Consonant confusions in white noise

Sandeep A. Phatak,[a] Andrew Lovitt, and Jont B. Allen
ECE Department and the Beckman Institute for Advanced Science and Technology, University of Illinois at Urbana-Champaign, Illinois 61801

(Received 27 September 2007; revised 24 March 2008; accepted 3 April 2008)

The classic [MN55] confusion matrix experiment (16 consonants, white noise masker) was repeated by using computerized procedures, similar to those of Phatak and Allen (2007). ["Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2316]. The consonant scores in white noise can be categorized in three sets: low-error set {/m/, /n/}, average-error set {/p/, /t/, /k/, /s/, /ʃ/, /d/, /g/, /z/, /ʒ/}, and high-error set {/f/, /θ/, /b/, /v/, /ð/}. The consonant confusions match those from MN55, except for the highly asymmetric voicing confusions of fricatives, biased in favor of voiced consonants. Masking noise cannot only reduce the recognition of a consonant, but also perceptually morph it into another consonant. There is a significant and systematic variability in the scores and confusion patterns of different utterances of the same consonant, which can be characterized as (a) *confusion heterogeneity*, where the competitors in the confusion groups of a consonant vary, and (b) *threshold variability*, where confusion threshold [i.e., signal-to-noise ratio (SNR) and score at which the confusion group is formed] varies. The average consonant error and errors for most of the individual consonants and consonant sets can be approximated as exponential functions of the articulation index (AI). An AI that is based on the peak-to-rms ratios of speech can explain the SNR differences across experiments.
 [DOI: 10.1121/1.2913251]

PACS number(s): 43.71.An, 43.71.Es, 43.66.Dc, 43.72.Dv [MSS] Pages: 1220–1233

I. INTRODUCTION

Masking experiments play a crucial role in understanding the perceptual features of elemental speech sounds. Masking one or more of these features, defined as *events*, leads to a perceptual confusion (Régnier and Allen, 2008). Events are different from, though related to, other commonly used categories of speech features such as articulatory features (place, manner, etc.) or acoustic features (spectrum, temporal modulations, etc.). These features, which are extracted from the signal by auditory system, form the basis for perception of different speech sounds. Events, and their acoustic correlates, can be identified by directly comparing the perceptual confusions with the corresponding masked speech stimuli, on an utterance by utterance basis (Régnier and Allen, 2008). Such comparisons require a quantitative analysis of both the perceptual confusions and the speech stimuli.

We use the confusion matrix (CM), which is an important analytical tool for quantifying the results of closed-set recognition tasks, to characterize the nature of perceptual confusions (Allen, 2005a). The classic Miller and Nicely (1955) [MN55] study, which used CMs for measuring consonant confusions for noise-masked and filtered speech, has inspired many subsequent noise-masking CM experiments (Wang and Bilger, 1973); Dubno and Levitt, 1981; Gordon-Salant, 1985; Grant and Walden, 1996; Sroka and Braida, 2005). Phatak and Allen (2007) [denoted here as PA07[1]] used confusion patterns and confusion thresholds, first defined by Allen (2005b), in a quantitative analysis of the CM. PA07 employed large numbers of talkers and listeners to take advantage of the large natural variability in speech production and perception. Many questions raised in PA07 remained open due to large dimensionality ($16C \times 4V \times 18$talkers $\times 10$listeners $\times 6$SNR) and relatively low CM row sums (N). For example, are the differences between PA07 and MN55, such as the asymmetric voicing confusions of PA07 solely due to different noise spectra or due to procedural differences? Are talker and listener variations in perceptual confusions systematic or random? Do these variations, if present, correlate with the variations in speech stimuli?

To answer these and other outstanding questions, a CM experiment was conducted by using procedures similar to PA07, but with a white noise masker, as in MN55. We will refer to this experiment as MN05. One of the main purposes MN05 was to to verify whether the results of the classic MN55 study can be reproduced, which can be considered a validation of the PA07 procedures. To achieve this, the procedures of MN05 were designed to match as close as possible to MN55 procedures by making the least possible changes to PA07 procedures. A three-way comparison among MN55, MN05, and PA07 will let us estimate the effect of the noise spectrum on consonant perception by ruling out the effects of procedural differences, such as the use of recorded speech stimuli, male talkers, digital filters, and computerized presentations. Table I lists the relevant details of the three experiments.

The MN05 data also allows analyses that were not possible with MN55 or PA07 data. For example, unlike MN55 the speech stimuli are now available in MN05 to compare with the consonant confusions in white noise (WN). Such

[a]Authors present address: Army Audiology and Speech Center, Walter Reed Army Medical Center, Washington, D.C. 20307

TABLE I. Experimental details for Miller and Nicely (1955) [MN55], Phatak and Allen (2007) [PA07], and the current experiment [MN05]. The details include number of consonants (C), number of vowels (V), number of talkers (T), noise spectrum, and the speech database used.

Experiment	C	V	T	Noise spectrum	Speech stimuli
MN55	16	1	5	White (WN)	Live talkers
PA07	16	4	18	Speech-weighted (SWN)	LDC2005S22
MN05	16	1	18	White (WN)	LDC2005S22

correlations are crucial in establishing the noise-robust acoustic correlates of the perceptual features of speech (Régnier, 2007). Furthermore, the MN55 data were pooled over listeners and talkers, which averages out the possible talker and listener variations that are important for finding noise-robust features. Phatak and Allen (2007) also showed that the articulation index (AI), based on the peak-to-rms ratios of the speech corpus, can be used to parametrize consonant errors. Such analysis can be tested for WN with the present experimental data, but not with the masking data of MN55 due to unavailability of the stimuli.

The long-term goal of our studies is to determine the noise-robust features of basic speech sounds. PA07 and MN05 are the first two experiments in a series of data-collection experiments intended toward achieving this goal. Identifying the perceptual features quantitatively requires comparing the perceptual data collected in these experiments with the corresponding stimuli [Régnier and Allen (2008)]. This paper presents ways to quantify the perceptual data, which is the first step toward such comparisons.

II. METHODS

The testing procedures from the study of PA07 were modified to optimally match the methods of MN55. The speech stimuli were CV syllables with the 16 MN55 consonants followed by vowel /ɑ/, from the LDC2005S22 corpus (Fousek *et al.*, 2004). The syllables used in this study were spoken in isolation by 18 talkers (ten males and eight female). All talkers were native speakers of U.S. English, but three talkers were bilingual and had a part of their upbringing outside the U.S./Canada. MN55 used only female subjects, with one serving as talker, while the other four served as listeners. Since no significant talker-gender differences were observed by PA07, both male and female talkers were used in MN05.

The CV tokens were normalized such that each talker had the same average rms level. Random WN was added to the speech at five different signal-to-noise ratios (SNRs), viz., −12, −6, 0, 6, and 12 dB. When a listener had consonant scores significantly above chance level at −12 dB, then those consonants were presented to that listener at −15 dB, and again at −18 and −21 dB SNRs, if required. Data indicate that all listeners reached −15 and −18 dB SNRs, but rarely reached −21 dB SNR. The SNR was set for each token by using VUSOFT, a software VUmeter (Lobdell and Allen, 2007). The peak value of the VUSOFT output was used to define the speech level for each CV syllable. The speech and noise were filtered to have a bandwidth of 200–6500 Hz to match that in the MN55 experiment. Additionally, the CVs were presented in the quiet condition (i.e., no noise masker) as a control condition.

The stimuli were diotically presented to listeners through headphones (Sennheiser HD280). The listener reported the heard sound by clicking the appropriate choice on a computer screen. Unlike PA07, MN05 did not involve vowel recognition. Therefore, the MATLAB graphic user interface used in PA07 was modified to provide only 16 consonant choices. Consistent with MN55, presentations were randomized over consonants, but not over talkers or SNR. Thus, 18 CVs spoken by the same talker were successively presented at a fixed SNR. The set of 18 CVs for each talker consisted of 16 possible CVs, plus two of those randomly chosen, to limit the possibility of guessing by listener. The talker and the SNR for the next block of trials were then randomly chosen.

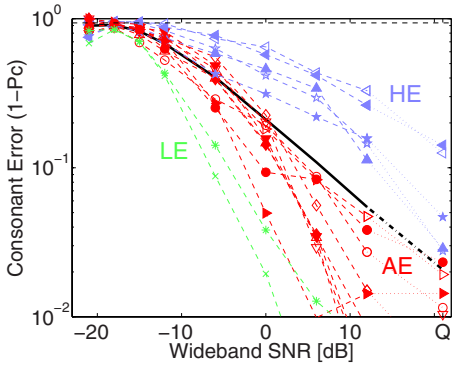
24 listeners (16 males and 8 females) having English as their primary language completed the experiment. The listeners were normal-hearing adults with no history of hearing problems. Three listeners had ages of 36, 45, and 50 years, while the remaining listeners were in the age group of 18–28 years (μ=21.57 yr, σ=2.32 yr). 21 listeners were born and broughtup in U.S. and self-reported to have a mid-western accent. The remaining three listeners had a part of their upbringing in India, South Korea, and China and reported to have South Asian, Southern U.S., and Chinese accents, respectively. However, no significant differences were observed in consonant scores and confusions of these three listeners and those of other listeners, and hence their responses were included. Each listener was trained for about 1 h in the quiet condition with visual feedback.

