Perceptual Confusions Among Consonants, Revisited—Cross-Spectral Integration of Phonetic-Feature Information and Consonant Recognition

Thomas U. Christiansen and Steven Greenberg, *Senior Member, IEEE*

***Abstract*—The perceptual basis of consonant recognition was experimentally investigated through a study of how information associated with phonetic features (Voicing, Manner, and Place of Articulation) combines across the acoustic-frequency spectrum. The speech signals, 11 Danish consonants embedded in Consonant + Vowel + Liquid syllables, were partitioned into 3/4-octave bands ("slits") centered at 750 Hz, 1500 Hz, and 3000 Hz, and presented individually and in two- or three-slit combinations. The amount of information transmitted (IT) was calculated from consonant-confusion matrices for each feature and slit combination. The growth of IT was measured as a function of the number of slits presented and their center frequency for the phonetic features and consonants. The IT associated with Voicing, Manner, and Consonants sums nearly linearly for two-band stimuli irrespective of their center frequency. Adding a third band increases the IT by an amount somewhat less than predicted by linear cross-spectral integration (i.e., a compressive function). In contrast, for Place of Articulation, the IT gained through addition of a second or third slit is far more than predicted by linear, cross-spectral summation. This difference is mirrored in a measure of error-pattern similarity across bands—Symmetric Redundancy. Consonants, as well as Voicing and Manner, share a moderate degree of redundancy between bands. In contrast, the cross-spectral redundancy associated with Place is close to zero, which means the bands are essentially independent in terms of decoding this feature. Because consonant recognition and Place decoding are highly correlated (correlation coefficient $r^2 = 0.99$), these results imply that the auditory processes underlying consonant recognition are not strictly linear. This may account for why conventional cross-spectral integration speech models, such as the Articulation Index, Speech Intelligibility Index, and the Speech Transmission Index do not predict intelligibility and segment recognition well under certain conditions (e.g., discontiguous frequency bands and audio-visual speech).**

***Index Terms*—Consonant recognition, cross-spectral integration, information theory, phonetic features, speech perception.**

Manuscript received January 26, 2010; revised June 30, 2010, January 21, 2011, and May 13, 2011; accepted May 15, 2011. Date of publication June 09, 2011; date of current version November 09, 2011. This work was supported by Forskningsrådet for Teknologi og Produktion (TUC), the Technical University of Denmark (SG), and the U.S. Air Force Office of Scientific Research (SG). This work was performed while S. Greenberg was a visiting professor with the Technical University of Denmark. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hui Jiang.

T. U. Christiansen is with the Centre for Applied Hearing Research, Department of Electrical Engineering, Technical University of Denmark, DK-2800 Kgs. Lyngby, Denmark (e-mail: tuc@elektro.dtu.dk).

S. Greenberg is with Silicon Speech, Kelseyville, CA 95451 USA (e-mail: steveng@silicon-speech.com).

Digital Object Identifier 10.1109/TASL.2011.2159202

List of Abbreviations

AI	Articulation index.
CV	Consonant + vowel.
CVC	Consonant + vowel + consonant.
F2	Second formant.
IT	Information transmitted.
IT_r	Relative amount of information transmitted.
MN55	Miller and Nicely (1955) [24].
PoC	Principle of complementarity.
PoE	Product of errors.
PF	Phonetic feature.
STFFT	Short-time fast Fourier transform.
SIQ	Spectral integration quotient.
SII	Speech intelligibility index.
SIM	Spectral integration metric.
SNR	Signal-to-noise ratio.
STI	Speech transmission index.
SyR	Symmetric redundancy.
VOT	Voice onset time.

I. Introduction

How speech information is processed and combined across the frequency spectrum has been the focus of numerous studies—so much so that it can be fairly stated that spectral integration forms a pervasive theme in modern-day speech research. The Articulation Index (AI; [1], [2]), SII [3], and STI [4] are functional models upon which much of our current understanding of human speech recognition is based.

Using low- and high-pass filtered speech, Fletcher and colleagues developed the AI as a principled framework for quantifying listeners' ability to process and decode the speech signal (e.g., [5]). The AI attempts to predict phonetic-segment recognition based on a weighted sum of the signal-to-noise ratios (SNRs) associated with the long-term average speech spectrum and the long-term average noise spectrum across acoustic-frequency channels. The AI was later extended to include corrections for elevated hearing thresholds, vocal effort, and dif-

ferent linguistic materials, and renamed the Speech Intelligibility Index (SII; [3]). Another extension to the AI was developed by Steeneken and Houtgast [4] to predict intelligibility in room environments. The Speech Transmission Index (STI) computes the low-frequency (2 to 12 Hz) modulation spectrum in a number of frequency channels rather than the SNR per se. It assumes that intelligibility depends on specific properties of the low-frequency modulation spectrum. However, the modulation spectrum is also an indirect measure of SNR. This is because the magnitude of the modulation spectrum, whose peak lies between 3 and 6 Hz for pristine, undistorted speech, is highly correlated with conventional SNR estimates. The higher the SNR, the higher the peak-to-valley ratio of the waveform modulation; in turn, this translates into a higher peak magnitude in the modulation spectrum. In this sense, the modulation spectrum reflects something akin to what the AI and SII were designed to measure.

All three spectral integration metrics (SIMs)—AI, SII, and STI—assume that the acoustic signal is decomposed by the auditory system into a series of frequency channels and that each is processed independently of the others. Information extracted from each channel is subsequently integrated across auditory frequency channels, which are assumed to be independent (the STI's most recent version [6] includes a redundancy-correction factor—see below). This classical framework has been the prevailing one for quantifying phoneme recognition and intelligibility for many decades.

