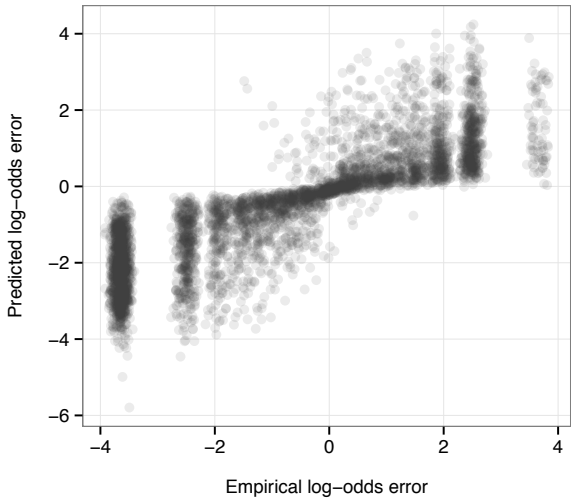

Fig. 6. Scatter plot showing proportion of listeners' errors and predicted model responses for the regression model that accounted for differences in token-level error thresholds (SNR_{50}). Data points are jittered slightly to prevent overplotting.

The lower R^2 value is due to the additional tokens for which we could not obtain an SNR_{50}. These sounds had errors that were either consistently near 100% at all SNRs or near 0% at all SNRs. Thus, the model does not explain responses to these sounds very well. This analysis does, however, provide a more complete picture of the amount of variance that can be explained in natural speech (including speech sounds with atypical response patterns).

On the second step of the regression, consonant and its interaction with SNR were added to the model. These factors correspond to the individual consonant means shown in Figure 1. Because consonant is a categorical variable (i.e., each consonant in the experiment represents a discrete category), *treatment coding* (Kleinbaum, Kupper, Muller, & Nizam, 1998) was used to create variables corresponding to the consonant coefficients (i.e., if the stimulus for a particular condition was a /b/, the variable corresponding to /b/ was coded as 1 and all other variables were coded as 0). This can be represented as an $n \times p-1$ matrix of treatment codes:

$$\mathbf{C} = \begin{bmatrix} C_{1,1} & C_{1,2} & \cdots & C_{1,p-1} \\ C_{2,1} & C_{2,2} & \cdots & C_{2,p-1} \\ \vdots & \vdots & \ddots & \vdots \\ C_{n,2} & C_{n,p-1} & \cdots & C_{n,p-1} \end{bmatrix} \quad (9)$$

where p=16, the number of consonants,[5] and n=5376, the number of stimuli in the experiment. These variables correspond to the mean error rate for each consonant. The slope of each consonant's error function is given by the interaction between SNR and **C**. This is coded by multiplying the SNR for each condition (n) by each of the consonant

[5] One less than the total number is needed to code all the consonants, since the model coefficients for one consonant (the reference category) are estimated by the overall slope and intercept as a function of SNR.

treatment codes:

$$\mathbf{SNR}\times\mathbf{C}_n=\left[SNR_n\right]\circ\left[C_{n,1}\cdots C_{n,p-1}\right]=\left[SNR\times C_{n,1}\cdots SNR\times C_{n,p-1}\right] \quad (10)$$

Including these factors yields the following regression equation for a particular speech sound (n):

$$L(e_n)=\left(\beta_0+\beta_1 SNR\right)+\left(\sum_{i=3}^{17}\beta_i\mathbf{C}_n+\sum_{j=18}^{32}\beta_j\mathbf{SNR}\times\mathbf{C}_n\right) \quad (11)$$

Here, β_i and β_j are coefficients for the intercept and slope, respectively, for each consonant, for a total of 32 coefficients in the model (15 for consonant intercepts, 15 for consonant slopes, 1 for the overall intercept, and 1 for the overall slope). This model accounted for an additional 30.8% of listeners' errors (ΔR^2=0.308, p<0.001). Thus, together, SNR and consonant explain 53.7% of the variance in listeners' responses.

