The influence of stop consonants' perceptual features on the Articulation Index model

Riya Singh[a)]
Mathworks, 3 Apple Hill Drive, Natick, Massachusetts 01760

Jont B. Allen
University of Illinois at Urbana-Champaign, 2061 Beckman Institute, MC-251, 405 North Mathews, Urbana, Illinois 61801

(Received 3 November 2010; revised 10 January 2012; accepted 12 January 2012)

Studies on consonant perception under noise conditions typically describe the average consonant error as exponential in the Articulation Index (AI). While this AI formula nicely fits the average error over all consonants, it does not fit the error for any consonant at the utterance level. This study analyzes the error patterns of six stop consonants /p, t, k, b, d, g/ with four vowels (/ɑ/, /ɛ/, /ɪ/, /ae/), at the individual consonant (i.e., utterance) level. The findings include that the utterance error is essentially zero for signal to noise ratios (SNRs) at least −2 dB, for >78% of the stop consonant utterances. For these utterances, the error is essentially a step function in the SNR at the utterance's detection threshold. This binary error dependence is consistent with the audibility of a single binary defining acoustic feature, having zero error above the feature's detection threshold. Also 11% of the sounds have high error, defined as ≥20% for SNRs greater than or equal to −2 dB. A grand average across many such sounds, having a natural distribution in thresholds, results in the error being exponential in the AI measure, as observed. A detailed analysis of the variance from the AI error is provided along with a Bernoulli-trials analysis of the statistical significance.
 [DOI: 10.1121/1.3682054]

PACS number(s): 43.71.An, 43.71.Gv, 43.72.Dv, 43.70.Mn [TD] Pages: 3051–3068

I. INTRODUCTION

The question *How do humans process and recognize speech?* (Allen, 1994) remains open because we do not yet understand the precise nature of the errors made by human listeners. This study directly addresses this question with a detailed look at human speech recognition (HSR) errors. In addition, we address two fundamental questions about the inner workings of Harvey Fletcher's 1921 Articulation Index (AI) theory (Allen, 1996): (a) why is the log-error (i.e., $P_e \equiv 1 - P_c$ on a log scale) linear in the AI and (b) what determines the minimum error (i.e., error when the AI = 1).

About the same time that Fletcher's 50 year revolution of speech telephone research at The Bell Telephone Laboratories was winding down, Claude Shannon began a second revolution with his *Theory of Communication* (Shannon, 1948). Shannon's key addition was his *source-channel model* of communication, which included the *confusion matrix* and *mutual entropy* to characterize the transmission of information, as described and used, for example, by Miller and Nicely (1955). George Miller's many classic studies of speech and its confusions are widely recognized as fundamental as they were the first to apply Shannon's source-channel model to speech perception. However, Shannon's very general theory did not lead to new insights into the nature of acoustic speech features (i.e., the nature of the speech code). Here we show how this connection may now be made. The analysis and results described here support binary perceptual cues, and leads to insights into the inner workings of the AI.

To understand the HSR code, and explain speech's natural robustness, as measured by the score as a function of the signal to noise ratio (SNR), we must account for the variability due to talker, accent, masking noise, listener, etc. Synthetic speech cannot be used to characterize natural variations in speech because by design, it does not have the natural variations of human speech. Furthermore in early experiments, synthetic speech was typically of very low quality, frequently leading to ambiguous, or at least complex, research conclusions. One can only identify and characterize features by inducing errors by the use of noise on natural speech, produced by large numbers of talkers, as recorded by a large number of listeners (i.e., trials), at many SNRs, analyzed at the utterance level.

This study is about the natural variability of speech and its impact on consonant perception errors. To understand the natural robustness of human speech, we have chosen to retain the natural variability of speech (thus its features), as produced naturally by the vocal apparatus, by a large numbers of talkers and listeners. Unlike previous studies, we do not average across utterances (i.e., talkers). Fortunately normal hearing listeners are similar, making it feasible (given some care) to average across the listener dimension, thus raising the number of trials per condition, giving increased analysis power.

A. Source-channel theories of HSR

1. Syllable errors

The first studies to characterize the information-bearing frequency dependent regions of speech, using real speech,

a)Author to whom correspondence should be addressed. Electronic mail: riyasingh87@gmail.com

with large numbers of listeners, began with the 1910 telephone research of George Campbell, followed by the life-long work of Harvey Fletcher, who in 1921 created the *AI model* of speech perception (Allen, 1994, 1996). Fletcher modeled maximum entropy (nonsense), consonant (C) vowel (V), VC, and CVC syllable recognition in terms of the average nonsense phone recognition score. On the basis of psychoacoustic experiments with many thousands of trials, Fletcher and his colleagues defined the *average nonsense phone articulation score s* for CVC syllables as $S_3 \equiv c^2v = s^3$, where c and v are consonant and vowel articulation scores. Likewise, average *CV* and *VC* syllable scores were accurately modeled as $S_2 \equiv cv \approx s^2$. The details of *Fletcher's methods of recognition*, with the precise definition of s, are documented in Allen (1994, 1996).

2. The AI model of average speech errors

Following the success of the *average phone score model*, Fletcher extended his syllable analysis to account for the effects of filtering the speech into bands (Allen, 1996). This method later became known as the *AI model*, which in 1969 became the ANSI AI standard (ANSI, 1969), loosely based on the French and Steinberg (1947) version of the AI.

The full-band speech error e is divided into $K = 20$ error bands,

$$e \equiv 1 - s = e_1 e_2 \cdots e_K, \tag{1}$$

where $e = 1 - s$ is the model average full-band phone error, s is the model full-band average articulation (i.e., the score for maximum entropy speech), and e_k is the error defined by the kth band. The total articulation error is the product of the band articulation errors over the K bands. Thus, the band errors are modeled as independent. Although the value of $K = 20$ was chosen empirically, it was later shown that each of these 20 articulation bands corresponds to approximately 1 mm along the basilar membrane [between the 0.2 and 7.5 kHz place (Allen, 1996)] defining the articulation density per critical band (also known as the *band importance function*), which was found to be constant in Fletcher's theory (Allen, 1994, 1996).

The multiband product rule [Eq. (1)] is also known as *the additive law of frequency integration* (it is additive in the exponent, as discussed in the following text) and is the foundation of the ANSI standard for the Speech Intelligibility Index (SII) (ANSI, 1997). This rule works not only for the average nonsense syllable score, but also fits the individual scores for more than half of the Miller-Nicely consonants, namely /p, k, f, ʃ, b, d, g, z, m, n/, as shown in several studies (Allen, 1994, 2005a; Phatak and Allen, 2007; Phatak *et al.*, 2008; Li and Allen, 2009).

Based on this assumption of independent articulation bands, French and Steinberg (1947) devised an empirical method to calculate the band error e_k based on the average critical-band speech to noise ratio (in dB) (Allen, 1994). They extended Fletcher's original formulation by providing a formula for relating the band error to the normalized critical band signal-to-noise ratio for that band (SNR_k dB). The band SNRs lead to band errors [Fletcher, 1950, Eqs. (1) and (5)], and thus the total error (normalized by $P_{chance} = 15/16$), and is

$$e = e_1 \cdot e_2 \cdots e_K = e_{\min}^{(1/K)\sum_k SNR_k} = e_{\min}^{AI}, \tag{2}$$

where $e_{\min}$ is defined as the minimum error under ideal conditions (when AI = 1) with $AI = 1/K \sum_{k=1}^{K} SNR_k$. The full details of computing the normalized SNR in each band (i.e., SNR_k) are provided in French and Steinberg (1947), Allen (1994, 2005a), Phatak and Allen (2007), and Phatak *et al.* (2008). From Eq. (2) we see that $e_{\min}$ is a key parameter of the AI model.

Several variations of the AI model are used to predict hearing-impaired speech perception (Dubno *et al.*, 1989; Pavlovic *et al.*, 1986; Humes *et al.*, 1986; Ching *et al.*, 1998) to characterize SNR-loss (Killion and Christensen, 1998) and for hearing-aid fitting (Rankovic, 1991). While it is widely recognized that the AI model characterizes the *average* score, little is known as to why and how it works. When isolated bands are removed, AI model predictions fail (Kryter, 1962). While this is a key question, it will only be addressed here qualitatively (Li and Allen, 2011; Kapoor and Allen, 2012). It is notable that there are no models that predict specific consonant confusions or that successfully address the large variance of the AI prediction, e.g., due to consonant and vowel dependence (Allen, 2005a,b; Phatak and Allen, 2007; Phatak *et al.*, 2008).

3. Capacity and error

Allen (2004) likened the AI model to Shannon's (Shannon, 1948) concept of *channel capacity* and suggested this similarity is a fundamental information-theoretical basis for the empirical success of the AI theory. According to Shannon's *channel-capacity theorem*, the error goes to zero while operating below capacity (he proved there is a loss-less transmission of information, but the coding can take an infinite amount of time). From a theoretical perspective, it is interesting to know if speech is operating below channel capacity. We show that under very specific conditions that speech has zero error transmission, consistent with the conclusion that human speech communication operates below the channel capacity.

B. Aims of this study

This study is a reformulated analysis of Phatak and Allen (2007) (aka, PA07), which used a database having a large number of talkers (14) and listeners (25). The aim of PA07 was to characterize consonant and vowel confusions in speech-weighted noise (SWN). For this purpose, PA07 selected "low error utterances" (CVs with less than 20% error in quiet) and the top 10 "high-performing" listeners. High error sounds were removed so that the impact of noise on the low-error consonants could be quantified. In the present study, we reanalyzed the data from PA07. This new analysis includes *all* the errors. We form a per-utterance analysis (i.e., we do not average over utterances) of the errors made in

"low-noise" (defined here as SNRs ≥ -2). We show that a large fraction of these utterances are essentially zero error and have a step function in the error, going from zero to chance, over a 6 dB change in the SNR, at an SNR that is utterance dependent. This is consistent with binary speech features and speech operating below channel capacity.

There are two driving motivations for this study. The main aim for probing in such detail is to analyze, and thus explain, the nature of the idiosyncratic (heterogeneous) errors. We show that for a large percentage of utterances for SNRs ≥ -2 dB, the error is essentially zero. Previous studies report a base error (in quiet) of 1–2% (Fletcher, 1929; French and Steinberg, 1947; Miller and Nicely, 1955; Allen, 2005a; Phatak and Allen, 2007).

Our second motivation is to understand speech loss in hearing impaired ears. To reach this goal requires a much better understanding in normal hearing ears. Ears having even a slight *hearing loss* (HL) experience significant and systematic consonant errors on these very same zero-error sounds. In our experience, any two ears having the same hearing loss, as characterized in terms of the *pure tone average* (PTA) or *speech reception thresholds* (SRT), never have similar errors (Phatak *et al.*, 2009; Yoon *et al.*, 2012; Han, 2011). Our several studies of consonant errors, in both normal hearing and hearing impaired ears, show that average scores fundamentally mischaracterize this idiosyncratic consonant speech loss (Phatak *et al.*, 2009; Han, 2011). This observation leads to many difficult yet important questions, such as: Why are /pa/'s from some of the talkers confused with /ta/, while others are rarely confused and why are certain consonant utterances more robust to masking noise than others. These questions have also been addressed in recent publications (Allen and Li, 2009; Li and Allen, 2011; Kapoor and Allen, 2012).

Key questions that remain unanswered are:

(1) What is the source of speech errors as a function of SNR [i.e., Eq. (1)]?
(2) Why does the AI model [Eq. (2)] fit so well for certain specific classes of nonsense syllables?
(3) What is the nature of speech errors humans make in small amounts of noise, i.e., what determines $e_{\min}$ [Eq. (2)]?
(4) Is the error zero above some threshold, as suggested by Shannon's channel capacity theorem, or does it go exponentially to a constant, as found by Fletcher's AI model?
(5) What is the magnitude (and source) of the variance from the average error?

This study will empirically address these five questions by reanalyzing the database of 25 normal hearing subjects responding to nonsense Miller-Nicely CV syllables (PA07), at various levels of speech-weighted noise, at the utterance level (no averages over consonants).

II. METHODS

As stated in Sec. I, the data to be analyzed include *all* the utterances and listeners of PA07.

As explained in Appendix B, the search for cues in speech has historically been limited by using:

(1) Artificial speech
(2) No masking noise
(3) A small number of talkers or listeners (the natural variability is not captured)
(4) High context (meaningful) sounds (subjects report what they understand rather than what they hear)
(5) Conditions that are inappropriately averaged together.

In our studies, we have carefully avoided these five conditions in our experimental design.

A. Stimuli

The experimental corpus is the same as that reported by Phatak and Allen (2007) and is called MN64 [MN because it is based on the classic Miller and Nicely experiment (Miller and Nicely, 1955), and 64 because the database has 16C × 4V]. MN64 used a subset of isolated CV sounds from the LDC2005S22 corpus (Fousek *et al.*, 2004), recorded by the Linguistic Data Consortium (University of Pennsylvania), as the speech database. This subset had 14 talkers speaking CVs composed of one of the 16 Miller-Nicely (Miller and Nicely, 1955) consonants (/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, / ð/, /z/, /ʒ/, /m/, /n/), followed by one of the four vowels (/ɑ/, /ɛ/, /ɪ/, /æ/). These vowels were chosen because they have similar formant frequencies, so as to make them more confusable. In the figures and tables, these vowels are referred to using the Darpabet symbols /a/, /e/, /I/, and /@/, respectively, due to the lack of IPA symbols in MATLAB, the software used to analyze the data and make the charts.