As useful as the AI, SII, and STI SIMs are, several unresolved issues remain. In order to explain the logic and motivation of our own study, which uses signals and analyses quite different from the traditional approaches, we first discuss the three most important unresolved issues.

First, the approach used by these classical SIMs relies on the long-term spectra of the signals. The speech studies conducted by Fletcher and his colleagues [5] used analog filters through which a signal was either high-pass or low-pass filtered or band-limited (i.e., both high- and low-pass filtering). As a consequence, the AI (and other models with comparable SIMs, such as the SII and STI) models articulation (and by extension, intelligibility—see below) as a function of signal bandwidth and frequency-weighted "importance." This approach does not accurately predict recognition performance for non-contiguous frequency spectrum signals, as was shown by Kryter [7] and others over the intervening years [8]–[12]. In his study, three 500-Hz-wide bands centered at 500 Hz, 1500 Hz, and 2500 Hz, provided much higher intelligibility than predicted by the AI, based on a single 1500-Hz-wide band with appropriate spectral weighting.

Second, the band-independence assumption (see [13]) does not hold under all conditions. Steeneken and Houtgast determined that to adequately model intelligibility [4] and phoneme-specific recognition [14] it is necessary to assume that a spectral region much broader than the classical critical band is involved in spectral integration of speech. Their estimate for this speech-specific bandwidth is 3/4 to 1 octave [4], [14], an estimate close to the integration bandwidth proposed by [15] for vocalic material and similar to that used in the current STI standard [6]. A "redundancy-correction factor" was used in [14] to model the recognition of CVC syllables presented over a range of SNRs. This correction factor—proportional to the amount that pairs of neighboring spectral bands contribute to overall speech transmission—significantly improved their model predictions and roughly approximated Kryter's [7] data. These authors suggested that the spectral resolution around 2 kHz should be modeled as finer-grained than in other frequency regions because of the relatively low degree of cross-spectral redundancy in this part of the spectrum and its greater "importance" relative to other frequencies.

Several models beside the AI, SII, and STI also assume channel independence in speech processing. Ronan and colleagues [16] evaluated several of these with respect to their ability to predict consonant-recognition accuracy, (i.e., percent correct). These models—known as "Pre-labeling" [17], "Post-labeling" [17], and Fuzzy-logic [18]—were originally designed to account for cross-modal integration in audio-visual speech recognition (e.g., [19]). Given subjects' responses to audio- or visual presentation alone, the models try to predict the responses to concurrent audio-visual presentation. The Pre-labeling model assumes responses are optimally selected prior to integration of the sensory streams (i.e., independent of the other modality). The Post-labeling model assumes this optimization occurs after integration of the signal modalities. The Fuzzy-logic model assumes optimal integration is based on a Euclidean centroid in a Fuzzy-logic response space. In [16], subjects were asked to identify consonants passed through either single, band-pass filters (whose center frequencies ranged between 700 and 2400 Hz) or through a combination of these pass-bands. The three models' predictions for the combined band-pass conditions were based on subjects' responses to the single, band-pass-filtered consonants. Although the models were able to satisfactorily account for subsets of the data, consonant-recognition accuracy associated with conditions that included high-frequency bands were not well predicted, suggesting that the band-independence assumption does not entirely hold for non-adjacent frequency bands. Other studies (e.g., [7] and [20]) have reached similar conclusions.

Third, one of the cornerstones of the AI, SII, and STI is their reliance on a frequency-weighting factor to model the differential "importance" of various parts of the acoustic spectrum (although the AI, as described in [2], does not explicitly include weighting functions for different linguistic materials, several subsequent publications describe such functions (e.g., [21]–[23]). The theoretical basis of this weighting has never been adequately explained, nor has the variation in frequency weighting required to adequately model different types of linguistic material.

Miller and Nicely [24] (henceforth, MN55) approached cross-spectral integration of speech information from a different perspective. They computed consonant-confusion matrices for high- and low-pass filtered CV syllables presented in a background of Gaussian noise and analyzed the errors in terms of distinctive, articulatory-acoustic (henceforth, "phonetic") features (PFs) [25]. Their study showed that the associated information is not distributed across the acoustic-frequency

spectrum in the same way for all features. For example, the amount of information transmitted (IT) for "Place of Articulation" (henceforth, "Place") increased almost linearly as the speech bandwidth widened (their Fig. 3). In contrast, the corresponding information functions for "Voicing" (MN55—their Fig. 3) and "Manner of Articulation" (henceforth "Manner") (MN55's Fig. 4) reached asymptote at relatively narrow bandwidths (ca. 1 octave). MN55's study is important because it focused on three aspects of speech processing that previously had been ignored—1) decomposition of consonants into structural primitives (i.e., PFs), 2) detailed error analyses derived from confusion matrices, and 3) the use of an information-theoretic analysis based on 1) and 2) that provided insight into the perceptual processes underlying consonant recognition.

Moreover, MN55's data point to some interesting properties of cross-spectral speech processing that have not received as much attention as they deserve. Their data imply that the information associated with Place is organized very differently than Manner and Voicing, and this difference may have significant implications for consonant recognition and speech perception in general, as discussed in Section IV.

One of the missing elements in MN55's study is that what they measured was probably not cross-spectral integration per se, but rather the distribution of phonetically relevant information across frequency channels. Although related, distribution of information and integration of information are not the same.