On the third step of the analysis, we included coefficients corresponding to individual token intercepts (again, as treatment-coded factors). Each token is defined as a combination of consonant, talker, and vowel. This is represented by the interaction between consonant and each of the 56 tokens for that consonant. First, we created treatment codes for the 56 tokens:

$$\mathbf{T}=\begin{bmatrix} T_{1,1} & T_{1,2} & \cdots & T_{1,q-1} \\ T_{2,1} & T_{2,2} & \cdots & T_{2,q-1} \\ \vdots & \vdots & \ddots & \vdots \\ T_{n,2} & T_{n,p-1} & \cdots & T_{n,q-1} \end{bmatrix} \quad (12)$$

where q=56, the number of tokens for each consonant. These were then multiplied by the consonant treatment codes to create 825 variables (15 consonant variables × 55 token variables per consonant) corresponding to the individual tokens in the experiment:

$$\mathbf{C}\times\mathbf{T}_n=\left[C_{n,1}\cdots C_{n,p-1}\right]\circ\left[T_{n,1}\cdots T_{n,q-1}\right]=\left[C\times T_{n,1}\cdots C\times T_{n,(p-1)(q-1)}\right] \quad (13)$$

Thus, for each condition in the experiment, the treatment code for that consonant's intercept ($C_{n,1}\cdots C_{n,p\text{-}1}$) is multiplied by the corresponding variable for one of its 56 tokens ($T_{n,1}\cdots T_{n,q\text{-}1}$). This yields the following regression equation:

$$L(e_n)=\left(\beta_0+\beta_1 SNR\right)+\left(\sum_{i=3}^{17}\beta_i\mathbf{C}_n+\sum_{j=18}^{32}\beta_j\mathbf{SNR}\times\mathbf{C}_n\right)+\left(\sum_{k=33}^{87}\beta_k\mathbf{T}_n+\sum_{l=88}^{912}\beta_l\mathbf{C}\times\mathbf{T}_n\right) \quad (14)$$

where β_k and β_l are the coefficients corresponding to the intercepts for specific talker-vowel combinations (k) and specific tokens (l). This model explained an additional 23.8% of the variability in listeners' errors (ΔR^2=0.238, p<0.001), for a total of 77.5%. Thus, by accounting for the effect of SNR across consonants and the mean error within consonants, we can explain a considerable amount of the variability in listeners' errors.

The final model accounted for the small differences in the slopes of the individual token error functions, as seen in Figure 5. Again, this is represented using a set of treatment codes, specifically corresponding to the three-way interaction between SNR, consonant, and token:

$$\mathbf{SNR}\times\mathbf{C}\times\mathbf{T}_n=\left[SNR_n\right]\circ\left[C_{n,1}\cdots C_{n,p-1}\right]\circ\left[T_{n,1}\cdots T_{n,q-1}\right]=\left[SNR\times C\times T_{n,1}\cdots SNR\times C\times T_{n,(p-1)(q-1)}\right] \quad (15)$$

This leads to the final regression equation:

$$L(e_n) = (\beta_0 + \beta_1 SNR) + \left(\sum_{i=3}^{17} \beta_i \mathbf{C}_n + \sum_{j=18}^{32} \beta_j \mathbf{SNR} \times \mathbf{C}_n \right) + \left(\sum_{k=33}^{87} \beta_k \mathbf{T}_n + \sum_{l=88}^{912} \beta_l \mathbf{C} \times \mathbf{T}_n \right) + \left(\sum_{m=913}^{967} \beta_m \mathbf{SNR} \times \mathbf{T}_n + \sum_{n=968}^{1792} \beta_n \mathbf{SNR} \times \mathbf{C} \times \mathbf{T}_n \right) \quad (16)$$

where β_m and β_n are the coefficients corresponding to the slopes for specific talker-vowel combinations (m) and specific tokens (n). This model accounted for an additional 8.1% of the variance (ΔR^2=0.081, $p<0.001$), for a total of 85.6%. The remaining 14% of the variance is likely attributable to small differences between listeners (since the listener is the only other factor in the experiment that was not included in the model) and to measurement error.

Thus, four factors, SNR, consonant, token mean, and token slope, can explain the vast majority (85.6%) of listeners' errors. Moreover, by simply adjusting the SNR based on differences SNR_{50}, as we did in the first analysis, we can explain 62.4% of the variability. Composite measures like the AI and SRT completely miss this information, since they average across speech sounds and only account for differences in SNR. Although SNR is an important factor, it only accounts for 23-29% of listeners' errors. As a consequence, speech tests based on these measures are missing most of the information that causes listeners to make errors.

Discussion

The results of this experiment demonstrate that there is a great deal of variability in listeners' errors (both across- and within-consonants) that is not captured by the wide-band SNR. Adjusting for differences in token-level error thresholds, as measured by the SNR_{50}, allows us to explain more than twice the variability in listeners' errors (62.4%) than we can explain via SNR alone (29.0%). Composite measures, such as the AI and SRT, fail to capture the large natural variability in listeners' errors. While these measures quantify the *aggregate* effect of SNR, they fail to capture the fact that more errors are driven by differences across and within consonants.[6] In order to develop a speech test that accurately characterizes listeners' errors, we must take SNR_{50} thresholds into account.

