Consonant and vowel confusions in speech-weighted noise[a)]

Sandeep A. Phatak[b)] and Jont B. Allen
ECE, University of Illinois at Urbana-Champaign, Beckman Institute, 405 N. Mathews Avenue, Urbana, Illinois 61801

(Received 7 April 2006; revised 30 October 2006; accepted 20 January 2007)

This paper presents the results of a closed-set recognition task for 64 consonant-vowel sounds (16 C × 4 V, spoken by 18 talkers) in speech-weighted noise (−22, −20, −16, −10, −2 [dB]) and in quiet. The confusion matrices were generated using responses of a homogeneous set of ten listeners and the confusions were analyzed using a graphical method. In speech-weighted noise the consonants separate into three sets: a low-scoring set C1 (/f/, /θ/, /v/, /ð/, /b/, /m/), a high-scoring set C2 (/t/, /s/, /z/, /ʃ/, /ʒ/) and set C3 (/n/, /p/, /g/, /k/, /d/) with intermediate scores. The perceptual consonant groups are C1: { /f/-/θ/, /b/-/v/-/ð/, /θ/-/ð/ }, C2: { /s/-/z/, /ʃ/-/ʒ/ }, and C3: /m/-/n/, while the perceptual vowel groups are /ɑ/-/æ/ and /ɛ/-/ɪ/. The exponential articulation index (AI) model for consonant score works for 12 of the 16 consonants, using a refined expression of the AI. Finally, a comparison with past work shows that white noise masks the consonants more uniformly than speech-weighted noise, and shows that the AI, because it can account for the differences in noise spectra, is a better measure than the wideband signal-to-noise ratio for modeling and comparing the scores with different noise maskers. © *2007 Acoustical Society of America.*
[DOI: 10.1121/1.2642397]

PACS number(s): 43.71.An, 43.71.Gv, 43.72.Dv [ADP] Pages: 2312–2326

I. INTRODUCTION

When a perceptually relevant acoustic feature of a speech sound is masked by noise, that sound becomes confused with related speech sounds. Such confusions provide vital information about the human speech code, i.e., the perceptual feature representation of speech sounds in the auditory system. When combined with a spectro-temporal analysis of the specific stimuli, this confusion analysis forms a framework for identifying the underlying *perceptual features* or the *events* (Allen, 2005a). Events are defined as the features, extracted by the human auditory system, which form the basis for perception of different speech sounds. It is these events which make human speech recognition highly robust to noise, as compared to machine recognition (Lippman, 1997). Thus, the use of events should increase the noise robustness of a speech recognition system, and should improve the functionality of hearing aids and cochlear implants.

It is our goal to identify these events by directly comparing the sound confusions with the corresponding masked speech stimuli, on an utterance by utterance basis. We wish to identify the acoustic features in speech which become inaudible when a masked speech sound is confused with other sounds.

Towards this goal we have performed a series of perceptual experiments that involve noise masking, time truncation, and filtering of speech. We employed large numbers of talkers and listeners, to take advantage of the large natural variability in speech production and perception. This paper presents the analysis of the confusion data for one of these noise-masking experiments.

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). Each entry in the CM, denoted $P_{s,h}(SNR)$, is the empirical probability of reporting sound h as heard when sound s was spoken, as a function of the *signal-to-noise ratio* (SNR). A Bayesian average of the diagonal entries $[P_{s,s}(SNR), h=s]$ gives the conventional "Recognition Score" or "Performance Intensity" (PI) measure $P_c(SNR)$. However, such an average obscures the detailed and important information about the nature of the sound confusions, given by the off-diagonal entries.

The confusion matrix was first used for analyzing speech recognition by Campbell (1910). CMs have been used to analyze confusions among vowel sounds in English [Peterson and Barney (1952), Strange *et al.* (1976), Hillenbrand *et al.* (1995)]. Miller and Nicely (1955) used the CM to analyze the consonant confusions for consonant-vowel (CV) sounds with 16 consonants and one vowel, presented at different levels of white masking noise. In 1955, Miller and Nicely (denoted MN55) collected data with five talkers and listeners, at six SNR levels and 11 filtering conditions. This classic confusion analysis experiment inspired many related and important noise-masking studies, such as Wang and Bilger (1973), Dubno and Levitt (1981), Grant and Walden (1996), and Sroka and Braida (2005).

The MN55 study clearly demonstrated that at low SNR, their consonants form three basic clusters of confusable sounds: Unvoiced, Voiced (non-nasals), and Nasals. As the SNR is increased, the first two clusters split into two

[a)] Parts of this analysis were presented at the ARO Midwinter Meeting 2005 (New Orleans), the Aging and Speech Communication 2005 Conference (Bloomington, IN) and the International Conference on Spoken Language Processing 2006 (Pittsburgh, PA).

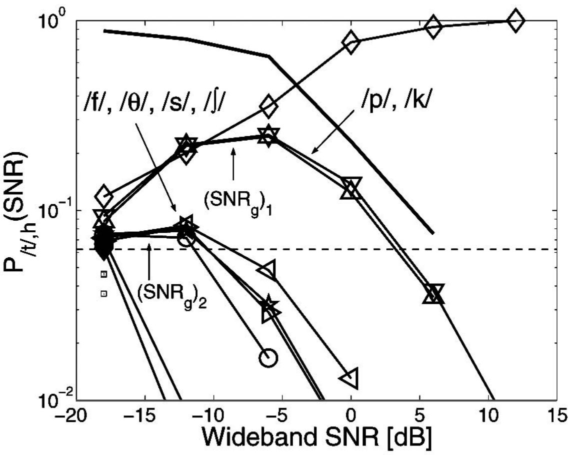
[b)] Author to whom correspondence should be addressed. Electronic mail: phatak@uiuc.edu subgroups—plosives and fricatives. Wang and Bilger (1973) extended the CM analysis to more consonants and vowels, but unfortunately their published CM data are pooled over all SNRs, thereby reducing the utility of their database for analyzing perceptual grouping. Dubno and Levitt (1981) compared the acoustic features of syllables with CM data, but they used only two SNR values and did not find common acoustic features that correlate with the confusions at their SNR levels. Grant and Walden (1996) measured the confusions of 18 consonants with auditory and visual cues, but did not provide a confusion analysis, as the primary goal of their study was to investigate the articulation index (AI). Sroka and Braida (2005) measured CMs at several SNRs and filtering conditions for humans and automatic speech recognizers (ASRs), however only one talker was used (either male or female, depending on the syllable), as the primary purpose of their study was to compare the human performance with that of ASRs.

In the consonant CM tables from MN55, the order of the consonants is crucial when viewing or analyzing the formation of such clusters. With a different order of consonants, the perceptual clusters of consonants are not obvious. An alternate clustering method, multi-dimensional scaling, does not depend on the order of consonants but is not stable and does not guarantee a unique solution (Wang and Bilger, 1973). A *confusion pattern* (CP) analysis, defined by a graphical representation of a row (particular value of s) of the CM as a function of SNR, is a simple tool that overcomes all of these difficulties (Allen, 2005b). In this report we use the CPs to further study human speech coding.

A. Confusion patterns

Figure 1 shows the CPs for sound $s=/\text{t}\alpha/$ from MN55. Each curve corresponds to a particular column entry (h) for the /t/ row, plotted as a function of SNR, namely $P_{/t/,h}(SNR)$. The diagonal entry $P_{/t/,/t/}(SNR)$, denoted by ◊, increases with SNR. As the SNR decreases, confusions of /t/ with /p/ (△) and /k/ (▽) increase and eventually become equal to the 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 $(SNR_g)_1 \approx -8$ dB, where $(SNR_g)_1$ is the point of local maximum in $P_{/t/,/p/}(SNR)$ and $P_{/t/,/k/}(SNR)$ curves. When the SNR is decreased below $(SNR_g)_2 \approx -15$ dB, consonant group [/f/, /θ/, /s/ and /ʃ/] merges with the [/t/, /p/, /k/] group, forming a super group. Since $(SNR_g)_2 < (SNR_g)_1$, consonants [/p/, /k/] are perceptually closer to /t/, and thereby form a stronger perceptual group with /t/ than the consonants [/f/, /θ/, /s/, /ʃ/]. Thus we use the *confusion threshold* SNR_g as a quantitative measure to characterize the hierarchy in the perceptual confusions.

At very low SNRs, where no speech is audible, all the sounds asymptotically reach the chance performance of 1/16, shown by the dashed line. The remaining nine off-diagonal entries are never confused with the target sound /t/, and as a result never exceed chance (e.g., the small squares).



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

FIG. 1. Confusion patterns (CPs) for $s=/\text{t}\alpha/$ from MN55. The thick solid line without markers is $1-P_{s,s}(SNR)$, which is the sum of off-diagonal entries. The horizontal dashed line shows the chance level of 1/16. The legend provides the marker style used for consonants. These markers will be used throughout the paper.

B. Experiment UIUCs04

The speech stimuli for the previous CM experiments either do not exist in recorded format, or were not publicly available. Without these speech wave forms, it is not possible to determine the acoustic, and thus the corresponding perceptual features. Thus a number of MN55 related closed-set confusion matrix experiments were conducted at the University of Illinois, using a commercially available database (LDC-2005S22) composed of nonsense sounds having 24 consonants, 15 vowels and 20 talkers. The first of these experiments, reported here and denoted “UIUCs04,” used 64 context-free consonant-vowel (CV) sounds (16Cs × 4Vs). Our first goal was to analyze consonant confusions. The purpose of choosing multiple vowels was to analyze the extent of the effect of vowels on the listener’s consonant CPs (i.e., the coarticulation effects).

The long-term goal of our data collection exercise is to identify perceptual features by the use of masking noise. Specifically, we wish to determine the acoustic features that are masked near the confusion thresholds. We have also used the natural variability in the confusion thresholds across utterances to identify the acoustic features and events. The analysis in the present paper is limited to consonant and vowel confusions, but not events. We compare our results with past work, and show how the consonant confusions in speech-weighted noise are different from those in white noise. We also show that the observed consonant groups are related to the spectral energy in the consonant above the noise spectrum, and show that the Articulation Index (AI),

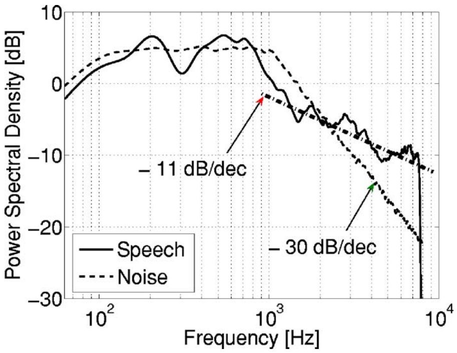


FIG. 2. (Color online) The power spectral densities (PSD) of average speech (solid) and noise (dashed) for UIUCs04 at 0 dB wideband SNR. The PSDs for both speech and noise were calculated using the *pwelch* function in MATLAB, with a hanning window of duration 20 ms (i.e., 320 samples) with an overlap of 10 ms and a fast Fourier transform length of 2048 points.

derived from the speech and noise spectra, is a better metric than the wideband SNR to characterize and compare the consonant scores.

