Consonant recognition and the articulation index

Jont B. Allen
ECE Department and the Beckman Institute, University of Illinois, Urbana, Illinois 61801

(Received 1 August 2004; revised 14 December 2004; accepted 15 December 2004)

The purpose of this paper is to provide insight into how speech is processed by the auditory system, by quantifying the nature of nonsense speech sound confusions. (1) The Miller and Nicely [J. Acoust. Soc. Am. **27**(2), 338–352 (1955)] confusion matrix (CM) data are analyzed by plotting the CM elements $S_{i,j}(\mathrm{SNR})$ as a function of the signal-to-noise ratio (SNR). This allows for the robust clustering of perceptual feature (event) groups, not robustly defined by a single CM table, where clusters depend on the sound order. (2) The SNR is then re-expressed as an articulation index (AI), and used as the independent variable. The normalized log scores $\log(1-S_{i,i}(\mathrm{AI}))$ and $\log(S_{i,j}(\mathrm{AI}))$, $j\neq i$, then become linear functions of AI, on log-error versus AI plots. This linear dependence may be interpreted as an extension of the band-independence model of Fletcher. (3) The model formula for the average score for the finite-alphabet case $P_c(\mathrm{AI},\mathcal{H})=\Sigma_{i=1}^{N}S_{i,i}/N$ is then modified to include the effect of entropy $\mathcal{H}$. Due to the grouping of sounds with increased SNR (and AI), the sound-group entropy $\mathcal{H}_g$ plays a key role in this performance measure. (4) A parametric model for the confusions $S_{i,j}(\mathrm{AI},\mathcal{H}_g)$ is then described, which characterizes the confusions between competing sounds within a group. © *2005 Acoustical Society of America.*
[DOI: 10.1121/1.1856231]

PACS numbers: 43.71.Gv, 43.71.An, 43.72.Ne [DOS] Pages: 2212–2223

I. INTRODUCTION

The articulation index (AI) and the confusion matrix (CM), denoted $\mathcal{C}_{i,j}$, are two important measures frequently used to characterize human speech recognition. This paper explores the merging of these two measures by expressing the CM as a function of the AI. The natural grouping of sounds, observed by Miller and Nicely in their classic 1955 experiment, is then explained in terms of a change in each group's entropy $\mathcal{H}_g$, which is also a function of the AI. A parametric model of the symmetric form of the CM confusions $S_{i,j}(\mathrm{AI},\mathcal{H})$ is then developed. These parametric models lead to a new formulation of the average score $P_c(\mathrm{AI},\mathcal{H})$, which accounts for chance. This model provides an accurate fit to the raw data, and provides new insight into Fletcher's *band independence* formulation for the average articulation score (Allen, 1994).

A fundamental building block in the theory of speech communication is the articulation index (AI) model of speech sound recognition (French and Steinberg, 1947; Fletcher and Galt, 1950; Allen, 1994). The AI is a speech audibility measure, averaged across many cochlear frequency bands, of a specific speech-to-noise ratio measure, that takes into account the effects of masking and cochlear filtering (Allen, 1994). The masking can be due either to external noise, when noise is added to the speech at the source, or internal noise, present in the cochlea and auditory nerve. The AI measure is weighted to account for both cochlear critical bands and those frequency regions containing important speech information. Fletcher developed the AI to predict the average nonsense-phone articulation score $P_c(\mathrm{SNR})$, as a function of the speech-to-noise ratio (SNR) (specified over 20 frequency bands), as a way of avoiding expensive and time-consuming listening tests on speech communication equipment. The AI also offers important insights into human speech perception. In this paper the AI is used to provide a new and insightful view of the confusion data of Miller and Nicely (1955).

Many procedures have been developed which claim to predict the phone performance score for noisy and/or filtered channels, several of which are called the ''articulation index.'' Following the first proposal in 1921 by Fletcher, there is the Bell Labs procedure of FS47 (key abbreviations are provided in Table I), followed by the more extensive version by Fletcher and Galt (1950). Then, Kryter published his simplified method in 1962 (Kryter, 1962a, 1962b), soon followed by the ANSI version. More recently, the STI was proposed to extend the AI procedure when room reverberation is present (Houtgast and Steeneken, 1973; Steeneken and Houtgast, 1980). This was then followed by the new ANSI procedure SII (S3.5-1997, 1997).

The focus on the articulation index provided here is not on predicting the performance of speech communications systems, but rather in understanding and modeling the perception and recognition of human speech sounds. The focus

TABLE I. Table of abbreviations.

Abbreviation	Definition
CV, CVC	Consonant+vowel sounds
intelligibility	Recognition of phonemes
articulation	Recognition of nonsense phones
CM	Confusion matrix
AM	Articulation matrix (CM of nonsense speech)
AI	Articulation index
SNR	Speech-to-noise ratio
rms	Root-mean-squared
events	Perceptual features
MN55	Miller and Nicely (1955)
FS47	French and Steinberg (1947)

here is not on procedures for predicting phone intelligibility, but in gaining leverage from the well-validated AI predictions, to provide insight into nonsense-phone identifications.

At least four fundamental questions are addressed in this presentation.

(1) What is the relation between the AI and the CM?
(2) How can one more accurately determine the groups that are present in Miller and Nicely's CM data at certain SNRs?
(3) What are the limits of Fletcher's *band independence* assumption?
(4) Can the AI be used to predict the phone score for closed sets (i.e., what are the limits of the AI procedure)?

Subsequent to the Bell Labs articulation studies, and following up work done during WWII, Miller took up the study of speech articulations in much greater detail. As detailed in Miller's (1951) book *Language and Communication*, information and communication theory are the basis for understanding the speech code. One of the basic tools of information theory is the *channel*, the mathematical characterization of a communication link (i.e., a noisy pair of wires, with codecs attached), in terms of discrete input and output symbols from some *alphabet*. One method for characterizing the human speech communication channel is the nonsense-phone confusion matrix (CM), which characterizes the probabilities of nonsense speech sound (the symbols) transmission errors. It is expected that an error analysis of this *articulation matrix* (AM), as a function of the SNR, can give important insight into the speech code.

A second key concept from information theory is that of entropy $\mathcal{H}$, which is a measure of the compactness of a probability distribution, which in the case of speech represents the distribution of the sound confusions. In his classic 1951 study, Miller *et al.* (1951) (MHL51) showed the effects of symbol alphabet size, and thus the entropy, on word recognition. Four years later, in a second classic study, Miller and Nicely (MN55) showed that as the wideband SNR is raised from -18 to $+12$ dB, *the sounds form perceptual groups*. The formation of a group, as a function of the SNR, also results in a change (reduction) in $\mathcal{H}$.

In analyzing their confusion matrices, MN55 quantified the grouping effect using a mutual-information (MI) analysis, on assumed groups. An natural advantage of the MI analysis method is its insensitivity to bias, as defined by the skew-symmetric form of the CM. This is at the same time a weakness of the MI, since it may be beneficial to remove the effects of subject bias prior to modeling the confusions. A major disadvantage of mutual information, as used in MN55, is that it gives no insight into the formation of the groups being analyzed.

This paper explores the limits and applicability of Fletcher's band-independence model, applied to the 1955 closed-set consonant articulation data of Miller and Nicely (MN55). Such data do not meet the usual assumptions of the AI audibility measure, of a large, high-entropy open-set corpus. By plotting the confusions as a function of SNR, it is possible to identify the groups in a systematic, logical way, without assuming any predetermined sound ordering. This leads to an accurate method of identifying the natural sound groups, and allows one to display the complex body of CM data at all SNRs, in a single figure. Once the groups have been identified, one is then free to further explore the relationship of the AM to the AI and $\mathcal{H}$.

The definitions of mathematical symbols has been summarized in Table II.

II. REPRESENTATIONS OF THE CONFUSION MATRIX (CM)

Figure 1 shows a typical MN55 consonant–vowel (CV) *confusion matrix* or *count matrix* for wideband speech (0.2–6.5 kHz), at a *speech-to-noise ratio* (SNR) of -6 dB (Miller and Nicely, 1955, Table III). The 16 consonants were presented along with the vowel /a/ as in father (i.e., the first three sounds were [/pa/,/ta/,/ka/]). After hearing one of the 16 CV sounds as labeled by the first column, the consonant that was reported is given as labeled along the top row. This array of numbers form the basic CM, denoted $\mathcal{C}_{s,h}$, where integer indices s and h (i.e., ''spoken'' and ''heard'') each run between 1 and 16. For example, /pa/ was spoken 230 times (the sum of the counts in the first row), and was reported heard 80 times ($\mathcal{C}_{1,1}$), while /ta/ was reported 43 times ($\mathcal{C}_{1,2}$). For Table III the mean row count was 250, with a standard deviation of 21 counts.

