Assessing the efficacy of hearing-aid amplification using a phoneme test

Christoph Scheidiger,[1,a)] Jont B. Allen,[2] and Torsten Dau[1]
[1]*Hearing Systems Group, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Lyngby, Denmark*
[2]*Beckman Institute, University of Illinois at Urbana-Champaign, Champaign, Illinois 61801, USA*

(Received 4 July 2016; revised 10 November 2016; accepted 22 January 2017; published online 13 March 2017)

Consonant-vowel (CV) perception experiments provide valuable insights into how humans process speech. Here, two CV identification experiments were conducted in a group of hearing-impaired (HI) listeners, using 14 consonants followed by the vowel /ɑ/. The CVs were presented in quiet and with added speech-shaped noise at signal-to-noise ratios of 0, 6, and 12 dB. The HI listeners were provided with two different amplification schemes for the CVs. In the first experiment, a frequency-independent amplification (flat-gain) was provided and the CVs were presented at the most-comfortable loudness level. In the second experiment, a frequency-dependent prescriptive gain was provided. The CV identification results showed that, while the average recognition error score obtained with the frequency-dependent amplification was lower than that obtained with the flat-gain, the main confusions made by the listeners on a token basis remained the same in a majority of the cases. An entropy measure and an angular distance measure were proposed to assess the highly individual effects of the frequency-dependent gain on the consonant confusions in the HI listeners. The results suggest that the proposed measures, in combination with a well-controlled phoneme speech test, may be used to assess the impact of hearing-aid signal processing on speech intelligibility.

[http://dx.doi.org/10.1121/1.4976066]

[MAH] Pages: 1739–1748

I. INTRODUCTION

Most day-to-day communication between humans is based on speech. Deficits in speech communication, e.g., as a result of a hearing impairment, can have strong effects on a person's quality of life and personal success. Hearing aids can help to regain the ability to hear speech, e.g., by compensating for the audibility loss. However, aided hearing-impaired (HI) listeners typically perform worse in speech understanding tasks than normal-hearing (NH) listeners. In particular, hearing-aid users commonly experience difficulties in challenging acoustical environments, such as noisy and/or reverberant spaces. In contrast, speech communication over a noisy transmission channel in NH listeners is typically robust.

Speech recognition can be limited by internal noise and external noise. External noise describes interfering acoustical signals that may mask or distract from the target signal. Internal noise characterizes the limitation and probabilistic nature of a listener's auditory system. A hearing loss may be viewed as an increase of internal noise. According to Plomp (1986), the internal noise can be further divided into an audibility component and a distortion component. The typical measure of the audibility component is an audiogram or a speech reception threshold in quiet (SRTq). The SRTq is defined as the speech level at which the recognition score equals the error score ($p_c = p_e$). While the SRTq is linked to the speech reception threshold in noise in Plomp's model, the audiogram and speech intelligibility in noise are not directly linked. Several studies have tried to link pure-tone thresholds to speech intelligibility of both NH and HI listeners (Humes *et al.*, 1986; Zurek and Delhorn, 1987; Pavlovic, 1988; Mueller and Killion, 1990). Mueller and Killion (1990) proposed the "Count-the-dots" method to calculate the articulation index, which can be transformed to a speech intelligibility score. Their method assesses how much of the long-term average speech spectrum (LTASS) is audible, i.e., above the pure-tone thresholds. This has become a widely used method to numerically quantify the benefit of a hearing instrument.

Speech intelligibility in noise may be measured with different speech materials. Phonemes (e.g., consonant-vowels, CVs) represent one class of speech materials. Phoneme identification experiments record which phoneme out of the phoneme set used in the experiment was chosen by a listener in response to a presented stimulus. The recorded responses are often presented in the form of a confusion matrix (CM), wherein each cell corresponds to one of the stimulus-response pairs. The stimuli are usually denoted as rows and the responses as columns. The diagonal of the matrix represents the counts of the correct responses and the row sum equals the total number of presentations for a given stimulus.

Phoneme perception research has a long history and started with the classical studies by French and Steinberg (1947) and Miller and Nicely (1955). French and Steinberg

a)Electronic mail: csche@elektro.dtu.dk

(1947) based their analysis on recognition scores only, i.e., the CM diagonal, and proposed a model to predict the percent correct value of phoneme pairs or triplets based on the individual phone scores. Later, Miller and Nicely (1955) applied an information theoretical analysis to their recorded CMs. Their entropy measure, which quantifies the randomness of responses, represents an approach to describe the communication process beyond pure recognition scores. In the case of a phoneme which is always misclassified (i.e., 100% error), this phoneme could be always confused with one specific other phoneme, which would correspond to an entropy of 0 bits. Alternatively, the phoneme could be confused with many other phonemes (instead of only one specific phoneme), in which case the entropy would be close to its maximum $\log_2(J)$ bits with J representing the number of possible response choices.

Entropy is powerful in quantifying the randomness of responses, but is insensitive to the kind of confusions. Two different phonemes might produce the same randomness in terms of observed responses but the individual confusions can be very different. Allen (2005) used confusion patterns (CPs) to visualize the individual confusions along with the recognition score. CPs show the response probabilities for all response alternatives as a function of the signal-to-noise ratio (SNR) for a given stimulus, i.e., they depict normalized CM rows as a function of SNR and thereby illustrate at which SNRs the recognition score drops and which confusion(s) was/were chosen instead of the correct response. If the response probabilities are shown on a logarithmic scale, confusions with low probabilities are clearly represented.

However, in order to use CV experiments to assess a HI listener, the perceptually relevant factors that underlie consonant perception need to be known and the CV experiments need to be designed accordingly. Despite the extensive research and elaborate analysis methods, only a few studies have revealed the effect of acoustic stimulus variability on consonant perception in individual listeners (Li and Allen, 2011; Phatak and Allen, 2007; Kapoor and Allen, 2012; Singh and Allen, 2012; Toscano and Allen, 2014; Zaar and Dau, 2015). This variability may be particularly relevant in studies with HI listeners (Trevino and Allen, 2013). Consonant perception has been demonstrated to be strongly affected by a high-frequency sensorineural hearing loss (e.g., Owens, 1978), reflecting the importance of high-frequency information contained in consonants (Li *et al.*, 2010; Li and Allen, 2011). Several studies thus proposed to control for the variance in the stimuli (e.g., Bilger and Wang, 1976; Boothroyd, 1984) as well as the variability across the HI listeners to reduce the variability in the CM data (e.g., Owens, 1978; Dubno *et al.*, 1984; Zurek and Delhorne, 1987; Trevino and Allen, 2013).