All talkers were native speakers of English. Ten talkers spoke all 64 CVs, while each of the remaining eight talkers spoke different subsets of 32 CVs, such that each CV in MN64 was spoken by 14 talkers. Thus the experiment had 56 (14 talkers × 4 vowels) utterances of each CV at each SNR. In the current study, we analyze the stop consonants (/p/, /t/, /k/, /b/, /d/, /g/).

For the experiment, the wideband noise RMS level was adjusted according to the RMS level of the CV sound to be presented to achieve the required SNR. While calculating the RMS level of a CV utterance, the onset and offset samples more than 40 dB below the largest sample (in magnitude) were removed (Phatak and Allen, 2007).

B. Testing paradigm

The full test procedures, described in Phatak and Allen (2007), are summarized here. The listeners were asked to identify the C and the V in the presented CV syllable by selecting one of 64 software buttons on a computer screen, arranged in a 16 × 4 grid. The isolated speech sounds were played at six SNRs (−22, −20, −16, −10, and −2 dB) and *Q* (quiet), in SWN (French and Steinberg, 1947), the spectrum of which is described in Phatak and Allen (2007). A "noise only" button was provided for when the participant heard only noise without hearing any speech sound; when scoring for the consonant, such responses were treated as chance errors and distributed uniformly among the 16 possible

responses ($P_{chance} = 15/16$). Based on the total number of trials of the stop consonants across the 25 listeners, the percentage of "noise only" responses was 0.03%, 0.03%, 0.15%, 4.4%, 29.2%, and 46.8%, respectively, for Q, −2, −10, −16, −20, and −22 dB SNR. Thus, this button was rarely used (0.03%) at Q and −2 dB. Listeners heard the stimuli binaurally via headphones (Senheiser, HD-265) at his/her most comfortable level (MCL). The listener was allowed to replay the CV sound as many times as desired before entering their response. Such repetition helped to improve the scores by eliminating the unlikely choices in the large 64-choice closed-set task and by allowing the listener to recover from common distractions during the long experiment. After the response button was clicked, the next sound was played after a short pause. The presentation of each CV sound was randomized over consonants, vowels, talkers, and SNRs. The total of 5376 presentations (16C × 4V × 14 talkers × 6 SNRs) were randomized and split into 42 tests, each with 128 sounds. Each listener was trained on the stimulus set using one or two practice tests with randomly selected sounds, presented in quiet, with visual feedback on the correct choice.

Each utterance was presented only once to a listener at each SNR, excluding the practice sessions. Because 14 listeners completed the task, the number of times a particular utterance was presented at a given SNR was at least 14. A few listeners did a few sessions more than once, thus they may have heard a subset of sounds more than once per SNR. On average, about 18–19 listeners heard a particular utterance (because the presentations are totally randomized, every listener who did not complete the task missed hearing a random set of utterances).

As reported by Phatak and Allen (2007), there is no systematic difference between scores ≥−2 dB SNR (in speech-weighted noise) for ≈80% of these six stop consonants (we will further support this observation in the analysis given in the following text). Thus the data from these two conditions are pooled, and SNR ≥ −2 is defined as the *low-noise environment*. Due to various factors, the number of times (N) a particular utterance was heard in the low-noise environment was utterance dependent but was on average ≈38 (±2). The actual value of N for each utterance is tabulated along with the utterance errors in Sec. III.

C. Listeners

In total, there were 25 normal hearing listeners with English as their first language (12 M and 13 F) having no known history of hearing impairment. As reported in the PA07 study, 14 listeners completed all the 42 sessions (5376 CV tokens). Of the remaining 11 listeners, 3 repeated a session, resulting in 5376 + 128 = 5504 responses. The remaining eight (11 − 3) listeners completed less than 42 test sessions (the minimum being 4 and the maximum being 23). The average number of trials per CV per SNR is about 1060. Ideally, it would have been 25 × 56 = 1400. Because there are 56 CV utterances, ≈19 = 1060/56 listeners heard a particular utterance at each SNR on average. As discussed in the appendix, this gives a significant number of trials per condition, providing the needed statistical power.

D. Analysis criteria and terminology

In this section, the terminology used in the study is explained and the error criteria, along with the rationale behind classifying the errors into groups, are discussed. Finally the normalized entropy, which extends the group error classification scheme, is defined.

1. Groups

Figure 1 proposes a grouping scheme for the case of consonant /p/, and Table I gives the details for the non-zero-error (NZE) sounds.

(1) An *utterance* is a single CV spoken by an individual. They are indicated as in f101pa, where f101 means female subject 101 speaking /pa/. A *per-utterance* analysis means at the utterance level.
(2) P_e is the *empirical error* (% units) at the utterance level
(3) $\mathcal{H}_{\mathcal{N}}$ is the *normalized entropy* of an utterance as defined in the next section. It is a robust measure of the relative randomness of the utterance confusions.
(4) There are 56 (14 talkers × 4 vowels) utterances for each consonant. The *low-noise environment* is defined as the SNR condition above −10 dB, i.e., −2 dB SNR and quiet. For 80% of all the utterances, there is no substantial difference between these two conditions [41 /p/ sounds have zero error ($P_e = 0$), and 11 more have a single error ($P_e < 3\%$), thus have a normalized entropy

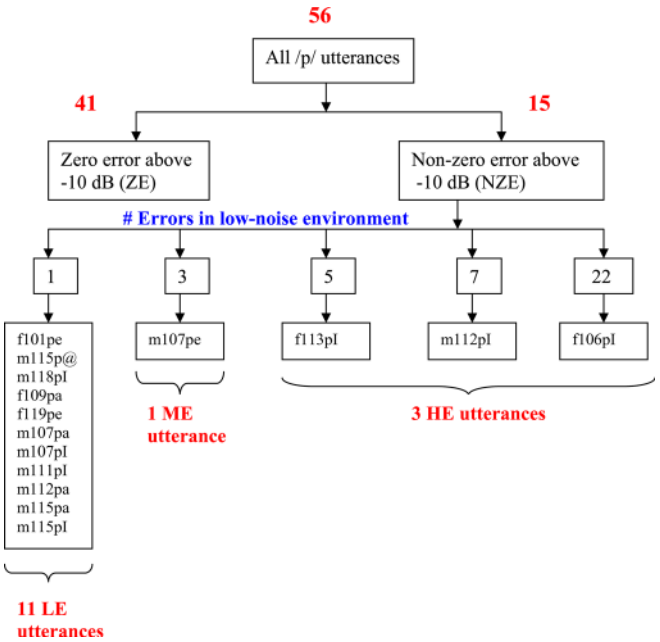


FIG. 1. (Color online) Error distribution of 56 /p/ utterances in the low-noise ($SNR \geq -2$ dB) environment: The total number of utterances as marked above the topmost block is 56 (14 consonants with 4 vowels). The zero-error (ZE) group is the leftmost and contains 41 of the 56 utterances as marked above the block. The number above a block gives the size of the group, i.e., number of utterances of 56 that belong to that group. Of the remaining 15 (56-41) utterances, the next level shows the number of errors made in the low-noise environment. From the figure, 11 utterances have 1 error (of 38 trials on average), forming the low error (LE) group. Four utterances (m107pe, f113pI, m112pI, and f106pI) have 3, 5, 7, and 22 errors, respectively. The first utterance (m107pe) belongs to the medium error group (ME), and the last three have an error greater than 12% (Table I), thus belong to the high error (HE) group.

TABLE I. Percentage error, N and SNR_{90} values for the 15 NZE utterances of /p/, shown in Fig. 1. The table is divided into three groups with horizontal lines. The top 11 utterances have exactly 1 error (<3%) thus $\mathcal{H}_{\mathcal{N}} = 1$ so we interpret these errors as random. The last three utterances (f113pI, m112pI, and f106pI) having more than 12% error thus belong to the high error (HE) group. Utterance m107pe is a lone member of the medium error (ME) group. The SNR_{90} (the SNR at which the score drops from 100% to 90%) is highly correlated with the acoustic feature threshold [Fig. 6a from Régnier and Allen (2008)] and is taken as an objective measure of the robustness of the sound. As seen from the tabulated values, ME and HE utterances have high ($\geq$ 2 dB) $SNR90$ thresholds. Thus they are easily confusable, even in the low-noise environment. In particular, f106pI has more than 50% error even in quiet, thus its SNR_{90} value is ∞. LE utterances have low values for $SNR90$ (< 2 dB) thus are *robust*. Therefore they should ideally be classified as in the zero error (ZE) group.

Utterance	P_e (%)	N	SNR_{90}
f101pe	2.70	37	−16
m115p@	2.78	36	−14
m118pI	2.78	36	−16
f109pa	2.78	36	−3
f119pe	2.56	39	−3
m107pa	2.70	37	−3
m107pI	2.70	37	−12
m111pI	2.86	35	−12
m112pa	2.56	39	−4
m115pa	2.70	37	−12
m115pI	2.70	37	−5
m107pe	7.69	39	5
f113pI	13.89	36	10
m112pI	18.92	37	15
f106pI	56.41	39	∞

of 1], hence the data are averaged across −2 dB SNR and quiet, to increase the utterance sample size N, as given in Table I.

(5) N is the total number of presentations of each utterance in the low-noise environment. Because, at a given SNR each listener hears the sound only once, N is equal to the number of subjects who heard the CV at −2 dB SNR and in quiet. The average value of N is ≈38. Given the average error P_e and N trials, one may calculate the variance of the mean μ (assuming *independent and identically distributed* (iid) Bernoulli trials) as $\sigma_\mu = \sqrt{P_e(1-P_e)/N}$ (see appendix).

(6) On the basis of errors made in the low-noise environment, the 56 utterances are divided into two groups: the zero error group (ZE), which contains *hits* (true-positive utterances) that have zero errors in the low-noise environment, and the NZE group, having at least one error in the low-noise environment (Fig. 1).

(7) Based on our error analysis, the NZE group is divided into three groups: low error (LE), medium error (ME), and high error (HE) (Fig. 1).

(8) The LE group contains utterances with $P_e \leq 3\%$ (i.e., a single error in $N \approx 38$ trials) in the low-noise environment. Observe that these *false-negative* errors are uncorrelated across listeners and vowels, hence appear random ($\mathcal{H}_{\mathcal{N}} = 1$). Later, we shall show that this is not precisely true as the LE rate depends somewhat on the difficulty of the task.

(9) The HE group contains utterances with $P_e \geq 12\%$ (i.e., more than 4 errors of 38 trials) in the low-noise environment with low entropy, which we denote *true errors* (i.e., true-negatives). We anticipate (and demonstrate) that these errors are due to poor articulation by the talker, explained by conflicting cues and timing errors. We show the HE utterances form a low-entropy confusion group, consistent with our view.

(10) The ME group contains the remaining utterances having $3\% < P_e < 12\%$. It is difficult to come to a precise conclusion about these utterances because there are so few of them, thus we will not analyze them further. A proper statistical analysis would require much more data. One would assume that the errors forming this group are due to a combination of many factors, such as random errors, listener biases, misarticulated utterances, and of course effects of noise.

(11) The utterances in the ME group are called "ambiguous" because these are due to poor articulation by the talker, easily identified by most listeners to be confusable within a low entropy (small) group. This group of sound is easily *primed*. The term *consonant-priming* is used as a test of the natural ambiguity of a phone as discussed in the text.

(12) The ZE and the LE group together define the *robust zero error* (RZE) group. The utterances in this group are called "robust" because they either have no errors or a single random error, inherent to any experiment using human subjects.

In summary, if $|\mathcal{G}|$ denotes the cardinality of a group $\mathcal{G}$, then $56 = |ZE| + |NZE|$, $|NZE| = |LE| + |ME| + |HE|$, the *robust sound cardinality* $= |RZE| = |ZE| + |LE|$ (hits + misses) and the *ambiguous sounds cardinality* $= |HE|$.

2. Normalized entropy $\mathcal{H}_{\mathcal{N}}$

A second useful tool to characterize consonant (or syllable) confusions is the normalized entropy $\mathcal{H}_{\mathcal{N}}$, defined as the *consonant entropy* $\mathcal{H}_s$ divided by the *maximum entropy* $\mathcal{H}_{\mathcal{M}}$, for a given error. When computing $\mathcal{H}_{\mathcal{M}}$, one spreads the errors uniformly over all possible alternatives and then computes the entropy. The consonant entropy (in bits) measures the average size of a confusion group, while the maximum entropy measures the maximum possible size of the confusion space. Thus $\mathcal{H}_{\mathcal{N}}$ is a useful measure of randomness of the error and is between 0 (ordered) and 1 (maximally random).

This measure is best illustrated by example. Because f101pe (spoken consonant s =/p/) has exactly 1 error of 37 presentations, there are two possible outcomes having probabilities $P_c = [36/37,\ 1/37]$, and 14 zero outcomes. In this case both $\mathcal{H}_s$ and $\mathcal{H}_{\mathcal{M}}$ are identical

$$\mathcal{H}_s \equiv -\sum_{h=1}^{16} P_c \log_2 P_c = -\left[\frac{36}{37}\log_2\left(\frac{36}{37}\right) + \frac{1}{37}\log_2\left(\frac{1}{37}\right)\right] = 0.1793,\ \text{bits},$$

resulting in $\mathcal{H}_{\mathcal{N}} = 1$. This is an important special case as it applies to the LE group.

