Evaluating hearing aid amplification using idiosyncratic consonant errors

Ali Abavisani[a] and Jont B. Allen
Department of Electrical and Computer Engineering, The Beckman Institute, University of Illinois at Urbana-Champaign, 405 North Mathews Avenue, Urbana, Illinois 61801, USA

(Received 6 September 2016; revised 8 November 2017; accepted 21 November 2017; published online 22 December 2017)

The goal of this study is to provide a metric for evaluating a given hearing-aid insertion gain using a consonant recognition based measure. The basic question addressed is how treatment impacts phone recognition at the token level, relative to a flat insertion gain, at the most-comfortable-level (MCL). These tests are directed at fine-tuning a treatment, with the ultimate goal of improving speech perception, and to identify when a hearing level gain-based treatment degrades phone recognition. Eight subjects with hearing loss were tested under two conditions: flat-gain and a treatment insertion gain, based on subject's hearing level. The speech corpus consisted of consonant-vowel tokens at different signal to speech-weighted noise conditions, presented at the subject's MCL. The treatment caused the average score to improve for 31% of the trials and decrease for 12%. An analysis method based on the accumulated error differences was devised to quantify the benefit each individual ear received from the treatment. Using this measure, the effect of the treatment could be evaluated, providing precise characterization of idiosyncratic phone recognition. This analysis directs the audiologist toward the most susceptible subject-dependent tokens, to focus in the process of fine-tuning the insertion gain of the hearing-aid. © *2017 Acoustical Society of America*.
https://doi.org/10.1121/1.5016852

[ICB] Pages: 3736–3745

I. INTRODUCTION

The primary purpose of hearing aids is to improve speech perception. Yet speech has not proven to be effective in evaluating hearing aid processing (Wilson *et al.*, 2007). The insertion gain of the majority of current hearing aids is designed to compensate for audiometric thresholds, as quantified by the hearing level (HL) (Steinberg and Gardner, 1940; Zurek and Delhorne, 1987), along with some assumptions about the wide band sound pressure level (SPL) required to detect speech features. However, there seems to be no consensus on the utility of HL as a metric. Some research has assumed that audibility is sufficient to characterize speech perception (Wang *et al.*, 1978; Zurek and Delhorne, 1987). Other research supports the hypothesis that HL, while a necessary factor, is not sufficient in accounting for speech perception in hearing impaired (HI) ears (Plomp and Mimpen, 1979; Plomp, 1986; Yoon *et al.*, 2012; Phatak *et al.*, 2009; Trevino and Allen, 2013b). A known problem has been finding objective metrics of Plomp's distortion factor, which is based on the *speech reception threshold*, a gross bi-syllabic word intelligibility metric. The recent observation denoted "hidden hearing loss" is an example of a condition, also known as *auditory neuropathy*, where audiometric thresholds can be normal, yet individuals cannot decode speech properly (Starr *et al.*, 1996). This condition is believed to result from the synaptic loss of high-threshold cochlear inner hair cells, following a single noise exposure resulting in a 50 dB temporary threshold shift (TTS) (Valero *et al.*, 2016).

Various insertion gain prescription methods have evolved, such as National Acoustics Lab (Revised) (NALR) (Dillon, 2001), in an attempt to map audiometric thresholds into an insertion gain. The assumption is that the optimal insertion gain will improve audibility and, as a result, speech intelligibility. This assumption is consistent with the results from Zurek and Delhorne (1987), which showed that audibility seems to work with reasonable fidelity, when averaged across a large variety of speech sounds. The persistence of speech loss, once audibility has been compensated, supports the possibility that there must be other factors, such as outer hair cell loss and auditory neuropathy, that are playing an important, if not a key role, in HI speech recognition.

One major problem with focusing on audibility is that there has been no fundamental understanding of the precise nature of the speech cues, namely, which speech features need to be audible? The popular view of speech cues are *distinctive features* such as *voicing, manner, place, and nasality* (Miller and Nicely, 1955). These broad-brush features are production rather than perception based, thus they do not account for the large within-class variability, because they show no correlation with token errors (Toscano and Allen, 2014). Acoustic features that are necessary for normal hearing (NH) listeners are also necessary for HI listeners, but they may not be sufficient (Trevino and Allen, 2013a). Consistent token-specific confusion groups between HI listeners support the hypothesis that HI ears use similar cues, despite the audiometric configuration (Trevino and Allen, 2013b). Thus, it has proven difficult to assign an audibility index to unspecified features.

The NALR method for finding the optimum gain rule is based on hearing thresholds and aimed at maximizing speech

[a]Electronic mail: aliabavi@illinois.edu

intelligibility, while making overall loudness of mid-level speech comfortable (Dillon, 2001). A speech metric that provides diagnostic information would be easily justified, but to date, such speech metrics have not been successful. Many non-articulatory approaches have been researched, such as *listening in the gaps* (Rhebergen *et al.*, 2006) and *speech fine-structure* (Apoux and Bacon, 2004). These methods are not feature based. Thus, it remains unclear how they can deal with the large idiosyncratic nature of HI phone recognition (Trevino and Allen, 2013b; Zaar and Dau, 2015). The best approach among these several methods is yet to be established.

The articulation index (AI) is a venerable feature based speech metric which depends on the signal to noise ratio (SNR [dB]) in human listeners' critical bands (Allen, 2005a, 1994). It has been demonstrated that while the AI is a valid identification metric for consonants averaged across a corpus of speakers, it massively fails at the individual token level (Singh and Allen, 2012; Toscano and Allen, 2014).

It was shown in a number of earlier studies that the errors HI subjects make depend on the token, not just on consonant or feature classes (Trevino and Allen, 2013a,b; Scheidiger *et al.*, 2017; Allen and Abavisani, 2016; Abavisani and Allen, 2017). These studies showed that our traditional view of class-average errors is misleading. At any amplification condition, there are numerous zero-error tokens along with a few high error tokens, and averaging hides the degree of error for individual errorful tokens, thus diminishing the estimate of received benefit from that amplification procedure. Token errors need to be statistically evaluated using a robust procedure. By focusing on token errors, we satisfy these requirements.

For example, Trevino and Allen (2013a,b) looked at idiosyncratic consonant errors for a given hearing impaired (HI) subject, and showed that the error specifically depends on individual tokens. Two different utterances of a same consonant, say /vɑ/, have different error patterns as a function of SNR (Trevino and Allen, 2013a, Fig. 4a). In fact, one /vɑ/ can (and does) have zero error, even at 0 dB, while a second /vɑ/ has 100% error. Following this observation, Trevino and Allen (2013a, Fig. 5) suggested sorting the tokens by errors, to parse out token errors.

This suggests the need to look deeper into individual differences, to get a better understanding of how HI ears recognize speech (Trevino and Allen, 2013b). For a given HI ear, there is no way to predict which tokens can be correctly recognized, and which cannot, as they are different for each ear. To advance understanding of this idiosyncratic deficiency of HI ears, a more sensitive test is required. Here, we further develop the concentration on token errors into a metric that is both robust and insightful.

In normal hearing ears each consonant becomes masked at a token dependent threshold, denoted SNR_{90} (Régnier and Allen, 2008). As the noise is increased from quiet (no noise), the identification of most sounds goes from less than 0.5% error to 10% error (at SNR_{90}), and then to chance performance (e.g., more than 90% error), over an SNR range of just a few dB (i.e., less than 10 dB) (Toscano and Allen, 2014). Hence SNR_{90} is an important token-specific threshold metric of noise robustness.