Miller and Nicely (1955) found that only a few of the possible response alternatives were chosen for a specific consonant, i.e., CM rows were sparse and the entropy thus small. Owens (1978) discussed a dependency of consonant perception on the specific selection of a consonant-vowel-consonant token, whereby a token represented a single phoneme recording. It was argued that the robustness and confusions obtained for individual tokens were specific to these tokens. The token dependency was later confirmed by Trevino and Allen (2013) who showed that the confusions in CV experiments became more consistent when the token variability was controlled for. Trevino and Allen (2013) analyzed confusions in HI listeners on a token basis and found that listeners with different audiograms showed similar confusions at the token level. This suggested that responses for a given CV token obtained across listeners can be more homogeneous than previously assumed. Furthermore, the authors found that different tokens of the same CV can result in different confusions in the same listener group. For example, the main confusion for a specific /ba/ token was /va/, whereas it was /da/ for another /ba/ token (Table II in Trevino and Allen, 2013). These results demonstrated the importance of considering consonant perception at the token level.

Dubno *et al.* (1984) reported a degraded CV recognition performance in HI listeners in the presence of noise, even in conditions when the speech was presented at high sound pressure levels, indicating that audibility alone was not sufficient to restore correct recognition. Furthermore, it was found that age had a detrimental effect on CV recognition in listeners with the same average hearing loss in terms of the audiogram. Zurek and Delhorne (1987) tested average consonant recognition scores both in HI and NH listeners. For the NH listeners, the phonemes were presented together with spectrally-shaped masking noise to simulate the sensitivity-related hearing loss of a matched HI listener. In contrast to the results from Dubno *et al.* (1984), Zurek and Delhorne (1987) found that matching NH ears to HI audiometric measures can result in a similar performance in terms of their average recognition errors. However, Zurek and Delhorne's conclusions were based on average recognition scores of their listeners and did not compare the confusions between the two listener groups, i.e., the off-diagonal elements of the CM, nor did they take the strong token dependence effect into account.

Trevino and Allen (2013) presented their stimuli to 16 HI ears at a comfortable overall loudness without a frequency-dependent gain to compensate for the audibility loss. They presented the CVs in quiet and at SNRs of 0, 6, and 12 dB in speech-shaped noise (SSN). It remained open if their observed consistency of the main confusions across listeners would also be observed if an individual frequency-dependent amplification was provided. For example, it is possible that the main confusion of /va/ observed in one token of /ba/ and the main confusion of /da/ observed in the other token of /ba/ would change if a frequency-dependent gain were provided.

The present study investigated phoneme perception on a token level in the same HI listeners as the Trevino and Allen (2013) study. In contrast to Trevino and Allen (2013), the listeners were provided with an individual frequency-dependent amplification to compensate for their audibility loss. It was tested how much the listeners improved in CV recognition as a result of the high-frequency amplification as compared to the earlier results obtained with flat (i.e., frequency-independent) amplification. The results were analyzed on a token basis using a response entropy measure to quantify the distribution of confusions as well as a vector space angular distance to evaluate how the specific nature of confusions changed between the two amplification conditions. It is argued that the two metrics together reveal a detailed picture of the relative efficacy of different

amplification schemes and could be used to assess strategies to improve speech intelligibility in general.

II. METHOD

A. Listeners

Eight HI listeners (16 HI ears) with a mean age of 74 years participated in the two experiments. All listeners reported American English as their first language and were regular users of hearing aids. They were paid to participate in the IRB-approved experiments. Tympanometric measures obtained before the start of the experiments showed no middle-ear pathologies (type A tympanogram). All 16 ears had a mild-to-moderate sensorineural hearing loss. Figure 1 shows the fitted pure tone threshold (PTT) functions of the individual listeners (Trevino and Allen, 2013). The audiograms were modeled as two piece-wise linear functions. These fittings were characterized by three parameters: the breakpoint f_0, the low-frequency loss h_0, and the slope of the high-frequency loss s_0. The break-point f_0 between the two linear functions indicates the frequency at which the sloping loss begins. At frequencies below f_0, the hearing loss was assumed to be constant over frequency (h_0). At frequencies above f_0, the audiogram was modeled by a linear function with a negative slope (s_0). The average root-mean-square error of the fitted curves over all audiogram frequencies (f = [125, 250, 500, 1000, 1500, 2000, 3000, 4000, 6000, 8000] Hz) was 5 dB (see the Appendix).

B. Stimuli

The CV syllables consisted of 14 consonants (six stops /p, t, k, b, d, g/, six fricatives /f, s, ʃ, v, z, ʒ/, and two nasals /m, n/) followed by /ɑ/. Two tokens (one recording of a male talker and one of a female talker) were selected per consonant from the Linguistic Data Consortium Database (LDC-2005S22; Fousek *et al.*, 2004). The tokens were chosen from those for which the recognition error was below 3% at a SNR of −2 dB in earlier experiments with NH listeners (Singh and Allen, 2012; Toscano and Allen, 2014). They were presented at 12, 6, and 0 dB SNR in SSN; a range in which NH listeners would not make any recognition errors. The CV tokens had previously been investigated using the three-dimensional-deep-search method (Li *et al.*, 2012) to identify perceptually relevant spectro-temporal cues in the stimuli. Furthermore, NH reference data with the same CV tokens had been collected in white noise as well as SSN (Phatak and Allen, 2007). Four of the male tokens (/f, n, s, ʒ/ + /ɑ/) had to be excluded from the analysis, as they were found to have been adversely affected by a stimulus preprocessing algorithm. The algorithm was intended to truncate all stimuli to the same duration by removing silent periods before and after the target token. Unfortunately, it truncated the weak bursts of these CV male tokens. The remaining 24 CV tokens were presented in two amplification conditions which were analyzed in the present study. The stimuli were presented to the listeners over an Etymotic Research (Elk Grove Village, IL) in-ear speaker (ER-2) in a single-walled sound booth in a room with the outer door closed.

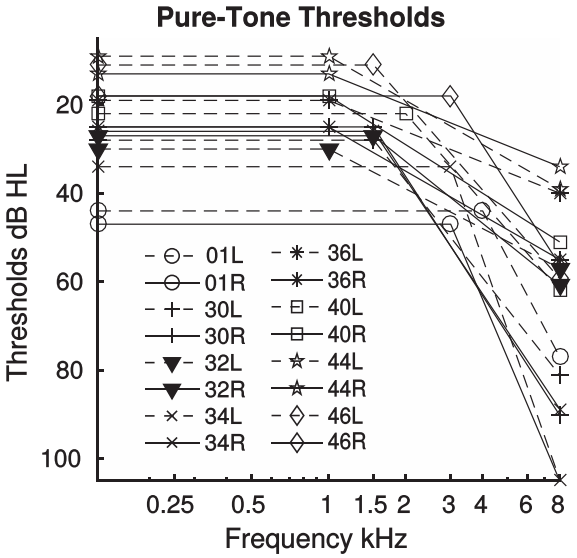

FIG. 1. Fitted pure-tone thresholds for all the listeners that participated in the study. All listeners had a steeply sloping hearing loss at high frequencies. The average root-mean-square error of the fitting was 5 dB (see the Appendix).