III. RESULTS

A. Listener and utterance selection

By following the analysis method of PA07, a post-hoc listener and utterance selection were carried out on the data of MN05. Listener selection is necessary to ensure that the listeners are attending to the task and that their scores are comparable to that of an average normal listener. The utterance selection is required to avoid misinterpreting errors due to mislabeled or mispronounced utterances as noise-induced confusions. The error thresholds used for this selection were same as those used in PA07.

23 of the 24 listeners had scores greater than 85% in the quiet condition and formed a homogeneous group (μ =92.4%, σ=3.5%). The responses of the one low-performance listener, who had 78.8% score in quiet, were removed from the dataset.



Consonant	Marker	Consonant	Marker
p	△	b	▲
t	◇	d	◆
k	▽	g	▼
f	☆	v	★
θ	◁	ð	◀
s	○	z	●
ʃ	▷	ʒ	▶
m	✳	n	✕

FIG. 1. (Color online) The left panel shows consonant errors $P_e(\mathrm{SNR})=1-P_c(\mathrm{SNR})$ on a log scale, plotted against the wideband SNR in decibels. The solid line shows the average consonant error, while the colored dashed lines with marker symbols are for individual consonants. The three consonant sets: low error (LE), average error (AE), and high error (HE) are color coded. The legend on the right lists the markers used for consonants. The quiet condition is denoted by Q and is plotted at +21 dB. The horizontal dashed line at the top is the chance level error of $1-1/16=15/16$.

Based on the responses of the final set of 23 listeners, the syllable error for each utterance, which is same as the consonant error in this case, was estimated in the quiet condition. 32 of the total of 286 utterances had more than 20% error in quiet and were therefore considered as "ambiguous" utterances. Accordingly, the responses to these utterances were removed from the database. Following this utterance selection, the one low-performance listener had a score of 83.8%, while all other listeners formed a tight group with mean score of 97.9% and standard deviation of 1.2%. Thus, the listener categorization was verified to be unaffected by the utterance selection.

B. Consonant Errors

Figure 1 shows the consonant errors $P_e(\mathrm{SNR})=1-P_c(\mathrm{SNR})$ as a function of SNR. These curves can be categorized into three sets—a low-error (LE) set {/m/, /n/}, an average-error (AE) set {/p/, /t/, /k/, /s/, /ʃ/, /d/, /g/, /z/, /ʒ/}, and a high-error (HE) set {/f/, /θ/, /b/, /v/, /ð/}. These three sets are different from the three consonant sets observed in PA07, due to different noise spectra. The HE set C1 ={/f/, /θ/, /b/, /v/, /ð/, /m/} of PA07 differs from the HE set only by one addition consonant /m/. However, the other two sets are quite different in the two experiments. Such distinct sets were not observed in the $P_e(\mathrm{SNR})$ curves of MN55, except for the nasals /m/, /n/, which had the lowest errors in MN55, consistent with the LE set.

For comparing our scores with the original Miller–Nicely experiment, we find the SNRs required to achieve the same score in the two experiments. These SNRs, i.e., $\mathrm{SNR}_{\mathrm{MN55}}(P_e)$ and $\mathrm{SNR}_{\mathrm{MN05}}(P_e)$, obtained for a range of P_e values, are plotted against each other to obtain the isoperformance SNR contours in Fig. 2(a). For those P_e values which fall between the measures P_e values, the SNRs are estimated by linearly interpolating the $P_e(\mathrm{SNR})$ curves. The dashed curves with markers represent individual consonants, while the thick dash-dotted curve corresponds to the average performance. The thin dashed "reference" line with a slope of 45° corresponds to identical performance in the two experiments. A consonant curve above this dashed line implies that the higher SNR was required in MN05 (ordinate) than in MN55 (abscissa) to achieve the same performance for that consonant. In other words, the consonants that have curves above the reference line have poorer performance in MN05 than in MN55, at a given SNR. A curve below the reference line indicates a better performance in MN05 than in MN55. The proximity of the average performance curve (thick dash dotted) to the reference line in Fig. 2(a) implies that the average consonant performance in MN05 was almost equal to that in MN55. On average, LE and AE consonants performed better, while HE consonants performed slightly worse in MN05 as compared to MN55.

Figure 2(b) shows a similar comparison between MN05 (WN) and PA07 (SWN). There is a 10–12 dB uniform difference between the average scores (dash-dotted line). All consonants have poorer performance in WN relative to SWN, but the precise difference in the performance depends on consonant and varies with SNR. The consonants in set C2

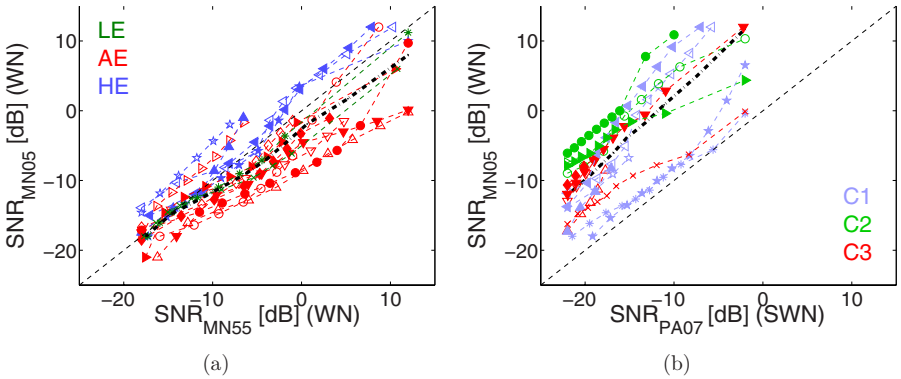


FIG. 2. (Color online) The isoperformance SNR contours for comparing MN05 scores with (a) MN55, and (b) PA07. In both panels, SNRs from MN05 form the ordinate. In (a), the individual consonant contours are color coded according to the three consonant sets LE, AE, and HE. In (b), the color scheme follows the three sets from PA07, i.e., the high-error set C1, the average-error set C3, and the low-error set C2.

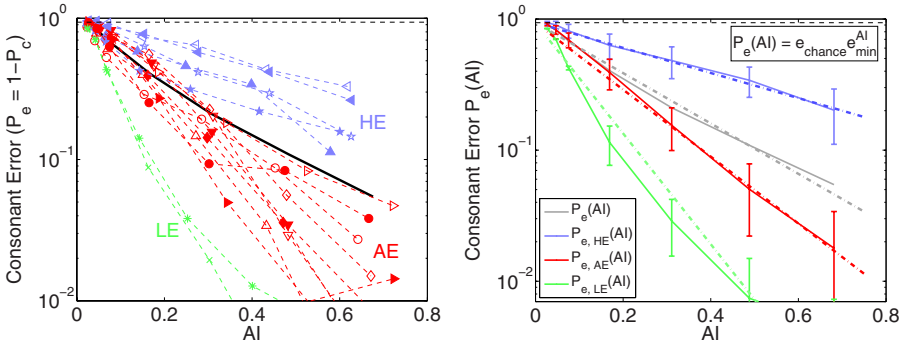


FIG. 3. (Color online) (a) The consonant errors $P_e(\mathrm{AI})=1-P_c(\mathrm{AI})$ on a log scale plotted as a function of AI. The AI values were calculated at each SNR, except for the quiet condition, where the exact SNR was not known. (b) The average consonant error (gray) and the average errors for LE, AE, and HE sets. The solid curves are the empirically measured errors, with errorbars indicating the standard deviations within the sets, while the dash-dotted lines are the predictions of the exponential AI model [Eq. (1)].

from PA07 (i.e., /s/, /ʃ/, /z/, /ʒ/, and /t/) have the largest decrease in performance in MN05 with respect to the PA07 performance. This is expected, because these consonants had highest scores in speech-weighted noise, due to high-frequency energy (PA07). The frequencies above 2 kHz have significantly higher masking in WN than in SWN, resulting in poorer scores for these high-frequency fricatives. On the other hand, consonants /v/, /m/ (both set C1), and /n/ (set C3) have the least decrease in performance, as most of the energy for these consonants is concentrated at low frequencies where the spectra of the two noises are not very different.

1. AI

Allen (2005b) showed that the consonant log errors [i.e., $P_e(\mathrm{AI})$ on log scale] for the MN55 data are linear functions of the AI, following the exponential model:

$$P_e(\mathrm{AI}) = e_{\mathrm{chance}} e^{\mathrm{AI}}_{\mathrm{min}}, \tag{1}$$

from Allen (1994), where e_{min} is minimum error (at AI=1) and e_{chance} is the chance performance error (at AI=0). In this case, $e_{\mathrm{chance}}=1-1/16=15/16$. Figure 3(a) shows that the average consonant log error (thick solid line) in MN05 also linearly decreases with AI, in accordance with the model.

Equation (1), which is based on Fletcher's *band-independence* theory, was defined only for average speech [Fletcher and Galt, 1950; Allen (2005a)]. However, the linearity of individual consonant curves in Fig. 3(a) demonstrates that the model also works for individual consonants, as previously observed by Allen (2005b) and PA07. It follows that the log-error curves for consonants can be expressed as

$$P_e(\mathrm{AI}, C_i) \approx e_{\mathrm{chance}} e^{\mathrm{AI}_i}_{\mathrm{min}_i}, \tag{2}$$

where e_{min_i}, and AI_i are the e_{min} and the AI values, respectively, for consonant C_i.