II. METHODS

A. Stimuli

A subset of isolated CV sounds from the LDC-2005S22 corpus (Fousek *et al.*, 2004), recorded at the Linguistic Data Consortium (University of Pennsylvania), was used as the speech database. This subset had 18 talkers speaking CVs composed of one of the 16 consonants (/p/, /t/, /k/, /f/, /θ/, /s/, /ʃ/, /b/, /d/, /g/, /v/, /ð/, /z/, /ʒ/, /m/, /n/) and followed by one of the four vowels (/ɑ/, /ɛ/, /ɪ/, /æ/). The vowels were chosen to have formant frequencies close to each other, with the goal of making them more confusable. All talkers were native speakers of English, but three talkers were bilingual and had a part of their upbringing outside the U.S./Canada. Ten talkers spoke all 64 CVs, while each of the remaining eight talkers spoke different subsets of 32 CVs, such that each CV was spoken by 14 talkers.

MN55 had five female talkers, who also served as the listeners. Because the power spectrum for average speech [Dunn and White (1940); Benson and Hirsh (1953); Cox and Moore (1988)] has a roll-off of about −29 dB/dec (≈−8.7 dB/oct) above 500 Hz, the white noise masks the high frequencies in speech to a greater extent than low frequencies. A noise signal that has a spectrum similar to the average speech spectrum would mask the speech uniformly over frequency. Such a speech-weighted noise, shown in Fig. 2, was used as masker in UIUCs04. The noise power spectrum was constant from 100 Hz to 1 kHz, with a roll-off of 12 dB/dec (≈3.6 dB/oct) and −30 dB/dec (≈−9.0 dB/oct) on the lower and higher sides, respectively. The noise was generated by taking the inverse Fourier transform of the magnitude spectrum, obtained from this power spectrum, combined with a random phase. The rms level of this noise was then adjusted according to the level of the CV sound to achieve the desired SNR. The average spectrum of CV sounds (Fig. 2) was found to have a different roll-off characteristic than the making noise spectrum. The roll-off of the average speech for this experiment was −30 dB/dec between 800 Hz and 1.5 kHz but then it reduced to about −11 db/dec (≈−3.3 dB/oct), resulting in a high-frequency SNR boost. The change in the slope above 2 kHz can also be observed in the speech spectrum from several studies [Byrne *et al.* (1994); Grant and Walden (1996)].

A new random noise with the desired spectral characteristics was generated for each presentation, and the wideband noise rms level was adjusted according to the rms level of the CV sound to be presented, to achieve the precise SNR. While calculating the rms level of a CV utterance, the samples below −40 dB with respect to the largest sample were not considered.

The CV sounds were presented in speech-weighted masking noise at six different signal-to-noise ratios (SNR): [−22, −20, −16, −10, −2, *Q*] dB, where *Q* represents the quiet condition. The sum of speech signal and masking noise was filtered with a bandpass filter of 100 Hz−7.5 kHz before presentation. The highest amplitude of the bandpass filtered output (i.e., speech plus noise) was scaled to make full use of the dynamic range of the sound card, without clipping any sample.

B. Testing paradigm

The listening test was automated using a MATLAB code with graphic user interfaces. The listener was seated in a sound booth in front of a computer monitor. The computer running the MATLAB code was placed outside the sound-treated booth to minimize ambient noise. The monitor screen showed 64 buttons, each labeled with one of the 64 CVs. The 64 buttons were arranged in a 16 × 4 table such that each row had the same consonant while each column had the same vowel. An example of the use of each consonant or vowel in an English word was displayed as the pronunciation key at the left of the rows and at the top of the columns. Listeners heard the stimuli via headphones (Sennheiser, HD-265) and entered the response by clicking on the button labeled with the identified CV. The listener was allowed to replay the CV sound as many times as desired before entering the response. Repeating the sound helped to improve the scores by eliminating the unlikely choices in the large 64-choice closed-set task. Repeating the sound also allows the listener to recover from the distractions during the long experiment. For each repetition, a new noise sample was generated. After entering the response, the next sound was played following a short pause.

In addition to the 64 buttons, the listener had an option of clicking another button, labeled "Noise Only," to be used only when the listener could not hear any part of the masked speech. The listeners were periodically instructed to use this button only when no speech signal was heard, and to guess the CV otherwise. The primary purpose to allow the Noise Only response was to remove the listener biases. The Noise Only responses for a CV were treated as "chance-level" responses and were distributed uniformly over the 64 columns, corresponding to 64 possible options, in the row of that CV.

Each presentation of CV sound was randomized over consonants, vowels, talkers, and SNRs. The total 5376 presentations (16C × 4V × 14 talkers × 6 SNRs) were random-

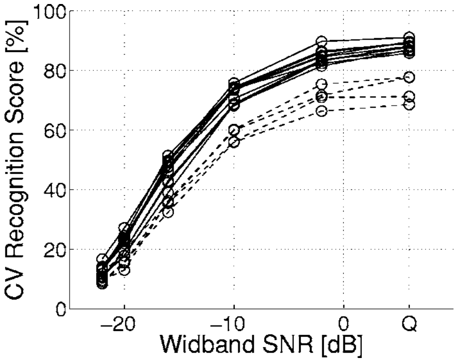


FIG. 3. The CV recognition scores of 14 listeners, as a function of SNR. Dashed lines show the four low performance (LP) listeners. The quiet condition is denoted by "Q."

ized and split into 42 tests, each with 128 sounds. Each listener was trained using one or two practice tests with randomly selected sounds, presented in Quiet, with visual feedback on the correct choice.

C. Listeners

Fourteen L1=English listeners (6M, 8F), ten having American accents and one with Nigerian accent, completed the experiment. All listeners, except one with age of 33 years, were between 19 and 24 years. They had no history of hearing disorder or impairment, and self-reported to have normal hearing. The listeners were verified to be attending to the experimental task, based on their scores, as described in the next section. The average time for completing the experiment was 15 h per listener.

III. RESULTS

Before analyzing the perceptual confusions, it is necessary to verify that the listeners attended to the required task and that the speech database was error free. We select a homogeneous group of listeners, based on their syllable recognition scores. In order to analyze the effect of noise on the perceptual confusions, we must verify that the utterances are heard correctly in the quiet condition, as a control. The mislabeled utterances can contaminate the perceptual confusions in noise. Therefore, based on the syllable errors in quiet, we select the low-error utterances that we use for analyzing perceptual confusion in noise. Following listener and utterance selection, we analyze the confusions of the CV syllables, as well as those of individual consonants and vowels. Finally, we compare our results with the past work from literature.

A. Listener selection

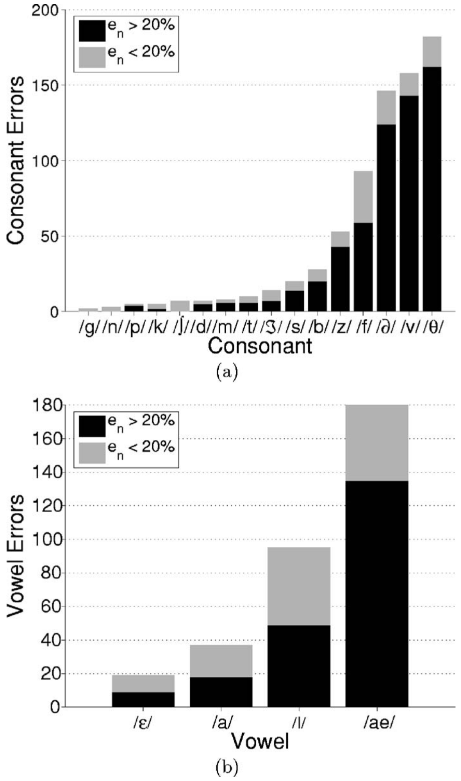
Ten "High Performance" (HP) listeners (i.e., listeners with scores greater than 85% in quiet, and greater than 10% correct at −22 dB SNR), shown by solid lines in Fig. 3, formed a homogeneous group. The scores of these HP listeners (5M, 5F) were comparable to the average score of NH listeners from other confusion matrix studies.[1] Responses of the four "Low Performance" (LP) listeners (dashed lines) were not considered for the subsequent analysis. All HP listeners had American accents (five Midwest, one New York, and four unspecified).

To investigate the sources of low scores for the LP listeners, their errors in the Quiet condition were further analyzed. All four LP listeners had 10–21% vowel errors, while three of the four LP listeners had 14–15% consonant errors in quiet. The average consonant and vowel errors for the HP listeners, in quiet, were 8% and 4%, respectively. For LP listeners, 61–72% of the consonant errors were for consonants /θ/, /v/ and /ð/, while the vowel errors were consistently high for vowel /æ/. The vowel sound /æ/ was mainly confused with /ɑ/. For the remaining consonants and vowels, scores of all 14 listeners were comparable. The pronunciation keys "/TH/ as in *THick*" and "/th/ as in *that*" and the labels TH and th were used for consonant sounds /θ/ and /ð/, respectively. It is possible that the four LP listeners confused the labels of these two consonants, which have the same spelling. However, the LP listeners confused /θ/ more with /f/ than with /ð/. Also, /ð/ was confused equally with /θ/ and /v/. Therefore, the most likely reason for the bad performance of the LP listeners is their inability to distinguish among consonants /f/, /θ/, /v/, /ð/, and between vowels /æ/ and /ɑ/.