As a second example, assume two identical errors, giving

$$\mathcal{H}_s = -\left[\frac{35}{37}\log_2\left(\frac{35}{37}\right) + \frac{2}{37}\log_2\left(\frac{2}{37}\right)\right] = 0.3034\,\text{bits},$$

while the maximum entropy is

$$\mathcal{H}_\mathcal{M} = -\left[\frac{35}{37}\log_2\left(\frac{35}{37}\right) + 2\cdot\frac{1}{37}\log_2\left(\frac{1}{37}\right)\right] = 0.3574\,\text{bits},$$

thus $\mathcal{H}_\mathcal{N} = 0.8489$. Note how we spread the two errors maximally over different outcome "bins."

As a third and final example, assuming there are 20 errors of 37 presentations, the probabilities used to compute $\mathcal{H}_\mathcal{M}$ are more difficult. The correct number is 17 responses, leaving 20 errors to be maximally spread out over the remaining 15 bins. Thus 5 bins would get 2 errors and the remaining 10, 1 each. Thus

$$\mathcal{H}_\mathcal{M} = -\left[\frac{17}{37}\log_2\left(\frac{17}{37}\right) + 5\frac{2}{37}\log_2\left(\frac{2}{37}\right) + 10\frac{1}{37}\log_2\left(\frac{1}{37}\right)\right] = 3.0612\,\text{bits}.$$

Assuming the 20 errors are identical

$$\mathcal{H}_\mathcal{S} = -\left[\frac{17}{37}\log_2\left(\frac{17}{37}\right) + \frac{20}{37}\log_2\left(\frac{20}{37}\right)\right] = 0.9953\,\text{bits}$$

thus $\mathcal{H}_\mathcal{N} = 0.3251$.

When a subject selects a sound from a two-group $\mathcal{H}_s = 1$ bit or from a three-group $\mathcal{H}_s = 1.5$ bits. The maximum entropy for 16 consonants is $\mathcal{H}_\mathcal{M} = 4$ bits. Thus for a two-group $\mathcal{H}_\mathcal{N} = 1/4$ and 3/8 for a three-group.

3. Terminology

As demonstrated by Li and Allen (2011), natural plosive and fricative consonants contain *conflicting cues*, defined as significant energy at frequency regions representative of non-target stop consonants. These conflicting cues explain most of the high error confusions. There are frequent examples in the LDC corpus, where the talker poorly pronounces the target utterance. As a result, for these utterances, the main perceptual feature (denoted *event*) is not robust, and a conflicting cue dominates, even at low levels of noise. In addition to conflicting cues, a small number of unvoiced stop consonant utterances have timing problems, where the cues (e.g., bursts) are closer than average to the start of the vowel, making the utterances susceptible to confusion with their voiced counterparts. Because of these misarticulations, a sound may not be robust, making it inherently ambiguous even at low-noise levels. These few high-error cases seem to be due to a variety of different sources.

The term *talker misarticulation* implies that the talker poorly articulated an utterance so that its perceptual feature is not robust, and consequently the scores are medium to high (i.e., well above chance). When an utterance is wrong 100% of the time and has a fixed and consistent error (small entropy), it is called *mislabeled* (a talker error). The term *consonant-priming* implies ambiguous situations, where a listener is forced to randomly guess between a small set of confusable sounds (as in Bernoulli trials). The term consonant-priming is not to be confused with an implicit memory effect, a definition widely used in psychophysics. Priming is defined here as the situation where one may mentally select one of several consonant as heard by making a conscious choice between several possibilities from a small set (Régnier and Allen, 2008). In typical priming situations, normal hearing subjects guess among a group of two or three confusable sounds (Allen, 2005a). As shown in several examples in the following sections, primable sounds are easily identified on the basis of their high error in low-noise conditions and low entropy (tight distributions of confusions), which can be explained by conflicting cues in the AI-gram (the AI-gram is a critical band spectrogram, normalized to the noise floor). As demonstrated by Régnier and Allen (2008) and Li *et al.* (2010) and exploited by Kapoor and Allen (2012), given a CV speech cues, one can calculate the thresholds of the primary and conflicting cues of an utterance from the AI-gram and can then reliably predict the SNR at which the utterance will be at a confusion boundary, thus perceptually ambiguous. Such is the power of precise knowledge of speech cues.

E. The AI model predictions

According to *the AI model* of speech perception, the *average* sound articulation error is given by Eq. (2). Hence, empirically the average error is an exponential function of the AI. For speech-weighted noise (MN64), the AI is proportional to the SNR (Allen, 2005a). This average is typically formed over consonants, vowels, talkers, and listeners. In the previous study on MN64 (Phatak and Allen, 2007), it was shown that the AI model fits the average error for three subsets of consonants: a low-scoring (high-error) set C1: (/f/, /θ/, /v/, /ð/, /b/, /m/), a high-scoring (low-error) set C2: (/t/, /s/, /z/, /ʃ/, /ʒ/), and set C3: (/n/, /p/, /g/, /k/, /d/) with intermediate scores. The respective $e_{\min}$'s for these three groups are 0.01 (1%), 2×10^{-5}, and 3×10^{-5}. Identifying these three subgroups accounts for a significant portion of the variance of Eq. (2).

Further shown in PA07, the AI model also works for 12 of 16 consonants using a refined expression for AI, further reducing the variance. For example, as shown in Fig. 3(d), the average /p/ error fits this form via linear regression on the log-error, giving

$$P_e(AI) = 0.035^{AI},$$

with $AI \equiv (SNR+21)/19$ (the AI = 1 for $SNR \geq -2$, and 0 for $SNR < -21$). The RMS error of this fit is 0.75%. because we do not know the actual SNR in the quiet condition, we cannot extend the total error of this model to Q.

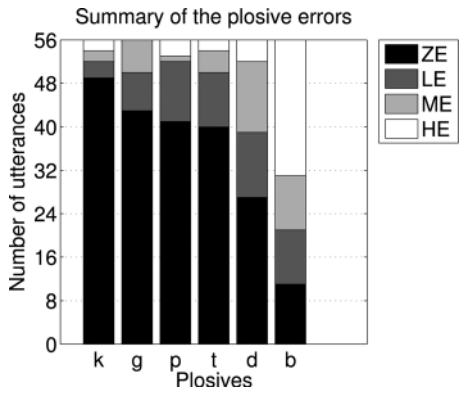


FIG. 2. Stacked bar-plots give the relative errors made by the six stop consonants in speech-weighted noise in the low-noise environment. The abscissa shows the six consonants, arranged in order of decreasing number of utterances in the ZE group (order of decreasing salience). The ordinate indicates the number of utterances of the consonant that falls into the ZE, LE, ME, and HE groups, respectively. The total is always 56. ZE is the zero-error group that contains utterances that all listeners gave correct responses at −2 dB SNR and quiet. LE is the low error group having low-grade random errors. ME is the medium error group with utterances having between 3 and 12% error. HE group utterances have errors greater than 12% and are primarily due to production errors. These are always ambiguous/primable utterances with high errors and low entropy. ZE and LE groups together form the robust zero error (RZE) group.

III. RESULTS

A. The analysis of individual utterance errors

The overall results of the grouping of errors across all the consonants, according to the method in Fig. 1, are summarized in Fig. 2. The visual per-utterance analysis of Fig. 1 across all the consonants categorizes the plosives into the ZE group (62.8%) and a NZE group. Using both a percentage error and entropy, we may further classify the NZE utterances into three subgroups, (1) a *low error* (LE) high entropy ($\mathcal{H}_{\mathcal{N}} = 1$ random) group (15.8%), which places these utterances into the ZE group, (2) a *high error* (HE) low entropy group (10.7%) (talker misarticulation), and (3) *a medium error* (ME) group (10.7%). The average errors for /p/,/t/,/k/,/b/,/d/,/g/ at −2 [dB] SNR are 1.8%, 2.3%, 0.8%, 11%, 2.2%, and 0.7%, respectively. Thus the errors are around 1%–2% with the notable exception of /b/, which has a much larger error by more than a factor of 5. Most (>78%) plosive utterances are in the RZE (ZE + LE) group (functionally ZE).

Once the errors are split into the RZE and HE groups, one comes to a very different understanding of the error than that provided by the AI model, which lumps all the errors as if they are homogeneous. One might view our groupings as form of factor analysis.

B. Error groups 354 for the unvoiced stop consonants

This section analyzes the three unvoiced stop consonants /p/,/t/,/k/ on an utterance-by-utterance basis using the methods developed in Sec. II. For /p/, we provide all the $P_e(SNR)$ curves, thus expanding on Fig. 1. In Sec. IIIC, we analyze the three voiced stop consonants /b/,/d/,/g/.

1. Error analysis for /p/

Figures 3(a) to 3(d) show $P_e(SNR)$ for each of the 56 different utterances for the syllable /p/, in terms of the groups of Fig. 1 (see Table I). In Fig. 3(a), we show $P_e(SNR)$ for the 41 of 56 ZE utterances, which are zero for $SNR \geq -2$ dB. Formally speaking, the ZE group is referred to as the "hits" (true positives), meaning they are "heard as /p/ given /p/."

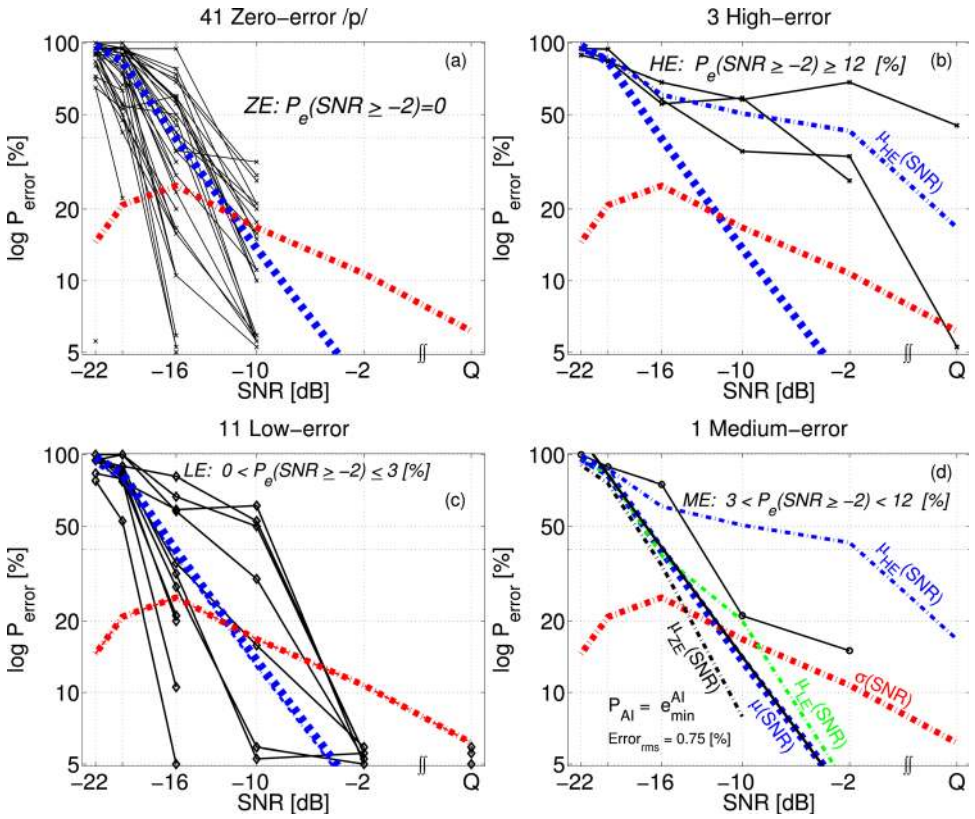


FIG. 3. (Color online) This figure shows the probability of error $P_e(SNR)$ for the 56 /p/ utterances, broken down into the four error groups as defined in Sec. II D 1. In each panel, the thick dashed curve is the grand-mean [$\mu(SNR)$] across all the 56 /p/ utterances while the thick dashed-dotted curve is the grand standard deviation [$\sigma(SNR)$], as labeled in (d). Here the *quiet* condition (indicated as Q) is arbitrarily assigned to 6 dB (Phatak and Allen, 2007). (a) shows the 41 ZE scores [$P_e(SNR \geq -2) = 0$]. (b) shows the 3 HE error sounds ($P_e \geq 12$ [%]), along with their mean [$\mu_{HE}(SNR)$] (thin dashed-dotted). (c) shows the 11 LE sounds ($3 < P_e < 12$ [%]). (d) Besides the one ME sound, also shown [solid line superimposed on the grand mean $\mu(SNR)$ thick dashed line] is the AI model error for /p/ computed from the AI error formula (lower-left), with $e_{\min} = 0.035$ (3.5%) and $AI \equiv (SNR + 21)/19$. The RMS error between $\mu(SNR)$ and the AI error formula is 0.75%. Also shown (thin dashed-dotted lines) are the means $\mu_{ZE}(SNR)$, $\mu_{LE}(SNR)$, and $\mu_{HE}(SNR)$, for the ZE, LE, and HE groups, respectively.