The three consonant sets are more obvious on an AI scale than on a SNR scale (Fig. 3). Figure 3(b) shows that not only average consonant log error (gray) but also the log errors for sets HE and AE are also linear functions of AI. The curvature in the $P_e(\mathrm{AI})$ for set LE is due to only one of the two consonants in that set, viz., /m/. The $\log[P_e(\mathrm{AI})]$ curves for 13 out of the 16 consonants can be matched to straight lines with very LE. This shows that the exponential AI model can be extended beyond the average consonant score, to consonant groups, and even individual consonants.

The average consonant error is a Bayesian sum of individual consonant errors and therefore, according to Eq. (2), can be expressed as a sum of exponential functions of AI, with different bases (i.e., the e_{min_i} values).

$$P_e(\mathrm{AI}) = \sum_{i=1}^{16} P_e(\mathrm{AI}, C_i) \approx e_{\mathrm{chance}} \sum_{i=1}^{16} e^{\mathrm{AI}_i}_{\mathrm{min}_i}. \tag{3}$$

Combining the two expressions for $P_e(\mathrm{AI})$ from Eqs. (1) and (3) results in

$$e^{\mathrm{AI}}_{\mathrm{min}} \approx \sum_{i=1}^{16} e^{\mathrm{AI}_{C_i}}_{\mathrm{min},c_i}. \tag{4}$$

A sum of exponentials cannot be an exponential, unless the bases are equal, but in this case the approximation fits well.

Figure 4 shows the consonant scores for the three different experiments, on SNR, and AI scales. The $P_c(\mathrm{AI})$ curves

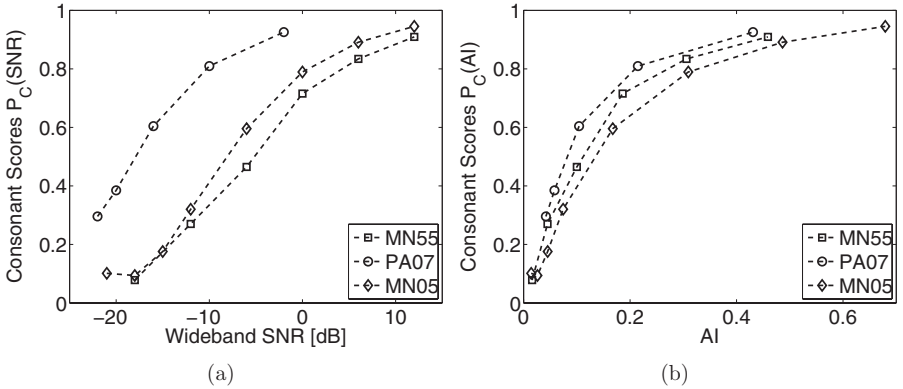


FIG. 4. A comparison of the consonant scores in the current study with those from [MN55] and PA07, plotted as a function of (a) SNR and (b) AI.

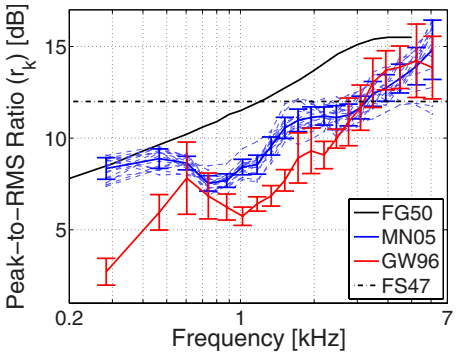


FIG. 5. (Color online) A comparison of the peak-to-rms ratios (r_k) used in different studies. The black solid curve (no errorbars) shows r_k values estimated by Fletcher and Galt (1950) [FG50]. The 16 dashed curves show r_k values for individual consonants, estimated from the CV tokens used in MN05. The means and standard deviations of r_k values for MN05 and Grant and Walden (1996) [GW96] stimuli are shown by the solid curves with error bars. The horizontal dash-dotted line shows the constant 12 dB peak-to-rms ratio used by French and Steinberg (1947) [FS47].

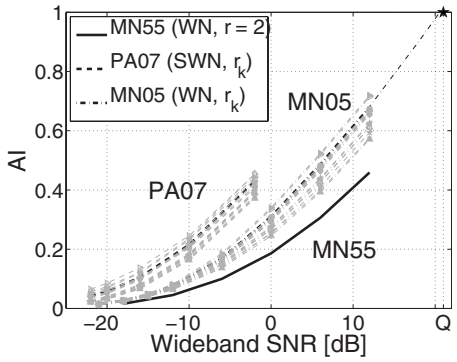


FIG. 6. The relationship between AI and wideband SNR for MN55 (solid line), PA07 (dashed), and MN05 (dash-dotted). The thin grayed-out lines are for individual consonants, while the thick black lines are for the average speech. The AI values for MN55 were estimated using $r=2$, while the frequency-dependent r_k values were estimated for average speech as well as for each consonant in case of PA07 and MN05. The present experiment curve, when extrapolated by using a third order polynomial fit, reaches AI $=1$ at about 21 dB SNR. The quiet condition (Q), which has AI$=1$ by definition, is plotted at this SNR.

are closer to each other than P_c(SNR) curves. This is because the differences in the noise spectra are accounted for by the AI, thus aligning the P_c(AI) curves for speech-weighted and WN. In spite of the same noise type, the AI values for MN55 and MN05 are not identical for a give SNR. The differences are due to the speech spectra and peak-to-rms ratios used in the AI calculation for the two experiments.

The AI values for MN05 were estimated by using the following PA07 formula.

$$\mathrm{AI} = \frac{1}{K}\sum_{k=1}^{K} \min\left[\frac{1}{3}\log_{10}(1 + r_k^2 \mathrm{snr}_k^2), 1\right], \tag{5}$$

where snr_k is the SNR and r_k is the peak-to-rms ratio (both in linear units, not in decibels) in the kth band, out of total $K=20$ articulation bands. The AI values for each consonant are estimated by using the average speech spectrum and average peak-to-rms ratios for that consonant. The details of calculating r_k values can be found in Appendix A of PA07. In this case, the peak-to-rms ratios varied from 2.19 ($\approx$6.89 dB) for /n/ in the 645–795 Hz articulation band to 6.65 ($\approx$16.45 dB) for /d/ in the 5720–7000 Hz band. Figure 5 compares peak-to-rms ratios (r_k) from the current study to those for the VCV syllables from Grant and Walden (1996) [GW96] study and with the r_k values reported by Fletcher and Galt (1950) [FG50]. The r_k values reported by FG50 were derived from the conversational speech data of Dunn and White (1940) and are frequently used as a standard for peak-to-rms correction in calculating AI (Pavlovic, 1984; Rankovic, 2002). The r_k values of MN05 and GW96 are lower than the FG50 values. The peak-to-rms ratios for GW96 are lower than those for MN05 below 3 kHz. This may be because GW96 stimuli (VCV) had two vowels per consonant, while MN05 stimuli (CV) had only one vowel per consonant. The steady and strong vowel formants, which dominate the envelope at these lower frequencies, significantly contribute to the rms value, but not so much to the peak value. The resultant would be lower peak-to-rms ratios for VCVs than for CVs. All three curves shows a significant variation in r_k over frequency, contrary to the claim by French and Steinberg (1947) that $r_k \approx 12$ dB, constant across frequency (horizontal dashed line). They made no claims regarding the variation in r_k across consonants. The dashed curves, which represent individual consonant, show that there is up to 5 dB variation in the r_k values over consonants, especially at higher frequencies.

Figure 6 shows a comparison of AI values, as a function of SNR, for the three experiments. The AI values for MN05 and PA07 were calculated using the spectra and peak-to-rms ratios that were directly estimated from the speech and noise stimuli. Since the speech and noise for MN55 was not available, the AI values for MN55 were calculated by using the straight-line approximation to the Dunn and White (1940) speech spectrum and a constant, frequency-independent peak-to-rms ratio of $r_k=2$ [Allen (2005b)]. At a given SNR, MN05 AI is higher that of the MN55 AI. This difference is predominantly due to the differences in peak-to-rms ratios, rather than the differences in the speech spectra. When the AI values for MN05 are calculated using a constant $r_k=1.7$, the average AI curve for MN05 coincides with the MN55 curve.