B. Utterance selection

The syllable error e_n for each of the 896 utterances $(1 \leqslant n \leqslant 896)$ was estimated from listener responses in the quiet condition. A syllable error occurs when a listener reports an incorrect consonant or an incorrect vowel, or both. These errors can be estimated from the CM as $e_n = 1 - P_{s,s}(n,\text{quiet}) = \Sigma_h^{s \neq h} P_{s,h}(n,\text{quiet})$, where $P_{s,s}(n,\text{quiet})$ is the diagonal element of the CM, representing correct recognition for utterance n. The syllable errors were calculated for all listeners, as well as for the 10 HP listeners. When responses of the 10 HP listeners were pooled, 59% of the total 896 utterances had no errors in quiet. These are the well-formed or "good" utterances. However, some utterances had very high errors; ten utterances had 100% error. Some of these high error utterances, which had $e_n > 80\%$, were consistently reported as another CV sound and therefore are better described as mislabeled. The responses to the other high error utterances were mostly incorrect and inconsistent. Such high error sounds, which were unaffected by listener selection, are inherent to the database.

Since the errors in the speech database (i.e., the high error sounds) could be misinterpreted as the listener confusions while analyzing the listener responses, they would contaminate the perceptual confusions in noise. Therefore, 146 utterances ($\approx 16\%$ of the 896 utterances), which had more than 20% errors, are defined as "confusable" utterances, and the responses to these utterances were removed before generating CMs. Removal of the confusable utterances improved the consonant recognition scores by greater margin $(91.6\% \rightarrow 98.0\%)$ than the vowel recognition scores $(96.0\% \rightarrow 98.2\%)$. The utterances with $0 < e_n < 20\%$ are defined as the "marginally confusable" utterances and were considered for analysis. Figure 4 shows the distribution of



(a)

(b)

FIG. 4. Histograms of the incorrect recognition of (a) consonants and (b) vowels, in the "confusable" ($e_n > 20\%$) and "marginally confusable" ($0 < e_n < 20\%$) utterances, where e_n is the syllable error for that utterance in the quiet. There are 560 responses for each consonant (4 vowels × 14 talkers × 10 HP listeners) while there are 2240 responses for each vowel (16 consonants × 14 talkers × 10 HP listeners) in the quiet condition.

the total number of errors per consonant (left panel) and per vowel (right panel), in the confusable utterances ($e_n > 20\%$) as well as for the marginally confusable utterances ($e_n < 20\%$). Most of these errors occurred for the five consonants /θ/, /v/, /ð/, /f/, /z/, and for the two vowels /æ/, /ɪ/ (Fig. 4). These consonants and vowels were also the ones for which the LP listeners performed very poorly relatively to the HP listeners (Sec. III A, last paragraph). This suggests that the reason for poor performance of the LP listeners was their inability to recognize the confusable sounds. It is possible that the LP listeners perform as well as HP listeners for the marginally confusable and good utterances, in which case, LP listener data would be useful. However, the score of LP listeners for marginally confusable utterances was found to be even lower than that of HP listeners. The $/\theta/ \rightarrow /f/$ and $/ð/ \rightarrow /v/$ confusions of the LP listeners decreased significantly after removing the confusable utterances, but $/\theta/ \leftrightarrow /ð/$ and $/æ/ \rightarrow /ɑ/$ confusions did not show a similar decrease. The recognition scores of the LP listeners for /θ/, /ð/ and /æ/ increased, but still remained lower than those of the HP listeners. The LP listeners had 4–8% consonant error and 7–19% vowel error after the utterance selection, as compared to the HP listeners, who had less than 2% consonant and vowel errors. All subsequent analysis uses 10 HP listener responses to the marginally confusable and good utterances.

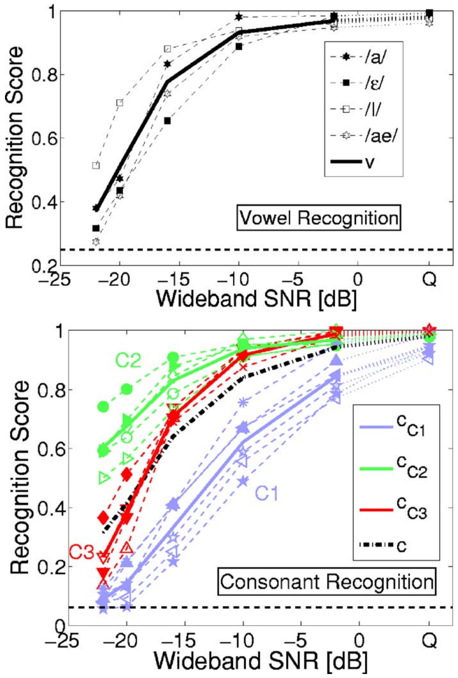


FIG. 5. (Color online) Recognition scores for vowels (*top*) and consonants (*bottom*), as a function of wideband SNR. In the top plot, the thick solid line is the average vowel recognition score (v), while in the bottom plot, the three solid lines represent the average scores for the three consonant sets and the dash-dotted line represents the average consonant score (c). The chance levels, 1/4 for vowels and 1/16 for consonants, are shown by the horizontal dashed black lines.

C. Recognition scores

The top panel in Fig. 5 shows the recognition scores for the four vowels, as well as the average vowel score (thick solid line), as a function of SNR. Except for the slightly higher scores of vowel /ɪ/ in presence of noise, the vowel scores are approximately equal. The previous studies show that /ɪ/ scores are relatively greater in a masking noise with speech-like spectrum, possibly due to a higher F_2 value (Gordon-Salant, 1985). The speech-weighted noise, therefore, seems to uniformly mask the four vowels.

One of the most interesting observations in this study is that the recognition scores of consonants show three groups (Fig. 5, bottom). One set of curves, shown in blue color, has relatively low scores, approaching the chance level of 1/16 below −20 dB. This set, which we call C1, contains consonants /f/, /θ/, /v/, /ð/, /b/ and /m/. In contrast, the consonants /t/, /s/, /z/, /ʃ/ and /ʒ/, which form set C2 (green lines) are high-scoring consonant and have scores greater than 50% even at −22 dB SNR. The remaining consonants (/n/, /p/, /g/, /k/ and /d/), grouped into set C3 (red lines), have relatively high scores, close to set C2 scores, above −10 dB SNR. However, the scores of C3 consonants drop sharply below −10 dB SNR, approaching the C1 scores at −22 dB.

The separation of the three sets of consonant curves is more evident in the vowel-to-consonant recognition ratio ($\lambda \equiv v/c$) plots shown in Fig. 6. The ratio λ was first used by Fletcher and Galt (1950) to compare the consonant and vowel performances. Figure 6 shows the values of $\lambda_i = v/c_i$, where v is the average vowel score and c_i represents the scores of individual consonants. The dash-dot line shows the

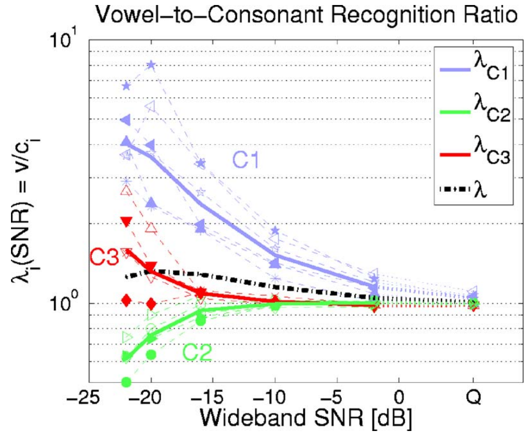


FIG. 6. (Color online) Vowel-to-consonant recognition ratio (on a log scale) as a function of SNR [$\lambda(SNR)$], for each consonant. The color and the markers denote the same information as that in the bottom panel of Fig. 5. The average $\lambda(SNR)$ for the consonant sets are shown by the thick solid lines, while that for average consonant score is shown by the thick, dash-dotted line.

average value of λ, which was just above unity. The C1 consonants have $\lambda_i(SNR)$ curves that well above 1 even for small amounts of noise, while those for set C3 stay close to unity for wideband SNRs $\geqslant -16$ dB, but rise sharply below that. The below-unity values of $\lambda_i(SNR)$ for C2 consonants in speech-weighted noise contradict the traditional assumption that the vowels are always better recognized in noise than consonants.

These consonant groups are also observed in the Grant and Walden (1996) data, which were collected in a speech-weighted noise masker, but not observed in the confusion data of Miller and Nicely (1955). Thus, while the white noise masks the 16 consonants almost uniformly, the speech-weighted noise has a nonuniform masking effect for consonants, masking set C1 more than set C2. This is further discussed in Sec. III G.

1. Articulation index (AI)

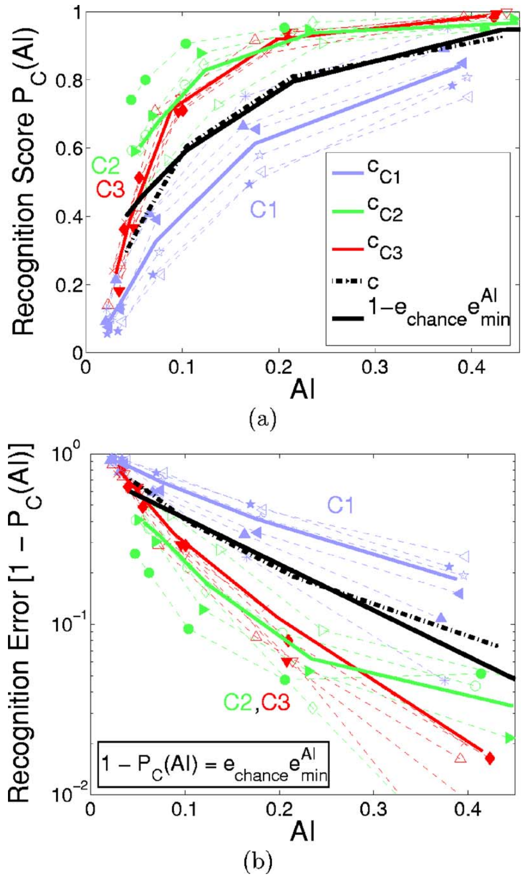
Allen (2005b) showed that the MN55 recognition scores for 11 of the 16 consonants, as well as the average consonant scores, can be modeled as

$$P_C(AI) = 1 - e_{\mathrm{chance}} e_{\mathrm{min}}^{AI}, \tag{1}$$

where AI is the articulation index, $e_{\mathrm{min}} = 1 - P_C(AI=1)$ is the recognition error at $AI=1$ and $e_{\mathrm{chance}} = 1 - 1/16$ is the error at chance ($AI=0$). Based on this relation, the log-error $\log(1-P_c(AI)) = AI \log(e_{\mathrm{min}}) + \log(e_{\mathrm{chance}})$ is a linear function of the AI. The AI, which is based on the SNRs in articulation bands, accounts for the shapes of signal and noise spectra (French and Steinberg, 1947). The articulation bands were estimated to contribute equally to the recognition context-free speech sounds (Fletcher, 1995). Allen (2005b) refined the AI formula to be

$$AI = \frac{1}{K}\sum_{k=1}^{K} \min\left[\frac{1}{3}\log_{10}(1 + r^2 snr_k^2), 1\right], \tag{2}$$

where snr_k is the SNR (in linear units, not in dB) in kth articulation band, $K=20$ is the number of articulation



(a)

(b)

FIG. 7. (Color online) (a) Consonant recognition scores $P_C(AI)$, and (b) consonant recognition error $1-P_C(AI)$ (Log scale), plotted as a function of AI. The dashed lines represent individual consonants, while the three colored solid lines represent average values for the three consonant sets. The average consonant score (thick dash-dotted line) is very close to that predicted by the AI model $P_C(AI) = 1 - e_{\mathrm{chance}} e_{\mathrm{min}}^{AI}$ (thick solid line). The data for the Quiet condition are not shown.

bands, and r is a factor that accounts for the peak-to-rms ratio for the speech.[2] The peak-to-rms ratios for the CV sounds used in UIUCs04, estimated using the method described in Appendix A, were found to vary over articulation bands. Therefore, a frequency-dependent value of r, denoted as r_k, was used for estimating AI. The resulting expression for the AI becomes

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

where r_k values are directly estimated from the speech stimuli (Appendix A).