We illustrate this distinction through an analogy with loudness models. Intensity is a physical measure, which can be represented in terms of auditory-nerve (and higher-level) neural-firing patterns. Loudness is intensity's perceptual correlate. Most models of loudness (see [26]) compute an excitation pattern, correlated with nerve-fiber firing patterns, grouped by frequency band. The widths and center frequencies of these bands are derived from psychophysical experiments and designed so that they contribute equally to loudness. The contribution from each frequency band can simply be summed across the frequency spectrum to yield the overall loudness (but cf. [27]).

The spectral distribution of information in such loudness models is implicitly given by the arrangement of frequency bands, which relates to peripheral processing. In contrast, the spectral integration of speech information ultimately depends on higher-level interpretation of this information by more central processes. For loudness, this process is often modeled by a simple summation across frequency channels. The aim of this paper is to show that a simple summation is not sufficient to model spectral integration of information for consonant recognition.

So, why were perceptual integration and information distribution conflated in MN55's study? It is because the low- and high-pass filtering employed in their study does not allow the two to be dissociated as readily as narrow, bandpass filtered signals would.

Most perceptual studies have investigated cross-spectral integration by degrading intelligibility via acoustic interference, usually white- or speech-spectrum-shaped noise (e.g., [24], [28], and [29]). Although noise-masking studies serve as the foundation of much of our current knowledge of human speech recognition, even this approach has certain drawbacks. For example, broadband, background noise may evoke certain nonlinear processes in the auditory pathway (e.g., [30]–[32]) that complicate the modeling of auditory speech processing. Moreover, the efferent system, providing feedback from the auditory cortex and brainstem into the cochlea via the olivo-cochlear bundle, is known to be active in the presence of broadband noise [33] and thereby modifies the speech signal's sensory representation. Additionally, under certain conditions background noise may actually enhance rather than degrade recognition (e.g., [34]). Another consideration in using noise masking is auditory scene analysis, wherein the listener tries to perceptually separate the foreground speech signal from the acoustic, interfering background [35]–[37]. Background noise may impede the listener's ability to fully attend to the speech signal independent of the background's energetic masking impact, thereby reducing recognition performance due to factors other than auditory processing and integration.

Using narrowband, spectral slits avoids most of these problems by eliminating background noise as part of the stimulus. The spectral limits of the signal are predefined and hence the portion of the spectrum used by the listener is known in advance.

A key innovation of MN55, one adopted in the current study, was to report results in terms of information transmitted rather than recognition accuracy in percent correct. IT takes both the correct and incorrect responses into account, providing a concise way of quantifying the ability of listeners to reliably distinguish each consonant from all others in the response set. The way in which MN55 used the IT metric went further because they also computed IT for the component PFs. This allowed them to determine which consonant properties are robust to distortion and which are not. One of the conclusions they drew was that Voicing and Manner cues are relatively robust to all but the highest amounts of background noise, and that Place cues are often fragile even in the presence of low-level interference. However, their study left several issues unaddressed, foremost of which is precisely why Place cues are so vulnerable to noise while Manner and Voicing are not. The most likely answer is that Place cues are more broadly distributed across the acoustic-frequency spectrum than the other features, but this is probably not the whole story; this explanation is inconsistent with studies showing that the second formant (F2) onset spectrum [38] and/or F2-transition pattern [39], [40] play an important role in consonant place-of-articulation decoding. Yet, if these acoustic cues were the sole source of Place information, the feature would not be so sensitive to speech and noise-masker bandwidth, given they are mainly confined to the F2 region (1200 to 2500 Hz). It is for such reasons that factors other than mere spectral distribution of phonetic cues are likely to play an important role in consonant recognition.

Since the publication of MN55, PFs have been used to investigate central processes underlying consonant recognition (e.g., [41]–[43], but see [13]). However, these studies did not directly investigate the cross-spectral *perceptual* integration of PFs, the focus of our paper.

II. Methods

A. Design

The contribution of three non-overlapping spectral regions ("low," "mid," and "high") to Danish consonant recognition and PF decoding was quantified. Roughly speaking, the low-frequency band, centered at 750 Hz, is close to, but not centered on, the first formant, the mid-frequency band (1500 Hz) within the range of the second formant, and the high-frequency band (3000 Hz) is close to the third formant. The bands ("slits") were presented individually and in combination with other slits, as described below. These slit patterns were used to generate consonant-confusion matrices from which the amount of relative information transmitted $(\mathrm{IT_r})$ associated with Consonant recognition and PF decoding could be computed. The $\mathrm{IT_r}$ patterns were used to compute a measure of cross-spectral integration and were also used to evaluate several previously published sensory-integration models. In addition, the confusion matrices were used to compute a measure of perceptual redundancy across spectral regions, and the results evaluated in conjunction with the spectral-integration measure.

B. Stimuli

Stimuli were Danish syllables recorded in a sound-insulated chamber [44]. Each stimulus presentation was a concatenation of a short, unfiltered carrier phrase "På pladsen mellem hytten og..." (English translation: "On the square between the cottage and...") and a test syllable. Each test syllable contained one of eleven consonants, [p], [t], [k], [b], [d], [g], [m], [n], [f], [s], [v], followed by one of three vowels, [i], [a], [u]. Each token concluded with the unstressed liquid + neutral vowel syllable [lə] (e.g., [tilə] [talə, [tulə]). Recordings of a female talker and a male talker, each enunciating the carrier sentence and the test syllables, were used.

A carrier phrase was used in order to: 1) focus the attention of the test subjects on a delimited point in time, 2) provide a relatively natural context in terms of sound level and talker, and 3) improve the listener's concentration.