C. Consonant confusions

We use *confusion patterns* (CPs) [Allen (2005a)] to analyze the consonant confusions. A CP for a speech sound is obtained by plotting the row of that sound in CM against the SNR. Unlike the tabular form of CM, the formation of consonant groups over a range of SNRs can be directly observed in the CP. The confusion groups are not obvious in a CM table without a specific order of rows and columns, while the CPs do not depend on row and column orders. For example, consider the CP for consonant /t/ from the MN55 data shown in Fig. 7. Each curve corresponds to a particular column entry (h) for the /t/ row, plotted as a function of SNR, namely, $P_{h|/t/}$(SNR). The horizontal dashed line indicates chance, defined as the probability of guessing, which is 1/16. The diagonal entry $P_{/t/|/t/}$(SNR), denoted by $\diamond$, increases with SNR. As the SNR decreases, confusions of /t/ with /p/ ($\triangle$) and /k/ ($\triangledown$) increase and eventually become equal to the

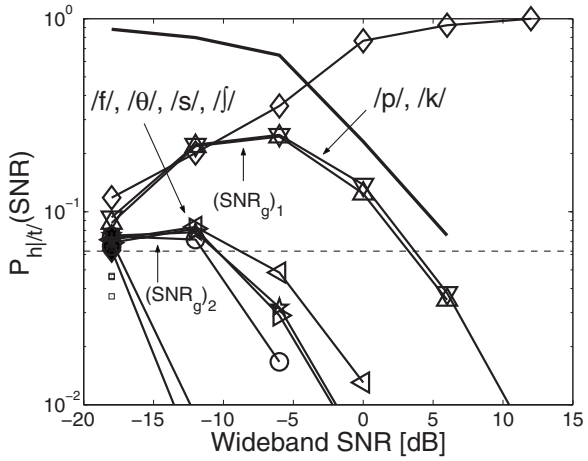
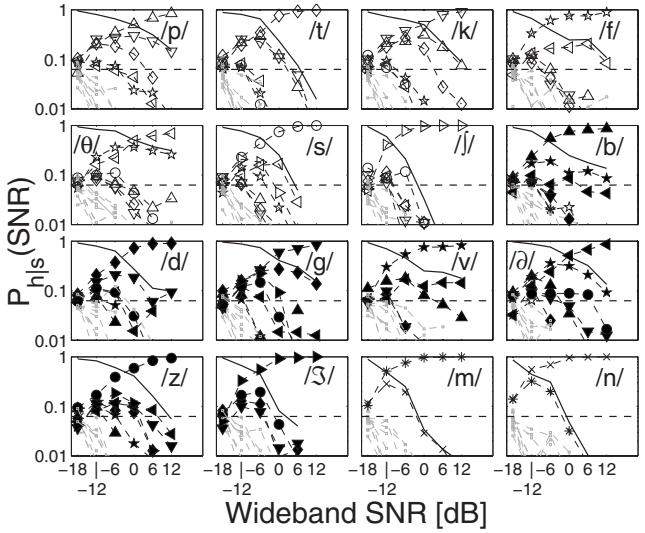


FIG. 7. Confusion patterns (CPs) for $s=/ta/$ from [MN55]. The thick solid line without markers is $1-P_{/b/|/t/}(\mathrm{SNR})$, which is the sum of off-diagonal entries. The horizontal dashed line shows the chance level of 1/16. Weak competitors, which do not exceed the chance performance, are shown by the gray square symbols.

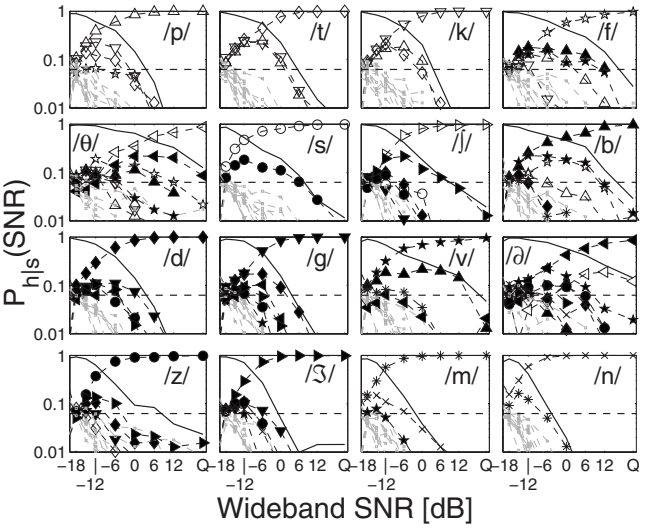
target for SNRs below −8 dB. We say that /t/, /p/, and /k/ form a *confusion group* (or *perceptual group*) at (or near) the *confusion threshold*, indicated by $(\mathrm{SNR}_g)_1 \approx -8$ dB, where $(\mathrm{SNR}_g)_1$ is the point of local maximum in $P_{/p/|/t/}(\mathrm{SNR})$ and $P_{/k/|/t/}(\mathrm{SNR})$ curves. When the SNR is decreased below $(\mathrm{SNR}_g)_2 \approx -15$ dB, consonant group [/f/, /θ/, /s/, and /ʃ/] merges with the [/t/, /p/, /k/] group, forming a supergroup. Since $(\mathrm{SNR}_g)_2 < (\mathrm{SNR}_g)_1$, consonants [/p/, /k/] are perceptually closer to /t/, and thereby form a stronger perceptual group with /t/ than the consonants [/f/, /θ/, /s/, and /ʃ/]. Thus we use the confusion threshold SNR_g as a quantitative measure to characterize the hierarchy in the perceptual confusions.

Figure 8(a) shows all 16 CPs for noise-masking data from MN55. Many confusion groups are not symmetrical. For example, the confusion of /θ/ with /f/ (☆ in second row, left panel) is significantly greater than confusion of /f/ with /θ/ (◁ in top right panel). Thus, the /f/-/θ/ confusion group is biased toward /f/. Allen (2005b) symmetrized the CMs, assuming these asymmetries to be insignificant. While the asymmetries for the /p/-/t/-/k/ and /m/-/n/ groups [the examples considered in Allen (2005b)] are negligible, for other confusion groups in MN55, such as the /f/-/θ/ group, these asymmetries are significant. These asymmetries are important for understanding the perceptual grouping of consonant under noisy conditions. Therefore, the CM should not be symmetrized.

Figure 8(b) shows the same 16 consonant CPs for MN05. These CPs are generated from the CM tables listed in the Appendix. The present experiment consonant confusions for plosives and nasals are very similar to those in MN55. The strong /p/-/t/-/k/, /d/-/g/-/z/ and /m/-/n/ confusion groups are common between the two experiments. However, the confusion thresholds in MN05 are at lower SNRs than those in MN55, indicating that the white noise has greater masking in MN55 than MN05. Part of this difference may be due to differences in the definition of SNR in the two experiments. To set the SNR in MN05, both speech and noise levels were



(a)



(b)

FIG. 8. The 16 consonant confusion patterns (CPs) for (a) the noise-masking data from [MN55] and (b) MN05. The horizontal dashed line shows the chance performance probability of 1/16. The weak competitors [i.e., $P_{h|s}(\mathrm{SNR}_g) < 1/16$] are grayed out for better visualization of the confusion groups. The quiet condition in MN05 (Q) is plotted at +21 dB SNR.

digitally measured by using a software VUMETER, whereas in MN55, the noise level (rms) was electrically measured and the speech level (peak) was measured by using a VUMETER instrument. Other possible factors, which cannot be tested with current data, could be the differences in speech stimuli (live talkers versus recorded) and familiarity of listeners with talkers in MN55.

A striking difference between the two experiments is observed in the fricative CPs. In MN55, consonants have negligible voicing errors, i.e., the unvoiced consonants have unvoiced competitors (hollow symbols) and the voiced consonants have voiced competitors (filled symbols). The unvoiced fricatives form /f/-/θ/ and /s/-/ʃ/ groups and their voiced counterparts have the corresponding /v/-/ð/ and /z/-/ʒ/ groups. These MN55 fricative groups are across place, but not across the voicing. In contrast, the fricatives in MN05 show significant voicing errors, but these voicing confusions

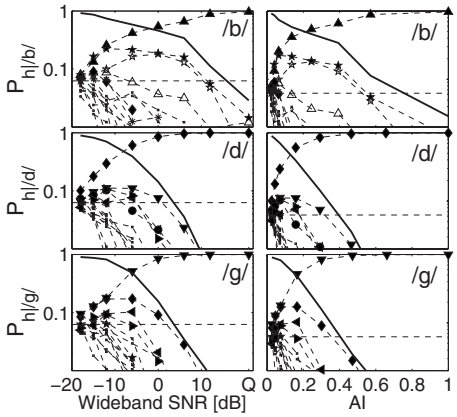


FIG. 9. The present experiment CPs for consonants /b/ (top), /d/ (center), and /g/ (bottom), as a function of SNR (left) and AI (right).

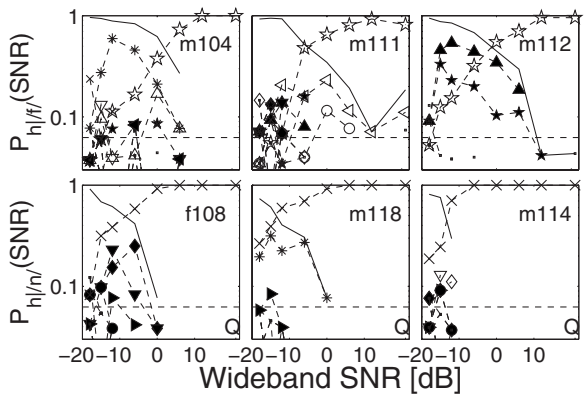


FIG. 10. Examples of confusion heterogeneity. The top row shows CPs for three utterances of /fa/: m104 (left), m111 (center), and m112 (right). The bottom row shows CPs for three utterances of /na/: f108 (left), m118 (center), and m114 (right). (m: male; f: female).

are biased in favor of voicing. That is, the strongest competitors for unvoiced consonants (hollow symbols) are voiced consonants (filled symbols), but not vice versa. For example, /s/ and /ʃ/ are only confused with /z/ and /ʒ/, respectively, but /z/ and /ʒ/ are hardly confused with any unvoiced consonant. The voiced fricative /v/ forms no confusion group with the unvoiced counterpart /f/ (in $P_{h|i/}$), but it is one of the strongest competitors of /f/ (in $P_{h|/f/}$). An interesting behavior is observed for consonant /ð/ in MN05. It is confused with /θ/, but only at SNR above 0 dB. At lower SNRs, it forms confusion groups with /v/ and /z/.