When the sounds are ordered as shown in Fig. 1, they form groups, identified in terms of hierarchical clusters of *articulatory features*. For example, the first group of sounds 1–7 correspond to unvoiced, group 8–14 are voiced, and 15, 16 are nasal (and also voiced).

At an SNR of -6 dB, the intraconfusions (within a group) are much greater than the interconfusions (between groups). For example, members of the group 1–7 (the unvoiced sounds) are much more likely to be confused among themselves, than between the voiced sounds (8–14), or the nasal sounds (15,16). The nasal are confused with each other, but rarely with any of the other sounds 1–14.

TABLE II. Table of mathematical symbols.

Symbol	Definition	Equation
$\mathcal{I}$	Intelligibility (meaningful sound recognition)	
$\mathcal{C}$	Confusion matrix	(2)
$\mathcal{A}$	Articulation matrix	(1)
S	Symmetric form of $\mathcal{A}$	(3)
A	Skew-symmetric form of $\mathcal{A}$	(4), (5)
P_c	Probability correct	(17), (16), (18)
$P_c^{(i)}$	Same as $\mathcal{A}_{i,i}$ and $S_{i,i}$	
$\mathcal{H}$	Entropy	
AI	Articulation index (AI)	(8)
AI_k	Specific AI in band k	(11)
e	Total error	(12), (13), (20)
e_i	Total error for sound i: $e_i \equiv 1-P_c^{(i)}=1-\mathcal{A}_{i,i}$	(7)
$e_{\min}$	Minimum error	(15)
e_{chance}	Chance error	(19)
ϵ_k	kth band error	(14)
$e_{\min}(i)$	Minimum error for sound i: $e_i\vert_{\mathrm{AI}=1}$	
S_{15}	Shorthand for $S_{15,15}$	
snr_k	Speech-to-noise rms ratio in band k	(9)

TABLE III. Confusion matrix for $S/N=-6$ db and frequency response of 200–6500 cps.

STIMULUS \ RESPONSE	p	t	k	f	θ	s	ʃ	b	d	g	v	ð	z	ʒ	m	n
p	80	43	64	17	14	6	2	1	1		1	1			2	
t	71	84	55	5	9	3	8	1				1	2		2	3
k	66	76	107	12	8	9	4					1			1	
f	18	12	9	175	48	11	1	7	2	1	2	2				
θ	19	17	16	104	64	32	7	5	4	5	6	4	5			
s	8	5	4	23	39	107	45	4	2	3	1	1	3	2		1
ʃ	1	6	3	4	6	29	195		3							1
b	1			5	4	4		136	10	9	47	16	6	1	5	4
d							8	5	80	45	11	20	20	26	1	
g					2			3	63	66	3	19	37	56		3
v				2		2		48	5	5	145	45	12		4	
ð					6			31	6	17	86	58	21	5	6	4
z					1	1	1	7	20	27	16	28	94	44		1
ʒ								1	26	18	3	8	45	129		2
m	1							4			4	1	3		177	46
n					4			1	5	2		7	1	6	47	163

UNVOICED VOICED NASAL

FIG. 1. Typical Miller–Nicely frequency of confusions, or count matrix $\mathcal{C}$, from Table III at −6-dB SNR. Each entry in the matrix $\mathcal{C}_{s,h}$ is the subject response count. The rows correspond to the *spoken* CVs, each row representing a different consonant, from $s=1,...,16$. The columns correspond to the *heard* CVs, each column representing a different consonant, from $h=1,...,16$. The common vowel /a/, as in "father," was used throughout. When the 16 consonants are ordered as shown, the count matrix shows a "block-symmetric" partitioning in the consonant confusions. In this matrix there are three main blocks delineated by the dashed lines, corresponding to unvoiced, voiced, and nasal. Within the voiced and unvoiced subgroups, there are two additional symmetric blocks, corresponding to affrication and duration, also delineated with dashed lines.

The MN55 articulatory feature classification scheme is far from perfect. For example, the nasals are voiced in the same sense as those labeled voiced; however, they clearly form a unique cluster. Thus, there is no unique simple articulatory label for sounds 8–14. Groups systematically depend on the SNR, and groups remain unidentified by this scheme. Using the example of Table III (Fig. 1), [/ba/,/va/,/ða/] form a group that is distinct from the nonfricative voiced subgroup. An improved order for sounds 8–14 would be [/ba/,/va/,/ða/], [/za/,/ʒa/,/da/,/ga/]. Of course, this example fundamentally breaks the MN55 articulatory feature classification scheme. In fact, the feature space cannot strictly be articulatory feature based.

The MN55 data have been the inspiration for a large number of studies. The sound grouping has been studied using multidimensional scaling, which has generally failed in providing a robust method for finding perceptually relevant groups of sounds, as discussed by Wang and Bilger (1973). Thus, the grouping problem has remained unsolved.

The data in the CM represent a psychological subject response, and therefore need to be represented in terms of *psychological variables* rather than physical (production) measures, as labeled by articulatory features. This could have been the role of *distinctive features*, had they been so defined. Unfortunately, there seems to be some confusion in the literature as to the precise definition of a distinctive feature. For example, are distinctive features production or perception quantities?

To avoid this confusion, I shall use the term *event* when referring to *perceptual features*. Since Miller and Nicely's confusion data are based on perception, they must be described by events. The precise nature of these events may be explored by studying the 15 plots $S_{i,j}$, $i\neq j$, as shown in the lower-left panel of Fig. 2 for the case of $i=2$.

A. The transformation from CM to AM

The term *articulation* is defined as the probability correct P_c of identifying nonsense-phone speech sounds (consonants and vowels), while *intelligibility* $\mathcal{I}$ is the probability of identifying meaningful speech sounds, such as words and sentences (Fletcher and Galt, 1950).

When normalized as a probability, the consonant confusion matrix is transformed to an *articulation matrix* (AM), denoted $\mathcal{A}$ (script A, Table II), with elements

$$\mathcal{A}_{s,h}\equiv\frac{\mathcal{C}_{s,h}}{\Sigma_h\mathcal{C}_{s,h}}. \tag{1}$$

This normalization, to an equal probability for each row, is justified because of the small standard deviation of the row sums (i.e., 250 ± 21).

The AM is the empirical conditional probability $P_c(h|s)$ of reporting sound h after speaking sound s, namely

$$\mathcal{A}_{s,h}\equiv P_c(h|s), \tag{2}$$

for integer labels s, h (i.e., spoken, heard). In another sense, $\mathcal{A}_{s,h}$ for $s\neq h$ is an error probability, since it is the probability of reporting the wrong sounds h after hearing spoken sound $s\neq h$.

Figure 2 shows the probability of responding that the sound $h=1,...,16$ was reported, following speaking /ta/ ($s=2$), as a function of the wideband SNR. The upper-left

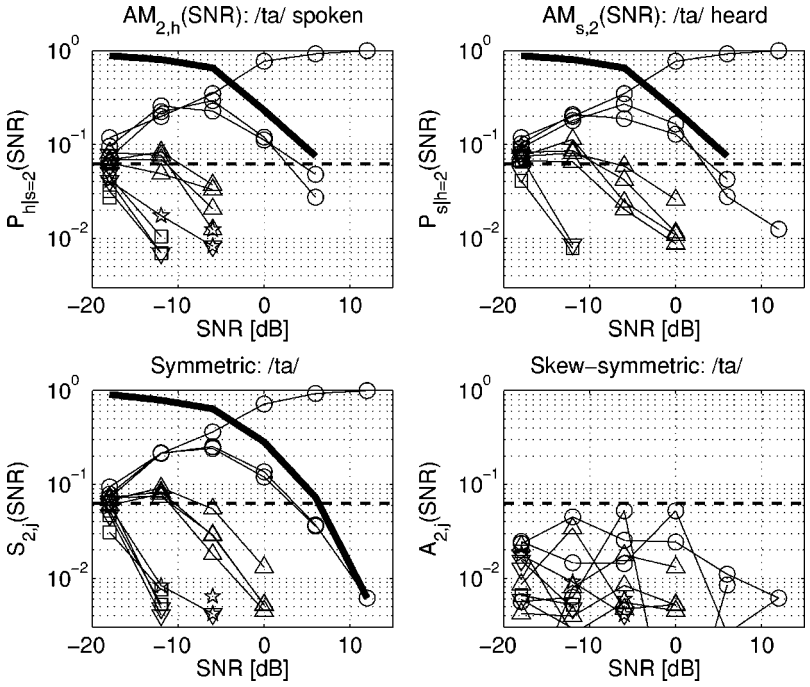

FIG. 2. This figure shows Miller and Nicely's 1955 wideband row-normalized confusion matrix data $\mathcal{A}_{s,h}(\mathrm{SNR})$ [Eq. (1)] for the sound /ta/ (sound 2) from MN55 Tables I–IV, as a function of the speech-to-noise ratio. The upper-left panel is a plot of the second row of the articulation matrix [$\mathcal{A}_{2,h}(\mathrm{SNR})$, $h=1,...,16$, corresponding to /ta/ spoken], while the upper-right panel is a plot of the second column [$\mathcal{A}_{s,2}(\mathrm{SNR})$, corresponding to /ta/ heard]. The matrix is not perfectly symmetric ($\mathcal{A}\neq\mathcal{A}^t$), which explains the small differences between these two plots. The lower-left panel is the symmetric form of the articulation matrix given by Eq. (3), which is the average of $\mathcal{A}$ and its transpose $\mathcal{A}^t$. The lower-right panel is the skew-symmetric form A [Eq. (4)]. The horizontal dashed line in each figure shows chance performance (i.e., 1/16).