The audio sample rate was 20 kHz. The signals were subsequently up-sampled to 44.1 kHz for stimulus presentation. The acoustic-frequency spectrum was partitioned into slits. The lowest-frequency slit was centered at 750 Hz, the middle slit at 1500 Hz, and the highest at 3000 Hz. All slit combinations were tested, yielding a total of 3 single slits + 3 two-slit combinations + 1 with all three slits = 7 slit configurations. An STFFT of the signal was performed using a cosine window comprising 256 sample points with 3/4-overlap. Bandpass filtering was achieved by zeroing frequency bins outside the nominal pass bands and then performing an inverse FFT with an identical cosine window. The resulting bandpass filters had a 3-dB bandwidth of 3/4-octave and nominal slopes of 120 dB/octave outside the passband.

The center frequency and bandwidth of the slits were chosen through extensive pilot experiments. In designing the signal's spectral properties, five criteria were met: 1) consonant-recognition accuracy with all spectral regions present were close to, but clearly below 100% to avoid ceiling effects; 2) consonant-recognition accuracy with two slits were clearly lower than with all three slits present but significantly higher than with only one slit present; 3) consonant-recognition accuracy for individual spectral bands presented alone were clearly above chance level in order to avoid floor effects; 4) consonant-recognition accuracy was roughly comparable across single-slit conditions, making it more straightforward to measure cross-spectral integration without significant disparities in baseline recognition performance; and 5) the slit bandwidths were the same in terms of geometric (i.e., octave) units.

In order to prevent listeners from using information outside the slit, many studies (e.g., [14] and [16]) add noise surrounding the pass-band. Background noise is often used in psychoacoustic studies to preclude "off-frequency" listening in signal-detection tasks. However, the task in the current study is not detection of a narrowband signal, where background noise that masks off-frequency listening is a valid method, but rather recognition and decoding of a speech signal. These speech-processing tasks rely on a large number of tonotopically distributed auditory neurons. Moreover, and as mentioned in Section I, it has recently been demonstrated that noise in the spectral gaps can actually enhance consonant recognition by means of "spectral restoration" [34]. For these and the reasons mentioned in Section I, background noise was not added to the speech signals in our study.

The seven conditions are listed, along with their consonant-recognition scores, in Fig. 1. They were presented once for each combination of consonant + vowel context and talker—$11 \times 3 \times 2 \times 7 = 462$ test presentations for each listener. Control conditions consisted of unfiltered combinations of all consonants, vowels and talkers—$11 \times 3 \times 2 = 66$ conditions, and were interleaved with the test conditions (details below).

MN55 used 16 English consonants [p t k f θ s $\int$ b d g v ð z ʒ m n], of which we excluded [θ $\int$ ð z ʒ], because [θ $\int$ z ʒ] do not have Danish counterparts and [ð] does not occur syllable-initially in Danish. The remaining 11 consonants are similar to their English counterparts. It is worth noting that because Danish [v] has no explicit frication it should be transcribed with the IPA-symbol for the approximant [υ]. Nevertheless, it still functions as a Danish consonant alongside the other Danish consonant phonemes /p t k b d g f s h m n l r j/, and is conventionally classified as a fricative [45].

The term "voicing" is conventionally used for distinguishing certain English stop consonants (e.g., [p] versus [b], [t] versus [d]). The term "aspiration" is the conventional phonetic term for distinguishing between their Danish counterparts. The single most important physical property distinguishing voiced stops from voiceless stops in English and aspirated stops from unaspirated stops in Danish is voice onset time (VOT); hence, the English convention, i.e., "voicing," is retained for the purpose of the present analysis and discussion.

C. Procedure and Subjects

The data associated with the seven different slit configurations were collected as part of a larger study comprising 83 slit configurations, where the presentations were divided into nine sessions, each lasting less than 2 hours, during which subjects were allowed to take short breaks. The data presented in

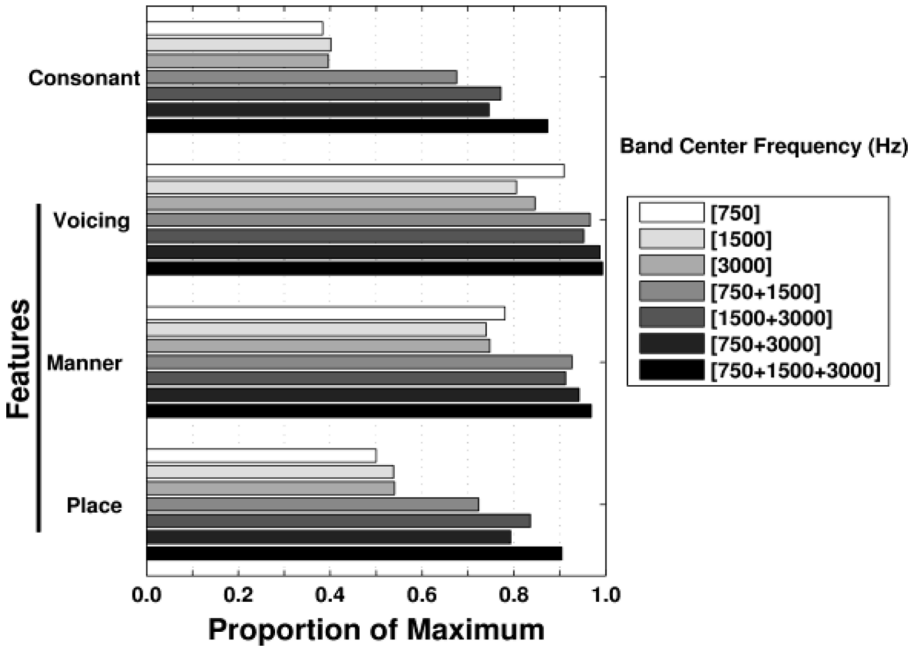

Fig. 1. Consonant-recognition accuracy and feature-decoding precision for each stimulus condition averaged across the six subjects. The coefficient of variation (i.e., standard deviation divided by the mean) was always less than 0.08 and usually below 0.03. 99.0% of the consonants were correctly recognized in the absence of band-pass filtering (i.e., the original, unprocessed signals).