C. Amplification schemes

The stimuli were presented in two different amplification conditions. These conditions were tested on separate days after verifying that the audiometric thresholds of the listeners had not changed since the last session. The listeners completed a 20-min long training session per amplification condition with separate tokens before starting the testing. In the first amplification condition (FG), a frequency-independent gain was provided. The gain was chosen by the listeners in a calibration run before the training session. The levels chosen by the listeners are indicated in the Appendix. The listeners were able to adjust the gain during the experiment. However, only listener 40L made use of this option (2 dB change).

For the second amplification condition (NAL-R), the CV stimuli were amplified with an NAL-R gain adjusted for each listener according to their audiogram (Byrne and Dillon, 1986). The goal of the NAL-R amplification scheme is to provide equal loudness in all frequency bands. The insertion gain prescription is based on the PTTs at the frequencies f = 0.25, 0.5, 1, 2, 3, 4, and 6 kHz. Also in this condition, the listeners were allowed to adjust the overall gain of the amplification. The corresponding chosen levels are represented in Table I.

D. Experimental procedure

A token could be repeated as many times as required to select one of the 14 response alternatives displayed on a computer screen. The display presented symbols from the International Phonetic Alphabet (as well as a common English word that started with the respective consonant. For each condition, SNR, and listener, a token was presented between 5 and 10 times. The data collection for each amplification condition was split into two sessions in which the stimuli were presented in a fully randomized order. The number

of stimulus presentations per SNR, ear, and token was four in the first session. In the second session, the number of presentations per SNR, ear, and token depended on the number of confusions in the first session. Zero or one confusion in the first session led to two more presentations in the second session. Two confusions led to five more presentations and more than two confusions led to six additional presentations. This resulted in 800–1000 trials per listener, with more presentations allocated to the CVs that were confused by the individual listeners. This helped in identifying specific problems of individual listeners at realistic SNRs with CV tokens that were known to be robustly recognized by NH listeners at the given SNRs.

E. Analysis

In the experiments, one CM per ear (16 ears), amplification condition (2 conditions), SNR (4 SNRs), and token (2 tokens) was obtained, resulting in a total of 256 CMs. In addition to the recognition scores (i.e., diagonal CM values), two measures were considered to analyze the data.

1. Entropy

In information theory, entropy describes the randomness of a communication process. In phoneme experiments, it can be used to quantify the randomness of responses. The CM cell $\mathrm{CM}(i,j)$ contains the counts of the listeners' responses with the response alternative $j = 1, \ldots, J$ when the stimulus $i = 1, \ldots, I$ was presented. The value $\mathrm{CM}(i,j)$ of the CM, normalized by the respective row sum $\mathrm{RS}(i) = \sum_j \mathrm{CM}(i,j)$, represents the response probability $p_{ij} = \mathrm{CM}(i,j)/\mathrm{RS}(i)$, whereby the overall sum of response probabilities for a row is one ($\sum_j p_{ij} = 1$). In terms of information theory, the observation of a listener responding with j when presented with stimulus i contains the information $\log_2(1/p_{ij})$, implying that a more likely response (e.g., the correct response $j = i$) carries less information than a rarely observed response. The response entropy $\mathcal{H}(i)$ is defined as the expected information from observing all responses to a stimulus

$$\mathcal{H}(i) = \sum_j p_{ij} \log_2\left(\frac{1}{p_{ij}}\right). \quad (1)$$

Entropy as defined with the log base 2 is measured in bits. If a listener were to only use one of the response alternatives, the entropy would be 0 bit, irrespective of whether or not the response used by the listener is correct. In contrast, if all 14 possible response alternatives were to occur equally likely ($p_{ij} = (1/14)$ for all j), the response entropy would reach its maximum value, $\mathcal{H}_{\max} = \log_2(J = 14) = 3.81$ bits. The higher the entropy, the more uncertain is the listener regarding his/her responses.

The entropy, as defined above, strongly depends on the recognition score (p_{ii}) as well as the distribution of the confusions. To use the entropy as a complementary measure to the recognition score, a measure *independent* of the recognition score is needed. The *confusion entropy* $\mathcal{H}_{\mathrm{Conf}}$ used in this study is obtained by replacing the normalized response vector p_{ij} by the normalized confusion vector p_{Conf} in Eq. (1). To obtain p_{conf} the count of correct responses is excluded from a CM row before normalizing it by the row sum, i.e., the vector only consists of counts representing confusions. The values in p_{conf} therefore express the probability of a confusion occurring given an error occurs.

2. Hellinger angle

A metric that is sensitive to changes in confusion probabilities was considered. Each CM defines a vector space, with each row $\mathrm{CM}(i)$ representing a vector in that space. The vector space is defined by the basis vectors (e_j), where each basis vector represents a possible confusion. In order to find the distance between two rows (e.g., two CVs or two tokens), a norm must be defined. Here, the Hellinger Distance was used (Scheidiger and Allen, 2013), which utilizes the square roots of the probability vectors $p_i = [p_{i1}, \ldots, p_{iJ}]$. All vectors defined by the square roots of the probabilities yield the same norm and therefore have the same length. Thus, the distance between two vectors can be expressed by the angle between the vectors. Via the Schwartz inequality, it is possible to calculate an angle θ_{kl} between any two response vectors p_k and p_l in the vector space

$$\cos(\theta_{kl}) = \sum_j \sqrt{p_{kj}}\sqrt{p_{lj}}. \quad (2)$$

The angle is a measure of how different the two vectors are. In addition to ensuring unit length of all vectors, the square-root transformation emphasizes less likely confusions and makes the metric more sensitive to small changes in the response vectors than correlation-based metrics. This angular distance measure was used in the present study to represent the difference between two confusion vectors obtained in the condition with frequency-dependent gain (NAL-R) and the flat-gain (reference) condition. A Hellinger distance of 0° between the normalized confusion vector (p_{conf}) of the flat-gain and the NAL-R condition implies that the same confusions were equally likely in the two conditions. In contrast, a Hellinger distance of 90° represents cases in which the confusions in one condition (e.g., flat-gain) were not present in the other condition (e.g., NAL-R). The Hellinger distance between confusion vectors is not defined and thus yields NaN (not a number), if one of the conditions does not exhibit any errors.