The AI values were calculated for all SNRs, except the quiet condition, using the same 20 articulation bands ($K=20$) as those specified by Fletcher (1995) and used by Allen (2005b). The AI for the quiet condition cannot be directly estimated, as the actual SNR for that condition is not known.

Figure 7(a) shows the individual consonant recognition scores of UIUCs04 data, plotted as a function of AI. The average recognition scores (c, dash-dotted line) match very closely with the predictions of the AI model $1 - e_{\mathrm{min}}^{AI}$, with $e_{\mathrm{min}} = 0.003$ (black solid curve). Following the transformation from the wideband SNR to the AI scale, the recognition scores for sets C2 and C3 nearly overlap. This is because the AI accounts for the spectral differences between sets C2 and C3 (see Sec. III D 1). However, the curve for C1 scores remains lower than the other two sets. We therefore conclude that, in addition to the spectral differences, there are other differences between sets C1 and C2, which cannot be accounted by Eq. (3).

Figure 7(b) shows that the log errors for all consonants, with the exception of four C2 consonants (green lines), are linear functions of the AI with different slopes. The slopes are given by the $\log(e_{\min})$ for each consonant. The C1 consonants, with the exception of /m/ (*), have significantly higher $e_{\min}$ values than the remaining consonants. The approximate $e_{\min}$ values for sets C1, C2, and C3 are 0.01, 2×10^{-5}, and 3×10^{-5}, respectively. This explains why the curves of C1 consonants do not overlap with those of C2 and C3 consonants. Note that C1 consonants were the most frequent among the confusable utterances (Fig. 4). Thus, the $e_{\min}$ for the C1 consonants would be even higher without utterance selection. High $e_{\min}$ values for the C1 consonants are also observed in our analysis of the Grant and Walden (1996) data (see Sec. III G).

The total recognition error $1-P_C$ (black dash-dotted line), which is the average of errors for the three sets (colored solid lines), can be expressed as

$$1 - P_C(AI) = \frac{1}{3}[e^{AI}_{\min,\mathrm{C1}} + e^{AI}_{\min,\mathrm{C2}} + e^{AI}_{\min,\mathrm{C3}}]e_{\mathrm{chance}} \tag{4}$$

$$= \frac{1}{3}[(0.01)^{AI} + (2 \times 10^{-5})^{AI} + (3 \times 10^{-5})^{AI}]e_{\mathrm{chance}}. \tag{5}$$

Since the total error is a sum of exponentials with different bases, it need not be an exponential. However, in this case, the exponential model $e_{\mathrm{chance}}e^{AI}_{\min}$ with $e_{\min}=0.003$ (solid black line) fits very closely to the average error $1-P_C(AI)$.

D. Confusion analysis

In this section, we analyze the individual confusions. Figure 8 shows the 64×64 row-normalized syllable CM at four different SNRs, displayed as gray-scale images. The intensity is proportional to the log of the value of each entry in the row-normalized CM, with black representing a value of unity and white representing the chance-level probability of 1/64. The rows and columns of the CM are arranged such that four CVs having the same consonant are consecutively placed with vowels /ɑ/, /ɛ/, /ɪ/, and /æ/, in that order. The consonants are stacked according to the sets C1, C3, and C2, separated by dashed lines.

For SNR$\leqslant -16$ dB, $\lambda < 1$ for the C2 consonants implies that the syllables with C2 consonants have more vowel confusions than the consonant confusions. This shows up in the CM images as the blocks around the diagonal for the CVs with C2 consonants. On the other hand, $\lambda > 1$ for the C1 consonants results in lines parallel to the diagonal in the C1 part of the CM images. The parallel-line structure is prominent at SNRs> -20 dB, however as the SNR decreases, vowel confusions appear, smearing the parallel lines.

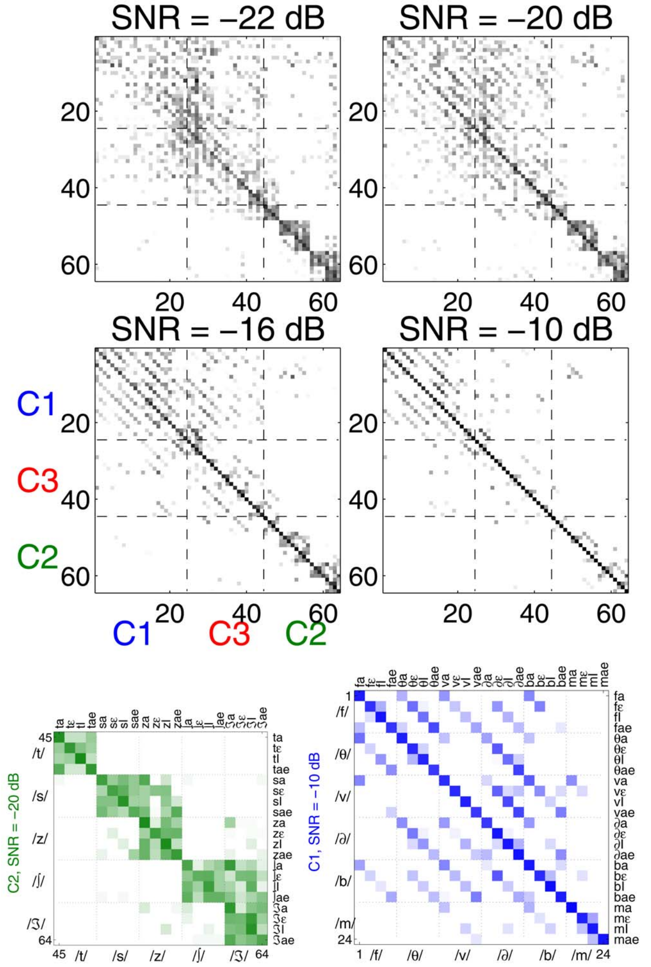


FIG. 8. (Color online) The four (2×2) small panels show the gray-scale images of the CMs at four SNR values. The gray-scale intensity is proportional to the log of the value of each entry in the row-normalized CM, with black color representing unity and white color representing the chance performance (1/64). Dashed lines separate sets C1 (Nos. 1–24), C3 (Nos. 25–44), and C2 (Nos. 45–64), in that order, from left to right and top to bottom. The two enlarged color panels at the bottom show set C2 at −20 dB SNR and set C1 at −10 dB SNR.

Sets C1 and C3 are confused with each other, while set C2 has negligible confusions with the other two sets. This correlates with the spectral powers of the consonants in the three sets (see Sec. III D 1). The asymmetry in the C1–C3 confusions, especially at −20 dB SNR, can be easily explained based on the recognition performances of C1 and C3 consonants. At −20 dB SNR, the C1 consonants are very close to chance level, while C3 consonants have scores between 20% and 50%. Thus, C1 consonants are confused with C3 consonants but not vice versa, which gives rise to the asymmetric confusions between sets C1 and C3.

Within set C2, there are asymmetric confusions between /s/-/z/ and /ʃ/-/ʒ/. This asymmetry is further investigated in Sec. III E. A few vertical lines can be observed in the CM images, suggesting some kind of bias towards certain CVs, however there is no consistent trend in terms of consonants or vowels in these lines.

1. Consonant PSD analysis

The nature of the confusions among the three consonant sets correlate with the SNR spectrum of the consonants. The

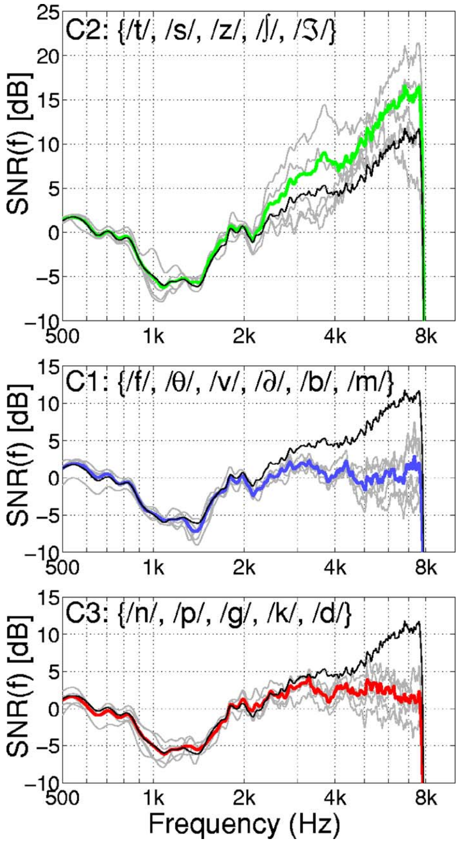


FIG. 9. (Color online) The SNR spectra $[\mathrm{SNR}(f)]$ for consonants in set C2 (*top*), C1 (*center*), and C3 (*bottom*). The thin gray lines show the SNR spectra for individual consonants while the thick colored line in each panel shows the average SNR spectrum for that set. Each panel contains the SNR spectrum for average speech (thin black line), estimated using the speech and noise spectra shown in Fig. 2.

SNR spectrum for a consonant is defined here as the ratio of power spectral density (PSD) of that consonant to the PSD of the noise. To estimate the PSD of a consonant, the PSD of all CV utterances with the given consonant were averaged. Such an average would practically average out the spectral variations due to the four different vowels and enhance the consonant spectrum.