SNR_{90} has recently been shown to be a useful tool for classifying HI ears (Trevino and Allen, 2013b). When HI ears are tested at SNR_{90} for each token, most tokens show errors around 10%, similar to those of NH ears. However, a small subset of tokens have errors much higher than 10%, even approaching chance at quiet. Thus, it is proposed that this subject-dependent subset of tokens quantifies the idiosyncratic variability between NH and HI ears.

This study proposes to use the speech tokens at four SNR levels well above SNR_{90} (i.e., $SNR_{90}+6$ dB). With such a scheme, a single error is highly statistically significant, since for the NH ear, one error in 32 presentations at $SNR_{90}+6$ dB is rare (Singh and Allen, 2012). Therefore, such a metric is highly efficient in characterizing each HI ear. Once high error sounds have been identified, one may seek the optimum treatment (insertion gain) to efficiently prevent increase of the token error relative to flat-gain condition. Presumably, with this strategy, a small subset of errorful tokens will not be worse than the flat-gain condition.

II. METHODS

There were two experiments, one with a flat insertion gain, and a second treatment gain that depended on each hearing impaired ear's hearing loss. The goal of the study was to quantify the error for each token as a function of signal to noise ratio (SNR) and insertion gain at each subject's most comfortable level (MCL).

A. Speech materials

Throughout these studies, the term *token* refers to one of 24 specific consonant-vowel (CV) sounds. The 14 consonants were one of /p, t, k, f, s, ʃ, b, d, g, v, z, ʒ, m, n/; the vowel was always /ɑ/ (i.e., as in /cot/).

The CV tokens were drawn from an earlier experiment that measured the confusions as a function of SNR for 30 normal hearing (NH) listeners (Li *et al.*, 2010). The tokens were restricted to be noise-robust, defined as having a recognition error as measured by 30 NH ears of less than 10% at SNR $= -2$ dB, with an average error of $< 3.1\%$ [i.e., less than 1 in 32 trials (Phatak and Allen, 2007; Singh and Allen, 2012; Toscano and Allen, 2014) at the four test SNRs (i.e., 0, 6, 12 dB and quiet)]. During the testing, speech shaped computer generated thermal noise was added to the token at one of the four SNRs.

Each token was naturally spoken as an isolated (i.e., no carrier phrase) consonant-vowel (CV) token, by an American English speaking talker, and available from the Linguistic Data Consortium Database (LDC-2005S22) (Fousek *et al.*, 2004). The sampling rate was 16 kHz.

After amplifying with the target insertion gain, the speech was presented at each subject's most comfortable level (MCL), as determined during initial trials used to familiarize the subjects with the task (Han, 2011). The subjects were allowed to subsequently adjust the presentation level at anytime during the experiments. However, none ever did this.

The speech samples were selected from a pool of six female talkers and five male talkers. The gender of the talker

was found to play no observable role in the identification of the consonant (Trevino and Allen, 2013a,b). Initially, the test included two talkers (one male and one female) for each consonant drawn from these 11 talkers, but after the experiment, it was discovered that four male tokens of /f, s, ʒ, n/ had been damaged when preparing the samples by zero-padding the onsets and codas, before adding the noise, such that the low-intensity fricative cues of these tokens had been partially removed. Therefore, the analysis is based on the remaining 24 tokens: 6 female (101, 103, 105, 106, 108, 109) and 3 male (111, 112, 118) talkers. A list of the specific tokens used in the current study, along with each token's SNR_{90}, is provided in Table I.

B. Subjects

Eight subjects with mild to moderate hearing loss, ranging from 65 to 84 years, were recruited from the Urbana-Champaign, IL, community. IRB approval was obtained from the University Review board. Subjects were paid.

Before running experiments, both ears were tested for hearing loss and type A tympanograms, consistent with no middle-ear pathologies. Based on the pure-tone thresholds, all ears had hearing loss greater than 20 dB for at least one frequency in the range from 0.25 to 4 kHz (see Fig. 1). The three frequency average (0.5, 1, 2 kHz) hearing loss for these eight subjects was 26 dB. All subjects identified American English as their first language. Subjects had been aware of their hearing loss for 16 years on average. All except one (subject 01), who had a 40 dB loss, were not regular users of hearing aids. Figure 1 illustrates subjects' audiograms.

As discussed in the Appendix, the hearing loss profile can be summarized by three parameters, h_0, f_0, and s_0, that represent the low-frequency hearing loss, the breaking frequency (intercept of the two lines), and the slope of the high-frequency hearing loss, respectively. These parameters are provided in Table II for each subject.

C. Insertion gain calculation

This study includes two separate experiments on the same subjects with the same set of tokens. Stimuli in experiment 1 were presented with flat-gain at the ear's most comfortable level (MCL_{Flat}). In experiment 2, stimuli were presented with a treatment (NALR) insertion gain, at the ear's most comfortable level (MCL_{Treat}). In order to determine the NALR gain, we used the real ear gain (REG) function that finds the frequency dependent insertion gain (REG_f) based on the pure-tone thresholds (PTT) for individual HI ears (Dillon, 2001). The function calculates REG_f in dB in two steps:

$$X = 0.15 \times (PTT_{0.5} + PTT_{1.0} + PTT_{2.0})/3 \quad (1)$$

and

$$REG_f = X + 0.31 \times PTT_f - C_f, \quad (2)$$

where PTT_f is the pure-tone threshold in dB at the measured frequencies $f = 0.25, 0.5, 1, 1.5, 2, 3, 4, 6$ kHz, and C_f is the specific gain 17, 8, 3, −1, −1, 1, 2, 2 dB, respectively.

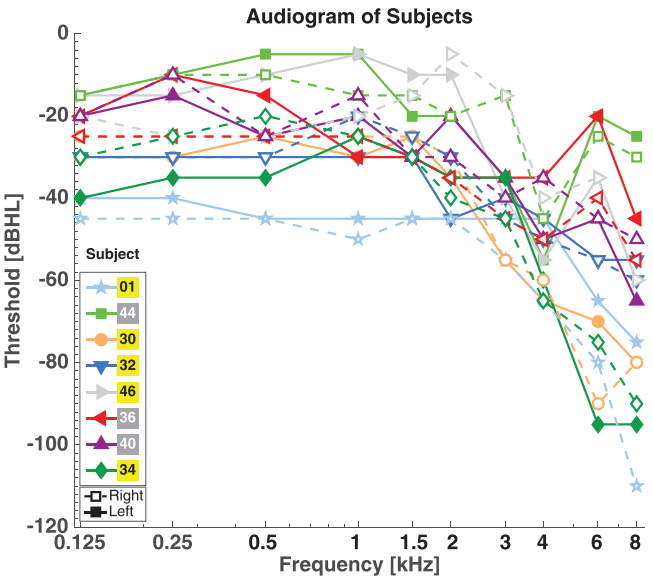

FIG. 1. (Color online) Audiograms (pure-tone thresholds) of all subjects participated in this study. Solid lines (closed symbols) indicate left ear and dashed-lines (open symbols) show right ears. All subjects had high frequency sloping hearing loss. One subject (44) had a notched loss at 4 kHz. Legend: subjects with average low error are labeled with gray background.

TABLE I. For each consonant-vowel (CV) token used in this study, the male (M) and female (F) talker labels are listed, along with the corresponding NH SNR_{90} values (in dB). As discussed in the text, several male talker fricative consonants (f, s, ʒ, n) were removed from this study, as indicated by a dash. Note: due to limitations in graphical software, vowel /ɑ/ is shown as /a/ and consonants /ʃ/ and /ʒ/ are shown as /S/ and /Z/ in figures throughout this manuscript.