this paper were collected during three sessions for each subject. The total number of presentations for each subject was 3×66 control conditions for the three sessions $+11 \times 3 \times 2 \times 7$ test conditions $= 660$. The test conditions were randomly distributed across the three sessions. The 66 control conditions were randomly distributed across each session. Their average consonant-recognition accuracy was 99.0%, and was always greater than 96.7%. The stimulus conditions excluded from this paper are associated with spectro-temporal manipulations that lie outside the scope of the present study.

The subject was seated in a double-walled sound booth. His/her task was to identify the initial consonant of the test signal by mouse-selecting it from the 11 consonant candidates displayed on a computer display. No response feedback was provided. Six native speakers of Danish (three males, three females) between the ages of 21 and 28 were paid for their participation. All reported normal hearing and no history of auditory pathology. The experiment protocol was approved by the Science-Ethics Committee for the Capital Region of Denmark; reference H-KA-04149-g. All subjects signed an informed consent form.

Stimuli were presented diotically over Sennheiser HD-580 headphones at a sound pressure level of 65 dB SPL using a computer running Matlab version R2006 under Windows XP with a RME Digipad 96 soundcard. The sound pressure level was calculated as the RMS-value of the given nonsense syllable after processing.

D. Analysis—Confusion Matrices, PFs, IT_r, and Symmetric Redundancy

When a consonant is identified correctly, all of its distinctive PFs are, by definition, decoded accurately. However, when a consonant is incorrectly identified, it is rare that all of its PFs are incorrectly decoded; one or two of the features are usually decoded correctly. The data were analyzed in this fashion for the features *Voicing*, *Manner*, and *Place*. Voicing is a binary feature, whereas Manner and Place encompass three classes for the Danish consonants used in this study, as shown in Table I.

TABLE I
DEFINITION OF PHONETIC FEATURES

	Voicing	Manner	Place
[p]	voiceless	stop	anterior
[t]	voiceless	stop	medial
[k]	voiceless	stop	posterior
[b]	voiced	stop	anterior
[d]	voiced	stop	medial
[g]	voiced	stop	posterior
[s]	voiceless	fricative	medial
[f]	voiceless	fricative	anterior
[v]	voiced	fricative	anterior
[n]	voiced	nasal	medial
[m]	voiced	nasal	anterior

Phonetic features for the 11 consonants used in the study. Voicing is a binary feature, while Manner and Place are ternary features.

Consonant-confusion matrices provide a straightforward means of analyzing error patterns associated with the consonant-recognition task in terms of constituent phonetic features. Table II shows an example of one such consonant-confusion matrix in Panel (a), where row values refer to the stimulus presented, while column values denote listener responses. Correct response counts are indicated in bold along the diagonal. The

TABLE II
(a) CONFUSION MATRICES, (b) DERIVED CONFUSION MATRIX FOR VOICING, (c) DERIVED CONFUSION MATRIX FOR MANNER, AND (d) DERIVED CONFUSION MATRIX FOR PLACE

(a)

Stimulus \ Response	p	t	k	b	d	g	s	f	v	n	m	Total
p	**7**	11	2	3	4	2	1	1	3	1	1	36
t	4	**15**	2	1	5	0	2	2	3	1	1	36
k	7	7	**12**	3	1	1	1	0	3	0	1	36
b	1	0	0	**14**	11	3	0	0	3	3	1	36
d	0	0	1	5	**24**	5	0	0	0	0	1	36
g	1	0	2	9	9	**7**	0	1	3	0	4	36
s	1	3	1	0	1	0	**13**	3	1	6	7	36
f	1	7	0	1	1	0	6	**9**	3	2	6	36
v	0	2	2	0	2	3	0	3	**16**	4	4	36
n	0	0	0	2	1	1	0	1	0	**24**	7	36
m	0	0	0	1	2	3	1	0	0	11	**18**	36
Total	22	45	22	39	61	25	24	20	35	52	51	396

Consonant-recognition accuracy: 159/396 = 40.2%

Example confusion matrices for the 1500-Hz slit condition summed across the six test subjects. Row values pertain to the stimulus presented, while column values are the listener responses. Panel (a) shows the raw responses to the consonant-recognition task. For example, test subjects reported hearing [t] 11 times when the stimulus was [p]. Correct response counts are indicated in bold. Consonant-recognition accuracy for this example is 40.2 percent. Panels (b), (c) and (d) show the confusion matrices derived from panel (a) for Voicing, Manner and Place, respectively, as described in Section II.D. The sum of all elements from the left-most sub-matrix in Panel (a) yields the number of correctly identified stop consonants (179), which in turn is the upper left-hand element in Panel (c). Similarly, the 23 stop consonants confused for fricatives in the right-most box of Panel (a), corresponds to the upper middle element of Panel (c).