panel shows the probability $\mathcal{A}_{2,h}(\mathrm{SNR})$ of each heard sound ($h=1,...,16$), given /ta/ was spoken. The upper-right panel shows the probability $\mathcal{A}_{s,2}$ of each sound spoken ($s=1,...,16$), given that /ta/ was heard. The curve that rises to 1 is the probability of correctly reporting /ta/ $\mathcal{A}_{2,2}(\mathrm{SNR})$, given that it was spoken (left), or spoken given that it was heard (right). The solid-thick curve is the total probability of error $e_2(\mathrm{SNR})\equiv 1-\mathcal{A}_{2,2}(\mathrm{SNR})$ of not reporting /ta/, given that it was spoken (left) or heard (right).

Symmetric and skew-symmetric decomposition

The lower-left panel of Fig. 2 is a plot of the second row $S_{2,j}$ of the symmetric form of the AM, defined as

$$S\equiv\tfrac{1}{2}(\mathcal{A}+\mathcal{A}^t), \tag{3}$$

where $\mathcal{A}^t$ is the transpose of $\mathcal{A}$, while the lower-right panel is the second row of $A_{i,j}$ of the *skew-symmetric* form of the matrix, defined as

$$A\equiv\tfrac{1}{2}(\mathcal{A}-\mathcal{A}^t). \tag{4}$$

It appears that the sampling error (statistical uncertainty) in the measurements, due to the sample size, is about 0.5% (0.005), which is where the measurements become scattered. This variability is determined by many factors, including the number of trials per sound, the smoothing provided by the symmetric transformation, the consistency of the talker, and the mental concentration and number of the observers (four in this case).

From the lower-right panel, it is clear that the AM is close to symmetric, since the skew-symmetric terms are small. A few terms of $A_{2,h}(\mathrm{SNR})$ are as large as 5%, but most are less than 1%. Since the MN55 data are close to symmetric, it is reasonable to force the symmetry, and then to study S and A separately, which is the approach taken here. Note that S is slightly smoother than $\mathcal{A}$, since each element $\mathcal{A}_{s,h}$ is the average of two similar terms, $\mathcal{A}_{h,s}$ and $\mathcal{A}_{s,h}$. Using the symmetric form simplifies the analysis of the matrix and gives us access to the skew-symmetric form.

Based on an analysis by Goldstein (1980), the interpretation of the skew-symmetric form is quite different from that of the symmetric form. The most likely explanation of the skew-symmetric matrix is that the subjects have a bias for one sound over another, and are therefore more likely to report the consonant for which they have the bias (Goldstein, 1980).

The largest skew-symmetric sounds in row 2 are /fa/, /θa/, /va/, and /ða/, which have errors approaching 5%, but are always less than chance (1/16). It seems significant that the skew-symmetric form always lies slightly below chance (Fig. 2, lower-right panel). For the rest of the sounds, the error patterns are similar in their nature to those of /ta/, with the largest errors of about 10% in a few places, but with most of the errors being a few percent or less.

There is an interaction between the row normalization [Eq. (1)], and the symmetry transformation Eq. (3), which requires that the row normalization and symmetric computations be iterated. This iteration always converges to the same result, and is always stable for all of the MN55 tables. An entry of "1" in $\mathcal{C}$ represents a single vote for the same utterance, from four listeners who heard that utterance. All 1's were deleted from the matrix before computing S. Once matrix S has been determined, A is computed from

$$A=\mathcal{A}-S. \tag{5}$$

Plotting the symmetric data $S(\mathrm{SNR})$ as a function of SNR, as shown in Fig. 2, provides a concise yet comprehensive summary of the entire set of measurements, and shows the hierarchical grouping, without a need to order the sounds. In the next section it is shown that if $S_{i,j}$ is described as a function of the AI, rather than the SNR, the same data may be quantitatively modeled, and the important effects of chance may be accounted for.

B. Grouping the sounds

Sound clustering in the CM was used by MN55 as the basis for arguing that the sounds break down into distinct groups, which MN55 identified as five discrete *articulatory features*, which they called *voicing, nasality, affrication, duration*, and *place*.

Each symbol in Fig. 2 labels a different articulatory feature. Sounds 1–3 (/pa/,/ta/,/ka/) are shown as circles, 4–7 (/fa/,/θa/,/sa/,/ʃa/) triangles, 8–10 (/ba/,/da/,/ga/) squares, 11–14 (/va/,/ða/,/za/,/ʒa/) upside-down triangles, while the nasal sounds 14 and 15 (/ma/,/na/) are labeled by 5-pointed stars.

The hierarchical clusters are seen as groups that peel away as the SNR increases. The symmetric /ta/ data shown in the lower-left panel of Fig. 2 are a great example: First, all the voiced sounds dramatically drop, starting from chance, as the SNR is raised. Next, the unvoiced-fricatives /fa/, /θa/, /sa/, /ʃa/ (triangles) peel off, after very slightly rising above chance at −12-dB SNR. Finally, the two main competitors to /ta/ (/pa/ and /ka/) peak around −6-dB SNR, and then fall dramatically, as /ta/ is clearly identified at 0-dB SNR and above. In the lower-left panel /pa/, /ta/, and /ka/ (○) are statistically indistinguishable below −6 dB, and approach chance identification of 1/16 at −18 dB. Above about −6 dB, /ta/ separates and the identification approaches 1, while the confusions with the other two sounds (/pa/ and /ka/) reach a maximum of about a 25% score, and then drop monotonically, as the SNR increases.

The MN55 sounds 4–7 (/fa/, /θa/, /sa/, and /ʃa/), like sounds 1–3 (/pa/,/ta/,/ka/), also form a group, as may be seen in the lower-left panel, labeled by △. This group also starts from chance identification (6.25%), rises slightly to a score of about 7% at −12 dB, and then monotonically drops at a slightly greater rate than sounds 1 and 3 (symbols ○).

The third group is the remaining sounds 8–16, labeled by the remaining symbols, which show no rise in performance; rather, they steeply drop from the chance level.

At the lowest SNR of −18 dB, the elements in the symmetric form of the AM approach chance performance, which for MN55 is 1/16, corresponding to closed-set guessing. Extrapolating the data of Fig. 2, chance performance corresponds to about −21-dB SNR.

Based on the clustering seen in the AM (e.g., MN55 Tables II and III), it was concluded by MN55 that the three sounds /ta/, /pa/, and /ka/ might be thought of as one group. These three sounds form the unvoiced, non-nasal, nonaffricate, low-duration group, having three different values of place. The details of these groupings depend on the SNR. A detailed analysis of these clusters show that the MN55 *articulatory features* (production feature set) do not always correspond to the *events* (perceptual feature set).

In fact, it would be surprising if it turned out any other way, given that production and perception are fundamentally different things. The details of a scheme that will allow us to make such an analysis of the optimal perceptual feature set, form the remainder of this paper.

1. Formula for the total error

The solid-thick curve in the top two, and bottom-left panels of Fig. 2, are graphs of the total error for /ta/

$$e_2(\mathrm{SNR}) \equiv 1 - S_{2,2}(\mathrm{SNR}). \tag{6}$$

Because each row of $S_{i,j}$ has been normalized so that it sums to 1, the total error for the ith sound is also the row sum of the 15 off-diagonal $(j \neq i)$ elements, namely

$$e_i(\mathrm{SNR}) = \sum_{\forall j \neq i} S_{i,j}(\mathrm{SNR}). \tag{7}$$

Since each error term is non-negative, e_i must bound each individual confusion $S_{i,j}$. For the data of Fig. 2, lower-left, the other two circle curves (/pa/ and /ka/), which compete with /ta/, and thereby form a 3-group, are nearly identical. All other error terms are much smaller. Thus, the solid-thick curve, $e_2(\mathrm{SNR})$, is approximately twice the size of the curves for /pa/ and /ka/. All the off-diagonal terms go to zero at +12-dB SNR so for that one point $e_2 = S_{2,j}$, a fluke of the small-number statistics.