a. LE utterances. The LE group (11 utterances) shown in Fig. 3(c) are referred to as *misses* (or false-negatives) with exactly one error (of $N = 38$ trials). Because for single LE utterances $\mathcal{H}_{\mathcal{N}} = 1$, we call these single error utterances "random errors." These "missed"' utterances are well-articulated because most listeners (i.e., $N - 1$ of N) get them right. If the experimental trials were repeated, we would expect this list of sounds to totally change, as they reflect the random error rate. As shown later, we estimate that for /p/, a listener makes a random error (miss) once every 190 trials or so, on average. Possible causes of these errors may be lack of attention, wrong button clicked, etc. Errors with $\mathcal{H}_{\mathcal{N}} = 1$ are expected and very difficult to control. LE utterances are not inherently ambiguous (cannot be primed), rather they have random low-grade errors, and we view them as belonging to the ZE (hit) group. This ZE ∪ LE group, defined as the robust zero error (RZE) group, contain 52 (41 + 11) of 56 /p/ utterances (92.8%).

b. ME utterances. The ME group Fig. 3(d) contains a lone utterance m107pe that has 3 errors of 39. These three errors are all at −2 dB SNR with confusions /f,g,ʒ/ ($\mathcal{H}_{\mathcal{N}} = 1$). We presently have no clear intuition about the underlying nature of these errors. However, the fact that the error goes to zero in Q implies that the supporting feature is mal-formed and weak. We cannot calculate SNR_{90} from the information we have (it must be greater than −2 dB SNR). More analysis based on much more data will be required to resolve the true nature of the ME group.

c. HE utterances. We shall show that the HE utterances shown in Fig. 3(b) that contain the three utterances f113pI, m112pI, and f106pI are *true-negatives* (i.e., true errors). The confusions for f113pI were /b,k,n,t,y/ ($\mathcal{H}_{\mathcal{N}} = 1$), for m112pI were /g,g,k,k,k,ð,ʒ/ ($\mathcal{H}_{\mathcal{N}} = 0.8536$) while all the errors for f106pI (22 of 39) were attributed to /t/ ($\mathcal{H}_{\mathcal{N}} = 0.314$). Thus f106pI is ambiguous and near a /p-t/ confusion boundary. However, the other two sounds, though high in error, are not consistent in their confusions, across listeners. It seems likely that f113pI belongs in the RZE group, but m113pI is not easily classified, but leaning toward an ambiguous /p,g,k/ three group.

Utterances with high errors and low normalized entropy are expected when given a talker misarticulation, which is heard by multiple listeners as confusable within a small confusion group. For example, the reason why most listeners (22 of 39 trials) reported /t/ when f106pI was presented can be easily explained by looking at the AI-gram of the utterance, Fig. 4. This utterance has significant energy above 4 kHz (rectangular box region), which is a /t/ cue (Régnier and Allen, 2008), rendering this utterance ambiguous, as either /p/ or /t/. When listening to this utterance, one can easily prime for /p/ or /t/ (but no other consonant). The SNR_{90} values for the NZE utterances are tabulated in Table I. Some sounds (e.g., f106pI, f103te, f101kI) never reach 90% score, even in quiet (i.e., $SNR_{90} = \infty$). The LE sounds have low values (error between 1/39 and 1/35 with $SNR_e < -2$ dB) at SNR_{90}. Hence, these sounds are robust in the low-noise environment, thus classified as being in the RZE group. ME and

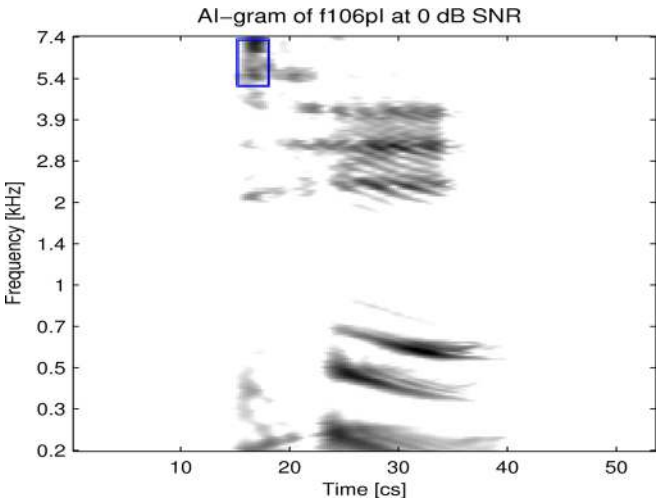


FIG. 4. (Color online) AI-gram of f106pI at 0 dB SNR. The conflicting cue is marked by a *solid box*. This clearly shows a high frequency conflicting /t/ burst (Régnier and Allen, 2008; Li and Allen, 2011). The utterance is primable as either /p/ or /t/. Correspondingly the error is 56%. The time axis is labeled in centiseconds [cs] (1 cs = 10 ms). Centisecond units are naturally relevant to speech perception.

HE utterances have high perceptual thresholds, hence have *true errors*, and are ambiguous.

If a sound was to have more than one event, the score would not drop rapidly (within a few dB). The very rapid drop in score below SNR_{90} demonstrates that there must be a single event, as is the case for /t/ (Régnier and Allen, 2008). We view such errors as binary. This view is consistent with our earlier study (Li *et al.*, 2010).

d. Variance of /p/ groups. The variance σ^2 (SNR) from the average error μ (SNR) is rarely studied. Looking across all the consonants, Phatak and Allen (2007) found three groups with low, medium, and high error, each of which followed the log-linear formula, thus accounting for a large portion of the variance. In Fig. 3(d), we look at a much finer level for /p/ and again see a very different picture: Shown in Fig. 3(d) as thin dashed-dotted lines are the means of the three other groups: $\mu_{ZE}(SNR)$, $\mu_{LE}(SNR)$, and $\mu_{HE}(SNR)$, and as in all the figures, the grand mean $\mu(SNR)$ and standard deviation $\sigma(SNR)$. These four means tell an interesting story about the error break down for /p/ that extends to all the consonants. Most of the error, and thus the variance, is in the RZE (ZE + LE) group but well below −2 dB SNR, as shown by σ (SNR). However, this variance is zero above −10 dB. Thus above −2 dB, all the error, and its variance, are due to only 8.7% (4 of 56) sounds. In Sec. IV, we shall further account for the RZE error variance in terms of each utterance's primary acoustic feature.

2. Error analysis for /t/

As in the previous case of /p/, we analyze /t/ at the utterance level. As seen in Table II, of 56 /t/ utterances, 40 have zero error in the low-noise environment (ZE group) and 10 are in the LE group, thus $|RZE| = 50$ (89%). Only two are HE utterances: m117te and f103te. Interestingly, all the errors in m117te (5/38) are /p/ ($\mathcal{H}_{\mathcal{N}} = 0.52$). This is a

TABLE II. Percentage error, N and SNR_{90} values for NZE utterances of /t/. Ten utterances in the topmost block with a single error (effectively less than 3% error) belong to the LE group, the next four in the middle block are ME utterances, while m117te and f103te are HE ambiguous utterances. The HE utterances have high SNR_{90} thresholds as seen in the table.

Utterance	P_e (%)	N	SNR_{90}
f109tI	2.70	37	−22
f119tI	2.56	39	−14
m114t@	2.70	37	−11
m120t@	2.78	36	−21
f106ta	2.56	39	−11
f108tI	2.63	38	−22
m104ta	2.70	37	−10
m114tI	2.70	37	−17
m118te	2.63	38	−17
m120ta	2.70	37	−22
f113tI	5.26	38	−11
m120tI	5.13	39	−22
f109t@	7.89	38	−16
m111t@	8.11	37	−4
m117te	13.16	38	18
f103te	68.42	38	∞

natural complement to the case of f106pI where all the /p/ errors were /t/. This again is predictable when one studies the AI-gram of m117te (not shown), having a significant low-frequency energy, which is a conflicting cue region for /p/ (Li *et al.*, 2010; Li and Allen, 2011). Utterance f103te (5 errors) is mostly confused with /d/ ($\mathcal{H}_{\mathcal{N}} = 0.41$) because the utterance has a very short time-gap between the burst feature and the vowel, as is characteristic of voiced /d/ (Li *et al.*, 2010).

3. Error analysis for /k/

Of 56 /k/ utterances (Table III) $|ZE| = 49$, $|LE| = 3$, thus $|RZE| = 52$ (93%). Only seven /k/ utterances are in error, and only two of these (f101ka and f101kI) show high errors. Both are confusable with only one other sound: /g/. Talker f101 is a poor articulator for /k/. Figure 5 shows the AI-grams of these two sounds. From the study by Li *et al.* (2010), the /ka/ cue is a mid-frequency burst around 2 kHz, articulated 5–7 cs before the vowel. On the other hand, /ga/,

TABLE III. Percentage error, N and SNR_{90} values for NZE utterances of /k/ in the low-noise environment. The NZE group is half that of /p/ and /t/. We interpret /k/ as having high *salience*, meaning it is easily articulated and easily identified (i.e., it is naturally robust). The top three utterances belong to the LE group, the next two to the ME group, and the last two are HE utterances (with high SNR_{90} values).

Utterance	P_e (%)	N	SNR_{90}
f103ke	2.56	39	−17
m115ke	2.56	39	−16
f119ka	2.63	38	−4
m112k@	5.13	39	−2
f119kI	7.89	38	−11
f101ka	13.89	36	18
f101kI	22.22	36	∞

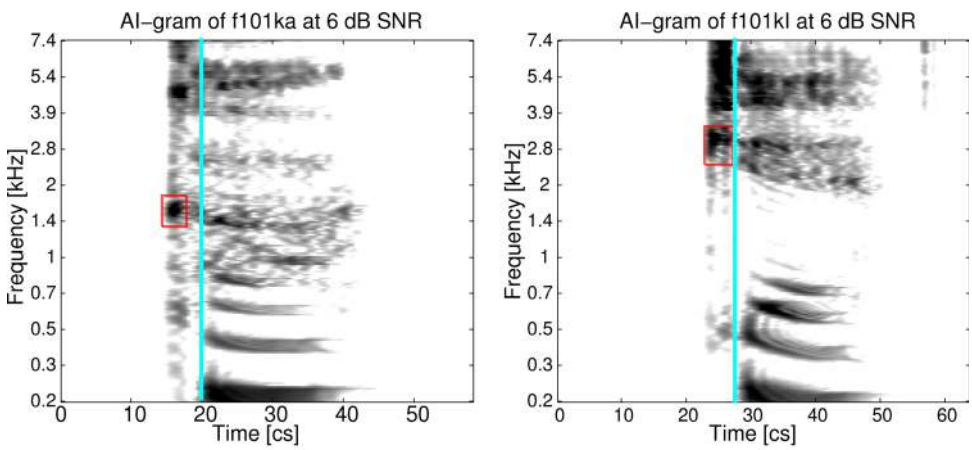
the voiced counterpart of /ka/, has a mid-frequency burst, typically followed by a F2 transition just before the start of sonorance. As seen from the AI-grams of Fig. 5, f101ka has its burst cue just before the vowel start and does not have the characteristic 5–7 cs gap before the onset of the vowel, typical of a clearly articulated /ka/. Similarly, f101kI is atypical because unvoiced stops do not have bursts close to the vocalic region. Hence, these two sounds are confused with /g/. Vowel onset is marked by a solid line, while the burst cue is boxed.

C. Error groups for the voiced stop consonants

As their unvoiced counterparts, the voiced stop consonants (/b/,/d/,/g/) also have utterances with different perceptual thresholds. /b/ is the lone stop consonant in the high error set (C1) of the PA07 study (Phatak and Allen, 2007). One might qualitatively describe /b/ as having *low salience*. However, robust ZE with low SNR_{90} thresholds still exist but are rare (11 of 56 utterances in this sample). For the voiced stops, the data is tabulated in the Tables IV, V, and VI.

1. Error analysis for /b/

Consonant /b/ is substantially different from the other five stop consonants used in the study, as it has an 11% error rate as compared to an average of ≈1.5% in quiet for the other consonants. Specifically, /b/ forms a confusion group



(a) AI-gram of f101ka. (b) AI-gram of f101kI.

FIG. 5. (Color online) AI-grams at 6 dB SNR. In both the AI-grams, the solid box is the /k/ feature while the start of the vowel is marked by a solid line. We see that the burst cue is very close to the beginning of the vowel, which is a characteristic of the /g/ feature (Li *et al.*, 2010), thereby explaining why these two /k/ utterances are highly confusable with /g/.

TABLE IV. Percentage error, N and SNR_{90} values for NZE utterances of /b/. The horizontal line is the demarcation between 10 low error (LE) utterances (above) and the 10 medium error (ME) utterances (below). The entire right column of the table is the HE utterances (25 in total of the 45 NZE utterances that have errors). Clearly, /b/ is a difficult sound compared to the other five stop consonants because a majority of its utterances have high errors. Such high errors are likely to be due to production errors as evidenced by the fact that one talker (f101) has no high error (just one utterance f101bI has a single random error). The 11 ZE sounds demonstrate that the listeners can hear a well articulated /b/. For most HE sounds, /b/ is confused with /v/ and /f/. These HE utterances have high thresholds and most do not reach 90% score, even in quiet.