III. RESULTS

Figure 2 shows the CPs of four listeners (30R, 32L, 36L, 40L) for the /bɑ/ token #1. The flat-gain condition is shown in the left panels, whereas the results obtained with NAL-R are shown on the right. The recognition score for /bɑ/ (black solid line), in general, dropped as the SNR decreased from the quiet condition (Q) to lower SNRs, i.e., at 12, 6, and 0 dB. For example, in the flat-gain condition, listener 30R (upper left panel) showed a recognition score for /bɑ/ of 63% in the quiet condition. At 12 dB SNR, the recognition score was 13% while the response probabilities for the /vɑ/ and /fɑ/ confusions increased from 0% in the quiet condition to 73% and 13%, respectively. At 6 dB SNR, listener 30R always indicated to have perceived /vɑ/. At 0 dB SNR, the confusion /vɑ/ still represented the dominating response, showing a probability of 60%, whereas the

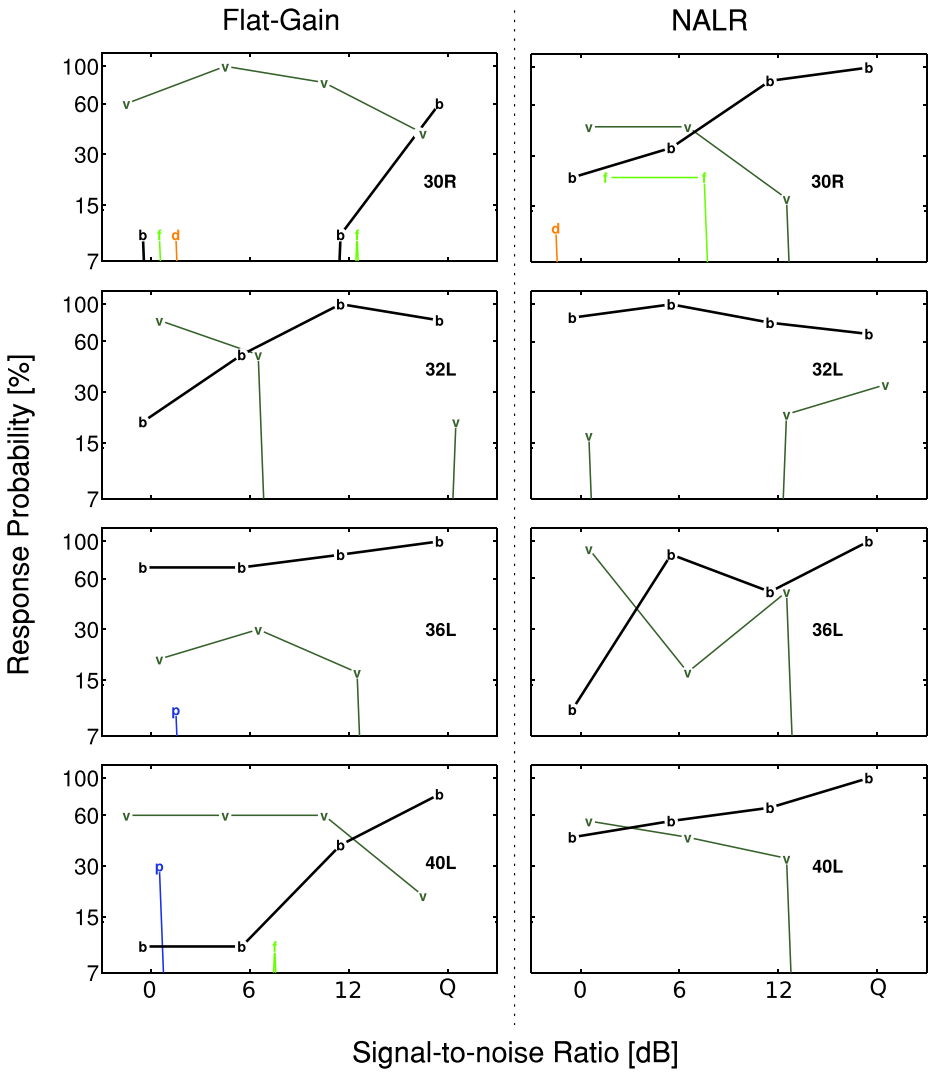

FIG. 2. (Color online) Confusion patterns for four of the subjects showing the response probabilities as a function of SNR for the token #1 of the CV /bɑ/. The left column shows the data with the flat-gain as also presented in Trevino (2013). The right column presents the data for the same listeners but with NAL-R gain. The main confusion with both gains is /vɑ/. A slight horizontal jitter was introduced to the data for better readability.

remaining responses were equally distributed over the correct response /bɑ/ and the two confusions /fɑ/ and /dɑ/.

When a frequency-dependent gain was provided using the NAL-R scheme (right column of Fig. 2), the obtained CPs differed. For example, in the case of listener 30 R, the recognition score became more robust to noise; the recognition score for /bɑ/ was at 100% in quiet, decreased to 85% at an SNR of 12 dB, and dropped to 30% at 0 dB SNR. However, despite the more robust recognition score than in the flat-gain condition, the /vɑ/ confusion was still also dominant in the NAL-R condition. With decreasing SNR, the response probability for /vɑ/ increased to 15%, 50%, and 50% at SNRs of 12, 6, and 0 dB SNR, respectively. For all four listeners shown in Fig. 2, the main confusion /vɑ/ observed in the flat-gain condition also represented the main confusion in the NAL-R condition. Less likely responses, such as /pɑ/ and /dɑ/, disappeared in the NAL-R condition. Despite the different audiograms and, therefore, different gains applied to the individual listeners in the NAL-R condition, the main confusions among the listeners remained the same. This finding is consistent with the observations reported in Trevino and Allen (2013), regarding their token-specific confusions.

Figure 3 shows the CPs obtained with the same listeners but for the other /bɑ/ token. As in Fig. 2, the recognition scores dropped as the SNR decreased. For listeners 30R, 36L, and 40L, the recognition scores with NAL-R gain were found to be more robust to noise than those obtained with flat-gain. The main confusions in the flat-gain condition for the second token were /gɑ/ and /dɑ/, in contrast to /vɑ/ in the case of the first token (Fig. 2). With the NAL-R gain (right panel), the /gɑ/ and /dɑ/ error patterns for /bɑ/ token #2 remained dominating. For example, for listener 30R (top panel), the recognition score of /bɑ/ became more robust to noise in the NAL-R condition and never dropped below 60%, but the main confusion, /gɑ/, also became more robust. For listener 32L, the NAL-R gain produced more prominent /gɑ/ confusions even at high SNRs, i.e., the presence of noise morphed the /bɑ/ into a /gɑ/.