CV	M-talker	SNR_{90}	F-talker	SNR_{90}	CV	M-talker	SNR_{90}	F-talker	SNR_{90}
bɑ	m112	−2	f101	−10	pɑ	m118	−14	f103	−17
dɑ	m118	−7	f105	−13	sɑ	—	—	f103	−13
fɑ	—	—	f109	−12	ʃɑ	m118	−16	f103	−15
gɑ	m111	−12	f109	−3	tɑ	m112	−17	f108	−14
kɑ	m111	−13	f103	−11	vɑ	m118	−3	f101	−10
mɑ	m118	−14	f103	−11	ʒɑ	—	—	f105	−17
nɑ	—	—	f101	−7	zɑ	m118	−17	f106	−18

TABLE II. Fitting parameters of the pure-tone threshold of HI ears using the flat low frequency loss h_0, the frequency f_0 at which sloping loss begins, and the sloping high frequency loss s_0. ϵ determines the root-mean-square error of the fit, and MCL shows the overall most comfortable level of amplification that is selected by the subject in each experiment. Subjects are ordered based on left and right ears, h_0 average values.

Subject		44		46		40		36		30		32		34		01	
Ear		L	R	L	R	L	R	L	R	L	R	L	R	L	R	L	R
h_0	[dB]	9	13	11	18	22	18	19	25	28	25	30	27	34	26	44	47
f_0	[kHz]	1	1	1.5	3	2	1	1	1	1.5	1.5	1	1.5	3	1.5	4	3
s_0	[dB/Oct]	10	7	20	27	20	11	7	10	22	27	9	14	50	26	33	41
ϵ	[dB-RMS]	11	7	9	7	5	5	8	4	3	5	3	3	6	4	2	4
MCL_{Flat}	[dB]	82	78	82	82	80	80	68	70	80	80	79	77	84	82	83	82
MCL_{Treat}	[dB]	77	77	85	86	80	80	75	75	79	79	81	78	85	85	88	89

Total insertion gain: In Fig. 2 we show the insertion gain difference between the two experiments. The treatment gain + MCL_{Treat} is corrected by the flat MCL_{Flat} gain, to give the total gain difference between the two conditions. This shows that the spread in the gain below 2 kHz has been reduced to have standard deviation less than 4.4 dB on average.

D. Experimental procedure

To investigate the effect of changing the speech amplification from flat to treatment insertion gain, presented at MCL, all the procedures in both experiments were the same, other than the insertion gain (REG_f). The speech signal was mixed with speech-weighted noise as described by Phatak and Allen (2007) to set the SNR to 0, 6, 12 dB and no-noise (quiet) condition. Presentation order was randomized over consonant, talker, and SNR. Each experiment was performed in two separate 30–40 min sessions for each HI ear, with a brief mid-session break. During the first session, each token was presented four times at each SNR, resulting in a total of 32 presentations for each consonant (2 talkers × 4 presentations × 4 SNRs). Based on the subject's performance, the second session was designed to repeat the errorful tokens of the first session in a random order. If the HI ear had no error for a given token, it was presented one more time. Otherwise, the token was presented six more times. The total number of presentations for each consonant thus ranged from N = 40 to 80 for each HI ear ($N_{total} = 5 - 10 \times 2$ talkers × 4 SNRs).

According to the Vysochanskij and Petunin (1980) inequality, ten presentations of a token are required for a statistically significant result, assuming a 95% confidence interval. This criterion was based on the null-hypothesis that a token with a probability of less than 10% error could be detected above chance when compared to a token having a 50% error. This method is an example of Fisher's exact method, analyzed via a Monte Carlo simulation. The details of this statistical analysis are described in the Appendix of Singh and Allen (2012).

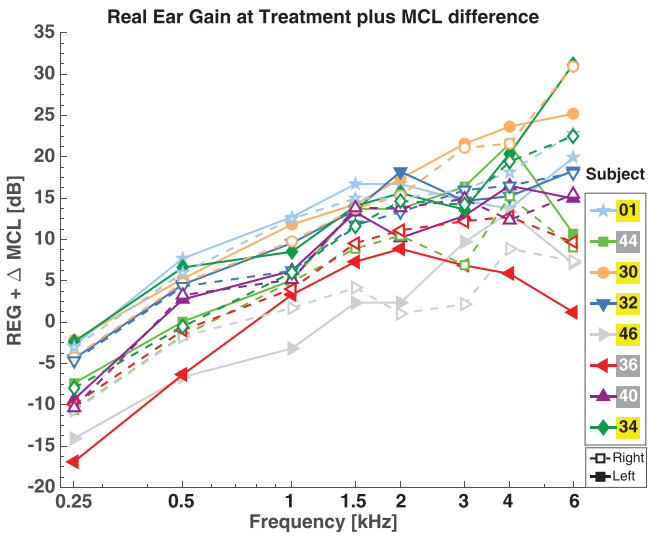

FIG. 2. (Color online) Difference between the total treatment gain less the flat-gain, including the MCL for the two conditions. The use of MCL decreased the spread of the compensation gain below 2 kHz (standard deviation <4.4 dB on average). Legend: subjects with average low error are labeled with gray background.

A MATLAB® graphical user interface was provided to run the experiments. All of the data collection sessions were conducted with the subject seated in a single-walled, soundproof booth with the door of the outer room closed. The speech was presented through an Etymotic ER-3 insert ear phone, one ear at a time. The contra-lateral ear was not masked or occluded. To familiarize the subjects with the testing paradigm, a practice session was run using non-test tokens. The MCL was determined during the practice session. Throughout the remaining sessions the tokens were randomized.

After hearing each token, the subject was instructed to choose the response from 14 possible consonants-vowel labeled buttons that were provided on the screen via a graphical interface. To get more precise results, subjects were allowed to play uncertain tokens up to two additional times before making their decision. To reduce fatigue, subjects were encouraged to take short breaks approximately every 20 min. A detailed description of these experiments is provided in Han (2011).

E. HI data analysis

The data collected by the experiments was the *confusion matrix* as a function of SNR. Since we conducted our study on 14 CV sounds, each of 16 HI ears resulted in $2 \times 2 \times 4 = 16$ confusion matrices of size 14×14: 2 experiments, each including 2 talkers and 4 SNRs conducted on 16 HI ears (total of 256 confusion matrices). Thus, each of the 24 tokens has an empirical probability distribution defined by a row of the count (unnormalized confusion) matrix. We refer to the ith token as CV_i, $i = 1, 2, ..., 24$. The probability of error of this token is

$$P_e(\mathrm{CV}_i, \mathrm{SNR}) = \sum_{j \neq i} P\{\text{heard CV}_j \,|\, \text{spoken CV}_i\}, \quad (3)$$

where $P_{ii} = 1 - P_e$ is the corresponding probability of correct response (diagonal element). For simplicity in notation, we may refer to $P_e(\mathrm{CV}_i, \mathrm{SNR})$ as P_e. Given the above probability of error for each of the 24 tokens, the average error for each ear is then

$$\overline{P_e}(\mathrm{Ear}, \mathrm{SNR}) = \frac{1}{24} \sum_{i=1}^{24} P_e(\mathrm{CV}_i, \mathrm{SNR}). \quad (4)$$