Utterance	P_e (%)	N	SNR_{90}	Utterance	P_e (%)	N	SNR_{90}
f101bI	2.63	38	−6	f103ba	13.51	37	18
f109ba	2.63	38	−11	f105bI	12.5	40	5
m107be	2.7	37	−6	m115b@	13.51	37	12
m120b@	2.7	37	−10	m115ba	13.89	36	11
f106bI	2.7	37	−6	f108ba	15.38	39	18
f113ba	2.78	36	−4	f105be	15.38	39	13
f113bI	2.86	35	−10	f108b@	16.22	37	14
m107bI	2.5	40	−4	m104be	14.63	41	12
m111bI	2.86	35	−11	m112b@	15.38	39	13
m112be	2.86	35	−4	m111ba	20.59	34	∞
m120be	5.71	35	−16	m104b@	18.92	37	18
m111b@	5.41	37	−4	f105b@	17.95	39	12
f109be	5.56	36	0	m118be	21.05	38	∞
m112ba	5	40	−3	m102ba	21.05	38	∞
m118ba	7.89	38	−3	m114b@	21.05	38	18
f105ba	7.89	38	−2	f119b@	23.68	38	18
f108be	8.33	36	6	f108bI	23.08	39	15
m107ba	10.81	37	7	m102be	24.39	41	18
m114be	10	40	7	m111be	24.39	41	15
m114ba	10.53	38	8	f109b@	28.21	39	∞
				f119bI	27.5	40	∞
				m118b@	32.43	37	∞
				f119ba	31.58	38	∞
				f119be	47.5	40	∞
				f103b@	60	40	∞

with the fricatives /v-f/ because the /b/ acoustic feature is not robust and is easily masked by noise.

Only 11 of 56 /b/ utterances have ZE in the low-noise environment. This breakup of the utterances into the two main error groups, and the distribution of the errors in the second group (NZE), is shown in Fig. 6 and tabulated in Table IV. Consonant /b/ forms a confusion group with /f/ and /v/. These three consonants have high errors even in low-noise environments (Miller and Nicely, 1955; Li *et al.*, 2010).

We suspect that the high /b/ error is mainly due to production errors as evidence by the 11 ZE utterances and that 13 of 14 talkers of /b/ are high error. Talker f101 has all its utterances in the RZE group. This proves that the listeners can do the task because they make no errors for this talker, who clearly enunciates the consonant /b/.

We conclude that consonant /b/ is more difficult to articulate and thus is more likely to be confusable (low salience). Unlike /t/ or /g/, it does not seem to have an easily identified single feature that makes it noise-robust (Li *et al.*, 2010).

TABLE V. Percentage error, N and $SNR90$ values for NZE utterances of /d/. The left four columns contain the 12 LE utterances. The horizontal line on the right four columns is the demarcation between the 13 medium error (ME) utterances (above), and the 4 high error (HE) utterances (below). The SNR_{90} values are well correlated with these three groups: LE sounds have low thresholds while HE sounds have high perceptual thresholds, even ∞ for sounds whose score does not reach 90% even in quiet.

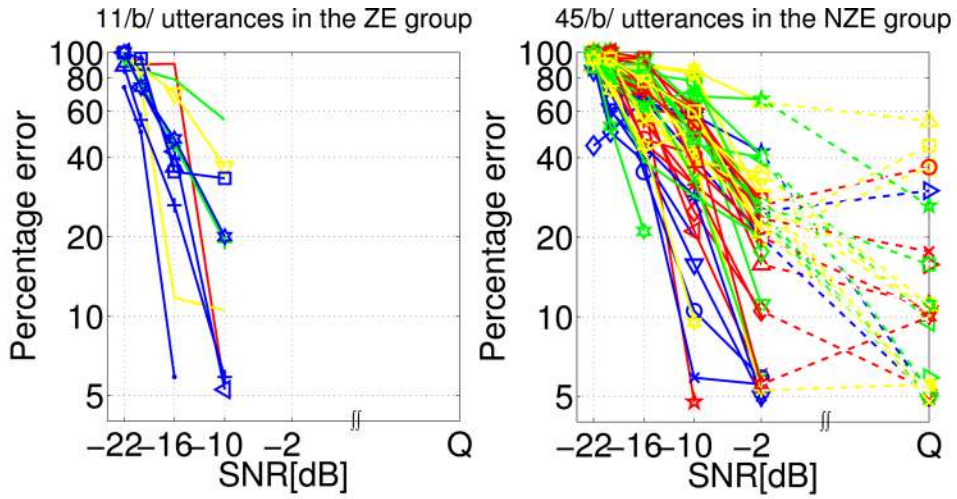
Utterance	P_e (%)	N	SNR_{90}	Utterance	P_e (%)	N	SNR_{90}
f101de	2.63	38	−21	m111d@	5.41	37	−4
f105dI	2.78	36	−17	f105de	5.13	39	−11
m118dI	2.63	38	−13	f119de	5.26	38	−10
f108dI	2.78	36	−21	f119d@	5	40	−20
f103de	2.44	41	−13	m107de	5.41	37	−11
f103dI	2.86	35	−20	m111de	5.26	38	−15
f108da	2.44	41	−11	m112de	5.26	39	−17
f119da	2.78	36	−20	m114de	5.13	39	−4
m112da	2.56	39	−10	m115da	5.13	39	−3
m114dI	2.7	37	−17	f108de	5.13	39	−2
m117da	2.63	38	−10	f109de	5.41	37	−20
m118de	2.56	39	−10	f113d@	5.56	36	−10
				m114da	8.57	35	−3
				m118d@	13.89	36	∞
				m102de	17.95	39	12
				m115dI	21.05	38	13
				m114d@	27.5	40	∞

As previously mentioned, we have assumed that the subjects form a homogeneous group. While this is a reasonable assumption for the other low-error stop consonants, it seems to break down when the task becomes difficult, e.g., for the perception of /b/. A difficult test naturally categorizes the listeners into performance groups. Given its very different nature, the analysis is extended to listener errors, as shown in Fig. 7.

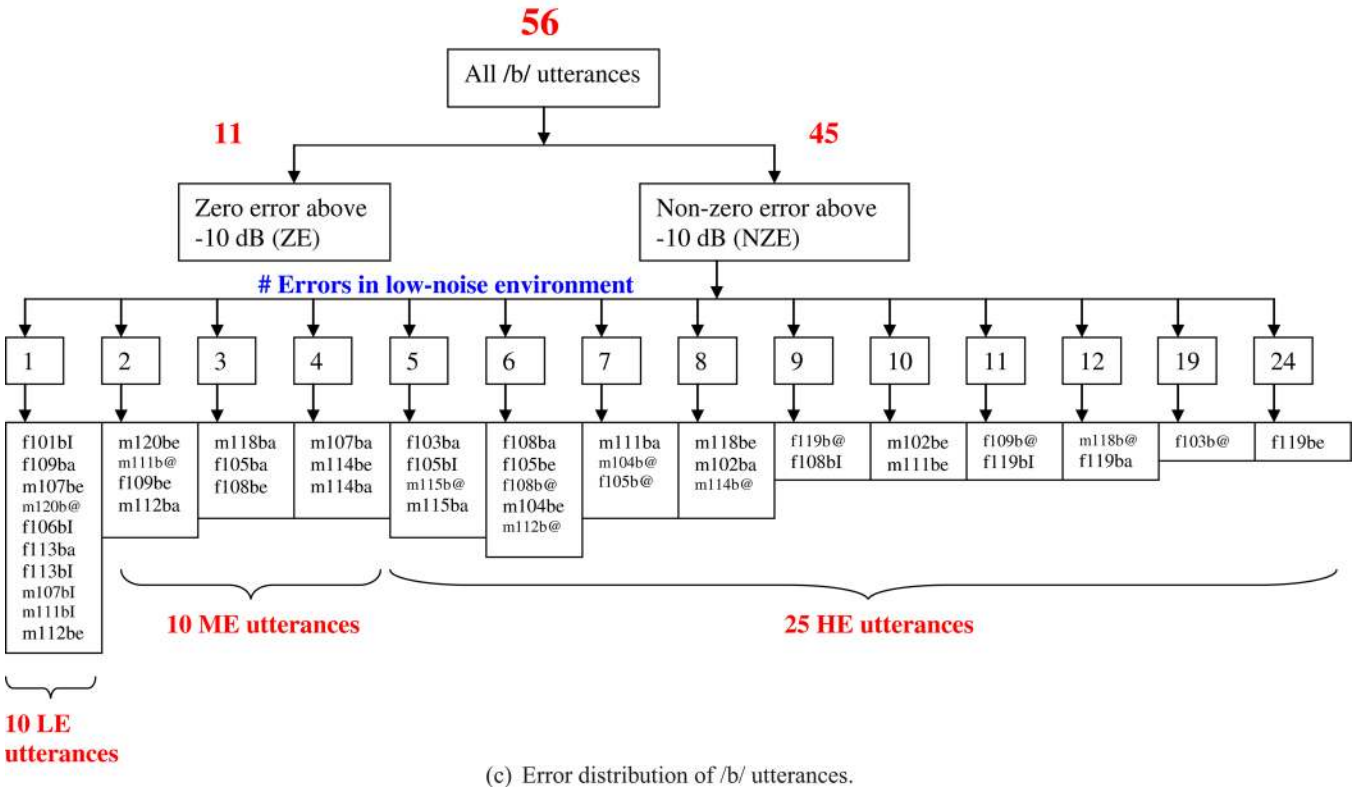
In PA07, four low-performance (LP) listeners, with scores less than 85% in quiet, were removed during analysis, and the top 10 high performance (HP) listeners were selected (Phatak and Allen, 2007, p. 2315). Each of these 14 (4 + 10) listeners completed the experiment (5376 tokens). Figure 7 shows the log-error versus SNR for consonant /b/ for these 14 listeners. The legend provides a two-letter listener ID. Of these, listener QN has the lowest error rate, except for quiet, suggesting a varying attention during the task. Subjects BH and LT have substantially higher error across SNR as

TABLE VI. Percentage error, N and SNR_{90} values for NZE utterances of /g/. All the 56 /g/ utterances used in the experiment are well-articulated and have no high errors. The utterances in the left four columns form LE group while the right three column utterances belong to the ME group. All NZE utterances have SNR_{90} threshold below −2 dB SNR.

Utterance	P_e (%)	N	SNR_{90}	Utterance	P_e (%)	N	SNR_{90}
m107ge	2.94	34	−11	f101g@	5.13	39	−13
m107gI	2.44	41	−12	m107g@	5.26	38	−13
m112ga	2.78	36	−11	f119ga	7.5	40	−7
m112ge	2.5	40	−12	m102ga	7.89	38	−3
m118g@	2.63	38	−13	m104gI	8.11	37	−10
f106g@	2.78	36	−5	m115ga	7.32	41	−3
f108ga	2.7	37	−3				



(a) Zero error /b/ utterances. (b) Non-zero error /b/ utterances.



(c) Error distribution of /b/ utterances.

FIG. 6. (Color online) This figure shows the distribution of errors of the 56 utterances of b. The colors in (a) and (b) indicate the four vowels. Quiet is arbitrarily marked at 18 dB and for (b) is joined to −2 by dashed lines. (a) Error vs SNR plot of the 11 ZE utterances. (b) Error vs SNR plot of the 45 NZE utterances. (c) Breaking down the errors in the low-noise environment, based on the absolute number of errors made. Twenty-two utterances are in the RZE group. Twenty-five (44%) utterances are HE utterances.

compared to the average. In quiet, the listeners at or above average error were AN, BH, LT, QN, CB, and SP. The four subjects removed from the PA07 analysis were AN, BH, LT, and QN. Thus with the obvious exception of QN, the poor performing listeners on average are also the poorer listeners of /b/. The other 11 listeners who completed varying number of trials are not shown in this figure. However, these listeners also naturally break down into performance groups. For an easy task, there is a smaller difference between the LP and HP listeners, but these groups clearly stand out once the task becomes difficult. As might be expected, most errors are attributed to these low-performance subjects.

2. Error analysis for /d/

Of 56 /d/ utterances, 27 have zero error in the low-noise environment. The distribution of errors is shown tabulated in Table V, which shows that /d/ has 12 utterances with random errors in the LE group, 13 in the ME group. Four utterances (m118d@, m102de, m115dI, and m114d@) are characterized by high error and low entropy and belong to the HE group. Of these, m118d@ and m114d@ have timing errors and are confusable with their voiced counterpart (/g/). m115dI has a conflicting cue of b and is confused 7 of 38 times with /b/ and once with /ð/. m102de is mainly confused

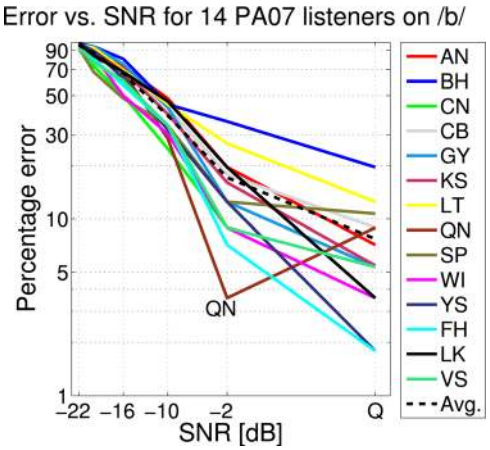


FIG. 7. (Color online) Log-error vs SNR for /b/ (average over 56 utterances) for the 14 listeners who completed the experiment (PA07). The grand average error over these 14 listeners is shown by a dashed line. The legend indicates each listener with a two-letter ID. In quiet, there were six listeners having greater than average error: AN, BH, LT, QN, CB, and SP. The four listeners removed from the PA07 analysis were AN, BH, LT, and QN (not CB and SP). We see from the figure that other than for quiet, QN was the best listener. For this figure Q was arbitrarily defined as 18 [dB] SNR.

with ð, perhaps because m102de is not articulated with sufficient "voicing."