When considering all results across all listeners, averaged across SNRs and the 24 tokens, the error rate (i.e., 1-recognition score) decreased from 20.1% in the flat-gain condition to 16.3% in the NAL-R condition. There was a significant relationship between the type of amplification and the correct recognition of the 14 phonemes [$\chi^2(1) = 56.1$, $p < 0.00001$]. The odds of a correct response with the

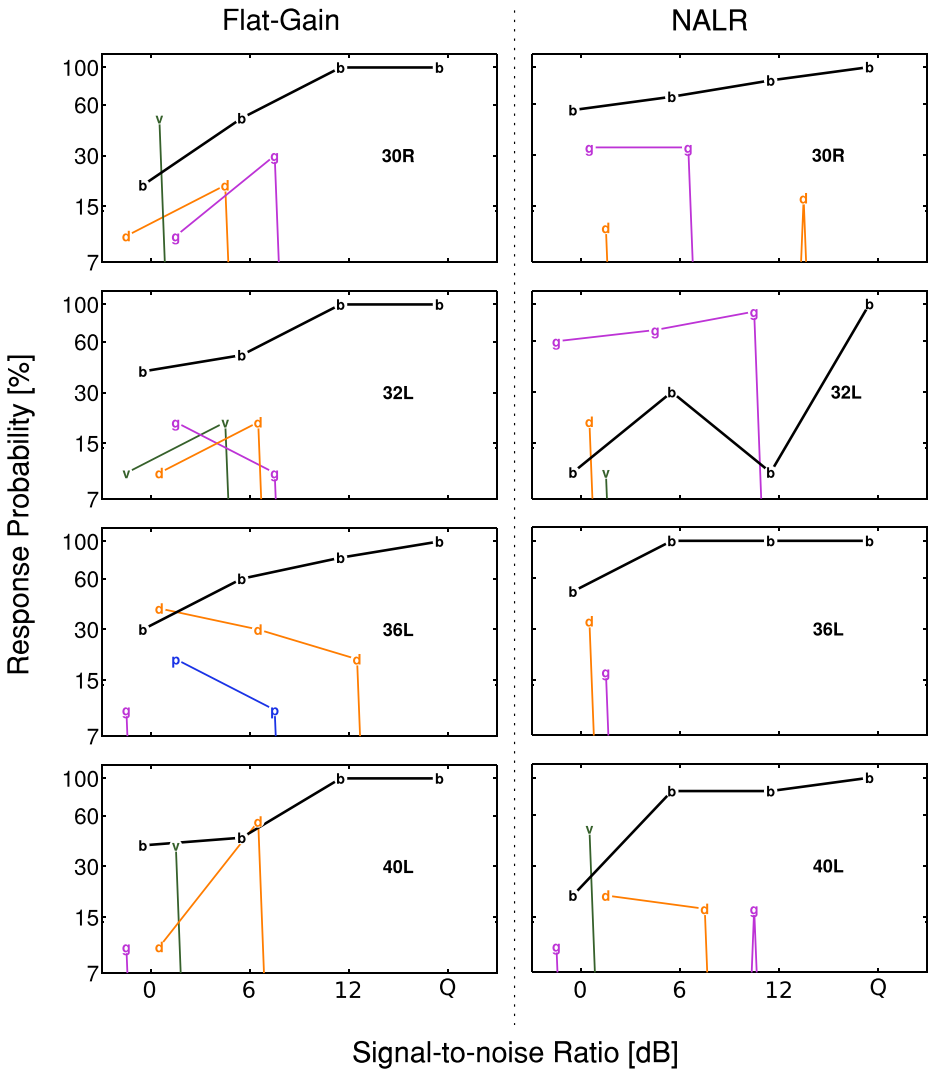

FIG. 3. (Color online) Confusion patterns for four of the subjects showing the response probabilities as a function of SNR for token #2 of the CV /bɑ/. The left column shows the data with flat-gain as also presented in Trevino (2013). The right column presents the data for the same listeners but with NAL-R gain. The main confusion with both gains is /dɑ/. A slight horizontal jitter was introduced to the data for better readability.

NAL-R amplification were 1.25 (1.18 1.33) times higher than with the flat-gain amplification. The average normalized confusion entropy ($\mathcal{H}_{\text{Conf}}$) decreased from 0.5 ($s = 0.1$) in the flat-gain condition to 0.3 ($s = 0.1$) in the NAL-R condition.

Figure 4 shows a more granular analysis of how error rates and normalized confusion entropies were affected by the two amplification conditions in an individual HI listener responding to a given token at a given SNR. For each listener-token pair, the error rate and confusion entropy ($\mathcal{H}_{\text{Conf}}$) at each SNR were calculated in the flat-gain condition and in the NAL-R condition. To compare the results obtained in the two different amplification conditions, the values in the flat-gain condition were considered as reference. The responses of the 16 HI ears to the 24 tokens at 4 SNRs resulted in 1536 response patterns for each condition.

The response patterns were divided into two categories: (i) $P_e = 0$, containing all 1044 (68%) patterns that showed maximally one erroneous response in either condition and (ii) $P_e > 0$, comprising the remaining 492 (32%) patterns which had more than one error in at least one condition. As consonant recognition was at ceiling for the $P_e = 0$ category, these response patterns were not considered in the subsequent analysis. In contrast, the $P_e > 0$ response patterns, which represent the critical/interesting cases, were further divided into three subcategories according to their error rates.

For 103 (21%) of the 492 considered token-listener pairs, P_e in the NAL-R condition increased by more than 10% as compared to the flat-gain condition (left branch in Fig. 4). In 74 response patterns (15%) the error rate did not change by more than 10% in either the positive or negative direction in the NAL-R amplification condition (middle branch). For the remaining 315 response patterns (64%), the error in the NAL-R condition decreased by at least 10% as compared to the flat-gain condition (right branch). Each of the three categories was in a last step subdivided into two subcategories according to how $\mathcal{H}_{\text{Conf}}$ changed in the NAL-R condition with respect to the flat-gain condition. The subcategories "more random" and "less random" contain the response patterns in which $\mathcal{H}_{\text{Conf}}$ in the NAL-R condition increased or decreased, respectively, compared to the flat-gain condition.