(b)

Stimulus \ Response	Voiced	Unvoiced	Total
Voiced	201	15	216
Unvoiced	62	118	180
Total	263	133	396

Decoding precision: 319/396 = 80.6%

(c)

Stimulus \ Response	Stop	Fricative	Nasal	Total
Stop	179	23	14	216
Fricative	25	54	29	108
Nasal	10	2	60	72
Total	214	79	103	396

Decoding precision: 293/396 = 74.0%

(d)

Stimulus \ Response	Anterior	Medial	Posterior	Total
Anterior	96	69	15	180
Medial	39	95	19	144
Posterior	32	18	22	72
Total	167	182	56	396

Decoding precision: 213/396 = 53.8%

corresponding PF confusion matrices for Voicing, Manner and Place are shown in Panels (b)–(d) of Table II.

The PF-confusion matrix is derived from the consonant-confusion matrix by first grouping the consonant-by-PF associa-

tions (see Table I). In Panel (c) of Table II, the groups are the Manner classes "stop" ([p], [t], [k], [b], [d], [g]), "fricative" ([s], [f], [v]) and "nasal" ([n], [m]). The sum of the consonants with "stop" as the manner of articulation identified correctly (179) is placed in the upper left-hand cell of the matrix [see Table II, Panels (a–c)]. The sum of the stop consonants with Manner wrongly identified as "fricative" (23) is placed in the cell immediately to the right, and so on.

The proportion of correctly identified elements in a confusion matrix (recognition score) can be calculated as the sum of the diagonal elements divided by the sum of all the elements, as shown in Table II(a)—where the sum of the elements along the diagonal is 159 and the total number of presentations is 396, resulting in a consonant-recognition score of 40.2%. Because the focus of this study is cross-spectral integration not the *specific pattern* of perceptual confusions, full confusion matrices are omitted. They are available from the authors upon request.

MN55 used IT as a bias-neutral way of quantifying patterns in confusion matrices. In their study, as in ours, IT was calculated not only for consonants, but also for PFs. In our view, IT provides a more transparent method of understanding the pattern of confusions than the conventional percent-correct metric. IT is particularly useful in the context of the present study in view of the 11-alternative-forced-choiced paradigm used.

Calculation of IT in bits for any given confusion matrix is shown as

$$I(S;X) = -\sum_{S,X} p_{SX} \log_2 \frac{p_s p_x}{p_{sx}} \tag{1}$$

where $I(S;X)$ refers to the mutual information (or IT) between S (stimulus) and X (response; i.e., the number of bits transmitted), p_{sx} is the probability of stimulus, s, co-occurring with response x, p_s is the probability of stimulus s occurring, and p_x is the probability of response x occurring.

Equation (1) can be applied to confusion matrices reflecting both consonant recognition and PF decoding. This measure of information transmitted provides useful insights about the distribution of errors (i.e., confusions).

In order to compare IT associated with Voicing, Manner and Place, as well as Consonant recognition, we (as do MN55) normalize the absolute IT values by a factor equal to the maximum observed information transmitted (i.e., the results are scaled to a normalized proportion of unity). The maximum information transmitted for a given probability distribution is equal to the entropy of the probability distribution as described as

$$H(X) = -\sum_{x} p(x) \log_2(p(x)) \tag{2}$$

where X is the feature variable, p is the probability mass function and x are the events (feature values or consonant-segment labels). For example, Voicing has the normalized value of $(5/11)\log_2(5/11) + (6/11)\log_2(6/11) = 0.990$ because, out of the eleven consonants, there are five with the value "Unvoiced" and six marked as "Voiced." Thus, the normalized indices are 0.990, 1.43, and 1.49 for decoding of Voicing, Manner, and Place, respectively, and 3.46 for Consonant recognition. Normalized in this way, the scores designate relative IT $(\mathrm{IT_r})$, which is not to be confused with relative entropy values, also known as Kullback–Leibler divergence.

We illustrate the advantage of using a PF-based, IT-metric analysis through the following example. Consider the consonant-recognition scores in terms of percent correct for the three single-slit conditions (Fig. 1). They are very similar. Compare these recognition scores with the $\mathrm{IT_r}$ for the same conditions (Fig. 3). In contrast to the consonant-recognition scores, the consonant $\mathrm{IT_r}$ for the 750-Hz slit is higher than the $\mathrm{IT_r}$ associated with the 1500-Hz and 3000-Hz conditions. This disparity between accuracy in percent correct and information-based metrics reflects the fact that the pattern of errors differs significantly across the three stimulus conditions, which is not reflected in the consonant recognition scores shown in Fig. 1. Hence, the ability to distinguish a consonant relative to others cannot truly be measured with percent-correct accuracy alone. A finer-grained analysis is required to quantify how well consonants are distinguished relative to each other. The use of PFs and an information-theoretic metric provides the capability of doing so.

The cross-spectral integration of phonetic information can be quantified using the "Spectral Integration Quotient" (SIQ) metric. The SIQ is defined as

$$\mathrm{SIQ}_m = \frac{\mathrm{IT}_{r,m}}{\sum_i \mathrm{IT}_{r,i}} \tag{3}$$

where $\mathrm{SIQ_m}$ is the spectral integration quotient for the multi-slit condition m, $\mathrm{IT}_{r,m}$ is the observed $\mathrm{IT_r}$ for multi-slit condition m, and $\mathrm{IT}_{r,i}$ is the $\mathrm{IT_r}$ for a given slit included in the multiple-slit condition m. If the cross-spectral integration were linear, the SIQ would be 1. SIQs less than 1 imply compression of spectral integration, whereas those much higher than 1 reflect expansion. The SIQ can be defined for consonants as well as for PFs.