Equation (7) says that the total error for the ith sound is linearly decomposed by the off-diagonal errors of the AM. This is a natural decomposition of the total error into its confusions that can help us understand the AI predictions in much greater detail.

For example: *Why does the probability of identification of sounds 1–3 and 4–7 increase even when these sounds are not spoken*? The initial rise for the two sound groups follows from the increase in chance performance due to the decreased entropy, which follows from the reduced size of the group. This conclusion follows naturally from Eq. (7). As the SNR increases, the size of the group exponentially decreases.

As the number of alternatives in a closed-set task decreases, the probability of guessing increases. Given 2 alternatives, chance is 1/2; given 16, chance is 1/16. Thus, grouping and the rise due to the confusion within the group are intimately tied together. In the same manner, as the SNR rises from −18 to −12, the MN55 sounds 4–16 are perceptually ruled out, increasing chance performance for sounds 1–3 from 1/16 to 1/3.

2. The nasals

In Fig. 3 $S_{i,j}(\mathrm{SNR})$ for $i = 15, 16$, corresponding to /ma/ and /na/, are presented. The two nasal sounds are clearly separated from all the other sounds, even at −18-dB SNR. As the SNR increases, the scores rise to ≈25%, peaking at or near −12-dB SNR, following with the identification rising and the confusion dramatically falling for SNRs at and above −6 dB.

Sounds 1–14 are solidly rejected, even at −18 dB. These scores exponentially drop as the SNR is increased. There is a slight (visual) hint of a rise of a few sounds for the case of /ma/, in some of the rejected sounds in the left panel, and some corresponding grouping, but the effect is small and it would be difficult to tease out. The rejected sounds in the right panel do not show any obvious grouping effect.

The subjects can clearly distinguish the two nasal sounds (sounds 15,16) from all the others (sounds 1–14),

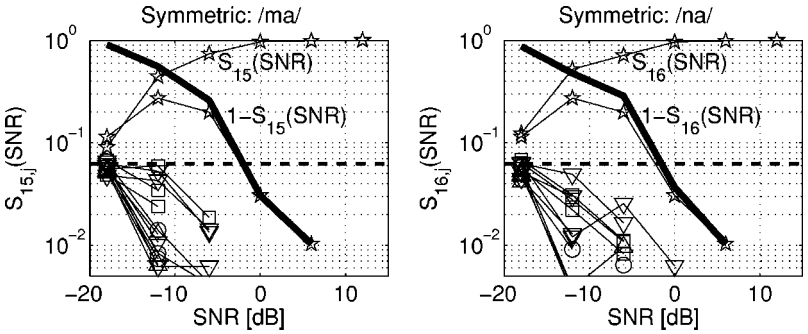

FIG. 3. Plots of the symmetric AM corresponding to the nasals /ma/ and /na/. The curve that rises to 1 is $S_{i,i}$(SNR) for $i=15$ (left) and $i=16$ (right). The solid thick curve in each panel is e_i [Eq. (7)]. The other curves represent confusions $S_{i,j}$(SNR) for the remaining sounds $j=1,...,14$.

even at the lowest SNR of −18 dB; however, they cannot distinguish between them until the SNR is greater than −12 dB. The subjects know the sound they hear is nasal, but the question is, which one? This identification of event-nasal leads to a significant increase in chance performance for SNRs between −18 and −6 dB, from 1/16 to 1/2.

One may also see this effect in the raw count data at −18 dB, where the confusions are approaching equal chance levels. For example, in MN55 Table I, the raw counts are [25,28;33,32]. At −12 dB, /ma/ and /na/ are significantly confused with each other, but rarely with the other sounds. For example, from MN55 Table II, /ma/ is heard 20 times when /ba/ is spoken [$S_{15,8}(-12)=6.72\%$ of the time], while /ba/ is heard 11 times when /ma/ is spoken (5.83% of the time).

III. TRANSFORMATION FROM THE WIDEBAND SNR TO THE AI

Miller and Nicely used the wideband SNR, in dB, as their measure of audibility. However, as discussed in the Introduction, there are reasons to believe that the AI(SNR) is a better audibility measure. We shall now demonstrate this for the MN55 data. Our approach is to transform MN55's wideband SNR into an AI, and then to plot the resulting $S_{i,j}$(AI).

To compute the AI for MN55 one needs to know the *specific* SNR, over articulation bands, denoted snr_k. This requires knowledge of the average speech spectra for five female talkers, and the noise spectra. The spectrum for five female talkers is shown in Fig. 4, while the noise spectra was independent of frequency (i.e., white). The procedure for computing AI(SNR) is described next.

A. Computing the specific AI

The AI is defined by FS47 [their Eq. (8)] as

$$\mathrm{AI}=\frac{1}{K}\sum_{k}^{K}\mathrm{AI}_k, \tag{8}$$

namely as a 20-band average over the *specific* AI, denoted AI_k. The specific AI is defined in terms of the speech-to-noise ratio

$$\mathrm{snr}_k\equiv\sigma_{s,k}/\sigma_{n,k}, \tag{9}$$

where the speech power is $\sigma_{s,k}^2$ [Watts/critical band] and the masking noise power is $\sigma_{n,k}^2$ [Watts/critical band], in the kth articulation band. When calculating $\sigma_{s,k}$, the average is over 1/8-s intervals. snr_k is the same as FS47's band sensation level E. The kth articulation band power-snr *speech detection threshold* may be modeled as

$$\frac{I+\Delta I}{I}\equiv\frac{\sigma_{n,k}^2+c^2\sigma_{s,k}^2}{\sigma_{n,k}^2}=1+c^2\mathrm{snr}_k^2, \tag{10}$$

where a frequency-independent *speech detection constant* c is determined empirically from data on the detection of speech in noise (Fletcher and Munson, 1937; French and Steinberg, 1947). The role of c is to convert the speech rms to the speech peaks, which are typically 12 dB above the rms speech level. When snr_k specifies the speech peaks, $c=2$.

Converting to decibels, and scaling by 30, defines the *specific* AI

$$\mathrm{AI}_k=\min(\tfrac{1}{3}\log_{10}(1+c^2\mathrm{snr}_k^2),1). \tag{11}$$

Relationship Eq. (11) follows from the detailed discussions of FS47 and Fletcher and Galt (1950), followed by the subsequent analysis by Allen (1994). [See especially (Fletcher, 1995, Eq. (10-3), page 167).]

Between 0 and 30 dB, AI_k is proportional to $\log(\mathrm{snr}_k)$ because the percent of the time the speech is above a certain level is proportional to the dB SL level (*re*: threshold sensation level) (French and Steinberg, 1947; Allen, 1994). The factor of 1/3 comes from the dynamic range of speech in a given articulation band (French and Steinberg, 1947, Fig. 4,

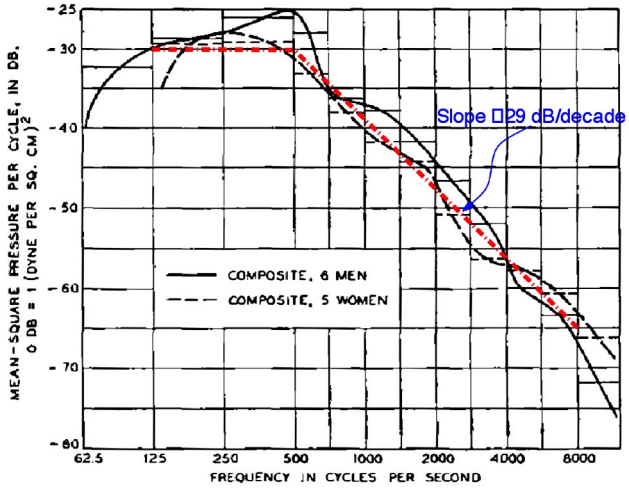

FIG. 4. This figure from Dunn and White, 1940 (their Fig. 10) shows the average power spectrum for six men and five women. The dashed curve, which approximates the power spectrum for the five women, has a slope of 0 from 125 to 500 Hz, and a slope of −29-dB/decade between 0.5 and 8 kHz.

page 95). As discussed extensively by FS47 (i.e., their Table 12), an empirical *threshold adjustment* must be made, labeled c in Eq. (10). The value of c is chosen such that the speech is just detectable when $\text{snr}_k=1$, in each cochlear critical band, corresponding to specific AIs of zero (i.e., $\text{AI}_k=0$). Equation (11) over-predicts the data of FS47 Table V by 3.2% ($c\equiv 1/2$, $\text{snr}_k=E$). A more precise estimation of c would require repeating Fletcher's critical ratio experiment using narrow bands of speech, with a white-noise masker and measuring snr_k at the detection threshold. The $\min(x,1)$ part of the definition limits the AI on the high end, since for an SNR above 30 dB, the noise has a negligible effect on the articulation (and intelligibility).