3. Error analysis for /g/

Of 56 /g/ utterances, 43 have zero error in the low-noise environment. The errors are tabulated in Table VI. /g/ is a robust (highly salient) sound and no utterance used in the PA07 experiment is misarticulated (i.e., no HE utterance), according to our criterion of $\geq$12% in the low-noise environment.

D. Error distribution across the four vowels

Of the total 336 utterances (6 stop consonants $\times$ 56 utterances of each) in the experiment, 125 belong to the NZE group (15 for /p/ + 16 for /t/ + 7 for /k/ + 45 for /b/ + 29 for /d/ + 13 for /g/). Broken down by the vowel, they are 33, 34, 31, and 27 for (/ɑ/,/ɛ/,/ɪ/,/æ/) respectively. This gives an entropy of 1.99 bits. Thus the error distribution over the vowels is almost uniform (uniform distribution would imply a maximum 2 bit entropy). This pattern of errors implies nothing about coarticulation effects, rather it simply shows the lack of correlation of misarticulated consonants with the following vowel. To study coarticulation effects one must look at the acoustic features for the zero error sounds, as a function of the vowel.

IV. SUMMARY AND DISCUSSION

Figure 2 summarizes the errors made by listeners on the six stop consonants. From the bar plot, /b/ has the largest number of utterances in the high error (HE) group. Hence, /b/ is a difficult sound (has low salience). The remaining five CVs have only a few utterances that fall into the HE group, and these represent a major component of $e_{\min}$. Some listeners have difficulty phonotactically identifying the difference between /d/ and /ð/, possibly due to insufficient early rigorous phonemic training.

By our definition, *robust utterances* are made up of the RZE group (|ZE| + |LE|) while *ambiguous* utterances compose the HE group and count (of 56) 3, 2, 2, 25, 4, and 0 for /p/, /t/, /k/, /b/, /d/, and /g/, respectively. As summarized in Fig. 2, the percentage of robust zero error (RZE) sounds (i.e., $100 \times |ZE + LE|/56$) for /p,t,k,b,d,g/, is 92.8%, 89.3%, 92.9%, 37.5%, 73.2%, and 89.3%, respectively (average 78.6%, which excluding /b/, approaches 90%). Averaged across all the six stop consonants, the percentage of utterances in the ZE, LE, ME, and HE group is 62.8%, 15.8%, 10.7%, and 10.7%, respectively.

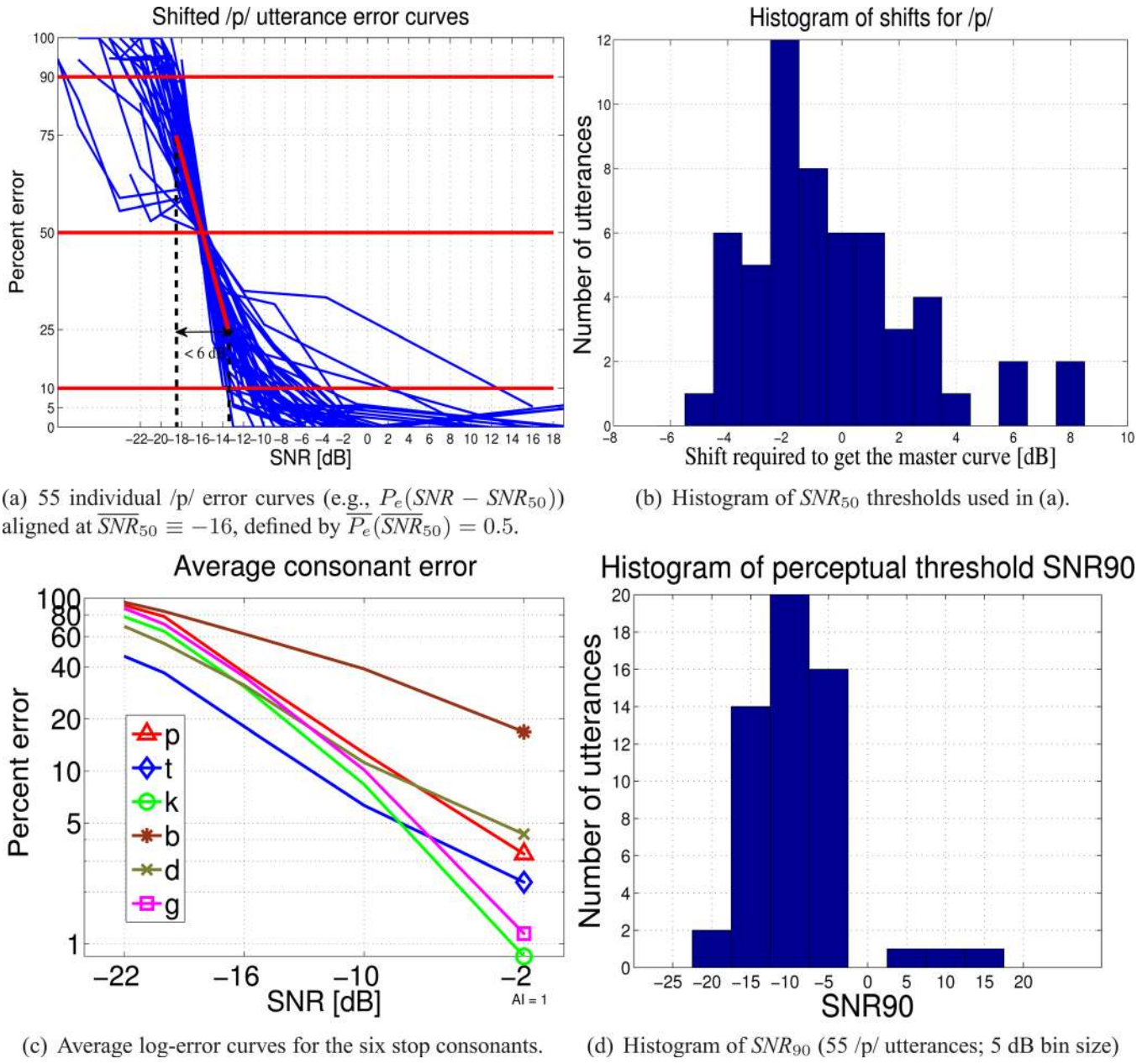
When the task is easy (i.e., for naturally low error utterances like /p/, /k/, /g/ etc., which have high salience), the *only* contributors to the error in low-noise environments (i.e., $e_{\min}$) are a small number of HE (ambiguous) utterances. These "errors" are not perceptual because these sounds are identifiably misspoken (everyone hears them otherwise).

For the highly confusable (low salience) stop consonants (i.e., /b/), there is a significant disparity across listeners. As shown by Fig. 7, average error for /b/ is primarily determined by four LP subjects (BH, LT, CB, and AN) because they form the at and above average-error subjects. The removal of these four subjects would reduce the /b/ errors dramatically (e.g., from 18% to 4%). This might make the /b/ errors similar to /d/ of Fig. 8(c).

A. Estimates of the random error rate

This section presents an estimate of the average number of trials needed by a listener before they make a low-level random error. The assumption is that all 25 listeners are homogeneous [this is of course not strictly true because some listeners are significantly poorer than others (Phatak and Allen, 2007)] or as in the case of /b/. Given that /p/ has a naturally low error (|RZE| > 92%), it seems reasonable to consider listeners as uniform for this task. In total, 2121 tokens of 56 /p/ utterances were presented in the low-noise environment (1059 at −2 dB SNR and 1062 in quiet). Thus N on average is $\approx 2121/56 = 37.88$. For these 2121 trials, the number of utterances with a single (random) error is 11, those with less than 3% error (see Table I). On average, a listener makes a random error every $2121/11 = 192.63$ trials. Hence, the *rate of random errors* is less than 1 of 190 (i.e., 0.53%). If random errors are assumed to be uncorrelated across utterances, other CVs should also have a similar error rate.

The corresponding value for the number of trials before a random error is made on average, for /t, k, b, d, g/ is 212, 710, 212, 150, and 303, respectively. The outliers are /k/ and /g/, which have much lower random error rate, specifically 0.14% for /k/ and 0.33% for /g/. The obvious question is: Why do /k/ and /g/ have this very low error (0.14% for /k/ vs 0.52% for /p/) in the low-noise environment? It must be that the random errors, as defined, are *not* totally uncorrelated across utterances, rather they are modulated by the difficulty of the task as in the case of /b/. It follows that some LE utterances, having a 1 in $N \approx 38$ error, may not be truly random. Likely they reflect a near-threshold feature, which some



(a) 55 individual /p/ error curves (e.g., $P_e(SNR - SNR_{50})$) aligned at $\overline{SNR_{50}} \equiv -16$, defined by $\overline{P_e}(\overline{SNR_{50}}) = 0.5$.

(b) Histogram of SNR_{50} thresholds used in (a).

(c) Average log-error curves for the six stop consonants.

(d) Histogram of SNR_{90} (55 /p/ utterances; 5 dB bin size)

FIG. 8. (Color online) (a) Individual /p/ error curves aligned at their 50% error values. The solid line shows the average "master error curve," which falls from 75% to 25% error over 6 dB. (b) Histogram of the shifts SNR_{50} for each /p/ utterance, required to shift to the average (i.e., the master curve). Individual error curves are aligned at their 50% error values at −16 dB (as defined by the solid line). (c) Average log-linear error curves for the six stop consonants, with AI = 1 marked at −2 dB SNR. Log-linear regression fits have correlation coefficients of 0.990, 0.997, 0.981, 0.996, 0.998, and 0.992 for /p/, /t/,/k/,/b/,/d/, and /g/, respectively. The average of these six curves is the thick dashed line labled $\mu(SNR)$ of Fig. 3(d). (d) Histogram of the perceptual thresholds SNR_{90} values for 55 /p/ utterances [utterance f106pI never reaches 100% score (i.e., $SNR_{90} = \infty$)]. If we ignore the three outliers having high (>0) threshold values, the remaining SNR_{90} values have a dynamic range of ≈ 20 dB. This is approaches the AI's 30 dB dynamic range, defined across all utterances (French and Steinberg, 1947).

listeners confuse. For example, of the 11 errors classified as random for /p/, 3 (f101pe, m115p@, m118pI) have their single error in quiet and are error-less at −2 dB SNR. The responses were /d/, /n/, /noise only/. Consonant /p/ is not expected to form a confusion group with these consonants (Li *et al.*, 2010; Li and Allen, 2011), and it is therefore reasonable to assert that the score in quiet will be higher than in noise. Hence, it is likely that these are truly random errors. The other eight LE /p/ sounds have their single error at −2 dB SNR and are confused with /f,k,k,θ,t,f,t,v/. Because /p-t-k/ is known to be a strong confusion group in noise (Li *et al.*, 2010; Li and Allen, 2011), it seems likely that these utterances, with such confusions, have a higher (e.g., 0 dB) threshold for their perceptual feature (i.e., they are less robust). The confusions suggest that these errors are not totally random and that the error rate is correlated with the difficulty of the task. Yet these utterances can still be termed as "robust" because they have such a very low error. Useful insight would likely be gained by studying the errors on these utterances at −10 dB in addition to −2 dB and quiet.

Our conjecture is that the true random error rate is actually less than 1/300 (0.33%), as for /k/ and /g/. Over

time we hope to discover improved methods of monitoring and controlling for these low-grade but significant random errors. While these errors are small, they are real, as humans are never *perfect* at any given task. It seems likely that percentage error may not be an adequate statistic (e.g., this percentage will be listener dependent). A more confident analysis might be stated on the basis of confusion groups, listener differences and/or difference between the -2 and Q SNRs.

The number of ZE utterances is of course a function of the number of presentations N. The probability of error as a function of N [$P_e(N)$], for large enough N, must become nonzero due to imprecision in human performance over large number of trials. For example, every simple task will have an error for sufficiently large N. Thus the concept of "zero error" seems essentially flawed as the number of ZE utterances will tend to zero as N becomes sufficiently large. However, we may still distinguish these true hits on the basis of their very low error and high entropy (low correlation of errors). This is because these sounds are inherently robust (not primable) and have a well-defined perceptual event that is not easily masked. A "zero error" sound implies an utterance for which the error (if any) will be of a random nature across thousands of trials, given low additive noise conditions. It is important to note that these sounds are common (63% of the sounds) in our sample of the MN64 database.

B. An analysis of the AI model

Next we wish to provide an insight into how Eq. (2) depends on individual utterance error curves, and why their average is typically exponential, with consonant-dependent values for $e_{\min}$. Consistent with AI theory, PA07 found that the exponential AI model [Eq. (2)] fits the data for the three consonant groups C1, C2, and C3 with group-dependent values for $e_{\min}$. Based on the Miller and Nicely (1955) data set, Allen (2005a) came to the same conclusion as did Li and Allen (2009) for stop consonant and fricatives.

From Fig. 3, for very low SNRs (<-20), all utterances approach chance error. In many situations, chance is either known or may be measured, allowing one to normalize-out this effect. From Figs. 3(a) to 3(c), at high SNRs (≥ -2 dB), *ZE* sounds have zero variance and *LE* utterances have low-level maximum-entropy (random) errors, again with zero variance. Thus in these two limits, the mean and normalized entropy are either 0 or 1 and the variance is zero. These two groups account for 93% of /p/ plosives (52/56) and $\approx$ 80% of all plosives studied here. Below -2 dB, the grand variance [i.e., $\sigma^2(SNR)$] is dominated by the RZE group as shown in Fig. 3(a). This is because the HE + ME group represents a small portion of the error and variance. We shall show next how even this grand variance, below -10 dB, may also be explained.