Such an approach seems extremely powerful. As a group, NH listeners are remarkably consistent in their ability to correctly identify speech sounds in noise. Given the SNR, consonant, and token, we can explain more than 85% of the variability in errors for NH listeners. Only the small residual variance (14% of the total) is attributable to individual differences between the 25 NH listeners.[7] This small difference between NH listeners implies that they represent a homogenous group for this task.

In addition, the results demonstrate that the log-linear relationship between error and SNR breaks down at the token (within-consonant) level for all the consonants,

[6] One could imagine calculating the AI for individual tokens, as Phatak and Allen (2007) did for individual consonants. However, as we show here, listeners' responses to individual tokens more closely resemble step functions, rather than the AI's log-linear dependence. Thus, this approach does not provide an accurate model of listeners' errors at the token level.

[7] Again, this is true for NH listeners, as we examined here. For HI listeners or situations where a speech test seeks to assess an individual listener, individual differences between listeners are critical.

consistent with results first demonstrated by Singh and Allen (2012) for stop consonants. Thus, as a function of SNR, each token's response (e_t, Fig. 3) is functionally binary over a very small range of SNRs (around 10 [dB] for most tokens; Fig. 4). Either NH listeners can hear the sound with nearly zero error, or they are at chance.

The results also show that the small error in quiet is due to a few high-error consonants. For example, /θ/ and /ð/ rarely drop below 40% error, while other consonants, such as /g/ and /k/, have ≈1% error above -2 [dB SNR]. These differences are likely driven by acoustic differences between the consonants that cause them to vary in their overall intelligibility and make them more or less robust to speech-weighted noise (Régnier & Allen, 2008; Li, Menon, & Allen, 2010; 2012). For example, /t/ sounds can be recognized on the basis of high-amplitude bursts in specific frequency regions. These bursts are generally very robust to noise (Li et al., 2010). In contrast, /v/ sounds, which have higher error rates, contain lower-amplitude frication cues that are more likely to be masked by noise (Li et al., 2012).

These results also fit with those of Singh and Allen (2012) who found that errors in quiet were driven by a few high-error stop consonant tokens. These tokens would have SNR_{50} thresholds above quiet. Thus, the error in the quiet condition is bimodal; most of the tokens have virtually no errors and a few have a very high error rate.

Assessment of hearing-impaired listeners' errors

What do these results suggest about how to assess effects of hearing loss on speech recognition? First, they demonstrate that we should not rely on composite measures that average across speech sounds. Second, they provide a useful baseline of speech recognition with NH listeners that could be used to create a test based on individual tokens with precisely known SNR_{50} values (from NH listeners). By examining responses to individual tokens, we can identify cases where a listener has difficulty with a particular consonant or talker, providing a fast, simple speech test for assessing hearing loss. If a listener's responses deviate from the pattern consistently observed for NH listeners, this tells us they have difficulty recognizing that sound.

Trevino and Allen (2013a; b) demonstrated the utility of this approach by showing that there are large individual differences in HI listeners' ability to recognize different phonemes. Unlike NH listeners, HI listeners (with even a slight hearing loss) can have significant errors above the SNR_{50} for a particular token. These individual differences between HI listeners are critical and must be measured in a speech test. There is considerable variability amongst HI listeners in their ability to identify specific tokens (whereas NH listeners identify them consistently), even though an individual listener is consistent in their responses (Trevino & Allen, 2013a; b). Thus, these two sources of variability (differences between tokens and differences between listeners) operate differently for NH and HI listeners. NH listeners show almost no individual differences in their error thresholds, while the errors vary considerably between tokens (as estimated by SNR_{50}). In contrast, HI listeners show large individual differences for the same tokens. Therefore, if we first account for token-level differences (e.g., by adjusting sounds by their SNR_{50} using data from NH listeners), we will be left with individual differences between HI listeners, providing the diagnostic information needed to assess speech recognition. Importantly, this approach also avoids ceiling effects (i.e., HI listeners do not correctly recognize

all the sounds, as can happen, for example, with HINT sentences in quiet; Gifford, Dorman, Shallop, & Sydlowski, 2010).