The band independence model of the total error

The average sound *articulation error* e (SNR), in terms of the average sound articulation $P_c(\text{SNR})$, is

$$e(\text{SNR})=1-P_c(\text{SNR}). \tag{12}$$

In 1921 Fletcher showed that the articulation error probability $e(\text{SNR})$ could be thought of as being distributed over K-independent articulation bands. The bandwidth of each of these articulation bands was chosen so that they contribute equally to e [the articulation per critical band is constant from 0.3–7 kHz (Fletcher and Galt, 1950; Allen, 1994, 1996)]. Assuming band independence, the total articulation error may be written as a product over K band articulation errors

$$e=\epsilon_1\epsilon_2\cdots\epsilon_k\cdots\epsilon_K. \tag{13}$$

This equation is called the *band independence model*.

Galt established that the articulation bandwidth is proportional to cochlear critical bandwidths (French and Steinberg, 1947, page 93), as measured by the *critical ratio* method and the frequency jnd (Allen, 1994, 1996). Fletcher then estimated that each articulation band was the equivalent of 1 mm of distance along the basilar membrane, thereby taking up the 20-mm distance along the basilar membrane, between 300 to 8 kHz (Allen, 1996). Thus, the AI [Eq. (8)] may be viewed as an average SNR, *averaged over dB units*, of a scaled specific SNR, defined over cochlear critical bands.

As first derived in Allen (1994), the probability of articulation error in the kth band ϵ_k may be written in terms of the specific AI as

$$\epsilon_k=e_{\min}^{\text{AI}_k/K}, \tag{14}$$

where the constant $e_{\min}$ is defined as the minimum error via the relationship

$$e_{\min}\equiv 1-\max_{\text{snr}}(P_c(\text{SNR})). \tag{15}$$

This constant $e_{\min}$ depends in general on the corpus, talkers, and subjects. For Fletcher's work, $e_{\min}$ was 1.5% ($\mathcal{H}\approx 11$, i.e., more than 2048 sounds). For the work reported here, a value of 0.254% ($\mathcal{H}=4$) was used, based on an extrapolation of the MN55 data to AI=1 and a minimization of the model parameters for a best fit to MN55 data.

It follows from the above relations that

$$P_c(\text{AI})=1-e_{\min}^{\text{AI}}. \tag{16}$$

The total error $e=e_{\min}^{\text{AI}}$ in Eq. (16) was represented by Fletcher as $e=10^{-\text{AI}/0.55}$. Both expressions are exponential in AI, differing only in the choice of the base ($e_{\min}=10^{(-1/0.55)}$). Equation (16) only applies to the case of nonsense phones, having the maximum entropy.

Figure 5, left, shows the relative spectrum and noise level corresponding to SNRs of -18 to $+12$ dB, for female speech with a white-noise masker. On the right one may see the resulting AI(SNR), based on the calculations specified by the equations presented in this section. The final values of the AI were determined with $c=2$ to be (starting from an SNR of $+12$): [0.459,0.306,0.186,0.1,0.045,0.016].

Because the spectrum of the speech and the spectrum of the noise are not the same, the AI(SNR) cannot be a linear function of SNR. Only for the case where the two spectra have the same shape will AI(SNR) be linear in SNR. For the case at hand, a white-noise masker, the high frequencies are progressively removed as the SNR decreases, as shown in the left panel of Fig. 5.

B. AM(SNR) to AM(AI)

The left panel of Fig. 6 shows the MN55 consonant identification curves $P_c^{(i)}(\text{SNR})\equiv S_{i,i}(\text{SNR})$, as a function of the SNR for each of the 16 sounds ($i=1,...,16$), along with their mean $P_c(\text{SNR})$ (solid curve with circle symbols)

$$P_c\equiv\frac{1}{16}\sum_{i=1}^{16}P_c^{(i)}. \tag{17}$$

It must be mentioned that Eq. (17) only applies to the case at hand, where the *a priori* probabilities of the sounds are equal. In the more general case, a Bayesian formulation would be required.

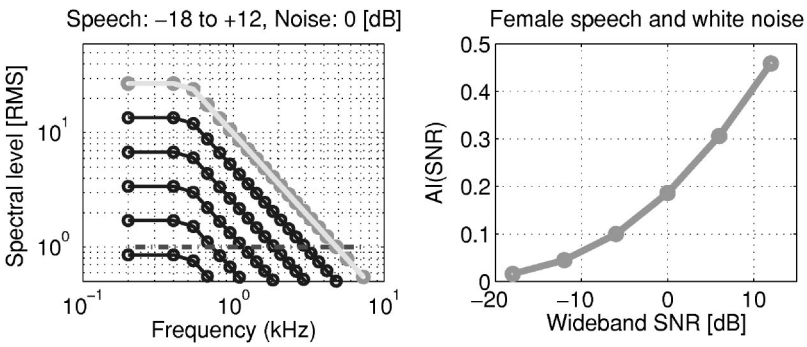

FIG. 5. Using the speech power spectrum given by the dashed line in Fig. 4, and assuming a uniform noise spectral level, the AI(SNR) was calculated. Each curve shows the relative spectral level of the speech having a peak level at the wideband SNRs used by Miller and Nicely [−18,−12,−6,0,6,12], in units of dB. The top curve shows the +12-dB speech spectrum. The dashed-dot line is the noise spectral level having an rms of 0 dB.

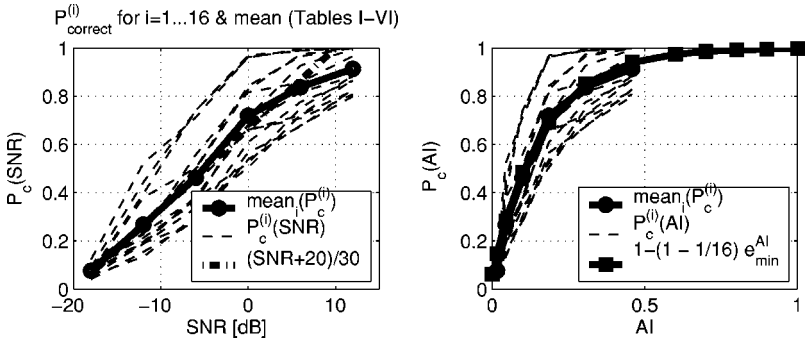

FIG. 6. The light dashed lines are $P_c^{(i)}$ for each of the 16 consonants. On the left the abscissa is the SNR in dB, while on the right, the AI is used as the independent variable. The solid-thick curve (circles) on both the left and right is the average score P_c, Eq. (17). The solid-thick curve (squares) on the right is the average phone prediction given by Eq. (18).

In the right panel the individual scores, along with the average, are shown as a function of the AI. To transform from SNR to AI the values shown in the right panel of Fig. 5 are used.

We also wish to compare the AI model prediction to the measurements shown in Fig. 6. However, it is necessary to modify Eq. (16) so that it accounts for chance (guessing) given by $P_{\text{chance}}=2^{-\mathcal{H}}$, when $\mathcal{H}=4$ and AI=0. This is done by again assuming independence of the error probabilities. Since chance error for guessing is $e_{\text{chance}}=1-P_{\text{chance}}$, the chance-corrected $P_c(\text{AI})$ formula is

$$P_c(\text{AI},\mathcal{H})=1-e_{\text{chance}}(\mathcal{H})e_{\min}^{\text{AI}}, \tag{18}$$

with

$$e_{\text{chance}}(\mathcal{H})\equiv 1-2^{-\mathcal{H}}. \tag{19}$$

Fletcher's formula Eq. (16) is the limiting case of Eq. (18) when $\mathcal{H}$ becomes large (Fletcher's $\mathcal{H}\approx 11$).

A plot of Eq. (18) is shown in the right panel of Fig. 6 (solid curve, square symbols), with $e_{\min}=0.254\%$, and $\mathcal{H}=4$. The fit of Eq. (18) to the average of the 16 MN55 curves is excellent.

Discussion

The left panel of Fig. 6 shows that there is an approximately linear relationship between $P_c(\text{SNR})$ and SNR over the range from -18 to 6 dB. The thick dashed-dot line is (SNR+20)/30. This line is useful as a simple reference.

The main deviation from the linear dash-dot curve is due to the strong saturation that occurs for the two nasal sounds and sound 7 [the three curves with the highest $P_c(\text{SNR})$]. Note that each of the sounds has a nearly linear $P_c^{(i)}(\text{SNR})$, with different saturation levels (if they are reached). The saturation point for $P_c(\text{SNR})$ occurs at an SNR of about 30 dB above the threshold, at -20 dB (thick solid line with circles). Note that since the relation $P_c(\text{SNR})$ depends on the noise spectrum, the linear relation observed in the left panel of Fig. 6 can only hold for the white-noise masker, since if the noise spectrum is changed, $P_c(\text{SNR})$ must change, and it is linear for the white-noise case.