From Fig. 8(a), each $P_e(SNR - SNR_{50})$ curve drops from a high error ($\approx$75%) to low error ($\approx$25%) within $\approx$6 dB, where SNR_{50} is defined by $P_e(SNR_{50}) = 0.5$. We show this by aligning the 55 /p/ utterances (excluding f106pI) at their 50% point, to define a *grand average master curve*. The shift required to align the 50% points to that of the average (at -16 dB SNR) is denoted SNR_{50}. The slope of this master error curve at the 50% point is $\approx$9%/dB. Thus we use the average (at -16 dB) as the reference point for /p/ to which we shift the individual curves as shown by the histogram in Fig. 8(b). By construction, the variance of the master curve is zero at -16 dB. We conclude that the variance in the RZE group is almost entirely due to the variable thresholds [Fig. 8(b)] (this would be exactly true if it were not for the finite 9%/dB slope of the curve). The SNR_{50} shift is a measure of the utterance's *perceptual threshold*. At a given SNR, most utterances are either at 100% or 0% error with very few utterances in the transition region (i.e., it is less than 6 dB). Each individual utterance error curve approximates a *step function at* SNR_{50}.

As shown in Fig. 8(c), the average error curves for the six stop consonants are also log-linear [consistent with Eq. (2)]. Note that /p/, /t/, and /d/ form a group with a similar log-error slope, as do /k/ and /g/, with comparable values of $e_{\min}$, while /b/ has a slope similar to /t,d/ but with a much larger $e_{\min}$. Hence, an exponential model fit the average error of these two groups because exponentials with the same log-error slope add. With this improved understanding of the *RZE* grouping, we see that the three groups (i.e., C1, C2, C3, defined in PA07) must also contain RZE sounds because they are the same data set.

From Figs. 8(c) we can also see how the *grand standard deviation* [$\sigma(SNR)$ of Fig. 3(d), dashed-dotted] is impacted by the large spread of consonant means. The grand mean error variance [$\sigma^2(SNR)$] is hierarchical: The first source is determined by the large scatter in the means of the individual consonants. The second source related to that shown shown in Fig. 3(b) for /p/, which is due to the distribution in SNR_{50} thresholds, as shown in Fig. 8(b). At high SNRs ≥ -2 dB, the mean and variance are determined by a small number of HE consonants.

We conclude that for normally articulated utterances, normal hearing speech perception is a *binary decision process* in which errors are essentially zero above their threshold. Individual utterances have different SNR_{90} thresholds as shown in Fig. 8(d). In every case, the group scores in quiet (the $e_{\min}$'s) depend on a small number of misarticulated utterances. The exponential nature of the average curve is therefore due to the threshold distribution and the few HE utterances. The RZE curves saturate at the ends of the AI range, as shown in the master curve, which is similar to Fig. 21 of French and Steinberg (1947).

This error model explains the AI model's characteristics, as given by Eq. (2). The exponential error is a consequence of the distribution of RZE thresholds over a large number of utterances with all but few utterances having no error in the low noise environment. Hence, for stop consonants, only a small number of utterances (HE + ME) contribute to $e_{\min}$.

Ronan *et al*. (2004) studied various combinations of five frequency bands having approximately equal articulation and attempted to fit the AI model to the recognition results of four (Exp I) and five (Exp II) listeners. It should now be clear why missing AI bands might not work within the AI model. When one removes a single band, the subset of

sounds having a feature in that band are converted to confusions due to the sound's conflicting cues. This has been carefully studied by Li and Allen (2011) and Kapoor and Allen (2012), where a specific feature (not a band), such as the burst of /k/ or of /t/, is removed. The change in the AI is not significant when a single acoustic feature is removed because an isolated burst contains such a small fraction of the speech energy—yet removing one feature dramatically alters the scores by activating the latent conflicting cues (Li and Allen, 2011; Kapoor and Allen, 2012).

There are many known limitations of AI theory. First, the AI was not designed to predict confusions of individual utterances. Furthermore it has a very large variance, a sort of mid-riff bulge, between −20 and 0 dB, where the variance from the mean error is huge. This variance is due to several factors. First is the large variance of the means across various consonants as may be inferred from the consonant means as displayed in Fig. 8(c) (Allen, 2005a; Phatak and Allen, 2007; Phatak *et al.*, 2008). Next is the variance due to the distribution in SNR_{50} thresholds for individual consonants as shown in Fig. 8(b). The corresponding bulge (spread) at the 10% error point is even larger as shown by Fig. 8(d). One would need to calculate these distributions, as a function of consonant class, to fully characterize the true nature of the AI's multifactor variance. One might conclude that while the AI has been a venerable and critically important research tool, it has many weaknesses. We believe that the present study provides deep insights into many of these imperfections.

V. IMPLICATIONS TO ASR

The key issue with automatic speech recognition (ASR) is its fragility due to noise (Lippman, 1997). It is the events that make HSR highly robust to noise as compared to machine recognition (Allen, 1994, 2005b). Scharenborg (2007) also provides a comprehensive argument in favor of using the knowledge from HSR research to improve ASR systems. A confusion matrix (CM) analysis by Sroka and Braida (2005) showed that ASR systems did a reasonable job in recognizing syllables degraded by low-pass and high-pass filtering. However, for syllables degraded by additive speech-shaped noise, none of the automated systems recognized consonants like humans. The phone classification accuracy in ASR systems is, at best, about 82% in quiet (Huang and Hasegawa-Johnson, 2008). For humans, the score in quiet is commonly assumed to be near 98%–98.5% (Allen, 2005a). But again, this is an average over a large number of utterances. Given our present results, we have raised the bar to match human performance. For HSR, when properly measured, the error is essentially zero for SNRs ≥ -2 dB. Given precise knowledge of human speech decoding, it must be possible to exploit this knowledge and build robust ASR front ends that are human-like in performance. Exactly how to do this is unknown.

These results are relevant to automatic speech recognition (ASR) because (a) the HE consonants are mainly production errors and (b) one talker had no production errors (all others had significant errors).

VI. LIMITATIONS AND FUTURE WORK

We believe that this study is the first to analyze normal hearing perception of individual utterances. This analysis provides important insights into the distribution of errors and thus explains why and how the AI theory works, for plosives. In the future, it would be useful to carry out a more extensive study, of the full nature of confusions of several other isolated syllables (fricatives, nasals, and vowels). We will need a more comprehensive analysis to fully characterize the utterances in the ME group. Confusion studies and normalized entropy seem to be the proper tools for such an analysis.

We also hope to build a better model of the AI that includes random errors, listener biases, and confusions. We must also characterize the underlying distribution of each consonant's set of perceptual thresholds (SNR_{50}) and more fully characterized by the confusion groups. Given the ease with which the subjective measure SNR_{90} is to estimate, it seems an excellent statistical measure of the quality (i.e., robustness) of each utterance.

VII. CONCLUSIONS

The key conclusions from this study are as follows:

(1) Most stop consonants have essentially ZE in low-noise environments, the summary of which is provided in Fig. 2. The consonant /b/ has the smallest ZE group (11/56).
(2) Normal hearing speech perception for salient syllables (RZE) is a binary decision making process (you either hear the cue or not) in which the errors are essentially zero when the syllable event is above threshold. This was first shown by Régnier and Allen (2008) for /t/ and is established here for other stop consonants based on this detailed utterance error analysis. The support for this claim is Fig. 8(a).
(3) Due to talker mispronunciation, HE group utterances can be separated from the LE and ME group utterances, based on their error ($P_e \geq 12\%$) and normalized entropy ($\mathcal{H}_{\mathcal{N}} < 1$).
(4) The source of errors in ambiguous HE stop consonants can almost always be easily explained, using the AI-gram, in terms of the robustness of their perceptual feature and the feature of the main confusion (conflicting cue) as shown in Figs. 4 and 5.
(5) The average error is exponential in SNR, expressed as Eq. (2). For $SNR < -2$ dB, this dependence follows from the underlying distribution of SNR_{50} [the utterance thresholds of Fig. 8(b)]. For $SNR \geq -2$ dB, the error is determined by $e_{\min}$.
(6) The minimum error ($e_{\min}$) under ideal conditions ($AI = 1$) is explained by errors in a small number of highly confusable tokens (ME + HE groups). These sounds may be characterized by their high SNR_{90} thresholds, typically >0 dB SNR, or even ∞ for utterances that never reach a 90% score.
(7) The average grand error mean has a large variance that may decomposed into the variability in consonant means [Fig. 8(c)] and talker variability (Fig. 7), but most important, variability due to the distribution in the event thresholds [Fig. 8(b)].

(8) Shannon's channel capacity seems to be obeyed because the error is essentially zero above −2 dB SNR (subject to some production errors). Thus humans transmit CVs below the channel capacity.

ACKNOWLEDGMENTS

The authors wish to express their special appreciation to Woojae Han and Sandeep Phatak. They also wish to thank other members of the HSR group at the University of Illinois, Urbana: Andrea Trevino, Abhinauv Kapoor, Anjali Menon, Roger Serwy, and Bob Cvengros for many helpful comments and discussions. We would especially like to thank Joseph Toscano for help with many aspects of the data analysis. This research has been supported by NIH under Grant No. RDC009277A and Phonak Hearing Instruments (with special thanks to Stefan Launer). This study represents, in part, the MS thesis of the first author.

APPENDIX A: BERNOULLI TRIALS AND SPEECH PERCEPTION

In this section, we deal with the problem of determining the number of trials required to quantify speech perception when building CV confusion matrices (or a count matrix). The problem may be simply stated: What number of Bernoulli trials N_t of a particular consonant-vowel sound is required to determine the probability $\mu = P_c$ with a specified confidence (e.g., 3σ), that the consonant is correctly heard.

To address this problem, one must make a minimum of two assumptions. The first is that the subject is consistent. In fact because the subjects are human and fall asleep, become bored, exhibit learning effects, or even play games during tedious psychological experiments in the booth, etc., one can never be sure that this is not violated. However, there are well known methods to keeping the subject attentive, such as frequent breaks and by monitoring the subject during the experiment. This may be a fragile assumption, but it is a necessary one. The second assumption is that we may model the outcomes using Bernoulli trials with binomial outcomes. Thus we limit ourselves to the binomial probabilities having weights "n choose k"

$$\frac{n!}{k!(n-k)!}P_c^k(1-P_c)^{N-k},$$

which are the probabilities of $N-k$ errors in N trials.

Given the preceding basic assumptions, we may apply well known results to compute estimates of confidence intervals for N_t as a function of P_c. We state these well known results in a series of three related statements.

(1) The best estimate of the true probability P_c given N_t Bernoulli trials is the mean

$$\mu = \frac{1}{N_t}\sum_{n=1}^{N_t} X_n,$$

where X_n is the random variable of binary outcomes of the nth trial, with $X_n = 1$ when $h = s$ (a *hit*) and 0 otherwise (a *miss*).

(2) The standard deviation of the above estimator of the mean μ is

$$\sigma_\mu = \sqrt{\frac{P_c(1-P_c)}{N_t}}.$$

(3) According to the well-know *Vysochanskij-Petunin inequality* (http://en.wikipedia.org/wiki/Vysochanskii-Petunin_inequality), the 95% confidence interval of this estimator is given by $3\sigma_\mu$.

Because for the RZE group our estimate of P_c is one error of $N = 38$ trials, or less than 3%, we find the 95% confidence bound to be $P_c < 0.97 - 3\sigma_\mu$, or <88.7%. This is close to 4 of 38 errors (10.5%). Thus all the sounds in the HE group fall in the significant range.

APPENDIX B: A BRIEF HISTORY OF SPEECH CUE RESEARCH

Our finding that the plosive cue must be binary raises the question about the novelty of such a finding, which we address in the following text.

Nearly all studies on *acoustic consonant features* refer back to the early Haskins studies using the *pattern-playback* synthesis technique (Liberman, 1957; Liberman *et al.*, 1967). There were many problems with these early studies. By design, artificial speech contains only those features that are synthesized. This has led to a fundamental uncertainty as to the nature (even the existence) of these basic consonant features (Cole and Scott, 1974; Dorman *et al.*, 1977; Blumstein and Stevens, 1980; Kewley-Port, 1982; Kewley-Port *et al.*, 1983; Allen, 2005a). *Second*, this early synthetic "speech" was quite primitive (Dorman *et al.*, 1977), leading to frequent perceptual errors. *Third* the confusion sets were typically over small closed sets of sounds, such as /pa, ta, ka/ or /ba, da, ga/. While these early techniques were successful in identifying several *candidates* for acoustic speech cues, given the synthetic quality of the speech, the lack of the natural cues and variability, and the very low-entropy of the small closed-set task, they could not resolve what these speech cues might actually be. For example, the early Haskins studies first claimed the features to be onset bursts (Cole and Scott, 1974), but later ruled out this possibility, emphasizing instead on a complex set of coarticulation effects including formant transitions (Dorman *et al.*, 1977)

Perceptual invariance of stop consonants were then analyzed in the classic studies of Stevens and Blumstein (1978) and Blumstein and Stevens (1980), which again required the use of synthetic syllables, so that the various assumed acoustic cues (e.g., F2 transitions) could be carefully controlled. Their later studies arguably ruled out formant transitions as primary cues and emphasized the diffuse spectral envelope of bursts of energy around the formant onsets, which is strongly related to the formant frequency (typically F2) onset, as first studied by Cooper *et al.* (1952). While providing some insight, they conclude that the basic questions of the speech code remained unresolved (Blumstein and Stevens, 1979). A few years later Kewley-Port (1982);

Kewley-Port *et al.* (1983) did another series of experiments to demonstrate that the diffuse burst spectrum was not a cue rather that the transitions are the cues. Benkí (2001) studied the effects of place of articulation and F1 transition on CV and VCV stimuli generated using the Klatt synthesizer (Klatt and Klatt, 1990). Other important studies on synthetic stop consonants include Lisker (1975), Sumerfield and Haggard (1977), and Massaro and Oden (1980).