In MN55, consonants /p/ and /k/ form a stronger confusion group with each other than with /t/. Comparatively, the /p/-/t/-/k/ group is more symmetrical in MN05 and the three consonants equally compete with each other. In MN05, /p/ forms a weak group with /f/, but not with /θ/. Thus, MN05 data show /p/-/f/ and /f/-/θ/ groups, but do not show the /p/-/f/-/θ/ group from MN55. Similarly, /b/-/v/ and /v/-/ð/ groups are observed in MN05, but not the /b/-/v/-/ð/ group from MN55. On the other hand, some place confusions from MN05, such as /f/-/b/ and /v/-/m/, are not observed in MN55.

1. AI

As shown in Sec. I, the log-error curves P_e(SNR) for individual consonants become linear on an AI scale [i.e., $\log[P_e(\mathrm{AI})] = \mathrm{AI}\log(e_{\min}) + \log(e_{\mathrm{chance}})$]. In this section, we investigate how the abscissa transformation from SNR to AI impacts the confusion patterns. Figure 9 shows the CPs for consonants /b/, /d/, and /g/, as a function of SNR (left) and AI (right). On the SNR scale, the consonant log errors (thick solid lines) have significant curvature. The slope of log-error curve changes as the number of competitors decreases with increasing SNR. The nonlinear SNR to AI transformation compresses the higher confusion regions into a small AI range. On AI scale, all confusion thresholds in the consonant CPs are compressed to $\mathrm{AI} \leqslant 0.2$. The remaining range of AI from 0.2 to 1 has only a small number of competitors, with linearly decreasing log confusions (i.e., log of the off-diagonal elements). The resultant is a linearly decreasing log error for the consonant.

D. Utterance variability

The availability of the confusion data for individual utterance allows us to analyze the utterance variability. In PA07, the utterance CPs were not analyzed because the row sums were too small to reliably analyze a 64×64 CM or even the 16×16 CM. In MN05, the number of listeners is more than twice the number of listeners in PA07, which gives relatively smoother CPs. The row sums for individual utterances in MN05 are slightly greater than the number of listeners (i.e., 23) because some utterances were presented more than once in a block to minimize listener guessing.

There is a significant variation in the recognition scores of different utterances of the same CV. Equally interesting and more complex are the variations observed in the distribution of confusions errors. We cannot directly attribute these variations to either the talker variability or to the within-talker utterance variability because only one utterance of a CV was available from each talker in the LDC database.

The variations in the utterance CPs can be broadly classified into two categories. First is the *confusion heterogeneity*, where the competitors in the confusion group vary from utterance to utterance. Second is *threshold variability*, where the confusion group remains the same, but the SNR and confusion probability at the confusion threshold are utterance dependent.

1. Confusion heterogeneity

The top row of Fig. 10 shows CPs for three different utterances of /fɑ/. In each case, /f/ is confused with different consonants. Talker m104's /f/ is confused mostly with /n/ and somewhat with /p/, while m111 /f/ is confused with /θ/ and /s/. Utterance m112 /f/ forms only one but strong confusion group with /b/ and /v/. The bottom row of Fig. 10 shows CPs for three utterances of /nɑ/. While utterance f108 /n/ forms confusion groups with /d/ and /g/, m118's /n/ is almost solely confused with /m/. Talker m114's /n/ is very robust with no errors at $\mathrm{SNR} \geqslant -6$ dB, and below that SNR it has no strong competitor, but many weak competitors that never exceed a 15% confusion probability.

Morphing. When a confusion significantly exceeds the recognition of the presented sound, such as that top left

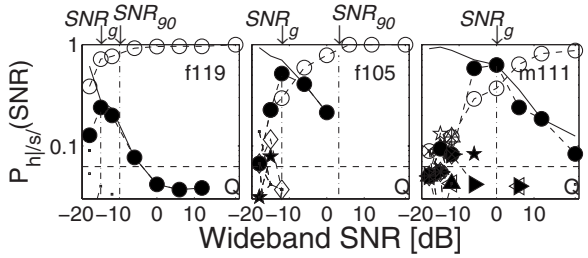


FIG. 11. Examples of threshold variability. CPs for three utterances of /sa/ by talkers f119 (left), f105 (center), and m111 (right). SNR_g denotes the /s/-/z/ group confusion threshold. SNR_{90} denotes the saturation point for /s/ recognition, where the diagonal score is 90%.

(m104 /fɑ/) and top right panels (m112 /fɑ/) of Fig. 10, we denote it as *morphing*. The target sound may be morphed into one or more other sounds by the noise masker. For example, while m104 /f/ is morphed to /m/, talker m112's /f/ has a double morphing toward /b/ and /v/. Not all utterances of a given consonant show morphing. Therefore, morphing is not observed in the average consonant CPs, shown in Fig. 8, which are obtained by pooling the data over utterances. Informal experiments show that at the crossover SNR, where the target and the competitor have equal scores, a listener can *prime* between two sounds.[2] If the target sound is presented in a meaningful word or sentence at the same SNR, then the priming would be resolved by context.

2. Threshold variability

On average, consonant /s/ is exclusively confused with /z/ in MN05 [see Fig. 8(b), second row, second column], with an average confusion threshold $SNR_g=-12$ dB and a confusion probability $P_{/z/|/s/}(SNR_g)\approx 20\%$. However, individual utterance CPs of /sɑ/ show a significant variation in the location of confusion threshold (Fig. 11). For talker f119, the threshold is at −12 dB with $P_{/z/|/s/}(SNR_g)\approx 25\%$, while for m111, it is at 0 dB with $P_{/z/|/s/}(SNR_g)\approx 65\%$. The confusion threshold for talker f105's utterance is between the two. Thus for the same confusion group, the threshold can vary from a minor confusion to a morph, depending on the utterance.

Noise robustness. Talker m111's /sɑ/ has more confusions, compared to f119 (Fig. 11). This means that m111 /sɑ/ is less robust to noise than f119 /sɑ/. To quantify the "robustness" of an utterance, we define a *saturation point*, denoted by SNR_{90}. This point forms a "knee" in the recognition score of the utterance, i.e., below SNR_{90}, the score rapidly decreases, while above SNR_{90}, the recognition score saturates. We quantify the saturation point SNR_{90} as the SNR where $P_c(SNR_{90})=90\%$. If the score for an utterance is always less than 90%, then $SNR_{90}=\infty$ is assigned to it. Thus, a lower SNR_{90} indicates a greater robustness to noise. In Fig. 11, f119 /sɑ/ ($SNR_{90}=-3.55$ dB) is more noise robust than f105's utterance ($SNR_{90}=4.25$ dB), while m111 /sɑ/ ($SNR_{90}=\infty$) is the weakest of the three.

The noise robustness of a sound depends on the masking noise spectrum. A quantitative analysis of SNR_{90} further supports the observations drawn from Fig. 2(b), that the consonants are more robust to SWN (PA07 dataset) than WN (MN05 dataset). Out of 192 common utterances, 174 utterances have $SNR_{90}(WN)>SNR_{90}(SWN)$, 14 have $SNR_{90}(WN)<SNR_{90}(SWN)$ and four have the same SNR_{90} in both experiments. In WN data of MN05, 17 utterances have $SNR_{90}=\infty$, compared to only eight in the SWN data of PA07. Four of these utterances (three /θ/ and one /ð/) have $SNR_{90}=\infty$ in both experiments.

IV. DISCUSSION

We have repeated the MN55 experiment using computerized techniques and a digitally recorded database. With few notable exceptions, the average consonant scores [Fig. 2(a)] and the consonant confusion patterns (Fig. 8) of MN05 closely match with those from the original [MN55]. This verifies that these "modern" computer based testing procedures can reliably reproduce the classic CM experiments. It also implies that the differences observed between PA07 and MN55 are due to differences in speech materials and noise spectra and not due to the procedural factors such as use of computers, digital filters, and a prerecorded database.

The consonant confusions of plosives and nasals in MN05 are virtually identical to those from MN55. However, the significant voicing confusions for fricatives, observed earlier in PA07 (SWN) were not present in MN55 (WN), but are present in MN05 (WN). Therefore, these confusions cannot be attributed to the differences in noise spectra between PA07 and MN55. These confusions were highly asymmetric, biased in favor of the voiced fricatives. Similar confusions are also observed in the Grant and Walden (1996) [GW96] acoustic-only data in SWN, and therefore cannot be attributed to our testing procedures and stimuli. These high voicing errors are responsible for the HE consonant sets, which contain fricatives /f/, /θ/, /v/, and /ð/, in the three experiments (PA07; MN05; and GW96), but not in MN55. One reason for low voicing errors by MN55 could be the familiarity of the listeners with the talker's voice. In MN55, the five listeners also served as the talkers, i.e., when one spoke the syllables, the other four listened and scored. There were no noticeable systematic differences in consonant scores, voicing scores, and consonant confusions for male and female talker utterances in MN05. Therefore, it is unlikely that the differences in MN05 and MN55 are due to the use of male talkers in MN05.

The isoperformance SNR contours (Fig. 2) are particularly useful when comparing performance across two different noise types. A comparison of WN (present experiment) and SWN [PA07] data shows that the difference in the noise spectra induces a constant SNR-loss of about 10–12 dB in WN, with respect to SWN [Fig. 2(b)]. The noise spectrum also impacts the distribution of consonant errors, resulting in different consonant sets in the two experiments. This further supports the conclusion of Phatak and Allen (2007) that consonants /s/, /ʃ/, /z/, /ʒ/, and /t/ have the greatest advantage in SWN. These consonants have lower scores in WN [present experiment; MN55]. The consonants /m/, /n/, and /v/ are almost equally masked by both types of noises. This is in agreement with the SNR-spectra analysis from PA07. The

three consonant sets observed in MN05 are the LE set {/m/, /n/}, the AE set {/p/, /t/, /k/, /s/, /ʃ/, /d/, /g/, /z/, /ʒ/}, and the HE set {/f/, /θ/, /b/, /v/, /ð/}.