In the right panel of Fig. 6 the extended AI model [Eq. (18)] is shown for MN55's data. Each of the 16 curves $P_c^{(i)}(\text{AI})$, $i=1,...,16$, is shown as the light-dashed curves. This average [Eq. (17)] is shown as the solid-thick curve with circles.

The solid-thick line with squares is the extended (chance-corrected) AI model, Eq. (18). The value of $e_{\min}$ of 0.254% is one-sixth that used by Fletcher (1.5%). The smaller size could be attributed to the larger amount of training the subjects received over such a limited set size $\mathcal{H}=4=\log_2(16)$.

As may be seen in the left panel of Fig. 5, since MN55 used white noise, the snr_k for frequency bands larger than about 0.7 kHz have an SNR of less than 30 dB, resulting in an AI of much less than 1. In fact, the AI was less than 0.5 for the MN55 experiment, corresponding to a maximum score of only 90%.

A most interesting and surprising finding is that the extended AI model [Eq. (18)] does a good job of fitting the average data. In fact, the accuracy of the fit over such a small set of just 16 consonants was totally unanticipated. This needs further elucidation.

C. Extended tests of the AI model

If one plots the total error probability $e(\text{AI})=1-P_c(\text{AI})$ in log coordinates, as a function of AI, such plots should approximate straight lines. This follows from the log of Eq. (18)

$$\log(e(\text{AI}))=\log(e_{\min})\text{AI}+\log(e_{\text{chance}}(\mathcal{H})), \tag{20}$$

which has the convenient form $y=ax+b$. The ordinate (y axis) intercept of these curves at AI=0 gives the log chance error [$b\equiv y(0)=\log(e(0))=\log(e_{\text{chance}}(\mathcal{H}))$], while the ordinate intercept of these curves at AI=1 defines the sum of the log-chance error and the log-minimum error, namely [$a+b\equiv y(1)$, thus $a=\log(e_{\min})$]. In Fig. 7 the log-error probabilities for each of the 16 sounds, along with the average and the AI model, are shown. The sounds have been regrouped so that the log-error plots have similar shapes. The shallow slopes are shown on the left and the steeper slopes on the right.

From Fig. 7, we shall find that the linear relationship [Eq. (20)] holds for 11 of the 16 sounds, with the free parameters $e_{\min}(i)$ and $e_{\text{chance}}(i)$, either depending on the sound, or on a sound group.

The upper two panels show the most linear groups, while the lower panels are the most nonlinear (nonstraight) log-error curves. The curves that are close to linear (the two top panels) are consistent with the AI model, due to Eq. (20).

This observation of a log-linearity dependence for the probability of error of individual sounds is rather astounding in my view. First, there was no *a priori* basis for anticipating that individual sounds might obey Fletcher's band-independence property, Eq. (18). Second, if individual

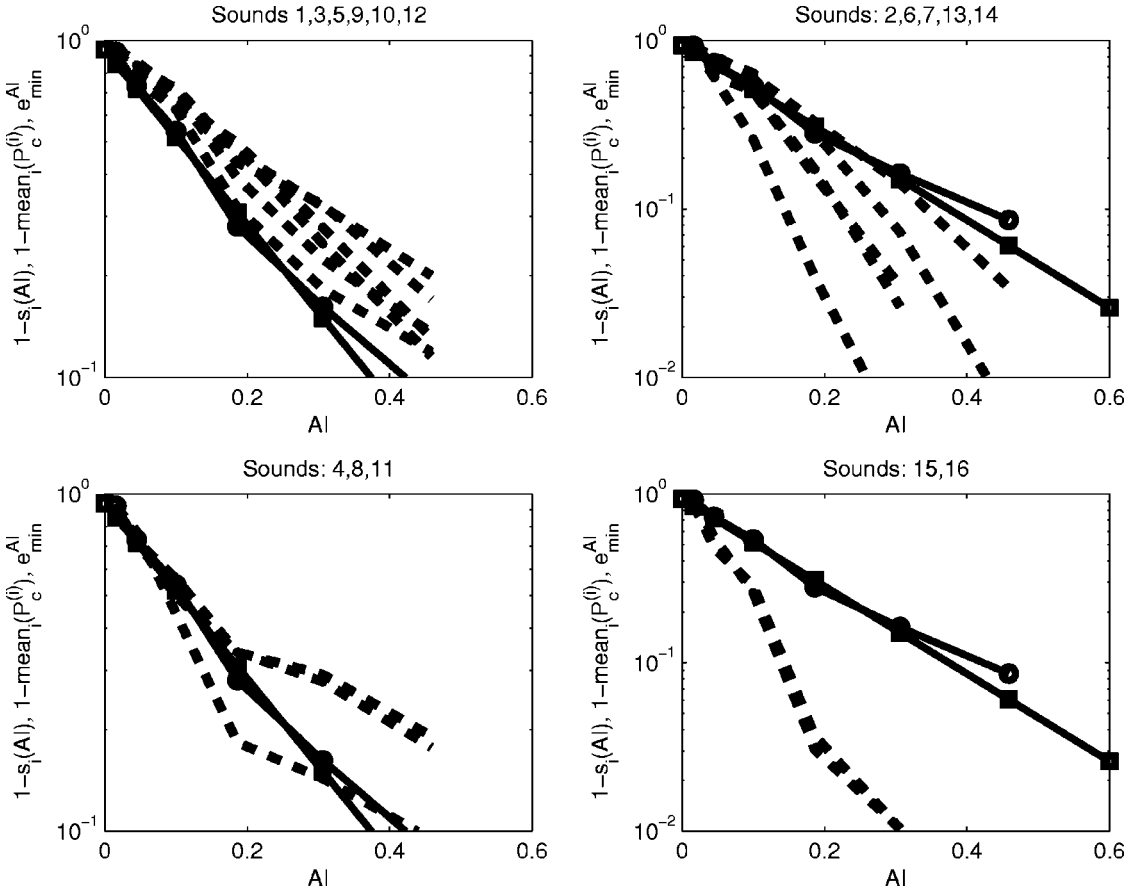

FIG. 7. This figure shows the probability of error for the ith sound, $P_e^{(i)}(\mathrm{AI}) \equiv 1 - P_c^{(i)}(\mathrm{AI})$, as a dashed curve. To reduce the clutter, the sounds have been sorted over the four panels, with the sound number indicated in each panel title. The top two panels are the cases where the individual sound-error curves are close to straight lines. The left-upper panel are those cases where the sound lies above the average, while the right-upper panel shows those cases where the sound lies below the average. The two lower panels correspond to the sounds that violate the exponential rule (are not straight lines on a log-error plot). For reference, each panel contains the average probability of error $P_e(\mathrm{AI}) \equiv 1 - P_c(\mathrm{AI})$, shown as the solid curve with circles, and the model error $e_{\min}^{\mathrm{AI}}$, shown as the solid line (squares).

sounds obey equations of the form of Eq. (8), then sums of such equations cannot obey Eq. (8), since the sum of many exponentials, each having a different base, is not an exponential.

The finding that individual CV recognition error is exponential in the AI (the basis of the band-independence hypothesis) therefore extends, and at the same time violates, Fletcher's original fundamental AI hypothesis that the average error is exponential.

It is therefore essential to understand the source of the deviations for the individual sounds from the average, and to critically assess the accuracy of the model for individual sounds. Five sounds (4,8,11,15,16) have a probability of error that deviates from linear, with the most nonlinear and the largest deviations from the mean, being the nasals (15,16), as shown in the lower-right panel. In the next section I explore the reasons for this.

1. Log-error for the nasals

In Fig. 8 the nasal data are shown using the same log-error linear decomposition used in Fig. 3, where the total error (solid-thick curve) is the sum of the errors of the competing sounds [i.e., Eq. (7)]. In the case of the nasals, the confusions for the other sounds is small, namely only /ma/ and /na/ significantly compete.

As a result of plotting the data as a function of AI, for $\mathrm{AI} > \mathrm{AI}_g = 0.045$ ($\mathrm{SNR} \geqslant -12$), the log-error curves become linear in AI, as predicted (modeled) by Eq. (18). This value of AI_g is shown in the plot with an arrow indicating the point of separation of the target sound from the competing sound. Extrapolating this linear region back to $\mathrm{AI}=0$, one finds the chance guessing probability of $1 - 2^{-\mathcal{H}_g} = 1/2$, corresponding to a nasal group entropy of $\mathcal{H}_g = 1$. This is shown on the graph by the dashed line superimposed on the corresponding error curve (stars). In the region $0 \leqslant \mathrm{AI} \leqslant \mathrm{AI}_g = 0.045$, $\mathcal{H}$ depends on AI, since it dramatically drops from 4 to 1.