Not all studies used synthetic speech. Meaningful real speech is called *redundant* due to *context effects*, whereas maximum entropy CV, VC, and CVC sounds (so called nonsense speech) are not considered to be redundant and thus are special because they minimize the powerful side-channel effects of context (i.e., real speech improves guessing). Many key studies provide examples of confusions between such real speech sounds, leading to further conjectures of various consonant features used by normal and hearing impaired listeners (Dubno and Levitt, 1981; Bell *et al.*, 1986; Dubno *et al.*, 1987). But again, no strong conclusions regarding consonant features could be reached. In fact, many studies have concluded that perhaps the long largely unsuccessful quest for invariant acoustic features implies they do not exist, or that they exist in complex forms, tangled with complex coarticulations (Flanagan, 1965; Dorman *et al.*, 1977; Greenberg, 1999; McMurray and Jongman, 2011). Chen and Alwan (2006) and Jiang *et al.* (2006) explored /p, t, k, b, d, g/ in the presence of three vowels (/a/, /i/,/u/). Both studies used natural speech produced by two male and two female talkers. While they classify the errors in terms of the gender of the talkers, they did not discuss the differences between the two talkers having the same gender. Both studies reported that many of the syllables had 100% correct responses in the absence of noise. Such a saturation in score is called a *ceiling effect*.

It would be fair to say that the many arguments regarding speech features are far from mature, given the long standing controversy: Cole and Scott (1974) proposed a model in which invariant and transitional cues were integrated to explain the perception of syllables while envelope cues were used to model perception of higher order units like words; Dorman *et al.* (1977) found strong coarticulations; Greenberg (1999) involved syllabic and lexical elements to understand pronunciation variation at a syllable level; Diehl *et al.* (2004) pointed out relationships between speech perception and production.

Recent studies by the authors have sought new robust ways to identify consonant features in CV sounds, via semi-automatic methods, derived from large amounts of psychoacoustic data, using natural CV sounds with masking noise and large numbers of talkers and listeners (Phatak and Allen, 2007; Phatak *et al.*, 2008; Allen and Li, 2009; Li *et al.*, 2010; Li and Allen, 2011). Régnier and Allen (2008) and Li *et al.* (2010) provided strong evidence that phonetic binary features *do* exist but in a different form (Allen and Li, 2009; Li and Allen, 2011) than previously suggested (e.g.: Delattre *et al.*, 1955; Dorman *et al.*, 1977; Blumstein and Stevens, 1980; Delgutte, 1997; Kewley-Port, 1982; Kewley-Port *et al.*, 1983).

Another way of classifying utterances is to compare the distribution of *acoustic thresholds* (a physical critical-band SNR acoustic-feature measure denoted SNR_e) first defined by Régnier and Allen (2008) and further developed in Appendix A of Li *et al.* (2010) to the perceptual *event thresholds* (a psychological measure, denoted SNR_{90}), defined as the SNR for which the score drops from 100% to 90%. For example, f101pe is a robust utterance with its SNR_{90} at −16 dB, while m107pe is a weak utterance ($SNR_{90} \approx 5$ dB).

As the noise increases, the acoustic feature is masked as measured by the AI-based SNR_e, thus the syllable becomes confusable with the loss of its primary feature. Correspondingly, below SNR_{90} the score abruptly falls to chance performance within 6 dB, as shown in Fig. 8(a) and predicted by SNR_e.

Allen, J. (**1994**). "How do humans process and recognize speech?" IEEE Trans. Speech Audio Process. **2**(4), 567–577.

Allen, J. (**1996**). "Harvey Fletcher's role in the creation of communication acoustics," J. Acoust. Soc. Am. **99**(4), 1825–1839.

Allen, J. (**2004**). "The Articulation Index is a Shannon channel capacity," in *Auditory Signal Processing: Physiology, Psychoacoustics, and Models*, edited by D. Pressnitzer, A. de Cheveigné, S. McAdams, and L. Collet (Springer Verlag, New York), pp. 314–320.

Allen, J. (**2005a**). "Consonant recognition and the articulation index," J. Acoust. Soc. Am. **117**(4), 2212–2223.

Allen, J. B. (**2005b**). *Articulation and Intelligibility* (Morgan and Claypool, LaPorte, CO), pp. 124.

Allen, J. B., and Li, F. (**2009**). "Speech perception and cochlear signal processing," IEEE Signal Process. Mag. **26**(4), 73–77.

ANSI. (**1969**). S3.5 *American National Standard Methods for the Calculation of the Articulation Index* (American National Standards Institute, New York).

ANSI. (**1997**). S3.5 *Methods for Calculation of the Speech Intelligibility Index (SII-97)* (American National Standards Institute, New York).

Bell, T. S., Dirks, D. D., Levitt, H., and Dubno, J. R. (**1986**). "Log-linear modeling of consonant confusion data," J. Acoust. Soc. Am. **79**(2), 518–525.

Benkí, J. (**2001**). "Place of articulation and first formant transition pattern both affect perception of voicing in English," J. Phonetics **29**, 1–22.

Blumstein, S. E., and Stevens, K. N. (**1979**). "Acoustic invariance in speech production: evidence from measurements of the spectral characteristics of stop consonants," J. Acoust. Soc. Am. **66**(4), 1001–1017.

Blumstein, S. E., and Stevens, K. N. (**1980**). "Perceptual invariance and onset spectra for stop consonants in different vowel environments," J. Acoust. Soc. Am. **67**(2), 648–662.

Chen, M., and Alwan, A. (**2006**). "On the perception of voicing in syllable-initial plosives in noise," J. Acoust. Soc. Am. **119**(2), 1092–1105.

Ching, T. Y. C., Dillon, H., and Byrne, D. (**1998**). "Speech recognition of hearing-impaired listeners: Predictions from audibility and the limited role of high-frequency amplification," J. Acoust. Soc. Am. **103**, 1128–1140.

Cole, R., and Scott, B. (**1974**). "Toward a theory of speech perception," Psychol. Rev. **81**(4), 348–374.

Cooper, F., Delattre, P., Liberman, A., Borst, J., and Gerstman, L. (**1952**). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. **24**(6), 597–606.

Delattre, P., Liberman, A., and Cooper, F. (**1955**). "Acoustic loci and translational cues for consonants," J. Acoust. Soc. Am. **27**(4), 769–773.

Delgutte, B. (**1997**). "Auditory neural processing of speech," in *The Handbook of Phonetic Sciences*, edited by W. Hardcastle and J. Laver (Blackwell, Oxford), pp. 507-538.

Diehl, R., Lotto, A., and Holt, L. (**2004**). "Speech perception," Annu. Rev. Psychol. **55**, 149–179.

Dorman, M., Studdert-Kennedy, M., and Raphael, L. (**1977**). "Stop-consonant recognition: Release bursts and formant transitions as functionlly equivialent, context-dependent cues," Percept. Psychophys. **22**(2), 109–122.

Dubno, J. R., Dirks, D., and Schaefer, A. (**1987**). "Effects of hearing loss on utilization of short-duration spectral cues in stop consonant recognition," J. Acoust. Soc. Am. **81**(6), 1940–1947.

Dubno, J., Dirks, D., and Schaefer, A. (**1989**). "Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss. II. Articulation index predictions," J. Acoust. Soc. Am. **85**(1), 355–364.

Dubno, J. R., and Levitt, H. (**1981**). "Predicting consonant confusions from acoustic analysis," J. Acoust. Soc. Am. **69**(1), 249–261.
Flanagan, J. (**1965**). *Speech Analysis Synthesis and Perception* (Academic, New York).
Fletcher, H. (**1929**). *Speech and Hearing* (D. Van Nostrand, New York).
Fletcher, H. (**1950**). "A method of calculating hearing loss for speech from an audiogram," J. Acoust. Soc. Am. **22**, 1–5.
Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (**2004**). "New nonsense syllables database—analyses and preliminary ASR experiments," in *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, pp. 2749–2752.
French, N. R., and Steinberg, J. C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90–119.
Greenberg, S. (**1999**). "Speaking in shorthand—A syllable-centric perspective for understanding pronunciation variation," Speech Commun. **29**, 159–176.
Han, W. (**2011**). "Methods for robust characterization of consonant perception in hearing-impaired listeners," Ph.D. thesis, University of Illinois, Champaign, IL.
Huang, J., and Hasegawa-Johnson, M. (**2008**). "Maximum mutual information estimation with unlabeled data for phonetic classification," in *Proc. Interspeech*, Brisbane, Australia (International Speech Communication Association), pp. 952–955.
Humes, L., Dirks, D., Bell, T., and Ahlstbom, C. (**1986**). "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners," J. Speech Hear. Res. **29**, 447–462.
Jiang, M., Chen, J., and Alwan, A. (**2006**). "On the perception of voicing in syllable-initial plosives in noise," J. Ac **119**(2), 1092–1105.
Kapoor, A., and Allen, J. B. (**2012**). "Perceptual effects of plosive feature modification," J. Acoust. Soc. Am. **131**(1), 478–491.
Kewley-Port, D. (**1982**). "Measurement of formant transitions in naturally produced stop consonant-vowel syllables," J. Acoust. Soc. Am. **72**(2), 379–389.
Kewley-Port, D., Pisoni, D., and Studdert-Kennedy, M. (**1983**). "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," J. Acoust. Soc. Am. **73**(5), 1778–1793.
Killion, M., and Christensen, L. (**1998**). "The case of the missing dots: Al and SNR loss," Hear. J. **51**, 32–47.
Klatt, D., and Klatt, L. (**1990**). "Analysis, synthesis, and perception of voice quality variations among male and female talkers," J. Acoust. Soc. Am. **87**, 820–857.
Kryter, K. D. (**1962**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**(11), 1689–1697.
Li, F., and Allen, J. B. (**2009**). "Additivity law of frequency integration for consonant identification in white noise," J. Acoust. Soc. Am. **126**(1), 347–353.
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.
Liberman, A. (**1957**). "Some results of research on speech perception," J. Acoust. Soc. Am. **29**(1), 117–123.
Liberman, A., Cooper, F., Shankweiler, D., and Studdert-Kennedy, M. (**1967**). "Perception of the speech code," Psychol. Rev. **74**(6), 431–461.
Lippman, R. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–15.
Lisker, L. (**1975**). "Is it VOT or a first-formant transition detector?" J. Acoust. Soc. Am. **57**(6), 1547–1551.
Massaro, D., and Oden, G. (**1980**). "Evaluation and intergration of acoustic features in speech perception," J. Acoust. Soc. Am. **67**(3), 996–1013.
McMurray, B., and Jongman, A. (**2011**). "What information is necessary for speech categorization? Harnessing variability in the speech signal by integrating cues computed relative to expectations," Psychol. Rev. **118**(2), 219–246.
Miller, G., and Nicely, P. (**1955**). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. **27**, 338–352.
Pavlovic, C., Studebaker, G., and Sherbecoe, R. (**1986**). "An articulation index based procedure for predicting the speech recognition performance of hearing-impaired individuals," J. Acoust. Soc. Am. **80**, 50–57.
Phatak, S., and Allen, J. (**2007**). "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**(4), 2312–2326.
Phatak, S., Lovitt, A., and Allen, J. (**2008**). "Consonant confusions in white noise," J. Acoust. Soc. Am. **124**(2), 1220–1233.
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.
Rankovic, C. (**1991**). "An application of the articulation index to hearing aid fitting," J. Speech Hear. Res. **34**, 391–402.
Régnier, M., and Allen, J. (**2008**). "A method to identify noise-robust perceptual features: application for consonant /t/," J. Acoust. Soc. Am. **123**(5), 2801–2814.
Ronan, D., Dix, A., Shah, P., and Braida, L. D. (**2004**). "Integration across frequency bands for consonant identification," J. Acoust. Soc. Am. **116**, 1749–1762.
Scharenborg, O. (**2007**). "Reaching over the gap: A review of efforts to link human and automatic speech recognition research," Speech Commun. **49**, 336–347.
Shannon, C. E. (**1948**). "A mathematical theory of communication," Bell Syst. Tech. J. **38**, 611–656.
Sroka, J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Commun. **45**, 410–423.
Stevens, K., and Blumstein, S. (**1978**). "Invariant cues for place of articulation in stop consonants," J. Acoust. Soc. Am. **64**, 1358–1368.
Sumerfield, Q., and Haggard, M. (**1977**). "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants," J. Acoust. Soc. Am. **62**(2), 435–448.
Yoon, Y., Allen, J., and Gooler, D. (**2012**). "Relationship between consonant recognition in noise and hearing threshold," J. Speech Lang. Hear. Res., Doi: 10.1044/1092-4388(2011/10-0239).