This approach should help a clinician to develop a profile of the specific speech deficits experienced by a listener with hearing loss, and it would allow them to determine how far above the SNR_{50} threshold the SNR must be in order for the listener to correctly recognize the sounds. Suppose, as quantified by Trevino and Allen (2013a; b), that a listener has difficulty recognizing /s/ sounds spoken by a specific talker, while they have no difficulty identifying other sounds. This could be indicative of difficulty hearing the high-frequency frication that provides a primary cue for recognizing this sound, and it could help the clinician generalize this deficit to similar speech sounds (e.g., they might focus their assessment on whether the patient also has difficulty with /z/ or /ʃ/, or whether they have difficulty with that particular talker). In contrast, a speech test that averages across different consonants and talkers would conclude that this listener only makes a few errors (since, on average, the error would be small). A test based on recognition of individual speech sounds with known SNR_{50} values (based on data from NH listeners) would provide the level of detail needed to assess the speech recognition deficit for this listener.

Types of confusions

Finally, although these analyses provide us with many useful insights about the factors that that cause normal-hearing listeners to make errors, they do not tell us anything about the nature of those errors, namely the particular confusions that listeners make. The error rate, by itself, does not tell us whether listeners consistently made the *same* confusion in cases where they made errors. Miller and Nicely (1955) found that certain consonant classes are much more likely to be confused with each other, suggesting that there is a great deal of information in the *types* of errors that listeners make. When listeners make errors, often they are not simply guessing. One way we could quantify this is to look at the entropy of listeners' responses, which provides a measure of how consistent listeners are in the type of error they make (Singh & Allen, 2012). If a token has a small entropy, listeners are highly consistent (only confusing the token with one or two other consonants). If the entropy is large, consistency is low.

This approach is useful when selecting tokens to be used in speech recognition tests. For example, some sounds can be said to be mispronounced in the sense that listeners agree on a particular response that is different from what the talker intended. These sounds are easily identified, as they typically have high error and low entropy. This is a useful way of identifying "mispronounced" sounds in a database and provides a way of quantifying the degree to which the sound is mispronounced (in terms of the number of different responses). This may also provide a means of restricting the set of likely responses to a much smaller subset of all the consonants (Allen, 2005a). Ongoing work is using this approach to examine the nature of listeners' errors in more detail.

Conclusions

The large acoustic differences between individual speech sound tokens are a major source of the variability in listeners' error rates. Only by carefully controlling for these token-level errors, can we hope to develop speech tests that effectively measure listeners' speech recognition abilities. The only way we know to quantify these errors is to directly measure them, token by token, with a cohort of NH listeners. This is what we have done here.

The results of this experiment demonstrate that: (1) there is large variability in listeners' error rates due to differences across consonants (Fig. 1); (2) there is additional variability due to within-consonant differences, which can explained by the token error threshold (SNR_{50}) and slope (Fig. 3); (3) NH listeners are remarkably good at recognizing speech at SNRs greater than SNR_{50}; (4) NH listeners are highly consistent with each other in this task; and (5) adjusting for differences in SNR_{50} allows us to account for 62% of the total variance in listener's errors. Such an approach can greatly improve the utility of speech testing for HI listeners (Trevino & Allen, 2013a; b). Together, these results suggest we must reconsider the widespread use of aggregate measures of speech recognition and develop new methods that take differences between individual tokens into account.

Acknowledgements

This research was supported by a Beckman Postdoctoral Fellowship to JCT.

References

Allen, J. B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*, *2*(4), 567-577.

Allen, J. B. (1996). Harvey Fletcher's role in the creation of communication acoustics. *Journal of the Acoustical Society of America, 99*(4), 1825-1839.

Allen, J. B. (2005a). *Articulation and intelligibility*. LaPorte, CO: Morgan and Claypool.

Allen, J. B. (2005b). Consonant recognition and the articulation index. *Journal of the Acoustical Society of America, 117*(4), 2212-2223.

Barr, D. J. (2008). Analyzing 'visual world' eye tracking data using multilevel logistic regression. *Journal of Memory and Language*, *59*, 457-474.

Bronkhorst, A. W., Bosman, A. J., & Smoorenburg, G. F. (1993). A model for context effects in speech recognition. *Journal of the Acoustical Society of America, 93*, 499-509.

Bronkhorst, A. W., Brand, T., & Wagener, K. (2002). Evaluation of context effects in sentence recognition. *Journal of the Acoustical Society of America, 111*, 2874-2886.

Dobie, R. A. (2011). The AMA method of estimation of hearing disability: A validation study. *Ear and Hearing*, *32(6)*, 732-740.

Dobie, R. A., & Sakai, C. S. (2001). Estimation of hearing loss severity from the audiogram. In *Noise Induced Hearing Loss: Basic Mechanisms, Prevention and Control* (pp. 351-363). London: Noise Research Network Publications.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *Journal of the Acoustical Society of America, 88*, 1725-1736.

Fletcher, H. (Ed.). (1929). *Speech and Hearing*. New York: D. Van Nostrand.