In order to quantify the amount of information redundancy contained in each band, conditions were pair-wise compared and the mutual information shared between the responses associated with each condition was computed. The resulting values were then normalized, with the average entropy of the responses calculated as

$$\hat{I}(X;Y) = \frac{2I(X;Y)}{H(X) + H(Y)} \tag{4}$$

where $\hat{I}(\mathrm{X};\mathrm{Y})$ is the Symmetric Redundancy (SyR) [46] of variables X and Y, $I(X;Y)$ is the mutual information shared between the responses X and Y, $H(X)$ is the entropy of X and $H(Y)$ is entropy of Y. The SyR can range between 0, signifying complete independence (i.e., no correlation among responses), and 1, denoting complete dependence (i.e., $(X;X) = 1$).

III. Results

We first discuss consonant recognition in terms of accuracy (i.e., percent correct), and also consider how this metric applies to PF-decoding. We then discuss how consonant recognition "translates" into the relative IT and SIQ metrics. These measures are used to evaluate five sensory-integration models as discussed below. Finally, we consider how the SyR metric (4) applies to cross-spectral integration of phonetic information.

A. Recognition Scores

Consonant-recognition accuracy is shown in Fig. 1. It is positively correlated with the number of concurrent slits and is roughly equivalent for single slits presented alone (38 to 40% correct), which is well above chance level ($\sim$9%). Although the interpolated, SII-frequency weights [3; Table B.2 NNS] predict lower recognition scores for the 750-Hz slit, it is unclear whether this inconsistency is due to differences in language (the SII was developed originally for English; Danish has a different consonant inventory), bandwidth (1/3-octave versus 3/4-octave) or some other factors.

In the two-slit conditions, consonants are more accurately recognized when the mid- and high-frequency slits are presented. For example, the 1500 Hz + 3000 Hz recognition scores are higher than those associated with the 750 Hz + 3000 Hz condition. These, in turn, are higher than the 750 Hz + 1500 Hz condition. Such a pattern of recognition performance is consistent with the interpolated SII-frequency weights. Consonant recognition is 88% percent correct for the three-slit (750 Hz + 1500 Hz + 3000 Hz) condition. This high, but not-quite-perfect level of performance allows the contribution of each slit to be quantified without confounding "ceiling effects," while maintaining a large dynamic range between the highest and lowest performance level.

Decoding scores were calculated for the features Voicing, Manner and Place, as described in Section II-D. The scores are included in Fig. 1. As for consonant recognition, decoding improves with the number of slits presented concurrently. There are frequency- and phonetic-feature-specific effects for spectral integration. The high-frequency slit improves Place decoding more than Voicing and Manner. In contrast, the low-frequency slit is associated with better decoding of Manner and Voicing. For Voicing, and to a lesser extent, Manner, the best decoding performance is associated with the low-frequency band for single-slit stimuli. These observations imply that consonant-recognition qualitatively resembles the general pattern observed for decoding Place. The correlation between PF decoding and consonant-recognition is illustrated in Fig. 2, where Voicing, Manner and Place decoding are plotted as a function of Consonant recognition.

The correlation between accuracy of Place decoding and Consonant recognition is nearly perfect ($r^2 = 0.99$). This nearly perfect correlation means that Place is rarely decoded correctly when the consonant itself is incorrectly recognized. In contrast, Voicing and Manner are often decoded correctly when the consonant is incorrectly recognized. This pattern is consistent with the lower correlation between Voicing and Manner decoding and Consonant recognition, as shown in Fig. 2. These results are also consistent with [47] in which a perceptual hierarchy for PF-decoding was proposed.

B. IT and Spectral Integration

Confusion matrices were computed for Voicing, Manner, and Place, and the amount of information transmitted calculated from these confusion matrices, as described in Section II-D. The results are shown in Fig. 3.

The specific pattern of information transmitted observed in Fig. 3 shows how phonetic cues are integrated across the

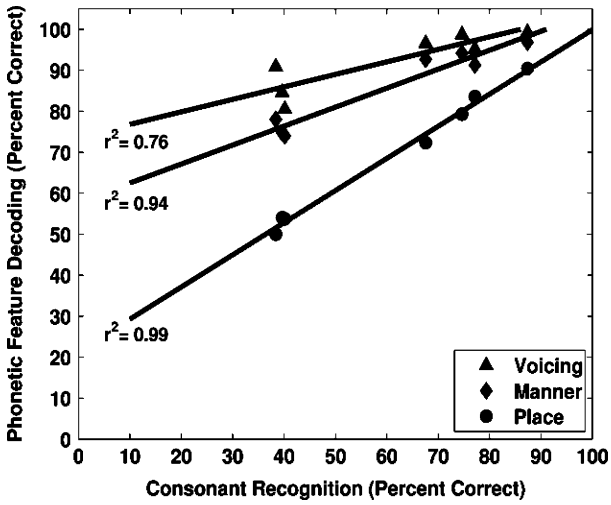

Fig. 2. Voicing, Manner, and Place decoding precision as a function of consonant-recognition accuracy for the same conditions and listeners as in Fig. 1. With each phonetic feature a best-fit linear regression and a correlation coefficient (r^2) are shown.

acoustic-frequency spectrum. For consonants, the IT associated with each slit sums almost linearly. For example, the IT_r associated with the two-slit condition 750 Hz + 1500 Hz is 0.67, close to the sum of the IT_r associated with single slits presented alone (0.61). However, the integration of phonetic information is not quite linear when considered from the perspective of the three slits. The addition of the 3000-Hz slit predicts an IT_r of $\sim$0.98. However, the IT_r observed for the 750 + 1500 + 3000 Hz condition is 0.83. Within the context of the AI and SII, such deviation from linearity would be expressed in terms of frequency-weighting or "importance" functions, where the contribution of a given frequency band is weighted differentially across the acoustic spectrum (e.g., [2], [5]). We discuss this issue in more detail below.