Two other measures are considered. The *confusion pattern* (CP) for a given token is a plot of one row of the confusion matrix [i.e., $P_{\text{heard}|\text{spoken}}(SNR)$], as a function of SNR (Allen, 2005b). This measure shows how the token score and confusions depend on SNR. Finally, we evaluate the treatment using error rate changes (ΔP_e) as a function of token ($\mathrm{CV}_i, i = 1, 2, ..., 24$) and SNR. Throughout this analysis we refer to the change in the error between the flat (experiment 1) and treatment gain (experiment 2), defined as

$$\Delta P_e(\mathrm{CV}_i, \mathrm{SNR}) \equiv P_e^{\mathrm{Treat}}(\mathrm{CV}_i, \mathrm{SNR}) - P_e^{\mathrm{Flat}}(\mathrm{CV}_i, \mathrm{SNR}). \quad (5)$$

When $\Delta P_e < 0$ (decrease the error), the treatment is said to *improve* the token score, whereas when $\Delta P_e > 0$ (increase the error), the treatment is said to *degrade* the token. Additionally, for each individual token, we use the accumulated error differences $\Sigma\Delta P_e(\mathrm{CV}_i)$ defined as

$$\sum \Delta P_e(\mathrm{CV}_i) = \sum_{\mathrm{SNR}} \Delta P_e(\mathrm{CV}_i, \mathrm{SNR})^{\mathrm{improved}} + \sum_{\mathrm{SNR}} \Delta P_e(\mathrm{CV}_i, \mathrm{SNR})^{\mathrm{degraded}} \quad (6)$$

to identify the overall benefit each subject received per token from the treatment gain. When $\Sigma\Delta P_e$ is negative for a token, we say that treatment gain degraded that token (increased the error). Furthermore, one can plot $\Sigma\Delta P_e(\mathrm{CV}_i)$ as a function of tokens $\mathrm{CV}_i, i = 1, 2, ..., 24$ and look into the area under each curve. This area, denoted by parameter A later in the paper, is a useful summary metric to determine whether a subject received benefit or harm from the hearing aid amplification.

III. RESULTS

This section discusses how the treatment insertion gain impacts the phone token recognition scores as a function of SNR, using the metric of changes in error rates [Eqs. (5) and (6)]. This metric allows us to address the large variability between HI ears phone recognition, and observe the idiosyncratic token confusions across different subjects.

A. Improvements and degradations due to the treatment

Figure 3 shows the overall performance change between the two gain conditions. As described in the methods section, to assure that the results are statistically significant, each token was presented between 5 and 10 times at each SNR. The goal was to determine how the treatment gain improves ($\Delta P_e \leq 0$) or degrades ($\Delta P_e > 0$) the token responses for each ear.

As shown in Fig. 3, 57% of the tokens had zero error for both flat and treatment insertion gains. For the remaining 43% of cases, the treatment changed the error. The ideal case is *improvements*, when the error reduces ($\Delta P_e \leq 0$). *Degradations* are when the error increases ($\Delta P_e > 0$) due to the treatment. We give a high priority to degradations, since

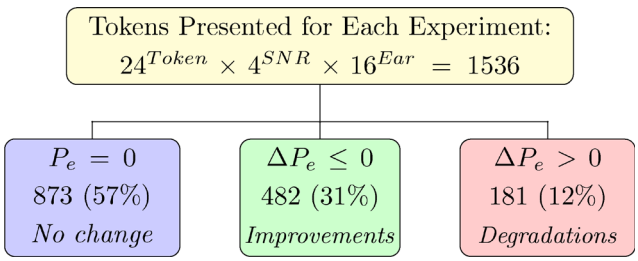

FIG. 3. (Color online) Number of *improvements* and *degradations* due to the treatment gain across the 24 tokens, 4 SNR conditions, and 16 ears (a total of 1536 presentations). Of these almost 2/3 (864 = 57%) had zero-error ($P_e = 0$) in both conditions. For the remaining 663 errorful tokens, nearly 1/3 (482 = 31%) *improved* (reduced error) and 1/8 (181 = 12%) *degraded* (increased error).

treatment increased the token error ($\Delta P_e > 0$). This happened 12% of the time.

Figure 4(a) illustrates the overall improvements vs degradations, broken down by consonants (not token), at the four different SNR conditions. The abscissa shows the 14 consonants, sorted by total number of improvements. The ordinate gives the number of tokens that either improved or degraded due to the treatment. The gray scale code (color online) separates the improvements and degradations by SNR. The sounds /vɑ/, /bɑ/, and /zɑ/ had the greatest improvements, while /vɑ/, /bɑ/ had the most degradations. Although we used robust tokens according to their SNR_{90} [see Table I and Phatak and Allen (2007)], the majority of tokens had some degradations, even at quiet. Recall that by design, NH subjects have no error in this task.

Figure 4(b) illustrates the improvements and degradations for each of the 16 ears, sorted by the total number of improvements. As a general rule, the treatment gain gave more improvement than degradation for every subject. Ear 44_{L} had the smallest error, and thus the least room for improvement. However, the number of degradations was nearly equal to the improvements, thus there was no significant net improvement. Ear 36_{L} had 10 improved tokens and only 2 degradations, thus received a net improvement from the treatment. Note how 30_{L} also gained a large 32 token improvement with only 5 degradations, thus 30_{L} received the largest net benefit from the treatment gain, slightly more than 34_{R} and 30_{R}.

There is a large variability in the improvements across both consonants and subjects. The degradations are also highly variable across token and ear, but they are always less than the improvements. In a few ears, degradations approach improvements (e.g., 32_{L}, 32_{R}, 44_{L}).

B. Error summary

Figure 5 compares the average probability of error $\overline{P_e}(Ear, \mathrm{SNR})$ for the 16 ears for the flat (left) and treatment (right) gains. On the left (flat-gain), the log-error $\overline{P_e}(Ear, \mathrm{SNR})$ is approximately linear as a function of SNR, with subject dependent slopes, consistent with the articulation index (Allen, 2005a; Toscano and Allen, 2014). This slope and intercept are key metrics of speech recognition loss for each ear.

Following the treatment, the log error [%] behavior collapses into two families of curves, such that the curves (subjects) at SNRs above 6 dB are divided into two groups. Half of the subjects do not strongly depend on SNR above 6 dB and (excluding subject 01) plateau around 10% error (*high-error group*). The other half continue to improve with SNR, reaching an error rate of less than 2% at quiet (*low-error group*). The high-error (10%) limit will be discussed in more detail in Sec. III C.