French, N. R., & Steinberg, J. C. (1947). Factors governing the intelligibility of speech sounds. *Journal of the Acoustical Society of America, 19*, 90-119.

Fousek, P., Svojanovsky, P., Grezl, F., & Hermansky, H. (2004). New nonsense syllables database—analyses and preliminary ASR experiments. *Proceedings of the International Conference on Spoken Language Processing*, 2749-2752.

Fowler, C. A. (1984). Segmentation of coarticulated speech in perception. *Perception and Psychophysics*, *36*, 359-368.

Gifford, R. E., Dorman, M. F., Shallop, J. K., & Sydlowski, S. A. (2010). Evidence

for the expansion of adult cochlear implant candidacy. *Ear & Hearing, 31*, 186-194.

Haskell, G. B., Noffsinger, D., Larson, V. D., Williams, D. W., Dobie, R. A., & Rogers, J. L. (2002). Subjective measures of hearing aid benefit in the NIDCD/VA Clinical Trial. *Ear and Hearing*, *23(4)*, 301-307.

Humes, L. E., Dirks, D. D., Bell, T. S., & Ahlstbom, C. (1986). Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners. *Journal of Speech, Language, and Hearing Research*, *29*, 447-462.

Kamm, C. A., Dirks, D. D., & Bell, T. S. (1985). Speech recognition and the Articulation Index for normal and hearing-impaired listeners. *Journal of the Acoustical Society of America, 77*, 281-288.

Kleinbaum, D. G., Kupper, L. L., Muller, K. E., & Nizam, A. (1998). *Applied Regression Analysis and Other Multivariable Methods*. Pacific Grove, CA: Duxbury Press.

Li, F., Menon, A., & Allen, J. B. (2010). A psychoacoustic method to find the perceptual cues of stop consonants in natural speech. *Journal of the Acoustical Society of America, 127*(4), 2599-2610.

Li, F., Trevino, A., Menon, A., & Allen, J. B. (2012). A psychoacoustic method for studying the necessary and sufficient perceptual cues of fricative consonants in noise. *Journal of the Acoustical Society of America, 132*, 2663-2675.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, *74*, 431-461.

Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception and Psychophysics*, *34*, 338-348.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338-352.

Nabelek, A.K., Freyaldenhoven, M.C., Tampas, J.W., Burchfield, S.B., Muenchen, R.A. (2006). Acceptable noise level as a predictor of hearing aid use. *Journal of the American Academy of Audiology, 17*, 626-639.

Pavlovic, C. V., Studebaker, G. A., & Sherbecoe, R. L. (1986). An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals. *Journal of the Acoustical Society of America, 80*, 50-57.

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

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *Journal of Speech and Hearing Research*, *29*, 146-154.

Plomp, R., & Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise level. *Journal of the Acoustical Society of America, 66*, 1333-1342.

Rankovic, C. M. (1991). An application of the Articulation Index to hearing aid fitting. *Journal of Speech, Language, and Hearing Research*, *34*, 391-402.

Régnier, M., & Allen, J. B. (2008). A method to identify the noise-robust perceptual features: Application for consonant /t/. *Journal of the Acoustical Society of America, 123*, 2801-2814.

Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Journal of the Acoustical*

Society of America, 92, 81-110.

Singh, R., & Allen, J. B. (2012). The influence of stop consonants' perceptual features on the Articulation Index model. *Journal of the Acoustical Society of America, 131*, 3051-3068.

Taylor, B. (2006). *Predicting real-world hearing aid benefit with speech audiometry: An evidence-based review*. Unpublished doctoral dissertation, Central Michigan University.

Toscano, J. C., & McMurray, B. (2012). Cue-integration and context effects in speech: Evidence against speaking-rate normalization. *Attention, Perception, & Psychophysics*, *74*, 1284-1301.

Trevino, A. C., & Allen, J. B. (2013a). Individual variability of hearing-impaired consonant perception. *Seminars in Hearing, 34*, 74-85.

Trevino, A. C., & Allen, J. B. (2013). Within-consonant perceptual differences in the hearing impaired ear. *Journal of the Acoustical Society of America, 134*, 607-617.

Walden, B. F., Schwartz, D. M., Williams, D. L., Holum-Hardegen, L. L., & Crowley, J. M. (1983). Test of the assumptions underlying comparative hearing aid evaluations. *Journal Speech and Hearing Disorders*, *48*, 264-273.

Ward, W. D. (1983). The American Medical Association/American Academy of Otolaryngology formula for determination of hearing handicap. *Audiology*, *22*, 313-324.