Figure 9 shows the SNR spectra for consonants (thin gray lines) in sets C2 (top), C1 (center), and C3 (bottom) at 0 dB wideband SNR. Each panel shows the average SNR spectra for that set (thick colored line), as well as the SNR spectra for the average speech (thin black line). The average speech PSD starts to roll-off at 800 Hz, while the noise PSD is flat up to 1 kHz (Fig. 2). The speech PSD crosses over the noise PSD at about 2 kHz. Correspondingly, the SNR spectra for average speech has a valley between 500 Hz and 2 kHz. Above 2 kHz, speech dominates the noise, resulting in the high-frequency boost in the SNR spectrum that is more than 10 dB above 6 kHz.

The C2 consonants have rising SNR spectra at high frequencies, while those of C1 and C3 either remain flat or slightly drop at higher frequencies, in spite of the high-frequency boost. The high-frequency energy makes the SNR spectra of C2 consonants significantly different from the other consonants, resulting in high scores and very few confusions for the C2 consonants. The SNR spectra of C1 consonants are indistinguishable from those of C3 consonants, which explains why C1 and C3 consonants are confused with each other, but not with C2 consonants.

2. Confusion matrices

There are 64 curves in each CV confusion pattern for the 64×64 CM, which makes it very difficult to analyze the confusions. Also, the row sums for the 64×64 CM are not large enough to obtain smooth curves in the confusion patterns. Therefore, we analyze the consonant and vowel confusions separately. We will also analyze the interdependence of consonant and vowel confusions.

To analyze the consonant confusions, the responses were scored for consonants only. This resulted in a 64×16, syllable-dependent consonant CM $P(C_h|C_sV_s)$. Averaging the rows of this CM over the spoken vowel gives a 16×16 vowel-independent consonant CM $P(C_h|C_s)$. Similar CMs can be generated to analyze vowel confusions. Five such CMs are listed in Table I, including the two CMs that are generated for vowel analysis, which will be discussed in Sec. III F.

E. Consonant confusions

The perceptually significant consonant confusions (i.e., those with a well-defined SNR_g) observed in UIUCs04 are /m/-/n/ (set C3), /f/-/θ/, /b/-/v/-/ð/, /θ/-/ð/ (set C1), /s/-/z/ and /ʃ/-/ʒ/ (set C2). Note that each of these confusion groups is within one of the three sets. At −2 dB SNR, more than 84% of the consonant confusions are within the three consonant sets. The consonant confusions across the three sets increase with decrease in SNR, but are mostly between sets C1 and C3.

The consonant confusions for C2 consonants do not depend on the following vowel, as $\lambda < 1$ for set C2. When C2

TABLE I. Mathematical expressions, sizes, and descriptions of the five basic types of CM used in this study. C_s and V_s indicate the spoken consonant and vowel, while C_h and V_h represent the consonant and vowel reported by the listener.

CM description	Size	Expression
Syllable (CV) confusions	64×64	$P(C_hV_h\|C_sV_s)=P_{s,h}(SNR)$
CVs scored on consonants	64×16	$P(C_h\|C_sV_s)=\Sigma_{V_h}P(C_hV_h\|C_sV_s)$
Consonant confusions	16×16	$P(C_h\|C_s)=\Sigma_{V_s}P(C_h\|C_sV_s)$
CVs scored on vowels	64×4	$P(V_h\|C_sV_s)=\Sigma_{C_h}P(C_hV_h\|C_sV_s)$
Vowel confusions	4×4	$P(V_h\|V_s)=\Sigma_{C_s}P(V_h\|C_sV_s)$

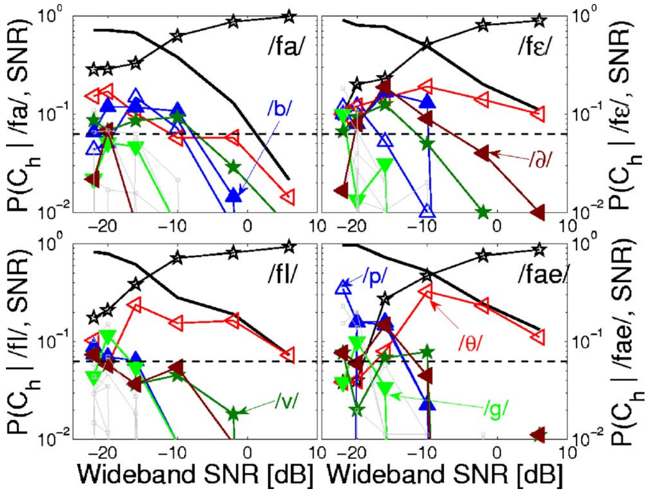


FIG. 10. (Color online) Consonant CPs $P(C_h|C_sV_s,SNR)$ (64×16) for C_s $=$/fɑ/ (top left), /fɛ/ (top right), /fɪ/ (bottom left), and /fæ/ (bottom right). $P(C_h|/\text{f}V_s/,SNR)$ are four rows of the 64×16 consonant CM that correspond to presentation of consonant /f/ and vowel V_h at a given SNR. The gray thin lines with square symbols in the CP figures represent the sounds that are not confused with the diagonal sound and hence do not cross above the chance level. In all CP figures, the quiet condition is plotted at +6 dB SNR for convenience.

consonants start to get confused $(\text{SNR}\leq -20$ dB), the vowels are hardly recognizable (Fig. 5) and are very close to being inaudible. The vowels can affect the consonant confusions only if they have high recognition when the consonants are being confused. Thus, only for consonants with $\lambda > 1$ (sets C1 and C3), the CPs can depend on the following vowel.

1. Vowel-dependent consonant confusions

The vowel-dependent 64×16 consonant CM $P(C_h|C_sV_s)$ showed that the CPs for some consonants depend on the spoken vowel V_s. Figure 10 shows the CPs for the four CV sounds with consonant /f/. The strongest competitor /θ/ (symbol ◁) stood out from the other competitors for the sounds /fɪ/ and /fæ/; it was closely accompanied by the secondary competitors (/b/, /ð/, and /v/) in case of /fɛ/, while it was buried as a secondary competitor of /fɑ/. Identical trends were observed for consonants /θ/, /v/, /ð/, and /m/ (all in C1). For /b/ and some C3 consonants, the CPs varied with V_s, but the variations had no specific identifiable trend.

2. Vowel-independent consonant confusions

Since the CPs for four C2 consonant /s/, /ʃ/, /z/, and /ʒ/ are independent of the spoken vowel they could be averaged across V_s. Figure 11 shows the corresponding four rows of the vowel-independent 16×16 consonant CM, $P(C_h|C_s)$. The /s/-/z/ and /ʃ/-/ʒ/ confusions are highly asymmetric (Fig. 8, set C2). The total error in recognizing unvoiced consonants /s/ and /ʃ/ can be accounted by the confusions with the voiced consonants /z/ and /ʒ/, respectively, whereas /z/ and /ʒ/ have multiple competitors that contribute to the total error. Thus the asymmetry is biased towards the voiced consonants /z/ and /ʒ/, i.e., these two are the preferred choices in /s/-/z/ and /ʃ/-/ʒ/ confusions in speech-weighted noise. The asymmetric parts of the confusion probability are as high as 0.13 and 0.14 for /s/-/z/ and /ʃ/-/ʒ/ confusions, respectively.

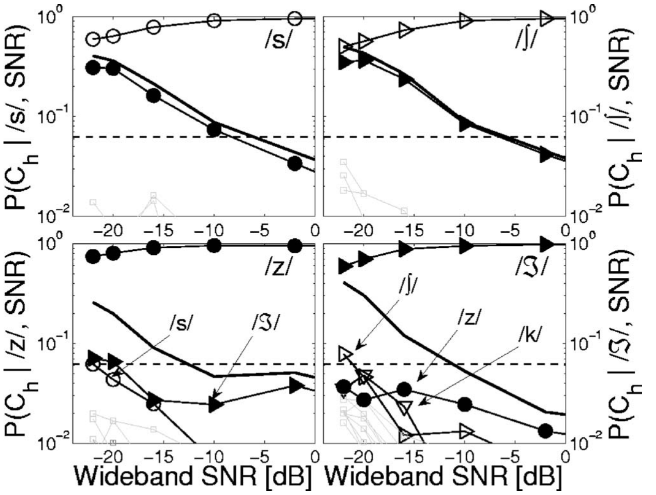


FIG. 11. Consonant CPs $P(C_h|C_s,SNR)$ (16×16) for consonants /s/ (top left), /ʃ/ (top right), /z/ (bottom left), and /ʒ/ (bottom right). The unvoiced consonants (top panels) have only one strong competitor which accounts for the total recognition error (thick solid line), while the voiced consonants have multiple competitors that contribute to the total error.

This asymmetry is slightly greater than the largest asymmetry found in the consonant CM of MN55, which was 0.1, but for a different set consonant confusion pair (Allen, 2005b).

F. Vowel confusions

The vowel confusions were analyzed using the 64×4 consonant-dependent vowel CM $P(V_h|C_sV_s)$ (Table I) and were found to be independent of the preceding consonant C_s. Therefore, the vowel confusions were averaged over C_s, giving the 4×4 vowel CM $P(V_h|V_s)$.

Figure 12 shows these consonant-independent vowel CPs $P(V_h|V_s)$. At very low SNR values, all entries in the 4 $\times 4$ vowel CM converge to the chance level performance for recognizing the vowels (1/4). The recognition score for each of the four vowels was greater than 30% at −22 dB SNR, not low enough to see clear groupings having a well-formed SNR_g, with the exception of $P(/ɪ/|/ɛ/)$ (top right panel, Fig. 12). However, the off-diagonal entries show some interesting

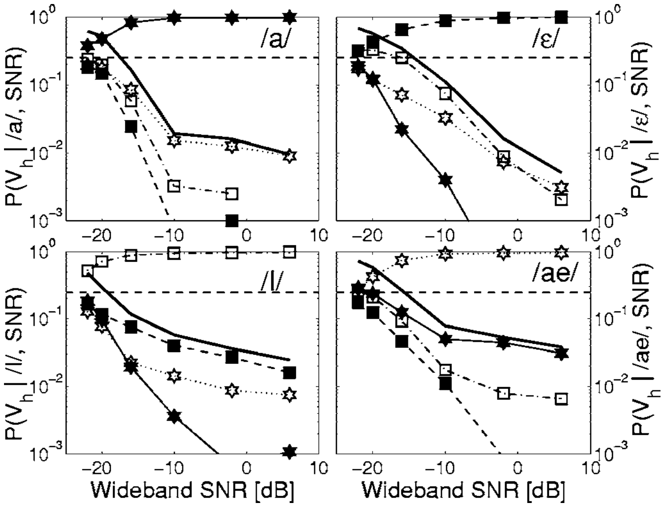


FIG. 12. Consonant-independent 4×4 vowel CPs $P(V_h|V_s,SNR)$ The legend for vowel symbols is given in Fig. 5, top panel.