A detailed comparison of $\overline{P_e}(Ear, \mathrm{SNR})$ shows that as the SNR increased, both the low-error and high-error ears received significant benefits from changing the gain from flat to treatment. However, high-error ears still have significant average error, even at quiet (e.g., 10%). Increasing the SNR at the treatment gain reduces $\overline{P_e}(Ear, \mathrm{SNR})$ for these

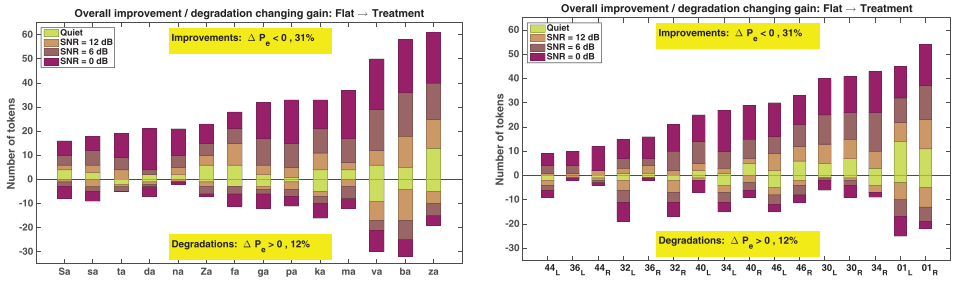

FIG. 4. (Color online) Overall number of token improvements ($\Delta P_e \leq 0$) and degradations ($\Delta P_e > 0$) due to the treatment. (a) Change in the number of errors as a function of the presented consonant + /ɑ/ (included all available tokens for such consonant). Note the usage of symbol "S" to identify consonant /ʃ/, symbol "Z" to identify consonant /ʒ/, and symbol "a" to identify vowel /ɑ/. (b) Change in the number of errors as a function of the individual ears. The positive direction indicates an improvement, while the negative direction indicates a degradation. The gray-scale code shows different SNRs (a darker shade represents higher noise). The highlighted percentage shows the net improvements (31%) and degradations (12%) across all presentations.

subjects down to the 10% *saturation limit* above SNR = 6 dB. Figure 5 highlights the utility of the logarithmic probability of error. On a linear scale, we would not see this grouping effect, nor would we see the 10% saturation limit.

While the average probability of error is a useful measure of overall performance, it hides the details of the immense variability in token recognition over ears, SNR and tokens. For example, observe that the left and right ears of subject 30 (●) become separated above SNR = 6 dB for the treatment gain, while they were similar for flat-gain. Another observation is for subject 32 (▼), who had 2% error at SNR = 12 dB and quiet for flat versus 7%–10% error for the treatment gain.

C. Individual differences emerging the idiosyncratic behavior

To understand what is going on, we need a more detailed analysis of the errors to determine the idiosyncratic contribution of individual tokens. Focusing on individual token errors, we further expand the improvement/degradations of Fig. 4 (a) to observe the error rate difference between improved and degraded tokens for each subject. Figure 6 illustrates the accumulated ΔP_e (total sum of ΔP_e) of individual tokens across SNRs for all subjects. The line in each panel indicates the overall difference between improved and degraded tokens' ΔP_e (i.e., $\Sigma\Delta P_e = \Sigma_{\langle \mathrm{SNR} \rangle} \Delta P_e^{\mathrm{improved}} + \Sigma_{\langle \mathrm{SNR} \rangle} \Delta P_e^{\mathrm{degraded}}$). The total area under this curve is shown at each panel as metric A. Subjects are sorted based on A, averaged across left and right ears. The value of A is a useful summary metric of the benefit the ear received from treatment gain. If A is negative (e.g., subject 32), the treatment is a net degradation. For subject 32, the speech recognition is degraded at 6, 12 dB SNR and quiet for both ears. A modest benefit was observed at 0 dB SNR.

Figure 6 shows that there is a large variability in degradations across subjects, tokens, and SNRs, and it is not trivial to generalize the error patterns for a given subject. To study the sources of degradations for each subject, we have examined the most overall degraded tokens at all SNRs (i.e., smallest $\Sigma\Delta P_e$ value) and generated the confusion patterns (CP) of the most degraded tokens. This way it becomes possible to observe the role of masking noise as well as the consistency of subjects for high error responses.

1. Summary confusion patterns

Figure 6 has isolated the cases where the treatment *increases* the error. According to Fig. 6, some subjects had significant degradations for a few tokens. To look deeper at these cases, Fig. 7 presents pairs of confusion patterns (CPs) for ten tokens in which the probability of correct response at quiet was reduced more than 20%, given the treatment. The CP is a plot of a row of the confusion matrix as a function of SNR. The row defines the target [see label at top left corner of flat-gain (right) CP; e.g., for subject 34_{L}, the target token is female /fɑ/ (F/fɑ/)], and the sounds confused with the target are displayed in terms of their response probability as a function of SNR. For each pair, the left CP corresponds to

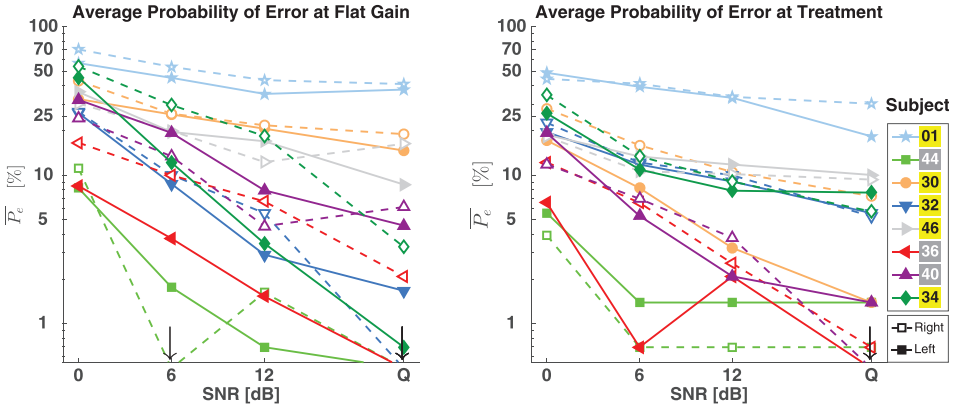

FIG. 5. (Color online) Average probability of error [$\overline{P_e}(Ear, \mathrm{SNR})$] for all ears in log-percent (%) versus SNR [dB] for the flat (left) and treatment (right) gain experiments. Each line represents an ear [solid lines (closed symbols): left ear, dashed lines (open symbols): right ear]. Wherever $\overline{P_e}(Ear, \mathrm{SNR})$ is zero, the curves are truncated at the bottom of the plot (shown by ↓).

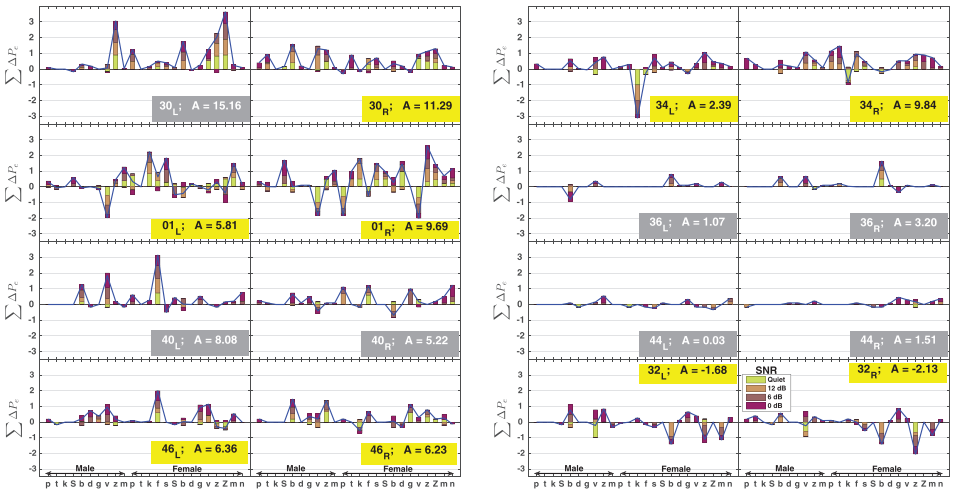

FIG. 6. (Color online) Accumulated error differences ($\Sigma\Delta P_e$) for each subject; the line shows the difference between improved and degraded tokens error. Abscissa shows the 24 male and female talker/consonants (the vowel /ɑ/ is omitted to save space, and consonants /ʃ,ʒ/ are shown as /S,Z/). (A) in each panel shows the area under the curve.

flat-gain, while the right CP corresponds to the treatment gain.