Thus, the reason that the nasal curves are not linear in Fig. 7 is that chance (the entropy factor) is dramatically changing between $0 \leqslant \mathrm{AI} \leqslant \mathrm{AI}_g$, due to the formation of the perceptual ''event-nasal'' group.

When the data are plotted as a function of SNR, as in Fig. 3, the log-error linearity is not observed. Also, the shape of the curve will depend on the spectrum of the noise. Clearly, the SNR to AI transformation is an important key to making sense of these data.

2. Log error for /pa/, /ta/, and /ka/

Finally, in Fig. 9 we return to the case of /pa/, /ta/, and /ka/. This 3-group generalizes the /ma/, /na/ 2-group conclusions of Fig. 8. In the middle panel it is clear that for small values of AI less than 0.045 $S_{2,2}(\mathrm{AI})$ for /ta/ is equal to the curves for /pa/ and /ka/ $[S_{2,j}(\mathrm{AI}), j=1,3]$. As the AI rises above about 0.1, the three curves (circles) split due to the identification of /ta/ and the rejection of /pa/ and /ka/. The shape and slope of the curves corresponding to the two rejected sounds are identical. The projection of the rejected curves back to AI=0 gives the probability of chance error for a group of 3 (i.e., 1−1/3), as shown by the dashed line in this middle panel. In the left-most and right-most panels, corresponding to /pa/ and /ka/, the two rejected sounds have very different log-error slopes. However, the two dashed curves still project back to the chance error probability for a group of 3 (1−1/3). This change in the slope for the two sounds shows that $e_{\min}(i,j)$ can, in general, depend on the sound in the group. This seems to reflect the more robust nature of /ta/ relative to /pa/ and /ka/ due to /ta/ having more high-frequency energy than its competitors.

Based on the small amount of the data shown in Fig. 8 and Fig. 9, it appears that the band-independence assumption [Eq. (13)] and the band error expression [Eq. (14)] model the individual sound confusions $S_{i,j}(\mathrm{AI})$ more accurately than they model the average band error [Eq. (13)]. The total sound error is more precisely the sum of these off-diagonal confusion terms, as given by Eq. (7). The implications of this model seem quite significant, but without more data it is unwise to speculate further at this time.

IV. CONCLUSIONS

The intent of this paper is to provide a theoretical analysis of the venerable 1955 Miller and Nicely confusion matrix data, which have been difficult to fully appreciate, due to inadequate analysis methods. Replotting the data as a function of the SNR, rather than as confusions at a fixed SNR, provides a novel way of robustly clustering the feature groups. This grouping, not robustly defined in a single CM, is easily determined in such SNR plots. When working with individual CM data at a single SNR, clusters depend on the sound ordering. When plotted as a function of SNR, sound order is irrelevant, and clusters depend instead on a smoothness, or continuity, across SNR.

A second contribution is the use of the AI as the independent variable. When the PI is plotted as a function of the SNR, the only structure observed are the clusters (Fig. 2). When these same data are plotted as a function of AI (Fig. 8), they become linear functions of AI (they form straight lines on log-error axes), consistent with the band-independence model of Fletcher [Eq. (13)], thereby corroborating Fletcher's AI model equation [Eq. (16)] for the case of single competing consonants. Plots of $S_{i,j}(\mathrm{AI})$ depend on the spectrum of the noise.

A third contribution is the extension (and verification) of Fletcher's articulation model equation for $P_c(\mathrm{AI})$, for the case of small set size [Eq. (18)], by introducing entropy $\mathcal{H}$ into the model [Eq. (16)], thereby accounting for chance (guessing).

As the SNR increases from chance levels (e.g., −21-dB SNR), sound groups form, forcing the entropy to decrease. The extended model [Eq. (18)] leads us to the conclusion that the entropy must depend on AI. This function, $\mathcal{H}(\mathrm{AI})$, decreases from its maximum value of 4 at AI=0, to $\mathcal{H}_g \equiv \log_2$ (group size) for $\mathrm{AI}=\mathrm{AI}_g$, where the group is fully formed. The entropy associated with these groups may be estimated from the clusters in $S_{ij}(\mathrm{AI})$, or from the intercept at AI=0 of the dashed lines of Fig. 8 and Fig. 9. For the nasal sounds the group size is 2 ($\mathcal{H}=1$), leading to an intercept of 1−1/2. For the [/pa/,/ta/,/ka/] group, the group size is 3 [$\mathcal{H}=\log_2(3)\approx 1.58$]; thus, the intercept is 1−1/3.

A. Parametric model

In summary, a parametric model of the confusion matrices for the sound groups 1–3 and 15, 16 has been established. Chance, defined by $e_{\mathrm{chance}}(\mathcal{H})$ [Eq. (19)], depends only on the experimental set (alphabet) size, characterized by $\mathcal{H}(0)$, not on the sounds themselves. Each sound may be described by three parameters. The sound-dependent parameters of Eq. (18) are $e_{\min}(i,j)$ $\mathcal{H}_g(i)$ and $\mathrm{AI}_g(i)$. The parameter $e_{\min}(i,j)$ appears to be a property of the individual

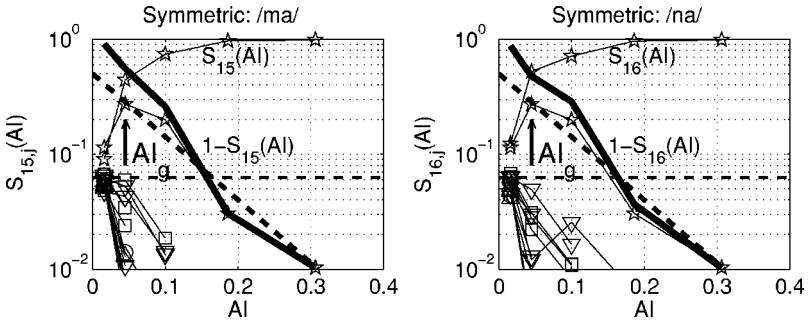

FIG. 8. Since the log-error plots for /ma/ and /na/ (see the lower-right panel of Fig. 7) show the greatest deviation from linear, they seem to be a "worst case" for the AI model. From this figure it is clear that the reason for the deviation from linear dependence is due to the migration of chance from 1/16 ($\mathcal{H}=4$) to 1/2 ($\mathcal{H}=1$), due to the nasal grouping. The rising nasal curves results from the robust grouping of the nasal, resulting in the increase in chance from 1/16 at AI=0 to 1/2 at AI≈0.045. The solid-thick curve is the sum of all the errors (and is $1-Pc$ for the spoken sound). A dashed line has been drawn from the point (0,.5) to (0.31,.01). This line fits the error curve (/na/ given /ma/, and /ma/ given /na/) with very little error, for $\mathrm{AI}=\mathrm{AI}_g>0.045$, and intercepts the ordinate at 1/2 for AI=0, as expected for a 2-group ($\mathcal{H}=1$). This further supports the band independence model Eq. (13).

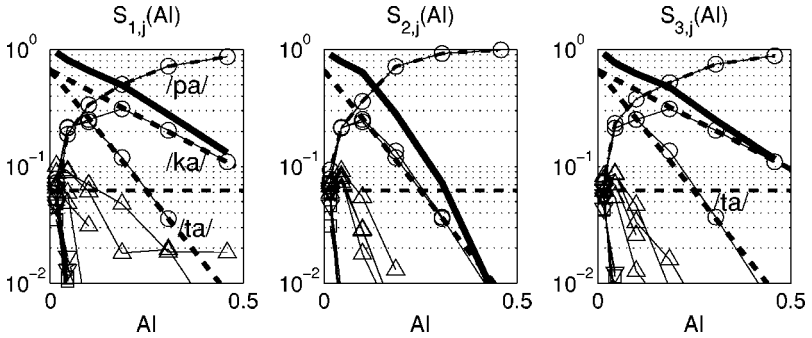

FIG. 9. This figure shows $S_{s,h}(\mathrm{AI})$ for $s=1$, 2, 3 corresponding to the sounds /pa/, /ta/, and /ka/. The dashed lines connect (0,1−1/3) with (0.48,0.1) and (0.442,0.01).

sounds in the group. For example, the value of $e_{\min}(i,j)$ for sounds 1 and 3 are the same, but the value for sound 2 is much smaller (Fig. 9).