The average recognition error in MN05 obeys the exponential AI model of speech recognition, given by Eq. (2). This model was introduced by Fletcher, and was shown to fit the average scores of isolate syllables [CV, VC, and CVC] [Fletcher and Galt (1950)]. Allen (1994) first expressed this relationship in terms of the minimum error $e_{\min}$ (i.e., the error at AI=1) and showed that the model can be extended to the individual consonant errors of the data of MN55 [Allen (2005b)]. The exponential AI model fits the error for individual consonants as well as for the three consonant sets in MN05 (Fig. 3). The log-error curves are linear in AI, relative to SNR, which can be partially explained by the CPs (Fig. 9). When plotted as a function of AI, the confusions in the consonant CPs are restricted to AI$<$0.2. As a result, the log of consonant error becomes more linear on AI scale than on SNR scale. As previously observed by Allen (2005b), the SNR-to-AI transformation also makes the log confusions [i.e., $\log(P_{h|s})$, the off-diagonal entries] more linear. Thus, it is possible to model the entire CM, not just the AE, in terms of AI.

Unlike the popular sigmoidal or ogive approximations of the performance-intensity curve P_C(SNR), the AI-model parametrization of the recognition performance has a solid theoretical psychoacoustic basis. Several standards for measuring speech quality, such the speech intelligibility index (SII) (ANSI-S3.5-1997, 1997) and the speech transmission index (Steeneken and Houtgast, 1980), are based on French and Steinberg's method of estimating AI. Allen (2005b) refined the original expression of French and Steinberg (1947) for estimating AI, to formulate a threshold correction to the AI, and showed that the AI is similar to the Shannon (1948) formula for channel capacity of a communication channel (Allen, 2004). However, this refined expression had a free parameter, which was later demonstrated by PA07 to be equal to the frequency-dependent peak-to-rms ratio of speech. The resulting AI expression [Eq. (5)] is explicitly computable from the speech and noise stimuli, and thus is completely independent of free parameters. This AI is equivalent to the *loudness, audibility, speech recognition* model of Studebaker *et al.* (1994), which is estimated from the peak spectrum of speech and rms spectrum of noise. As a result, the consonant recognition scores across experiments match better on the AI scale than on the SNR scale (Fig. 4). This is because the AI accounts for relative spectral shapes of speech and noise spectra, which are ignored in the wideband SNR calculation.

The CPs for individual utterances (Figs. 10 and 11) are not as smooth as the consonant CPs [Fig. 8(b)] due to lower row sums (N). After the utterance and listener selection, row sums for the consonant CPs range from 162 to 485 responses. In comparison, the typical N for utterance CPs is equal to number of listeners (i.e., 23) because each utterance was presented only once at each SNR to each listener. At low SNRs, when there are multiple competitors, the probability distribution in a row is multimodal. With a low N, the estimation error is relatively high, and thus multimodal distributions cannot be accurately estimated. However, at high SNRs, when there are a small number of competitors, the curves become relatively smooth. Thus, in spite of the low N, the confusion groups and the utterance variability analysis reveal useful results. All published CM data are pooled over listeners and talkers to reduce variance and to obtain an "average response." However, the variations across speech utterances and listener responses provide rich information, such as the morphing phenomenon, which is obscured by such averaging.

The utterance variability could not be analyzed either with MN05 data due to lack of stimuli or with PA07 data due to very low row sums. Thus, one of the aims for conducting MN05 was to investigate whether the utterance variations are random or systematic. The analysis of utterance confusion patterns shows that these variations are not only systematic but also can be quantitatively characterized into two types—confusion heterogeneity and threshold variability.

Morphed utterances provide a unique opportunity to better understand speech perception. The morphing phenomenon is also observed in a time-truncation experiment, where the CV syllables are gated from the consonant side (Régnier and Allen, 2008). For example, when a /sɑ/ utterance is truncated from consonant side, it first morphs into a /zɑ/. When truncated further, it first morphs to /dɑ/ and then to /ðɑ/, until only vowel is perceived. The truncation time at which an /s/ morphs to /z/ is consistent with the voice-onset time of a natural /z/. When a natural /zɑ/ utterance is truncated, it also morphs first to /dɑ/ and then to /ðɑ/, but it never morphs to /sɑ/. This is consistent with the asymmetric /s/-/z/ confusions observed in WN (present experiment) and SWN (PA07) masking data. This asymmetry suggests that the set of perceptual features or events that define consonant /z/ is a subset of those which define /s/. Thus, when the additional features in /s/ are masked or truncated, it is confused with, and in many cases morphs to, the consonant /z/, but /z/ is never confused with /s/. Comparing the /s/ utterance morphed into /z/ with a natural /z/ utterance can reveal these additional features in /s/ which distinguish it from /z/.

Consonants /p/ and /t/ form another pair that show this asymmetric morphing. Ten out of the 12 /tɑ/ utterances tested in the time-truncation experiment morphed to /pɑ/, but none of the /pɑ/ utterances morphed to /tɑ/ (Régnier, 2007). The individual utterance CPs for WN (present experiment) and SWN (PA07) masking data also significantly show more /t/ to /p/ morphing than /p/ to /t/ morphing, in terms of both the number of morphed utterances and the probability of morphing [i.e., $P_{h|s}(\mathrm{SNR}_g)$]. This results in the average $P_{/p/|/t/}(\mathrm{SNR}_g)$ [$\diamond$ in top left panel of Fig. 8(b)] to be lower than the average $P_{/t/|/p/}(\mathrm{SNR}_g)$ [$\triangle$ top row, second panel from left]. Thus, the event set for /p/ is a subset of the event set for /t/. Régnier (2007) show, by using time-frequency modification experiments, that the high-frequency release burst for /t/ is the event which separated /t/ from /p/. This result is consistent with the prediction by Heil (2003) that the envelope onset cues are critically important for speech intelligibility. This prediction is based on the neural data, which provides a physiological basis to the peak-to-rms ratio-based AI. The peak-to-rms ratios account for these perceptually

important temporal variations in speech, thus giving a temporal perspective to the otherwise spectral AI metric. An AI that accounts for the speech peaks can predict the recognition scores better than the ANSI-S3.5-1969 (1969) standard AI (Rankovic, 1998), and has led to the recent SII standard ANSI-S3.5-1997 (1997). The speech peaks are also critically important in extending the AI to predict speech intelligibility in fluctuating noise (Rhebergen and Versfeld, 2005).

Some utterances of a given sound are more robust to noise than others (Fig. 11). A quantitative analysis of the noise robustness, using the saturation point SNR_{90}, revealed that consonants are more robust to speech-weighted noise than WN. Noise-robustness analysis, combined with a spectrotemporal analysis of the stimuli, can lead us to the perceptual coding of speech. For example, Régnier and Allen (2008) found that SNR_{90} of /t/ utterances are highly correlated with the intensity of the transient in the release burst. Such a quantitative correlation would not be possible without the quantitative measures of CPs (i.e., SNR_g and SNR_{90}). Régnier and Allen (2008) also found that the event (i.e., the across-frequency coincidence of energy onset) is invariant, and it is the acoustic correlate (i.e., the intensity of the onset transient) which is responsible for the threshold variability. Similarly, it is likely that the confusion heterogeneity (Fig. 10) is due to differences in the relative intensities of the acoustic correlates of invariant events.

An alternative explanation for the heterogeneity is listener bias. If a listener narrows down the heard sound to a subset of possible choices, but is not confidant about the answer, then the response may be determined by the listener's bias for a specific answer. Such listener biases would dominate the responses in noisy conditions, where weak utterances are not clearly perceptible. This hypothesis can be easily tested by analyzing the consistency of listener responses to these utterances at low SNRs. However, such an analysis is not possible with the current data, as each utterance was presented only once or twice to each listener, at a given SNR. We have collected such data on listener consistency and this analysis is in progress.

V. CONCLUSIONS

The most important conclusions of this study can be briefly summarized as follows.

(1) The results of the classic Miller and Nicely (1955) can be reliably reproduced by using a recorded speech database and modern computerized testing procedures. The differences in the consonant data of Phatak and Allen (2007) (speech-weighted noise) and MN55 (white noise) are primarily due to different noise spectra.
(2) A normal-hearing listener's perception of consonants is more robust to speech-weighted noise than white noise. The noise robustness of an utterance can be quantified by using a saturation point (Fig. 11). Consonants /s/, /ʃ/, /z/, /ʒ/, and /t/ have the most disadvantage in white noise compared to speech-weighted noise, while consonants /v/, /m/, and /n/ are least affected by this difference in noise spectrum [Fig. 2(b)].
(3) An AI calculated from the specific speech and noise stimuli, by using PA07 AI formula [Eq. (5)], was verified to satisfy the exponential AI model [Eq. (1)] for consonant errors in white noise. The model can be extended to individual consonants as well as the consonant groups (Fig. 3).
(4) Masking noise cannot only reduce the recognition of a consonant, but also perceptually morph it into another consonant.
(5) In the presence of masking noise, fricatives show highly asymmetric voicing confusions, biased in favor of voiced consonants. MN55 data are an exception.
(6) There is a significant and systematic variability in the scores and confusion patterns of different utterances of the same consonant, which can be characterized as (a) *confusion heterogeneity*, where the competitors in the confusion groups of a consonant vary (Fig. 10), and (b) *threshold variability*, where confusion threshold (i.e., SNR and score at which the confusion group is formed) varies (Fig. 11).