This categorization provides a detailed picture of how the NAL-R amplification scheme affected the responses to the considered CVs on a token basis. If NAL-R had improved all listeners' performance, this would have resulted in a decrease in P_e along with a decrease or no

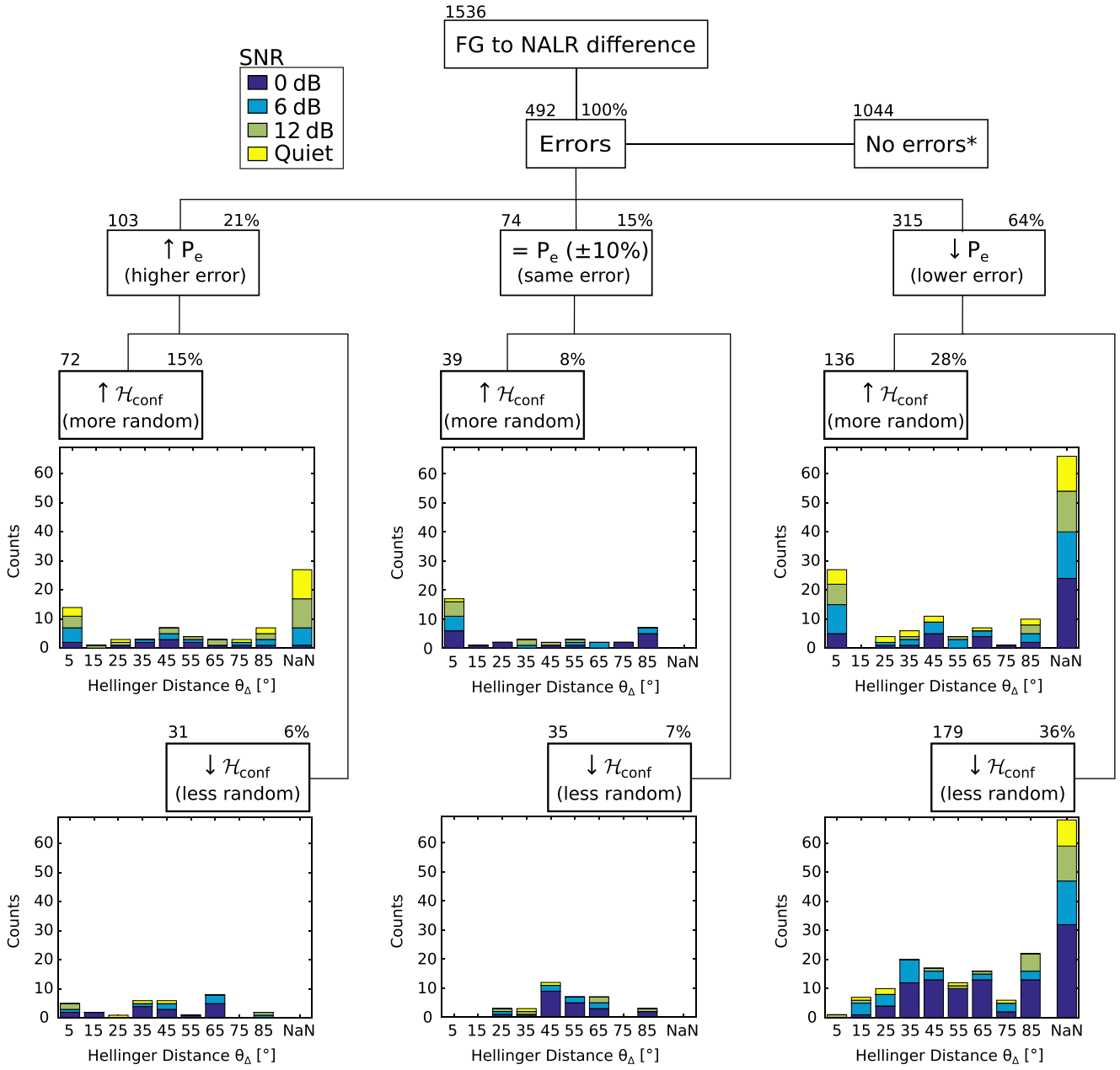

FIG. 4. (Color online) Categorization of the CV perception data for the 24 tokens, 16 listeners, and 4 SNRs. The category "No errors*," contains cases with just one or zero errors out of all trials. The response patterns with at least two errors in one of the conditions were divided into three categories according to how the error rate changed from the flat-gain condition to the NAL-R condition. Twenty-one percent of the erroneous response patterns had an increased error, 15% showed the same error (±10%), and the remaining 64% showed at least 10% fewer errors. These three categories were each further divided into two subcategories depending on how $\mathcal{H}_{\text{Conf}}$ changed in the NAL-R condition as compared to the flat-gain condition. For each subcategory, a count histogram of the Hellinger angles θ_Δ is shown on the bottom, the bins are 10° wide, and labeled by their center.

change in $\mathcal{H}_{\text{Conf}}$. However, only 36% of the considered response patterns fell into this category, while 28% showed a decrease in P_e along with more random response behavior (right branch in Fig. 4). Furthermore, 21% of the considered response patterns showed an increase in the error rate with the NAL-R amplification (left branch). Rather few response patterns were unaffected by NAL-R (15%; middle branch).

The error rate and normalized confusion entropy do not characterize the nature of confusions. Two response vectors obtained with two different tokens of the same CV might result in the same error rate and normalized confusion entropy; however, one token may show a different main confusion than the other (Trevino and Allen, 2013; cf. Figs. 2 and 3). To quantify specific confusions, the Hellinger distance was used to measure the angular distance between different response vectors. The two bottom rows of Fig. 4 show θ_Δ count histograms for each $\mathcal{H}_{\text{Conf}}$-$P_e$ subcategory. The bins of the histogram are 10° wide and are labeled by their center angle. Each response pattern is color-coded according to the SNR at which it was obtained (blue for 0 dB, turquoise for 6 dB, green for 12 dB, yellow for Quiet). It can be seen that response patterns at lower SNRs (blue, turquoise) mostly fall into the decreased error rate category (right branch in Fig. 4) and also that the cases in which NAL-R

increased both the error and the randomness of the error are dominated by quiet conditions (yellow).

The angular distance is undefined and thus yields NaN if no errors were recorded in one of the conditions. In the upper-left category ($\uparrow P_e$, more random), the 30 response patterns with $\theta_\Delta = \text{NaN}$ did not show any error in the flat-gain condition but showed errors in the NAL-R condition. These error rates were by no means small. The average error rate in the NAL-R condition for these cases was 45%, one-fifth of these cases showed error rates of >90%, indicating significant changes in the percept. Those cases can be referred to as "morphs," as NAL-R morphed them from a perceptually robust correct response into a robust confusion. For the 140 response patterns for which $\theta_\Delta = \text{NaN}$ in the $\downarrow P_e$-categories (right panel of Fig. 4), NAL-R reduced the error rate to zero. These can be referred to as "optimal" cases.

The two extreme bins of the θ_Δ histograms (centered at 5° and 85°) indicate listener-token pairs with the same or entirely different confusions in the two conditions, respectively. The 5° bin contains the cases for which the confusions and their proportions remain virtually unchanged irrespective of the amplification. In the case of the $\uparrow P_e$-categories (left panel in Fig. 4) they represent cases in which the flat-gain main confusions were chosen even more frequently in the NAL-R condition. In the $\downarrow P_e$ category (right panel in Fig. 4), they represent cases for which the error rate decreased but the main confusion remained the most likely confusion. A low θ_Δ indicates that the confusions in the flat-gain condition also dominated the response pattern in the NAL-R condition. The 5°-bin reflects the most prominent examples for this behavior but the same trend can also be observed for bins where $\theta_\Delta < 45°$. Considering a threshold of $\theta_\Delta = 45°$ to indicate whether the main confusion remained the same (<45°), the analysis reveals that in 63% of the cases the main confusions remained unchanged.