The value of e_i is the sum of all the errors of the corresponding off-diagonal values, namely

$$e_i(\mathrm{AI})=\sum_{j\neq i} S_{i,j}(\mathrm{AI}). \tag{21}$$

This equation is the same as Eq. (7) except for the independent variable. The total minimum error is given by the average of the row errors, evaluated at AI=1

$$e_{\min}=\frac{1}{16}\sum_i e_i(\mathrm{AI})\bigg|_{\mathrm{AI}=1}. \tag{22}$$

This follows from Eq. (17) and Eq. (7).

Parameter $\mathrm{AI}_g(i)$ characterizes the transition from the maximum alphabet entropy $[\mathcal{H}(\mathrm{AI}=0)]$ to the group entropy $\mathcal{H}_g$. We have not attempted to find an analytical expression for $H(\mathrm{AI})$, to describe the transition from entropy maximum $\mathcal{H}(0)$ to that of the group $\mathcal{H}(\mathrm{AI}_g)$. The hierarchies mentioned in the Introduction each have their own AI_g parameter, each group within the hierarchy having a smaller value of AI_g.

The change in the entropy, from $\mathcal{H}\rightarrow\mathcal{H}_g$, due to the formation of groups as the sounds start becoming identified, accounts for the deviations from linear of the log-error probability.

One might even view the ultimate identification of the sound, as AI→1, as a further reduction of the sound's row entropy to zero.

Since the parametric model is based on very little data, there is presently no clue as to how it will generalize.

B. Preprocessing of $\mathcal{C}$

The confusion matrix $\mathcal{C}$ was transformed to form the articulation matrix $\mathcal{A}$, by a normalization step [Eq. (1)], and then further transformed by symmetrizing [Eq. (3)]. These two transformations interact and must therefore be iterated to convergence. While these transformation steps are not essential, they are justified, since they reduce sampling noise and remove subject bias. Sampling noise and bias are interesting topics in their own right that deserve further analysis. For example, the skew-symmetric form $A(\mathrm{AI},\mathcal{H})$ [Eq. (4)] should be carefully considered in the future, to characterize and determine the precise nature of the subject bias. This bias should be considered when designing MN55-type experiments.

Once the CM has been preprocessed so that its rows sum to 1, an error decomposition is possible. The equation for this is Eq. (7), which says that the total sound identification error is the sum of the confusions. This expression is useful in those cases where the sounds group, as it uniquely decomposes the error into the group confusions. When plotted as log error (e.g., Fig. 9) one may characterize the sources of the sound errors in a quantitative way [i.e., using the parameters $e_{\min}(i,j)$ and $\mathrm{AI}_g(i)$]. This method seems superior to all previous analysis methods of such confusion matrices.

As shown in Fig. 7, the nasals appear to violate Fletcher's independence formula Eq. (13), since the total error is not a straight line on the log-error plot. However, when decomposed by Eq. (7), we see that *individual competing sounds obey band independence*. Thus the total error deviates from a linear log error, due to the dramatic change in $\mathcal{H}$ with AI, from 4 to 1, over the small range of AI between 0 and $\mathrm{AI}_g\approx 0.045$ (Fig. 8).

Furthermore, the projection of the straight lines that lie along the log-error curves, back to AI=0, gives e_{chance}, for the group. This is an important corroboration of Eq. (18). Two examples of these are seen for the nasals, where the group has entropy 1, and the dashed lines of Fig. 8 project back to 1–1/2, and the 3-group of Fig. 9, where the dashed lines project back to 1–1/3.

C. Calculating the AI

The procedure for calculating the AI, developed in Sec. III A, has some novel aspects as well. Rather than defining the specific AI in terms of the band SNRs, the modified function of Eq. (11) was used. The justification for Eq. (11) comes from the work of Fletcher as well as French and Steinberg, both of whom promote (but did not use) this detection formulation. The *speech detection constant c* is chosen to characterize the detection of the speech peaks when noise is added to the speech. Even though this formulation of the AI has some important advantages, and is more accurate, it is never referred to in the modern AI literature. This, I feel, is a mistake that needs rectification. Again, this approach was not studied in detail in this paper; however, there is a detailed analysis of Eq. (11) in both of the references, and a deeper analysis here is off topic. It was necessary to introduce Eq. (11) to get reasonable values of AI(SNR) when fitting the model $P_c(\mathrm{AI},\mathcal{H})$, as shown in Fig. 6. This is because the estimates of the SNR as a function of frequency [i.e., $\mathrm{snr}_k(\mathrm{SNR})$], in the left panel of Fig. 5, are strongly affected by this detection model, and on the specific choice of c in

Eq. (10). Without the use of this speech detection parameter ($c=2$), the band SNR values $\mathrm{snr}_k(\mathrm{SNR})$ would be unrealistic for small values of AI.

D. Band independence

Fletcher's band-independence assumption Eq. (13) has proven to be an important tool at the individual sound level. This should come as a surprise, as it was not anticipated by Fletcher's work, or any work following (that I am aware of). On the other hand, the fact that it works at all should lead us to the possibility that it could generalize. It would appear from the analysis provided here that Eq. (13) is more accurate in describing competing sounds than in describing the average probability correct $P_c(\mathrm{AI},\mathcal{H})$. This statement is supported by the very linear behavior of the off-diagonal confusion terms $S_{i,j}$ in Fig. 8 and Fig. 9. The partial errors are highly linear once the group has formed ($\mathrm{AI}>\mathrm{AI}_g$). It follows from Eq. (7) that the deviations from linear are a result of the groups, which depend on the noise spectrum. The influence of a group formation distorts this basic linear character, and therefore distorts the linearity of the sum over many error terms [i.e., $P_c(\mathrm{AI})$]. Based on the small amount of data we presently have (those shown in this paper), it would be reasonable to conclude that band independence is more a property of individual consonants than it is of the group means, as first proposed (derived) by Fletcher. Much more data and analysis are needed to verify this possibility, which is hardly proved at this time.

E. Implications to ASR

Automatic computer recognition of speech (ASR) could benefit from many of the same considerations as those of MN55. It would be interesting to run similar experiments on modern ASR systems, to characterize their $\mathcal{A}(\mathrm{AI},\mathcal{H})$ performance. In many cases this might not be practical, due to the limited performance of the ASR front ends, or if the confusion matrices turned out to be skew-symmetric. The ASR language model performance is inhibited when using nonsense speech, since most of these systems depend on some sort of language context for their performance.

ACKNOWLEDGMENTS

The inspiration for this work started with a question by David Nahamoo which I could not answer: ''What is the meaning of $e_{\min}$?'' I would like to thank my students Suvrat Budhlakoti, Bryce Lobdell, Andrew Lovitt, and especially Sandeep Phatak, and also thank Harry Levitt, for many important insights and critical discussion, and thank Anthony Watkins and two anonymous reviewers, for many insightful comments. Finally, I would especially like to thank George Miller for doing his original 1955 work, for reading the present manuscript, and for the personal encouragement he has provided.

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

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

ANSI S3.5-1997 (**1997**). ''Methods for calculation of the speech intelligibility index (SII-97)'' (American National Standards Institute, New York).

Fletcher, H. (**1995**). ''Speech and hearing in communication,'' in *The ASA Edition of Speech and Hearing in Communication*, edited by J. B. Allen (Acoustical Society of America, New York).

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

Fletcher, H., and Munson, W. (**1937**). ''Relation between loudness and masking,'' J. Acoust. Soc. Am. **9**, 1–10.

French, N. R., and Steinberg, J. C. (**1947**). ''Factors governing the intelligibility of speech sounds,'' J. Acoust. Soc. Am. **19**, 90–119.

Goldstein, L. (**1980**). ''Bias and asymmetry in speech perception,'' in *Errors in Linguistic Performance*, edited by V. A. Fromkin (Academic, New York), Chap. 17, pp. 241–261.

Houtgast, T., and Steeneken, H. (**1973**). ''The modulation transfer function in room acoustics as a predictor of speech intelligibility,'' Acustica **28**, 66–73.

Kryter, K. D. (**1962a**). ''Methods for the calculation and use of the articulation index,'' J. Acoust. Soc. Am. **34**(11), 1689–1697.

Kryter, K. D. (**1962b**). ''Validation of the articulation index,'' J. Acoust. Soc. Am. **34**(11), 1698–1702.

Miller, G. A. (**1951**). *Language and Communication* (McGraw Hill, New York).

Miller, G. A., and Nicely, P. E. (**1955**). ''An analysis of perceptual confusions among some English consonants,'' J. Acoust. Soc. Am. **27**(2), 338–352.

Miller, G. A., Heise, G. A., and Lichten, W. (**1951**). ''The intelligibility of speech as a function of the context of the test material,'' J. Exp. Psychol. **41**, 329–335.

Steeneken, H., and Houtgast, T. (**1980**). ''A physical method for measuring speech-transmission quality,'' J. Acoust. Soc. Am. **67**(1), 318–326.

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