ACKNOWLEDGMENTS

We thank all members of the HSR group at Beckman Institute, UIUC, for their inputs. We thank the three anonymous reviewers and the Associate Editor for their constructive comments and encouragement. This research was partially supported by a University of Illinois grant. The data-collection expenses were covered through research funding provided by Etymotic Research and by Starkey Labs.

APPENDIX

Tables II–IX show the consonant CMs, pooled over utterances and listeners, after the utterance and listener selection. The "only noise" responses, listed in the last column labeled ϕ, were considered to be chance performance responses and were distributed uniformly over the remaining 16 columns. These CMs were row normalized to have unity row sums for plotting the confusion patterns in Fig. 8(b). The SNR of −21 dB was rarely presented to the listeners due to their low scores at −18 dB SNR, resulting in very low row sums ($8 \leqslant N \leqslant 22$) and high variability at −21 dB SNR. Therefore, data at −21 dB SNR were not used in plotting the CPs and the corresponding CM is not listed here.

Since there were 18 talkers of each CV and 23 listeners, the row sums should be 414. However, as described in Sec. II, two randomly chosen CVs were repeated in each block (i.e., same SNR, same talker) to limit guessing by the listener. These extra presentations make the row sums greater than 414. Consonants /f/, /θ/, /v/, /ð/, and /z/ have row sums lower 414. This is because these consonants occurred most frequently in the ambiguous utterances and their row sums decreased after removing responses to ambiguous utterances.

[1]The experiment was called UIUCs04 by Phatak and Allen (2007), as it was conducted at the University of Illinois at Urbana-Champaign (UIUC) in the summer of 2004.
[2]For a priming condition between choices A and B, the listener will answer "Yes" with 100% probability to both questions-"Do you hear A?" and "Do you hear B?"

TABLE II. Consonant CM table. Quiet condition.

Quiet	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	457	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
/t/	0	469	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/k/	0	3	476	0	0	0	0	0	0	2	0	0	0	0	0	0	0
/f/	0	1	0	386	5	0	0	0	0	0	1	4	0	0	0	0	0
/θ/	0	0	0	5	202	4	0	0	0	0	0	20	0	0	0	0	0
/s/	0	0	0	0	1	429	0	0	0	0	0	1	3	0	0	0	0
/ʃ/	0	0	0	0	0	3	459	0	0	0	0	0	0	6	0	0	0
/b/	5	0	0	1	0	0	0	405	0	0	6	0	0	0	0	0	0
/d/	0	0	0	0	0	0	0	0	463	4	0	0	0	0	0	0	0
/g/	0	0	2	0	0	0	0	0	0	462	0	0	0	0	0	0	0
/v/	0	0	0	2	2	0	0	5	0	0	367	8	1	0	0	0	0
/ð/	0	0	0	0	19	0	0	0	0	0	3	133	0	0	0	0	0
/z/	0	0	0	0	0	0	0	0	0	0	0	3	378	6	0	0	0
/ʒ/	0	0	0	0	0	0	4	0	0	0	0	0	2	412	0	0	0
/m/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	468	0	0
/n/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	463	0

TABLE III. Consonant CM table. SNR=12 dB.

12 dB	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	463	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0
/t/	4	460	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0
/k/	0	0	474	0	0	0	0	0	0	2	0	0	0	0	0	0	0
/f/	5	0	0	341	3	0	0	27	0	0	22	0	0	1	0	0	0
/θ/	0	0	0	15	159	3	0	9	0	0	3	48	0	0	0	0	0
/s/	0	0	0	0	0	429	0	0	0	0	0	0	12	0	0	0	0
/ʃ/	0	0	0	0	0	0	445	0	0	0	0	0	0	22	0	0	0
/b/	4	0	0	20	0	0	0	369	0	0	23	0	0	0	0	0	0
/d/	0	0	0	0	0	0	0	0	472	2	0	0	0	0	0	0	0
/g/	0	0	2	0	0	0	0	0	2	467	0	0	0	0	0	0	0
/v/	0	0	0	2	0	0	0	57	0	20	332	1	0	0	0	0	0
/ð/	0	0	0	0	29	0	0	1	2	0	5	112	2	0	0	0	0
/z/	0	0	0	0	1	1	0	0	0	0	1	7	377	5	0	0	0
/ʒ/	0	0	0	0	0	0	2	0	0	1	0	0	3	413	0	0	0
/m/	0	0	0	0	0	0	0	0	0	0	1	0	0	3	465	2	0
/n/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	458	0

TABLE IV. Consonant CM table.SNR=6 dB.

6 dB	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	454	6	3	4	2	0	0	1	0	0	0	0	0	0	0	0	0
/t/	9	455	11	1	3	0	0	1	0	0	0	2	0	0	0	0	0
/k/	3	6	466	1	0	0	1	0	1	2	0	0	0	0	0	0	0
/f/	15	1	0	282	5	2	0	50	0	1	37	4	0	0	2	0	1
/θ/	2	0	0	22	133	4	0	15	4	0	4	51	0	0	0	0	0
/s/	0	1	0	1	4	398	2	0	0	0	0	1	28	1	0	0	0
/ʃ/	0	0	0	0	1	1	430	0	0	0	0	0	0	37	0	0	0
/b/	13	0	1	54	2	0	0	272	0	0	66	5	0	0	0	0	0
/d/	0	0	0	0	1	0	0	0	459	11	0	3	0	2	0	0	0
/g/	0	0	2	0	0	0	0	0	13	447	0	1	0	0	0	0	0
/v/	0	0	0	1	0	0	0	76	0	1	299	3	0	0	3	0	0
/ð/	0	0	0	0	26	0	0	0	6	2	14	88	9	0	0	6	0
/z/	0	0	0	0	1	9	0	0	6	0	1	7	351	8	0	0	0
/ʒ/	0	0	0	0	0	0	1	0	0	2	0	0	1	401	0	0	0
/m/	0	0	0	0	0	0	0	0	0	0	1	0	0	3	465	6	0
/n/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	473	0

TABLE V. Consonant CM table. SNR=0 dB.

0 dB	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	390	14	21	22	2	0	0	4	0	0	1	3	0	0	0	0	0
/t/	41	368	54	4	2	2	0	1	0	0	0	3	0	0	0	0	0
/k/	44	25	376	2	10	1	3	1	1	2	1	6	0	0	0	0	0
/f/	35	1	3	226	14	3	0	59	1	1	33	3	0	0	5	2	1
/θ/	3	5	4	33	84	11	0	27	8	4	7	52	0	0	0	0	0
/s/	0	0	0	4	14	364	2	1	0	0	7	8	50	1	0	0	0
/ʃ/	0	0	0	0	3	17	382	0	6	0	0	0	4	55	0	0	0
/b/	15	0	1	74	5	0	0	220	0	0	76	8	0	2	6	0	0
/d/	1	0	0	0	2	0	1	0	402	35	0	10	1	7	0	0	0
/g/	0	0	1	0	8	1	2	2	44	409	0	10	1	7	0	0	0
/v/	1	1	0	8	0	0	0	87	0	1	276	9	0	1	14	4	0
/ð/	0	2	0	1	16	0	0	2	17	4	25	69	15	7	0	4	0
/z/	0	0	0	0	1	12	0	0	6	3	0	5	350	9	0	0	0
/ʒ/	0	0	0	0	0	1	2	0	0	11	1	0	3	383	0	2	0
/m/	1	0	0	0	0	0	0	1	0	0	2	0	0	0	453	14	0
/n/	0	0	0	0	0	1	0	0	1	1	0	0	0	0	6	454	0

TABLE VI. Consonant CM table. SNR=−6 dB.

−6 dB	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	287	46	83	24	7	0	1	10	0	3	6	1	0	1	3	5	0
/t/	114	208	101	19	7	0	0	1	2	2	2	3	0	1	5	1	0
/k/	86	59	252	18	9	3	1	5	2	10	5	7	0	1	3	6	0
/f/	43	10	6	146	20	4	1	66	2	5	62	16	2	1	11	2	0
/θ/	5	15	14	30	63	5	2	21	10	16	18	34	3	1	1	0	0
/s/	0	2	1	10	18	313	3	8	5	1	7	10	56	6	0	0	0
/ʃ/	0	1	0	0	12	21	268	1	26	5	0	7	16	99	0	2	0
/b/	24	4	4	67	16	0	0	170	8	1	88	14	3	2	5	1	1
/d/	0	0	0	0	11	7	6	3	281	51	4	35	21	38	1	5	0
/g/	0	3	7	4	14	6	5	6	82	246	8	49	17	28	0	2	0
/v/	13	2	5	14	8	2	0	68	3	3	229	18	1	1	29	2	0
/ð/	1	7	0	0	8	3	1	4	8	6	34	34	15	15	5	9	0
/z/	0	4	0	1	5	13	1	0	14	7	2	8	287	39	0	3	0
/ʒ/	1	1	1	2	2	0	14	0	16	30	2	15	16	301	1	11	0
/m/	3	3	3	2	3	0	0	6	0	0	8	5	0	0	394	32	0
/n/	0	0	1	0	1	0	1	1	8	3	0	0	0	2	24	436	1

TABLE VII. Consonant CM table. SNR=−12 dB.