$\theta_\Delta = 90°$—contained in the bin centered at 85°—indicates that the confusions were different and that the response vector for the NAL-R condition did not contain the confusions in the flat-gain condition and vice versa. Thus, in these cases, NAL-R introduced new confusions that were not present in the flat-gain responses (morphs). In all but two θ_Δ-histograms, the 5°-bins exhibited larger counts than the 85°-bins, indicating that the main confusions in these patterns were unchanged.

IV. DISCUSSION

The results from the present study support the findings of Trevino and Allen (2013) that the confusions in CV experiments are token specific, even if a frequency-dependent gain (NAL-R) is provided. While NAL-R, on average, decreased the error rate in the listeners' responses, the occurrence of the main confusions often remained the same (Figs. 2 and 3 and <45° in Fig. 4), indicating that NAL-R alone does not effectively compensate for the deficits that cause the main confusion. The observation of small values for the normalized confusion entropy in both amplification conditions (0.5 bit in the flat-gain condition as compared to 0.3 bit in the NAL-R condition) suggests that the main confusion is a robust and consistent phenomenon caused by token-specific cues and deficits in the individual auditory system. The different main confusions for the two /bɑ/ tokens that are robust across the two amplification conditions, suggest that they are caused by the acoustic properties of the stimulus, i.e., by conflicting consonant cues (Kapoor and Allen, 2012). A stimulus that evokes responses with low entropy but a high error rate must have been chosen based on a robust auditory percept. This percept must therefore result from some distorted internal auditory representation of the stimulus which could be considered as reflecting a "supra-threshold" distortion (such as, e.g., a temporal and/or spectral auditory processing deficit). Such a distortion could affect the primary consonant cue and increase the perceptual salience of a secondary cue that then causes the main confusion. In the case of the 30 morphs observed in the results, the robust confusions resulted from supra-threshold deficits in the HI listeners' auditory processing in combination with the high-frequency amplification. An understanding of which specific cues were used by the HI listeners would require a closer analysis of the individual audiometric configuration, the applied amplification, and the specific cues of the confused tokens (Li *et al.*, 2010, 2012) which were not undertaken in the present study. In contrast to the conditions with low-entropy response patterns, conditions where the confusion entropy was large are not based on a robust percept and should be assessed differently. The high entropy in these responses indicates that the listener did not respond based on a robust cue, but instead selected the response randomly. Such randomness may be caused by the effect of "internal" noise or attention deficits of the listener.

To define the entropy threshold for a robust percept, the average size of the Miller and Nicely (1955) confusion groups (/p, t, k, b, d, g/; /f, t, s, ʃ/; /v, D, z, ʒ/; /m, n/) may be used. A listener is most likely guessing and therefore not responding based on a robust percept when confusions outside of the known confusion groups appear. The average size of confusion groups is three; thus, if more than three confusions occur, a decision-threshold for a robust percept could be defined in terms of the normalized confusion entropy which would be $\mathcal{H}_{\text{Conf}} = 0.43$ bit (3 equally likely confusions out of the 13 possible confusions). When assessing the flat-gain response vectors with this definition, only 268 out of the 1536 token-listener pairs (17%) would not qualify as robust percepts.

A robust auditory percept might also be more appropriate than the traditional PTT and LTASS (i.e., count-the-dots method) to assess the audibility of CV signals. In experiments such as the ones from the present study, the differentiating perceptual cues of CVs may be manifested as local energy bursts or spectral edges in the signal (Li *et al.*, 2010, 2012). These cues can be more intense than the LTASS in a critical band over several 10 ms (Wright, 2004), but have a negligible contribution to the LTASS which is dominated by the vowel energy. It has been shown that CV recognition on a token level in NH listeners can drop from 100% correct to chance level if the energy of the noise masker is increased by less than 6 dB (Singh and Allen, 2012; Toscano and Allen, 2014). This "binary"-like recognition supports the importance of specific acoustic speech cues. These cues are either detectable, in which case the CV can be recognized despite the presence of noise,

or are masked by the noise, in which case the listener might use a secondary cue or might start guessing. PTTs do not characterize a listeners' sensitivity to recognize these spectro-temporal consonant cues. Furthermore, if a different amplification scheme were chosen instead of NAL-R that aims at restoring audibility, e.g., a scheme as proposed in Reed *et al.* (2016), the specific confusions that exist after compensating for audibility can be used as an indicator of a supra-threshold distortion loss. To quantify the distortion loss based on CMs, the angular Hellinger distance measure could be used.

The response-patterns where both the error rate and the confusion entropy increased with NAL-R indicate the listener-specific phonemes for which the improvement strategy failed. These specific confusions could not be eliminated by NAL-R alone and should be addressed by alternative compensation strategies. Such strategies should take the token-specific consonant cues into account; the primary consonant cue should be amplified and conflicting secondary cues attenuated (Kapoor and Allen, 2012). For example, individually tuned frequency transposition algorithms may be able to transpose the spectro-temporal cues of the affected CVs to bands that are less affected by the distortion loss. Phoneme tests can help determine sensible limits for such frequency transposition algorithms to avoid further distortions (Schmitt *et al.*, 2016). Such phoneme tests should consist of several well-characterized tokens for each consonant. These tokens should be correctly perceived by NH listeners at the SNRs tested. The recognition results should be analyzed on a token-specific level taking confusions and not only recognition scores into account. Zaar and Dau (2015) emphasized that the additive noise should be frozen noise, i.e., one noise realization per token, to further decrease the within-listener variance in the responses.

V. SUMMARY AND CONCLUSION

CV perception in the same HI listeners as in Trevino and Allen (2013) was analyzed on a token level in two amplification conditions: a condition with frequency-independent amplification (flat-gain) and a condition with frequency-dependent amplification (NAL-R). The response patterns were analyzed in terms of their recognition scores, their confusion entropy, and an angular distance between the confusions in the two amplification conditions. The recognition score in the NAL-R condition was shown to be significantly higher than in the flat-gain condition. In a granular analysis (Fig. 4), the response patterns showed mixed results for the NAL-R condition, despite the overall increased recognition score.

Two measures were proposed to analyze the efficacy of speech intelligibility improvement strategies using a phoneme test, namely, the confusion entropy and an angular distance. The effect of a frequency-dependent gain was exemplarily investigated. The confusion entropy measure showed robust perception in all but 17% of the token-listener pairs in the flat-gain condition and thus demonstrated the validity of the results obtained at the most comfortable listening level. The proposed angular distance measure revealed that in 63% of the token-listeners pairs, the main confusions remained unchanged despite NAL-R, suggesting these are caused by acoustic properties of the chosen tokens rather than the amplification condition. The results suggest that a compensation strategy different than NAL-R would be needed to eradicate the main confusion. It was also observed that NAL-R in combination with the individual loss introduced new robust confusions in 30 cases.