As shown in Fig. 5, the low-error ears responded to improved SNR, down to quiet, whereas the high-error ears have a saturated error around 10%, independent of SNR, above 6 dB. The ten tokens of Fig. 7 are from the high-error group. Thus the degradations and the high-error tokens are highly correlated.

Interesting cases in Fig. 7 are subjects 32_L (row 3, columns 3–4) and 32_R (row 4, columns 3–4) for target token male /vɑ/. For the flat-gain cases, the scores start out at 100% in the quiet condition for both ears. At 12 dB SNR, the scores drop to between 65%–75%, with the main confusions being /nɑ/, /mɑ/. At 0 dB the scores for the target /vɑ/ are near chance, and the number of confusions is around 4. In the left ear the main response is 60% /fɑ/ and in the right ear /vɑ/ and /mɑ/ are tied at 33% each. Given the treatment, however, the left ear is reporting /fɑ/ 100% and the right ear 60% at quiet. At 0 dB SNR the target is reported correctly for both ears at 60%. In the left ear at 12 dB, subject 32 reports the target correctly 100%. This is highly significant to report the correct response for all 10 trials, but in quiet then to switch to /fɑ/ 10 out of 10 trials. Subject 32 is very systematic, but struggling with the perception of /vɑ/, due to confusions in the cues, as a result of applying the treatment gain and added noise. This subject has similar interactions with audible cues of target /bɑ/ (row 3, columns 1–2) and /zɑ/ (row 4, columns 1–2).

2. An example of manipulation of the insertion gain

Given the specific confusions introduced by the treatment, along with a precise knowledge of the speech cues used by normal hearing listeners (Allen and Li, 2009), we

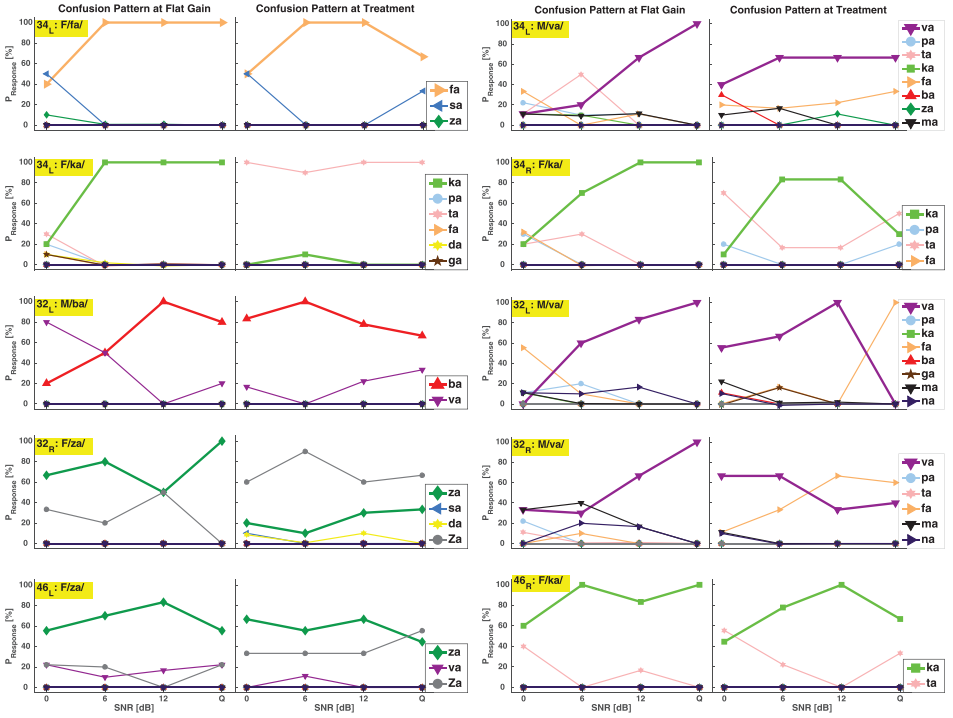

FIG. 7. (Color online) Comparison of confusion patterns (CP) for ten degraded tokens, that for the quiet condition, contribute the most to the net-degradation in summary metric A in Fig. 6. In each panel, the CP on the left is the flat-gain condition and the CP on the right is the treatment gain. The subject and target token are identified in the flat-gain CP, in the upper left corner (the first subject is 34_L and the first target token is female /fɑ/). By studying the nature of the confusions as a function of SNR it may be possible to understand what the subject's strategy was under that specific condition.

can begin to understand why some HI ears failed to benefit from the treatment. For example, in Fig. 7, subject 34_L in flat-gain (row 1, column 1) had perfect scores for /fɑ/ above 0 [dB], but the treatment (column 2) reduced the score in quiet to 60%. A second example is 32_L for target /vɑ/ (row 3, column 4), where the score in quiet dropped to zero (100% error). If we were to reduce the treatment gain in small steps, and remeasure the few errorful tokens, we would presumably find an optimum gain. The same strategy should apply to 34_R (row 2, column 4), and other similar cases. Of course, by changing the treatment gain, one may introduce some new error for tokens that previously had no error, but if that happens, it would provide even greater insight. The procedure could achieve an optimal gain rule that settles on the best treatment gain (lowest error rate among all test tokens). Use of the $\Sigma\Delta P_e$ analysis would greatly reduce the testing time for the fitting process, due to real-time monitoring of a relatively small number of token errors.

IV. DISCUSSION

A major complication of the HI phone recognition analysis is the idiosyncratic variability across tokens for each subject. Consonant confusions are complex functions of ear, token, SNR and treatment gain. The key question addressed here is how an audiologist can identify and prioritize tokens based on phone recognition experiments, to prescribe the best treatment gain. It seems likely that plasticity will play a role. By taking advantage of the small numbers of significant errors and idiosyncratic subject variability in consonant recognition scores, we can classify subjects based on a token speech metric such as $\Sigma\Delta P_e$, to ultimately improve the treatment gain protocol. An efficient tool could change the clinical landscape.

Clearly, too much treatment gain can be harmful for some tokens. We have tested in flat-gain at MCL, and found that audibility is not an issue. Note that audibility/detectability of a stimulus differs from correct recognition of that stimulus. For example, the token can be audible for the subject, but at the same time, parts of its spectrum, needed to identify the token, remain masked. Trevino and Allen (2013b) found large differences in error for the same consonant spoken by two different talkers. It is unlikely that audibility is the issue, if the two sounds are at the same RMS and SNR level, and one is clear and the other 100% error (all these sounds are recognizable to NH ears). Token dependent conflicting cues are likely the explanation of such anomalous cases (Li and Allen, 2011; Kapoor and Allen, 2012; Cole, 2017).

For the majority of subjects, more than half of the sounds have zero error, across all SNRs. About 1/3 of the corpus have large *systematic* errors (Fig. 4). Given the bimodal nature of the errors in Fig. 5 (right panel), averaging the few errors with the large number of zero-error tokens, distorts the statistics (Trevino and Allen, 2013b; Toscano and Allen, 2014). There are large variations in degradations at different SNRs among ears. It seems clear that based on the degradations, we need to tune the gain formula to minimize them. It is also likely that these degradations are sounds that could be learned given time and feedback.

When we investigated individual token errors, which define the improvement and degradations, as illustrated in Fig. 4, we found a large variability across individual consonants and subjects, making it difficult to draw conclusions about the impact of the treatment gain without well defined metrics.