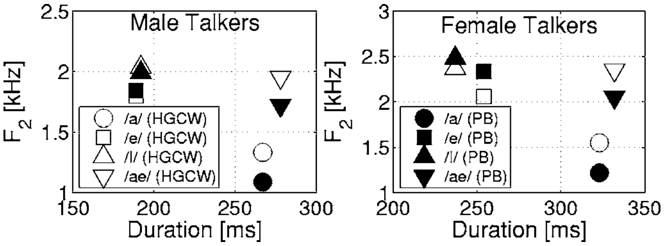


FIG. 13. Plots of the average values of the second formant frequency (F_2) of vowels vs the vowel durations for male (left panel) and female talkers (right panel). The values of the duration are from Hillenbrand *et al.* (HGCW), while the values of F_2 are from HGCW (hollow symbols) as well as Peterson and Barney (PB) (filled symbols), estimated using isolated /hVd/ syllables.

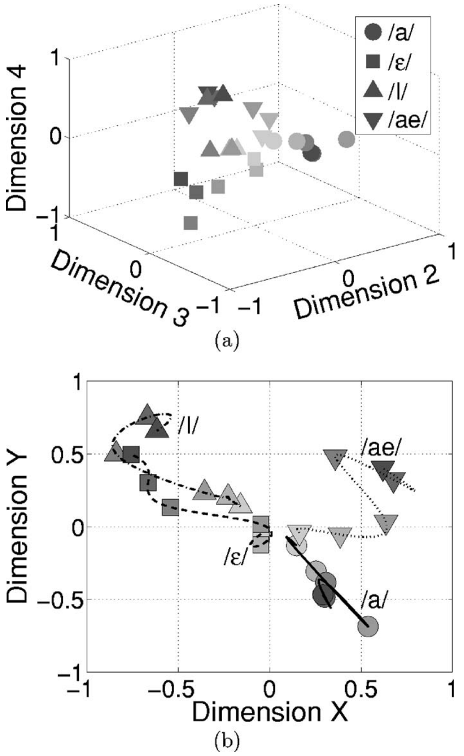


FIG. 14. (a) Vowel clustering in 3D eigenspace (dimensions 2–4) of the 4 ×4 vowel CM $[P(V_h|V_s)]$. The gray-scale intensity of the symbols corresponds to the six SNR levels (i.e., the lightest≡−22 dB SNR and the darkest≡Quiet). (b) Two-dimensional projection of the vowel clusters. The projection matches the clean speech clustering with the vowel distribution in the left panel of Fig. 13. The lines indicate paths traced by the vowels in the 2D plane of projection, as the SNR decreases.

behavior, at scores that are an order of magnitude smaller than the chance level. Due to the very large row sums (1700–2000 responses), the data variability is relatively small, resulting in well-defined curves, even at such low values.

At very low SNR, each vowel seems to be equally confused with the other three vowels, except for /ɛ/, which clearly formed a group with /ɪ/ ($SNR_g \approx -20$ dB, top right panel of Fig. 12). But as the SNR is increased, /ɛ/ becomes equally confused with /ɪ/ and /æ/, though the total number of confusions decrease. For the other three vowels, the curves of the off-diagonal entries separate, showing a clear rank ordering in the confusability. Above −10 dB, /æ/ and /ɑ/ emerge to be the strongest competitors of each other (top left and bottom right panels), with /ɪ/ being the next stronger competitor and /ɛ/ being the weakest competitor for both vowels. The vowel /ɛ/ is the strongest competitor of /ɪ/ above −20 dB (bottom left panel), with /æ/ as the second strongest competitor.

Thus, the four vowels seem to fall into two perceptual groups: {/ɑ/, /æ/} and {/ɛ/, /ɪ/}. These two groups correlate with the vowel durations (Hillenbrand *et al.*, 1995), i.e., /ɑ/-/æ/ are long, stressed vowels, while /ɛ/-/ɪ/ are short and unstressed. Vowel /æ/ is a stronger competitor than /ɑ/ for the short vowels /ɛ/ and /ɪ/ at SNR ⩾−16 dB. This relates to the second formant frequencies of the vowels [Peterson and Barney (1952); Hillenbrand *et al.* (1995)], which would be audible at higher SNRs. Figure 13 shows the vowel durations measured by Hillenbrand *et al.* (HGCW) versus the second formant frequencies measured by HGCW and Peterson and Barney (PB) for our four vowels, categorized by the talker gender. The vowel group /ɛ/-/ɪ/, which was the only vowel group with a clear SNR_g, is much more compact than the /ɑ/-/æ/ group in the Duration-F_2 plane.

1. Vowel clustering

A principal component analysis was performed on the 4×4 vowel CM $[P(V_h|V_s)]$ to analyze the grouping of vowels. The four dimensions of the eigenvectors were rank ordered from 1 to 4 in the decreasing order of the corresponding eigenvalues. The highest eigenvalue was always unity since the vowel CM was row normalized and the coordinates of the four vowels along the corresponding dimension (i.e., Dimension 1) were identical (Allen, 2005a). Figure 14(a) shows the four vowels in the vector space of the remaining three dimensions. The gray-scale intensities of the symbols show the six SNR levels, with the lightest corresponding to −22 dB SNR and the darkest corresponding to the quiet condition. The clustering of the vowels in the three-dimensional (3D) eigenspace, when projected on a specific plane in the eigenspace, is very close to the graph of vowel duration versus the second formant frequencies (Fig. 13). The procedure used for obtaining the two-dimensional (2D) projection [Fig. 14(b)] is described in Appendix B. The dimensions of the 2D projection are abstract and hence the axes are labeled *Dimension X* and *Dimension Y*. However, the dimensions *X* and *Y* are closely related to the vowel duration and the second formant frequency, respectively. The projection coefficients indicate that dimension *X* is almost identical to Dimension 2 (see Appendix B), which is associated with the largest eigenvalue of the 3D subspace. This suggests the vowel duration was the most dominant acoustic cue for the perceptual grouping of the four vowels.

The addition of masking noise reduces the perceptual distance among the vowels and draws them closer in the eigenspace. The vowel /ɛ/ is perceptually closer to /ɪ/ for an SNR as low as −16 dB. However, below −16 dB, /ɛ/ makes a large shift towards /ɑ/ and /æ/, and becomes equally close to the three vowels in the eigenspace. This is consistent with the vowel CPs for /ɛ/ (Fig. 12, bottom left panel). The vowel /ɪ/ is the most remote in the presence of noise, which is consistent with its highest scores (Fig. 5).

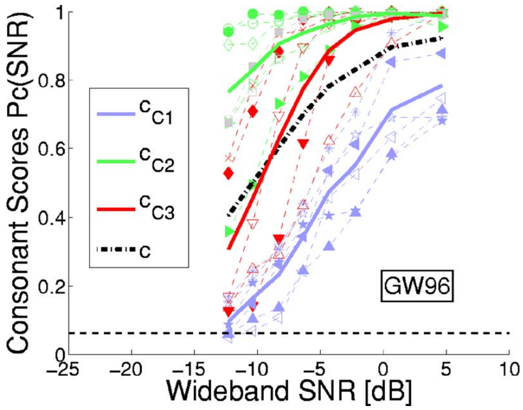


FIG. 15. (Color online) Consonant recognition scores for the 18 consonants used by Grant and Walden (1996). The hollow square and the opaque square represent the consonants /tʃ/ and /dʒ/, respectively.

G. Comparison with the past work

1. Grant and Walden (1996)

The Grant and Walden (1996) [GW96] consonant recognition scores (Fig. 15), measured in speech-weighted noise, also show consonants grouped into the same three sets. However, set C3 is not as tightly formed in GW96 as in UIUCs04. At 50% score the SNR spread of C3 consonants in UIUCs04 is about 2 dB while that in GW96 is about 7 dB. The average consonant recognition in GW96 is smaller than UIUCs04, primarily because of the low scores of C1 consonants in GW96. Note that our utterance selection (Sec. III B) removed mostly C1 consonant syllables. Without utterance selection, the C1 scores in the two experiments are closer in quiet. However, in the presence of noise, the C1 scores and hence the average consonant recognition in UIUCs04 still remain significantly greater than that in GW96. There could be several reasons for these differences. For example, the average speech spectrum in GW96, which was for a single female talker, had a greater roll-off than that in UIUCs04, which had 18 talkers. Also the noise spectrum in GW96 was a better match to the average speech spectrum than in UIUCs04. In spite of these differences, the consonant confusions in GW96 (not shown) are very similar to those in UIUCs04.

2. Miller and Nicely (1955)

Unlike MN55, the consonant groupings observed in UIUCs04 and in GW96 were not correlated with the *production* or the *articulatory* features (sometimes known as the *distinctive* features) such as voicing or nasality. In fact, a large number of voicing confusions such as /s/-/z/ and /ʃ/-/ʒ/ were observed in UIUCs04. Furthermore, the stop plosives { /p/, /t/, /k/ } and { /b/, /d/, /g/ } did not form perceptual groups in speech-weighted noise, as they do in MN55 data.

If a noise masker masks the consonants uniformly, then the consonant scores should be almost identical at a given SNR, with very little spread. The maximum spread in the UIUCs04 consonant scores is at −20 dB SNR (Fig. 5, bottom panel), with consonant /v/ (★) at 6% and consonant /z/ (●) at almost 80%. The consonant scores GW96 show even greater spread at −10 dB SNR, with a similar distribution of individual consonant scores. In comparison, the highest spread of the consonant scores in white noise [i.e., MN55 data, shown in Fig. 6 of Allen (2005b)] is 50–90% at 0 dB SNR, which is almost half of the maximum spread observed in UIUCs04. Nasals have very high scores in MN55 data. Unlike UIUCs04 and GW96, the consonant scores in MN55 form a continuum, rather than distinct sets. Thus, white noise masks consonants more uniformly relative to the speech-weighted noise, which implies that the events for the consonant sounds are distributed uniformly over the bandwidth of speech. The events important for recognizing the C2 set consonants are at the higher frequencies that are relatively less masked by the speech-weighted noise. The events for nasals are located at low frequencies, which are masked more by speech-weighted noise than white noise.