Phoneme recognition tests and methods that analyze confusions on a token-level, as the ones used in the experiments presented here, may be useful in the evaluation process of hearing-instrument algorithms. The tests could be conducted with selected robust tokens that have been shown to be correctly identified by NH listeners at the SNRs used in the test. Knowing the token-specific consonant cues and using a test that is focused on natural speech without context, a detailed diagnosis of an individual listener's speech loss seems possible and appropriate. A carefully constructed speech test could be used as a diagnostic tool where individual CPs of well characterized tokens may provide detailed information about a listener's hearing loss beyond what PTTs reveal.

ACKNOWLEDGMENTS

The authors are grateful for helpful input by Johannes Zaar and the HSR group at the University of Illinois. The work leading to this deliverable and the results described therein has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement No. PITN-GA-2012-317521.

APPENDIX

Additional information about the listeners who participated in the study.

TABLE I. Information about all the listeners participating in the experiments. The columns contain the following information: (i) label for each listener and the identifier for the left or right ear, (ii) age of the listener, (iii) the pure-tone average of the audiogram of the ear, (iv) the root means square error of the fitted audiogram, (v) the overall presentation level chosen by the listener in the FG experiment, and (vi) the overall presentation level chosen by the listener in the NAL-R experiment.

HI ear	Age	PTA	RSME	FG	NALR
44L	65	10	11	82	77
44R	65	15	7	78	77
46L	67	8.3	9	82	85
46R	67	16.6	7	82	86
40L	79	21.6	5	79,81	80
40R	79	23.3	5	80	80
36L	72	26.6	8	68	75
36R	72	28.3	4	70	75
30L	66	30	3	80	79
30R	66	26.6	5	80	79
32L	74	35	3	79	81
32R	74	26.6	3	77	78
34L	84	31.6	6	84	85
34R	84	28.3	4	82	85
02L	82	45	2	83	88
02R	82	46.6	4	82	89
(*m*,*s*)	(74,7)	(29,15)	(5,2)	(79,4)	(81,5)

Allen, J. B. (**2005**). "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**, 2212–2223.
Bilger, R. C., and Wang, M. D. (**1976**). "Consonant confusions in patients with sensorineural hearing loss," J. Speech Hear. Res. **19**, 718–748.
Boothroyd, A. (**1984**). "Auditory perception of speech contrasts by subjects with sensorineural hearing loss," J. Speech Hear. Res. **27**, 134–144.
Byrne, D., and Dillon, H. (**1986**). "The National Acoustic Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," Ear Hear. **4**, 257–265.
Dubno, J. R., Dirks, D. D., and Morgan, D. E. (**1984**). "Effects of age and mild hearing loss on speech recognition in noise," J. Acoust. Soc. Am. **76**, 87–96.
Fousek, P., Grezl, F., Hermansky, H., and Svojanovsky, P. (**2004**). "New nonsense syllables database-analyses and preliminary asr experiments," in *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, 2004, p. 29.
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sound," J. Acoust. Soc. Am. **19**, 90–119.
Humes, L. E., Dirks, D. D., Bell, T. S., Ahlstrom, C., and Kincaid, G. E. (**1986**). "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **29**, 447–462.
Kapoor, A., and Allen, J. B. (**2012**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am. **131**, 478–491.
Li, F., and Allen, J. B. (**2011**). "Manipulation of consonants in natural speech," IEEE Trans. Audio, Speech, Lang. Process. **19**(3), 496–504.
Li, F., Menon, A., and Allen, J. B. (**2010**). "A psychoacoustic method to find the perceptual cues of stop consonants in natural speech," J. Acoust. Soc. Am. **127**, 2599–2610.
Li, F., Trevino, A., Menon, A., and Allen, J. B. (**2012**). "A psychoacoustic method for studying the necessary and sufficient perceptual cues of fricative consonants in noise," J. Acoust. Soc. Am. **132**, 2663–2675.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Mueller, H. G., and Killion, M. C. (**1990**). "An easy method for calculating the articulation index," Hear. J. **43**(9), 14–17.
Owens, E. (**1978**). "Consonant errors and remediation in sensorineural hearing loss," J. Speech Hear. Disord. **43**, 331–347.
Pavlovic, C. V. (**1988**). "Articulation index predictions of speech intelligibility in hearing aid selection," ASHA **30**(6–7), 63–65.
Phatak, S. A., and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2326.
Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception-threshold of the hearing-impaired," J. Speech Hear. Res. **29**, 146–154.
Reed, C. M., Desloge, J. G., Braida, L. D., Perez, Z. D., and Agnès, C. L. (**2016**). "Level variations in speech: Effect on masking release in hearing-impaired listeners," J. Acoust. Soc. Am. **140**, 102–113.
Scheidiger, C., and Allen, J. B. (**2013**). "Effects of NAL-R on consonant-vowel perception," in *4th International Symposium on Auditory and Audiological Research* (*ISAAR-2013*), Nyborg, Denmark.
Schmitt, N., Winkler, A., Boretzki, M., and Holube, I. (**2016**). "A phoneme perception test method for high-frequency hearing aid fitting," J. Am. Acad. Audiol. **27**, 367–379.
Singh, R., and Allen, J. B. (**2012**). "The influence of stop consonants perceptual features on the articulation index model," J. Acoust. Soc. Am. **131**, 3051–3068.
Toscano, J. C., and Allen, J. B. (**2014**). "Across- and within-consonant errors for isolated syllables in noise," J. Speech Lang. Hear. Res. **57**, 2293–2307.
Trevino, A. (**2013**). "Techniques for understanding hearing impaired perception of consonant cues," Ph.D. thesis, University of Illinois at Urbana-Champaign.
Trevino, A., and Allen, J. B. (**2013**). "Within-consonant perceptual differences in the hearing impaired ear," J. Acoust. Soc. Am. **134**(1), 607–617.
Wright, R. (**2004**). "A review of perceptual cues and cue robustness," in *Phonetically Based Phonology*, edited by B. Hayes, R. Kirchner, and D. Steriade (Cambridge University Press, Cambridge), pp. 34–57.
Zaar, J., and Dau, T. (**2015**). "Sources of variability in consonant perception of normal-hearing listeners," J. Acoust. Soc. Am. **138**, 1253–1267.
Zurek, P. M., and Delhorne, L. A. (**1987**). "Consonant reception in noise by listeners with mild and moderate sensorineural hearing impairment," J. Acoust. Soc. Am. **82**, 1548–1559.