Studying the average error of each ear at two different insertion gains, seemed to be a widely accepted measure to analyze benefits received from the treatment gain (Fig. 5). What is missing in average error analysis, is a way to parse out subjects by quantifying the degree of speech loss. It has been widely assumed, that the average error could be a reasonable measure of speech loss. However, we now know this assumption is leading to a weakened metric (score or average probability correct), since the average error can be poorly correlated with the error of a particular token (Kapoor and Allen, 2012; Trevino and Allen, 2013b; Han, 2011). To reduce the test time in clinic, those few tokens having large confusions, need to be identified. As shown in Fig. 6, only a small number of sounds have significant error, with a complex dependence to SNR. The exact nature of this dependence is shown in the confusion patterns of Fig. 7.

To address the large variability of Fig. 4, we expanded the improvement/degradations into the accumulated error differences for each token (Fig. 6). Given the token error distribution, we were able to directly compare the token recognition for flat and treatment gain, for all ears. For low-error ears, $\Sigma\Delta P_e$ either remains close to zero, for ears that had good performance at both flat and treatment gains (e.g., subjects 44, 36, and 40_R), or has large improvement spikes for some tokens (e.g., ears 30_L and 40_L). High-error ears had large errors at both flat and treatment gain, thus no specific conclusion can be made through net-improved tokens. However, audiologists can focus on largest degraded tokens for high-error ears when adjusting the hearing aid amplification. The process could continue until settling down on a gain rule that minimizes the total degradations.

The accumulated error differences ($\Sigma\Delta P_e$) shown in Fig. 6, informs us on how tokens responded to the treatment. The effectiveness of the treatment gain is quantified by the area under the curve. If the $\Sigma\Delta P_e$ area (A) is negative, it shows that the treatment is too aggressive.

Since we tested at subject's MCL gain at both flat and treatment experiments, the applied insertion gain (REG) adjusted to MCL contained some attenuation at low frequencies for some subjects (Fig. 2). The MCL adjustment was necessary since these subjects found the treatment gain too loud, and adjusted their MCL accordingly. One might assume that some of degradations happened because of this attenuation at low frequencies. But this seems unlikely since most of the hearing loss of subjects is above 2 kHz (see Fig. 1) and most of the errors are high frequency sounds such as /k,v,z,f/. While low frequency attenuation could be a reason for some degradations, additional scrutiny on degraded subject-token pairs prevents generalization of such an argument. Examples include subjects 36_L and 46_L who had highest low frequency attenuation in Fig. 2; while 36_L was one of

the lowest error subjects, 46_L was one of highest error subjects in both experiments.

According to Fig. 6, subject 46_L had the most degradations for female /zɑ/ and /ʒɑ/ tokens, and 36_L had the most degradations for male /bɑ/ token. These three sounds, have primary cues in low (/bɑ/) or high (/zɑ,ʒɑ/) frequencies. We can conclude that low frequency attenuation provided by REG-MCL could not be the primary factor of such degradations for ear 46_L, while in case of 36_L it can be a factor. However, token male /bɑ/ was found to be fragile even for NH listeners (Phatak and Allen, 2007).

About half of the subjects have some tokens with $\Sigma\Delta P_e < -1$. In such cases there is a clear need for optimizing the treatment gain. The small number of missed sounds, and the possibility that they are related to some internal distortion, is reminiscent of Plomp's distortion measure. If the phone error is zero at flat-gain, the ear has good audibility, without any frequency-dependent treatment. In such cases, the treatment should be focused on reducing the error on that small subset of high error tokens. If changing the treatment gain introduces new token errors, sorting tokens based on degraded $\Sigma\Delta P_e$ directs to the most significant token errors. Eventually, one must be obliged to compromise over the degree of error for a few tokens in order to reduce the overall error. At this fork, plasticity will likely play a role.

An asymmetry between left and right ears in Fig. 6 can be interesting. As an example, for subject 34, for most of the tokens the right ear received benefit from the treatment gain at 0, 6, and 12 dB SNR, while the left ear received significant benefit at 0 dB and slight benefit at 6 dB (see Fig. 6, first row, columns 3–4, and also Fig. 7, row 2). For both left and right ears at quiet, tokens are either gain-independent or show a degradation. The three parameter hearing loss profile for this subject (Table II) shows that left and right ears have a large difference in the hearing loss profile, yet this difference only shows up in $\Sigma\Delta P_e$ at 12 dB SNR.

Once the degraded consonants have been identified, confusion patterns become the tool of choice, since CPs provide insight as to why the Treatment led to large degradations. Sample CPs in Fig. 7 show that the error can become worse at quiet with treatment. It is also interesting to look at the confusions that are created by the treatment. Examples include /vɑ/ → /fɑ/ (voicing), /vɑ/ → /pɑ/ (voicing and affrication) /zɑ/ → /ʒɑ/ (place), and /kɑ/ → /tɑ/ (place).

Conflicting cues: Several relevant studies have noted how token dependent conflicting cues can produce non-monotonic errors as a function of SNR (Li and Allen, 2011; Kapoor and Allen, 2012). Examples of this are seen in the CP of subject 32_L (row 3, columns 3–4 of Fig. 7) at quiet, where /vɑ/ is reported correctly in flat-gain, but with treatment, subject reported /fɑ/ 100% of the time. Subject 34_R (row 2, columns 3–4) had a similar reversal for target /kɑ/, which with treatment were reported as /pɑ/ and /tɑ/ in quiet. Such non-monotonic errors were reported by (Kapoor and Allen, 2012) which were shown to be due to the masking of audible conflicting cues. Cole (2017) studied plosives and found out that HI ears use the same primary cues as NH ears do, and conflicting cues explain most of the confusions in HI ears.

V. SUMMARY

(1) While the treatment gain largely results in improvement, it dramatically fails for a small subset of ears and tokens. As summarized in Fig. 3, the treatment gain helped to improve the phone recognition of HI ears in 31% of presented tokens, while it led to degrading the phone recognition in 12%. The majority (57%) of tokens had zero-error for both flat and treatment gains, thus are independent of the gain treatment. These many tokens do not need to be considered by the hearing aid fitting procedure, as long as the modified gain falls between the flat and treatment gains. The question of what speech cues HI ears are listening for is the subject of continuing efforts in Human Speech Recognition group.

(2) Figure 4 shows that when we compare ears based on the improvements and degradations, we see a large variability across both consonants and subjects, making clear the importance of viewing each ear and token independently.

(3) SNR is a major factor in HI consonant confusions since the treatment gain is more effective at lower SNRs (0 dB) than higher SNRs (quiet) (Fig. 4). It is also possible that a token becomes more error-prone as the SNR is increased. In such cases, the masked conflicting cues become more audible, thus more confusions. Such confusions might be avoided over time, given appropriate feedback and training.

(4) As shown in the left panel of Fig. 5, the average error at quiet is nearly uniformly distributed on a log scale, from less than 1% to 50%. Following the treatment (right panel), the average error bifurcates into two groups having errors of 1% and 10% in quiet. The treatment separated the subjects above 6 dB SNR, where the noise has reduced masking. While the low-error ears receive more benefit from the treatment as the SNR increases, the high-error ears saturate at 10% error above 6 dB SNR. To explain this bifurcation, one must focus on errors at the token level.