−12 dB	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	183	57	107	31	8	3	0	16	2	3	12	5	3	1	26	7	0
/t/	103	122	116	25	10	6	1	12	8	15	17	8	3	4	12	9	0
/k/	81	100	126	19	18	7	1	19	12	10	16	14	3	3	18	25	4
/f/	51	23	30	65	15	11	3	71	8	11	50	14	4	6	21	8	1
/θ/	23	19	26	45	18	7	3	21	11	13	22	14	5	2	3	4	1
/s/	5	11	14	23	15	212	15	5	10	5	11	12	82	16	5	5	3
/ʃ/	3	22	18	13	21	26	110	7	47	20	7	23	41	91	3	10	9
/b/	36	8	14	41	10	9	1	106	25	12	92	12	10	5	27	3	0
/d/	5	15	12	3	14	11	13	25	135	51	17	29	47	49	14	18	0
/g/	5	21	16	11	24	20	10	24	81	85	30	45	32	43	5	22	0
/v/	21	9	6	17	7	10	3	72	15	12	124	25	12	7	32	15	3
/ð/	8	8	6	10	5	2	0	11	13	9	22	20	15	11	10	13	1
/z/	4	12	5	5	8	21	11	10	41	24	16	24	145	53	4	6	1
/ʒ/	6	7	6	6	7	9	22	12	44	41	25	16	44	120	12	31	2
/m/	23	10	13	13	2	1	2	25	7	10	24	11	4	4	263	51	0
/n/	6	7	12	3	6	4	3	4	22	22	12	16	4	14	57	274	1

TABLE VIII. Consonant CM table. SNR=−15 dB.

−15 dB	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	101	63	88	28	10	7	4	31	5	15	22	10	6	5	23	12	44
/t/	67	77	84	22	18	15	15	27	8	14	20	10	5	7	28	23	37
/k/	53	71	73	26	8	7	8	22	12	23	20	7	7	5	32	34	62
/f/	41	26	25	26	11	12	8	55	13	19	44	12	12	2	26	15	41
/θ/	19	19	26	24	12	3	4	26	12	16	24	9	8	6	7	8	23
/s/	10	20	16	19	10	132	19	16	10	7	16	13	58	9	8	8	74
/ʃ/	9	24	19	18	10	31	35	16	38	35	25	20	29	44	9	17	91
/b/	34	16	25	31	8	14	6	53	35	20	66	17	13	5	26	12	30
/d/	15	16	16	14	16	12	13	28	81	42	20	36	37	30	25	22	50
/g/	16	18	29	16	18	14	18	23	57	58	32	26	34	37	6	31	41
/v/	16	12	19	30	14	14	3	49	20	26	73	16	28	12	27	22	26
/ð/	8	12	8	7	0	3	0	12	7	8	17	6	12	10	14	12	19
/z/	3	16	7	12	7	26	12	15	35	23	28	11	57	48	12	19	61
/ʒ/	10	19	10	10	8	17	8	7	36	29	31	15	38	59	20	34	58
/m/	26	28	20	28	8	7	3	25	6	12	34	7	10	7	136	67	49
/n/	8	24	14	1	8	11	9	9	25	17	20	11	19	19	73	138	66

TABLE IX. Consonant CM table. SNR=−18 dB.

−18 dB	/p/	/t/	/k/	/f/	/θ/	/s/	/ʃ/	/b/	/d/	/g/	/v/	/ð/	/z/	/ʒ/	/m/	/n/	ϕ
/p/	37	25	35	12	3	6	7	22	11	13	11	6	13	5	17	28	216
/t/	28	26	39	13	4	15	11	14	9	15	18	9	14	11	15	28	198
/k/	27	31	28	16	6	18	12	20	14	16	12	6	4	2	24	30	207
/f/	18	21	12	13	6	5	14	18	17	15	20	9	12	6	21	24	157
/θ/	9	11	7	9	3	4	5	9	12	16	14	3	4	2	13	6	102
/s/	13	20	12	6	4	43	16	5	8	7	8	10	19	6	10	10	238
/ʃ/	12	16	17	9	8	13	8	16	17	22	9	10	6	17	13	13	273
/b/	22	21	18	15	11	10	3	22	18	18	22	6	7	13	17	16	175
/d/	13	24	12	11	9	15	9	25	35	19	17	12	22	17	17	16	198
/g/	8	29	14	7	9	10	16	12	26	31	16	10	13	12	14	27	210
/v/	13	21	15	13	4	11	6	29	12	21	27	3	8	8	21	8	171
/ð/	11	7	7	0	4	3	4	3	6	6	10	3	2	3	5	12	68
/z/	8	17	12	8	8	7	10	8	12	11	10	11	13	6	8	17	220
/ʒ/	12	11	10	2	6	12	12	13	18	18	17	6	15	16	13	16	202
/m/	18	16	10	12	3	10	5	16	11	10	18	7	11	5	58	47	215
/n/	8	19	14	3	6	11	8	10	14	16	12	8	10	11	29	60	241

Allen, J. B. (**1994**), "How Do Humans Process and Recognize Speech?" IEEE Trans. Speech Audio Process. **2**, 567–577.

Allen, J. B. (**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), Chap. Speech, pp. 314–320.

Allen, J. B. (**2005a**), in *Articulation and Intelligibility*, Synthesis Lectures in Speech and Audio Processing, edited by B. H. Juang (Morgan and Claypool, USA).

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

ANSI-S3.5-1969 (**1969**), "American National Standard methods for the calculation of the articulation index" (American National Standards Institute, New York).

ANSI-S3.5-1997 (**1997**), "American National Standard methods for calculation of the speech intelligibility index" (American National Standards Institute, Inc., New York).

Dubno, J. R., and Levitt, H. (**1981**), "Predicting consonant confusions from acoustic analysis," J. Acoust. Soc. Am. **69**, 249–261.

Dunn, H. K., and White, S. D. (**1940**), "Statistical Measurements on Conversational Speech," J. Acoust. Soc. Am. **11**, 278–287.

Fletcher, H., and Galt, R. H. (**1950**), "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89–151.

Fousek, P., Svojanovsky, P., Grezl, F., and Hermansky, H. (**2004**), "New Nonsense Syllables Database—Analyses and Preliminary ASR Experiments," in Proceedings of International Conference on Spoken Language Processing (ICSLP), pp. 2749–2752.

French, N. R., and Steinberg, J. C. (**1947**), "Factors Governing the Intelligibility of Speech Sounds," J. Acoust. Soc. Am. **19**, 90–119.

Gordon-Salant, S. (**1985**), "Some perceptual properties of consonants in multitalker babble," Percept. Psychophys. **38**, 81–90.

Grant, K. W., and Walden, B. E. (**1996**), "Evaluating the articulation index for auditory-visual consonant recognition," J. Acoust. Soc. Am. **100**, 2415–2424.

Heil, P. (**2003**), "Coding of temporal onset envelope in the auditory system," Speech Commun. **41**, 123–134.

Lobdell, B., and Allen, J. B. (**2007**), "Modeling and using the vu-meter (volume unit meter) with comparisons to root-mean-square speech levels," J. Acoust. Soc. Am. **121**, 279–285.

Miller, G. A., and Nicely, P. E. (**1955**), "An analysis of perceptual confusions among some english consonants," J. Acoust. Soc. Am. **27**, 338–352.

Pavlovic, C. V. (**1984**), "Use of articulation index for assessing residual auditory function in listeners with sensorineural hearing impairment," J. Acoust. Soc. Am. **75**, 1253–1258.

Phatak, S. A., and Allen, J. B. (**2007**), "Consonant and vowel confusions in speech-weighted noise," J. Acoust. Soc. Am. **121**, 2312–2316.

Rankovic, C. M. (**1998**), "Factors governing speech reception benefits of adaptive linear filtering for listeners with sensorineural hearing loss," J. Acoust. Soc. Am. **103**, 1043–1057.

Rankovic, C. M. (**2002**), "Articulation index predictions for hearing-

impaired listeners with and without cochlear dead regions," J. Acoust. Soc. Am. **111**, 2545–2548.
Régnier, M. (**2007**), "Perceptual features of some consonants studied in noise," Master's thesis, University of Illinois at Urbana-Champaign, Urbana, IL.
Régnier, M., and Allen, J. B. (**2008**), "A method to identify noise-robust perceptual features: application for consonant /t/," J. Acoust. Soc. Am. **123**(5), 2801-2814.
Rhebergen, K. S., and Versfeld, N. J. (**2005**), "A speech intelligibility index-based approach to predict the speech reception threshold for sentences in fluctuating noise for normal-hearing listeners," J. Acoust. Soc. Am. **117**, 2181–2192.
Shannon, C. E. (**1948**), "The mathematical theory of communication," Bell Syst. Tech. J. **27**, 379–423.
Sroka, J., and Braida, L. D. (**2005**), "Human and machine consonant recognition," Speech Commun. **45**, 401–423.
Steeneken, H. J. M., and Houtgast, T. (**1980**), "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.
Studebaker, G. A., Taylor, R., and Sherbecoe, R. L. (**1994**), "The effect of noise spectrum on speech recognition performance-intensity functions," J. Speech Hear. Res. **37**, 439–448.
Wang, M. D., and Bilger, R. C. (**1973**), "Consonant confusions in noise: a study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.