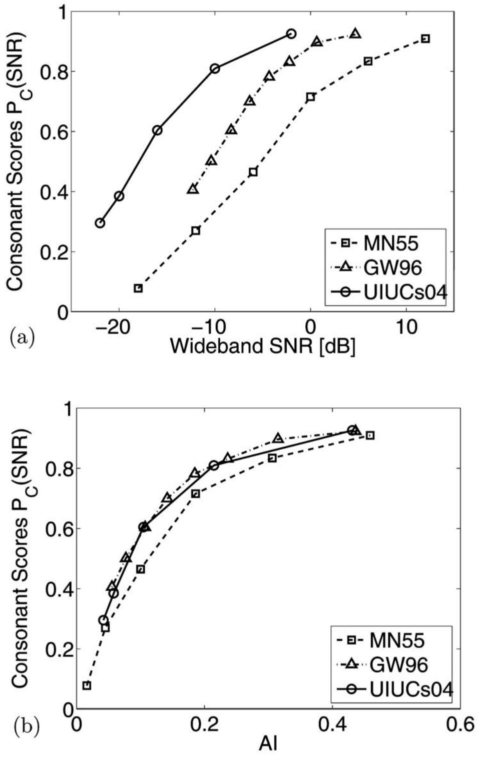


FIG. 16. A comparison of the consonant recognition scores for the current experiment [UIUCs04], Grant and Walden (1996) [GW96], and Miller and Nicely (1955) [MN55] as functions of (a) SNR and (b) AI.

3. AI

At a given wideband SNR, the recognition score of a consonant depends on the spectrum of the noise masker. Therefore, the wideband SNR is not a good parameter for modeling recognition scores and a parameter that accounts for the spectral distribution of speech and noise energy is required. As the AI accounts for the speech and noise spectra, it is a better measure for characterizing and comparing the scores across experiments.

Figure 16(a) shows the consonant scores $P_C(SNR)$ from UIUCs04, MN55, and GW96, as functions of wideband SNR. At 50% score, the SNR of GW96 is about 8 dB higher than UIUCs04, while the MN55 SNR is about 5 dB higher than GW96. The foremost reason for this SNR difference is the different noise spectra in the three experiments.

However, these consonant scores overlap when replotted on an AI scale [$P_C(AI)$, Fig. 16(b)]. The AI values for GW96 were calculated using the spectra and the peak-to-rms ratios (r_k) estimated from the GW96 stimuli. The original stimuli for MN55 are not available and therefore the AI values from Allen (2005a), which were estimated using Dunn and White spectrum and $r=2$, were used. This could be one reason why the $P_C(AI)$ curve for MN55 does not match as closely as the other two curves.

IV. DISCUSSION

In this research we have explored the perception of CV sounds in speech-weighted noise. Our analysis tried to control for the many possible sources of variability. For example, we do not want the bad or incorrectly labeled utterances to be misinterpreted as the perceptual confusions in noise. There is no gold standard for a correct utterance. We used the listener responses in the quiet condition as a measure to select the good utterances. It was therefore necessary to make sure that the listeners are performing the given task accurately, in quiet. Hence the responses of four LP listeners, having significantly lower scores than the ten HP listeners, were removed before the utterance selection.

The HP listener responses showed that although 59% of the utterances had zero error in the quiet, there were few utterances which had more than 80% error and very small response entropy (not shown). A low response entropy for a high-error utterance indicates a clear case of mislabeling, i.e., the utterance was consistently perceived, but the perception of the CV was different from its label. Such utterances must be either removed or relabeled. Other high-error utterances had high response entropy, which indicates that the listeners were unsure about these utterances. The syllable error threshold for separating the good utterances from the high-error utterances should be set according to the experimental design and aims. For our purpose, we selected a conservative threshold of 20%. Also, the results were not significantly different for a 50% threshold. The listener selection and the utterance selection are interdependent. However, we verified that the HP-LP listener classification was unaffected by the utterance selection.

Another source of variability is the primary language of the listeners and the talkers. In addition to the 14 L1 =English listeners, there were 6 L1 ≠ English listeners who completed the experiment. Three of the L1 ≠ English listeners had scores worse than the LP listeners while only one L1 ≠ English had scores comparable to that of the HP listeners. Since it has been shown that the primary language affects the consonant and vowel confusions [Singh and Black (1965); Fox *et al.* (1995)], the analysis in this paper was limited to only L1=English listeners. All the talkers from LDC2005-S22 database were native speakers of English and three of those were bilingual. The syllable errors were not different for bilingual and monolingual talkers, and therefore, all talkers were used.

Once the listeners and the utterances were selected, it was possible to reliably study the effects of noise. The speech-weighted noise masks the vowels uniformly, but has a nonuniform masking effect on the consonants, dividing them into three sets: low-scoring C1 consonants, high-scoring C2 consonants, and the remaining consonants, clubbed together as C3, and having intermediate scores. The predominant sets C1 and C2 are also observed in GW96. However, in case of MN55, no distinct consonant sets are observed and the spread in the consonant scores is much smaller in white noise (i.e., MN55) relative to the speech-weighted noise (UIUCs04 and GW96).

Analysis of the 64 × 64 CM (Sec. III D) shows two well-defined structures that relate to the consonant sets. The syllables with C1 consonants ($\lambda > 1$) show the parallel-line structure (i.e., consonant confusion but correct vowel) while those with C2 consonants ($\lambda < 1$) show the diagonal blocks (i.e., vowel confusion but correct consonant). Thus the vowel-to-consonant recognition ratio λ quantifies the qualitative analysis of CM images. It is also correlated with the vowel dependence of the consonant confusions. The CPs for consonants with $\lambda > 1$ (sets C1 and C3) are more likely to be affected by the following vowel than those for consonants with $\lambda < 1$ (i.e., set C2).

The consonant PSDs and therefore the SNR spectra (Fig. 9) are dominated by the vowels at low frequencies, but a clear difference can be observed at the high frequencies. The high SNR at high frequencies distinguishes C2 consonants from the other two sets. The PSDs of C3 consonants are indistinguishable from the C1 PSDs, which explains why C1 and C3 consonants are confused with each other, but not with the C2 consonants. The C2 scores are higher than C1 and C3 scores at a given SNR (Fig. 5) due to the high SNRs at high frequencies. This spectral difference is accounted by the AI, which makes the $P_C(AI)$ curves for C2 and C3 overlap on the AI scale [Fig. 7(a)]. However, the $P_C(AI)$ curves for C1 do not overlap with the C2 and C3 curves, due to higher $e_{\min}$ values. There may also be spectral differences between C1 and the remaining consonants at lower frequencies, which are dominated by the vowel energy. In such a case, the spectral differences are not detectable in the SNR spectra and therefore cannot be accounted for by the AI. Also, note that since most of the removed utterances had C1 consonants, these consonants are not only hard to perceive, but are also difficult to pronounce clearly.

Confusions within the C2 consonants are highly asymmetric and are biased in favor of the voiced consonants (Fig. 11). These asymmetric confusions are not observed in MN55. Therefore, it is possible that the speech-weighted noise, which has more energy at low frequencies, introduces a percept of voicing. Another explanation for the asymmetry is that the speech-weighted noise masks the voicing information (i.e., either presence or absence) at the low frequencies and in absence of this information, human auditory system assumes the voicing to be present, by default. Specific experiments would be required to test these hypotheses.

In several cases, there is a noticeable variation in the consonant confusions for different utterances of the same CV (not shown). This variation is obscured after pooling the responses to all utterances of a given CV. Some utterances show interesting phenomenon that we call *consonant morphing*, i.e., when confusion of a consonant with another consonant is significantly greater than its own recognition. The confusion threshold of an utterance depends on the intensities of various features in that utterance. This natural variability in speech could be used to locate the perceptual features. For that matter, the confusable sounds with high response entropy could be a blessing in disguise. Comparing spectro-temporal properties of such sounds with that of the nonconfusable sounds will provide vital information about the perceptual features.

The SNRs used in this study were not low enough to get clear perceptual grouping of vowels, as defined by SNR_g, in spite of having close formant frequencies. The four vowels are uniformly masked by the speech-weighted noise, resulting in practically overlapping recognition scores $P_C(SNR)$. However, based on the hierarchy of the competitors in the vowel CPs, vowels formed two groups—the long, stressed vowels (/ɑ/-/æ/) and the short, unstressed vowels (/ε/-/ɪ/). The eigenspace clustering of the vowels is strikingly similar to that in the Duration-F_2 space, with the Duration relating to the strongest eigenspace dimension. The vowel confusions were found to be independent of the preceding consonant. However, these observations should be verified with a larger set of vowels before generalizing.

Finally, we compare the consonant scores from UIUCs04 with the Grant and Walden (1996) and Miller and Nicely (1955) scores (Fig. 16). The $P_C(SNR)$ curves for the three experiments are neither close nor parallel to each other on the wideband SNR scale, due to different noise spectra. However, the $P_C(AI)$ curves practically overlap. Thus, we have shown that, in spite of different experimental conditions, the AI can consistently characterize and predict the consonant scores, for any speech and noise spectra.

V. CONCLUSIONS

The important observations/implications from this study can be briefly summarized as follows.