(5) As shown in Fig. 6, the accumulated error differences ($\Sigma\Delta P_e$) between the two gain treatments gives a simple display of how treatment might help each individual HI ear. Illustration of $\Sigma\Delta P_e$ for tokens shows the large variability across subjects, tokens and SNR. This illustration quantifies the number of tokens that degraded the most, thus need less gain treatment. It also shows whether the left and right ears of the same subject received different benefit from the treatment. The area under $\Sigma\Delta P_e$ curve is a useful summary metric to measure the net benefit from the treatment gain.

(6) Identifying the target tokens for confusion patterns of Fig. 7 has been simplified by using the accumulated error difference analysis.

(7) Accumulated error differences of Fig. 6 along with confusion patterns of Fig. 7 constitute the way to evaluate CV speech recognition in HI ears for tuning a treatment, to minimize speech loss.

ACKNOWLEDGMENTS

The authors thank Phonak and Stefan Launer, who funded this study. Most importantly, thanks to Woojae Han for collecting the data, with technical support from Riya Singh. The data collection was funded by an NIH R21.

APPENDIX: FITTING AUDIOMETRIC MEASUREMENTS

The majority of the ears in this study have slight-to-moderate hearing loss with high-frequency sloping configurations. According to Trevino (2013), the audiometric configuration of low-frequency flat loss with high-frequency sloping loss can be modeled as a piece-wise linear function of the form

$$PTT(f) = \begin{cases} h_0, & f \le f_0, \\ h_0 + s_0(\log_2(f/f_0)), & f > f_0, \end{cases}$$

where $PTT(f)$ is the approximated hearing loss in dB and f is frequency in kHz. The parameter f_0 estimates the frequency at which the sloping loss begins; h_0 estimates the low-frequency ($f \le f_0$) flat loss in dB; s_0 estimates the slope of the high-frequency loss in dB/octave. The parameters are fit to minimize the root-mean-square (RMS) error ϵ, in dB. The RMS of a fitted curve is calculated as

$$\epsilon = \sqrt{\frac{1}{N}\sum_{f=f_1}^{f_N}(PTT(f) - HL(f))^2},$$

where $N = 10$ is the number of measured frequencies in hearing loss profile, $HL(f)$ is the measured hearing loss at frequency f, $PTT(f)$ is the approximated hearing loss at frequency f, and $f_1, ..., f_{10} = [0.125, 0.25, 0.5, 1, 1.5, 2, 3, 4, 6, 8\,\text{kHz}]$. The resulting parameters and RMS ϵ values for each model fit, as well as the MCL values for each ear at either flat and treatment gains are shown in Table II.

Abavisani, A., and Allen, J. B. (**2017**). "How to improve a hearing aid fitting based on idiosyncratic consonant errors," J. Acoust. Soc. Am. **141**, 3633.
Allen, J. B. (**1994**). "How do humans process and recognize speech?," IEEE Trans. Speech Audio **2**(4), 567–577.
Allen, J. B. (**2005a**). *Articulation and Intelligibility* (Morgan and Claypool, LaPorte, CO).
Allen, J. B. (**2005b**). "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**(4), 2212–2223.
Allen, J. B., and Abavisani, A. (**2016**). "Normal and impaired hearing recognition of speech segments in noise," J. Acoust. Soc. Am. **139**, 2188.
Allen, J. B., and Li, F. (**2009**). "Speech perception and cochlear signal processing," IEEE Sign. Process. Mag. **26**(4), 73–77.
Apoux, F., and Bacon, S. (**2004**). "Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise," J. Acoust. Soc. Am. **116**(3), 1671–1680.
Cole, C. L. (**2017**). "On the effects of masking of perceptual cues in hearing-impaired ears," Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.
Dillon, H. (**2001**). *Hearing Aids* (Thieme, New York), pp. 239–242.
Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (**2004**). "New nonsense syllables database—Analyses and preliminary ASR experiments," in *Proceedings of International Conference on Spoken-Language Processing (ICSLP)*, pp. 2749–2752.
Han, W. (**2011**). "Methods for robust characterization of consonant perception in hearing-impaired listeners," Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.
Kapoor, A., and Allen, J. B. (**2012**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am. **131**(1), 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**(4), 2599–2610.
Miller, G. A., and Nicely, P. E. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**(2), 338–352.
Phatak, S., and Allen, J. B. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**(4), 2312–2326.
Phatak, S. A., Yoon, Y., Gooler, D. M., and Allen, J. B. (**2009**). "Consonant loss profiles in hearing impaired listeners," J. Acoust. Soc. Am. **126**(5), 2683–2694.
Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Lang. Hear. Res. **29**(2), 146–154.
Plomp, R., and Mimpen, A. M. (**1979**). "Speech reception threshold for sentences as a function of age and noise level," J. Acoust. Soc. Am. **66**(5), 1333–1342.
Régnier, M. S., and Allen, J. B. (**2008**). "A method to identify noise-robust perceptual features: Application for consonant /t/," J. Acoust. Soc. Am. **123**(5), 2801–2814.
Rhebergen, K., Versfeld, N., and Dreschler, W. (**2006**). "Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise," J. Acoust. Soc. Am. **120**(6), 3988–3997.
Scheidiger, C., Allen, J. B., and Dau, T. (**2017**). "Assessing the efficacy of hearing-aid amplification using a phoneme test," J. Acoust. Soc. Am. **141**(3), 1739–1748.
Singh, R., and Allen, J. B. (**2012**). "The influence of stop consonants' perceptual features on the Articulation Index model," J. Acoust. Soc. Am. **131**(4), 3051–3068.
Starr, A., Picton, T. W., Sininger, Y., Hood, L. J., and Berlin, C. I. (**1996**). "Auditory neuropathy," Brain **119**(3), 741–753.
Steinberg, J. C., and Gardner, M. B. (**1940**). "On the auditory significance of the term hearing loss," J. Acoust. Soc. Am. **11**, 270–277.
Toscano, J., 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., and Allen, J. B. (**2013a**). "Individual variability of hearing impaired consonant perception," Semin. Hear. **34**(2), 74–85.
Trevino, A., and Allen, J. B. (**2013b**). "Within-consonant perceptual differences in the hearing impaired ear," J. Acoust. Soc. Am. **134**(1), 607–617.
Trevino, A. C. (**2013**). "Techniques for understanding hearing-impaired perception of consonant cues," Ph.D. thesis, University of Illinois at Urbana-Champaign, Urbana-Champaign, IL.
Valero, M. D., Hancock, K. E., and Liberman, M. C. (**2016**). "The middle ear muscle reflex in the diagnosis of cochlear neuropathy," Hear. Res. **332**, 29–38.
Vysochanskij, D. F., and Petunin, Y. I. (**1980**). "Justification of the 3 rule for unimodal distributions," Theor. Probab. Math. Stat. **21**, 25–36.
Wang, M. D., Reed, C. M., and Bilger, R. C. (**1978**). "A comparison of the effects of filtering and sensorineural hearing loss on patterns of consonant confusions," J. Speech Lang. Hear. Res. **21**(1), 5–36.
Wilson, R. H., McArdle, R. A., and Smith, S. L. (**2007**). "An evaluation of the BKB-SIN, HINT, QuickSIN, and WIN materials on listeners with normal hearing and listeners with hearing loss," J. Speech Lang. Hear. Res. **50**(4), 844–856.
Yoon, Y., Allen, J., and Gooler, D. (**2012**). "Relationship between consonant recognition in noise and hearing threshold," J. Speech Lang. Hear. Res. **55**, 460–473.
Zaar, J., and Dau, T. (**2015**). "Sources of variability in consonant perception of normal-hearing listeners," J. Acoust. Soc. Am. **138**(3), 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**(5), 1548–1559.