1. Unlike the white noise, the speech-weighted noise nonuniformly masks the consonants, resulting in a larger spread in the consonant recognition scores. The C1 consonants (/f/, /θ/, /v/, /ð/, /b/, /m/) have the lowest scores while consonants C2 (/s/, /ʃ;/, /z/, /ʒ/, /t/) have the highest scores (Fig. 5, bottom). The remaining consonants have scores between the C1 and C2 scores and are grouped together as set C3.
2. Sets C1 and C3 are confused with each other with some degree of asymmetry, but set C2 is not confused with the other two groups (Fig. 8). This is consistent with the spectral power of the consonants above the noise spectrum (i.e., the SNR spectra, Fig. 9). The asymmetric confusions between sets C1 and C3 can be explained by the difference in their recognition scores.
3. The consonant confusion groups in speech-weighted noise are C1: { /f/-/θ/, /b/-/v/-/ð/, /θ/-/ð/ }, C2: { /s/-/z/, /ʃ/-/ʒ/ }, and C3: /m/-/n/ (Sec. III E). There is no across-set consonant group. Unlike the white-noise case (MN55), there are very high voicing confusions in the speech-weighted noise. The perceptual groups /s/-/z/ and /ʃ/-/ʒ/ are highly asymmetric, biased in favor of the voiced consonant in the presence of noise (Fig. 11).
4. The vowel-to-consonant recognition ratio λ is a quantitative measure of the confusions observed in the CM images, i.e., $\lambda > 1 \Rightarrow$ consonant confusions dominate, resulting in the parallel lines, while $\lambda < 1 \Rightarrow$ vowel confusions dominate, resulting in the diagonal blocks in CM images.
5. The confusions for set C1 ($\lambda > 1$) depend on the vowels, while those for set C2 ($\lambda < 1$) are independent of vowel. (Sec. III E.)
6. Vowels are uniformly masked by the speech-weighted noise (Fig. 5, top) and form two confusion groups, viz. /ɑ/-/æ/ and /ε/-/ɪ/. The eigenspace clustering of the vowels (Fig. 14) relates to the duration and the second format frequencies of the vowels (Fig. 13).
7. The recognition errors for 12 of the 16 consonants (dashed lines, Fig. 7) used in this study, as well as the average error (dash-dotted line) can be modeled with the exponential AI model [Eq. (1)] proposed by Allen (2005b). However, the model works better with a frequency-dependent peak-to-rms ratio r_k [Eq. (3)], than the frequency-independent ratio (Allen, 2005b).
8. The Articulation Index accounts for the spectral differences in the speech and noise spectra and is a better parameter than the wideband SNR for characterizing and comparing the consonant scores across experiments (Fig. 16).

ACKNOWLEDGMENTS

We thank all members of the HSR group at the Beckman Institute, UIUC for their help. We thank Andrew Lovitt for writing a user-friendly code that made data collection easier and faster. Bryce Lobdell's input was crucial in revising the manuscript. We are grateful to Kenneth Grant for sharing his confusion data, stimuli, and his expertise.

APPENDIX A

Traditionally, the peak level of speech is measured using the volume-unit (VU) meter. The peak level is given by the mean value of peak deflections on the VU meter ("dBA fast" setting) for the given speech sample (Steeneken and Houtgast, 2002). The peak deflections in the VU meter correspond to the peaks of the speech envelope, estimated in $T=1/8$ s intervals [Lobdell and Allen (2006), French and Steinberg (1947)].

The speech signal filtered through each of the K articulation bands has the same bandwidth B_k as that of the articulation band. Therefore, for estimation of the envelope with optimum sampling, according to the Nyquist criterion, the duration of intervals is selected to be

$$T_k = \frac{1}{2B_k} = \frac{1}{2(f_{U_k} - f_{L_k})}, \tag{A1}$$

where f_{L_k} and f_{U_k} are, respectively, the lower and the upper cutoff frequencies of the kth articulation band. The value of r_k is then calculated as

$$r_k = \frac{p_k}{\sigma_k}, \tag{A2}$$

where σ_k is the rms value of the speech signal filtered through kth articulation band and p_k is the envelope peak for the same filtered speech. The value of r_k increases with the center frequency of the articulation band, ranging from 3.3 ($\approx$10.4 dB, for /n/) in the lowest articulation band (200–260 Hz) to 11.2 ($\approx$21.0 dB, for /d/) in the highest articulation band (6750–7300 Hz). For the GW96 stimuli, the r_k values ranged from 1.2 ($\approx$1.56 dB, for /ʃ/) in the 200–260 Hz band, to 8.98 ($\approx$19.1 dB, for /d/) in the 6370–6750 Hz band.

APPENDIX B

The PCA or the eigenvalue decomposition of the 4×4 vowel CM $P(V_h|V_s)$ can be represented in matrix form as $P(V_h|V_s) = EDE^{-1}$, where

$$D = \begin{bmatrix} D_1 & 0 & 0 & 0 \\ 0 & D_2 & 0 & 0 \\ 0 & 0 & D_3 & 0 \\ 0 & 0 & 0 & D_4 \end{bmatrix}$$

is the rank-ordered eigenvalue (singularity) matrix, with D_1 being the largest eigenvalue and D_4 being the smallest eigenvalue, and

$$E = [\mathbf{E}_1\ \mathbf{E}_2\ \mathbf{E}_3\ \mathbf{E}_4] = \begin{bmatrix} e_{11} & e_{21} & e_{31} & e_{41} \\ e_{12} & e_{22} & e_{32} & e_{42} \\ e_{13} & e_{23} & e_{33} & e_{43} \\ e_{14} & e_{24} & e_{34} & e_{44} \end{bmatrix}$$

is the eigenvector matrix. Each eigenvector represents a dimension in the eigenspace. Because $P(V_h|V_s)$ is row normalized, D_1 is unity and the coordinates along the first dimension are identical (e_{1i}=0.5). Therefore, the vowel clustering in the eigenspace is plotted only along the dimensions 2–4 [Fig. 14(a)]. The 4×3 coordinate matrix for the four vowels along the three dimensions is

$$C = (E\sqrt{D})_{2-4} = \begin{bmatrix} \sqrt{D_2}e_{21} & \sqrt{D_3}e_{31} & \sqrt{D_4}e_{41} \\ \sqrt{D_2}e_{22} & \sqrt{D_3}e_{32} & \sqrt{D_4}e_{42} \\ \sqrt{D_2}e_{23} & \sqrt{D_3}e_{33} & \sqrt{D_4}e_{43} \\ \sqrt{D_2}e_{24} & \sqrt{D_3}e_{34} & \sqrt{D_4}e_{44} \end{bmatrix}.$$

Let

$$F = \begin{bmatrix} f_1 & d_1 \\ f_2 & d_2 \\ f_3 & d_3 \\ f_4 & d_4 \end{bmatrix}$$

be the feature matrix of the four vowels, which contains the values of the second format frequencies F_2 (f_i) and the vowel durations (d_i). The feature matrix F is normalized to have zero mean and unit variance along both dimensions f and d. The 2D projection of C that matches the F can be obtained by the linear transform

$$F = CA, \tag{B1}$$

where A is a 3×2 matrix that rotates C about the origin and orthogonally projects it on a 2D plane. The closed form solution for the minimum mean square estimate for A is

$$\hat{A} = [C^T C]^{-1} C^T F. \tag{B2}$$

The 2D projection in Fig. 14(b) was obtained by matching the eigenvectors for quiet condition to the normalized version of the 2D clustering in Fig. 13, left panel. The feature matrix F was obtained using the average of the second format frequencies (f) and the vowel durations (d) for male talkers. The matrix was then normalized by subtracting the means and dividing by the standard deviations along the f and d dimensions. The projection matrix in this case was

$$\hat{A} = \begin{bmatrix} 0.9932 & -0.7860 \\ 0.5512 & 0.4203 \\ 0.4363 & -0.0261 \end{bmatrix}.$$

The value of coefficient a_{11}, which is the projection of Dimension X on Dimension 2, is almost unity. This suggests that Dimension X is almost the same as Dimension 2, since the angle between the two dimensions is very close to zero $[\cos^{-1}(0.9932) = 6.68°]$.

[1] The "listener scores" are the CV syllable recognition scores, i.e., the scores of recognizing both consonant and vowel correctly. The average consonant scores (c) are equal to the average vowel scores (v) in quiet condition (see Sec. III C). Therefore, a threshold of 85% for CV recognition corresponds to a threshold of 92.2% for phone recognition.

[2] The relative levels of speech and noise spectra are set according to the wide-band SNR, which is calculated from the rms levels. The contribution of the articulation bands to the speech intelligibility is proportional to the peaks in the articulation band-filtered speech signal, that are above the noise floor, and therefore a correction for the peak-to-rms ratio of speech is necessary (French and Steinberg, 1947). French and Steinberg (1947) suggested a correction of 12 dB, for all articulation bands, which is consistent with the measured peak-to-rms ratios for speech (Steeneken and Houtgast, 2002), and approximately corresponds to r=4.

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

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

Benson, R. W., and Hirsh, I. J. (**1953**). "Some variables in audio spectrometry," J. Acoust. Soc. Am. **25**, 499–505.

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kholy, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G. I., Westerman, S., and Ludvigsen, C. (**1994**). "An internation comparison of long-term average speech spectra," J. Acoust. Soc. Am. **96**, 2106–2120.

Campbell, G. A. (**1910**). "Telephonic intelligibility," Philos. Mag. **19**, 152–159.

Cox, R. M., and Moore, J. N. (**1988**). "Composite speech spectrum for hearing aid gain prescriptions," J. Speech Hear. Res. **31**, 102–107.

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. (**1995**). *The ASA Edition of Speech and Hearing in Comunication*, edited by Jont B. Allen (Acoustical Society of America, New York).

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 the International Conference on Spoken Language Processing (ICSLP)*, October 4–8, Jeju, South Korea, http://www.isca-speech.org/archive/interspeech_2004. Viewed 3/26/07.

Fox, R. A., Flege, J. E., and Munro, M. J. (**1995**). "The perception of English and Spanish vowels by native English and Spanish listeners: A multidimensional scaling analysis," J. Acoust. Soc. Am. **97**, 2540–2551.

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, URL http://www.wramc.amedd.army.mil/departments/aasc/avlab/datasets.htm. Viewed 3/26/06.

Hillenbrand, J., Getty, L. A., Clark, M. J., and Wheeler, K. (**1995**). "Acoustic characteristics of American English vowels," J. Acoust. Soc. Am. **97**, 3099–3111.

Lippman, R. P. (**1997**). "Speech recognition by machines and humans," Speech Commun. **22**, 1–15.

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.

Peterson, G. E., and Barney, H. L. (**1952**). "Control methods used in a study of vowels," J. Acoust. Soc. Am. **24**, 175–184.

Singh, S., and Black, J. W. (**1965**). "Study of twenty-six intervocalic consonants as spoken and recognized by four language groups," J. Acoust. Soc. Am. **39**, 372–387.

Sroka, J., and Braida, L. D. (**2005**). "Human and machine consonant recognition," Speech Commun. **45**, 401–423.

Steeneken, H. J. M., and Houtgast, T. (**2002**). "Basics of STI measuring methods," *Past, Present and Future of the Speech Transmission Index*, edited by S. J. van Wijngaarden (TNO Human Factors, Soesterberg, The Netherlands).

Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). "Consonant environment specifies vowel identity," J. Acoust. Soc. Am. **60**, 213–224.

Wang, M. D., and Bilger, R. C. (**1973**). "Consonant confusions in noise: A study of perceptual features," J. Acoust. Soc. Am. **54**, 1248–1266.