When cross-spectral integration of IT_r is computed for each of the three PFs, an interesting pattern emerges. The cross-spectral integration of IT_r associated with Voicing appears linear for two slits (i.e., the IT_r nearly doubles for the two-slit conditions), but becomes highly compressive when a third slit is added, (i.e., the growth function saturates)—the IT_r increases only slightly over that associated with the two-slit conditions. The cross-spectral integration of Voicing is compressive for more than two concurrently presented slits. Manner behaves similarly to Voicing in that the IT_r for the two-slit conditions is approximately double that of single-slit stimuli. However, adding a third slit increases the IT_r only slightly more than for Voicing. In other words, cross-spectral integration is only slightly less compressive for Manner than it is for Voicing. In contrast, the cross-spectral integration function for Place is expansive rather than compressive. For Place, the two-slit IT_r increases by up to a factor of four relative to the single-slit IT_r. Adding a third slit increases the IT_r by a factor of six to eight relative to the single-slit IT_r. Hence, the pattern of cross-spectral IT_r integration is very different for Place. In multi-sensory perception, this form of integration is known as the "principle of inverse effectiveness," in which performance improvement is greater for individual signals that are poorly decoded when presented alone [48].

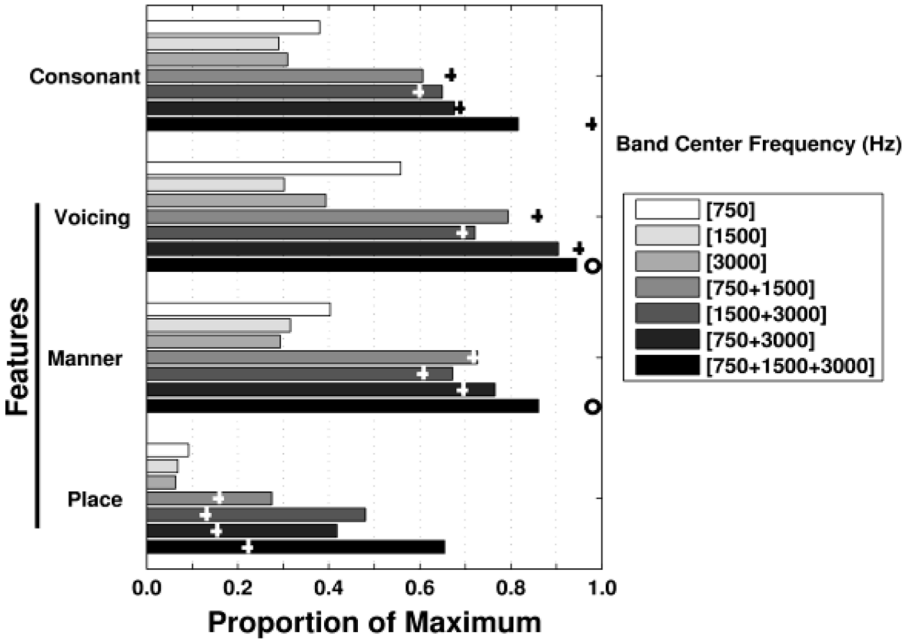

Fig. 3. Relative information transmitted for each stimulus condition across six subjects. The amount of information is calculated from the confusion matrices (see Section II-D for method). The analyses are based on the same confusion matrices as those used for Fig. 1. For multi-slit conditions crosses indicate the sum of relative information transmission from the corresponding single slits. White crosses indicate that the sum is less than the measured relative information transmission; black crosses indicate that the sum is more than the measured relative information transmission. Crosses are replaced by circles where the sums of single slit values fractionally exceed 1.

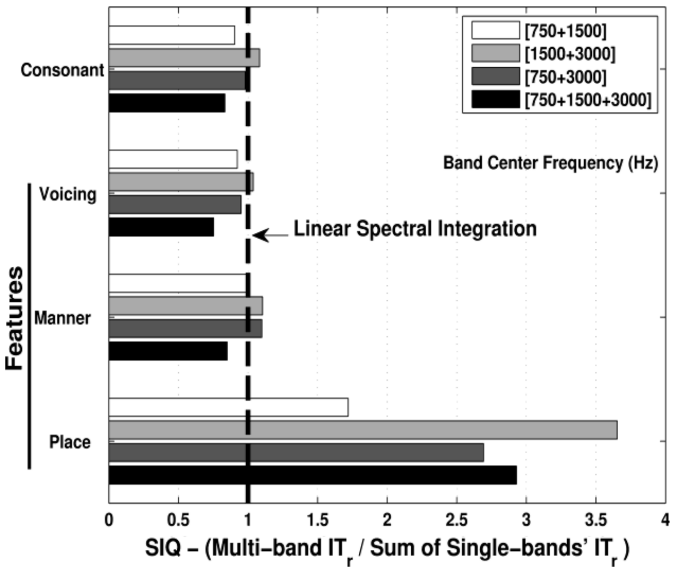

Fig. 4. Spectral integration quotients for the multi-slit conditions. The quotient is defined as the ratio between the observed $\mathrm{IT_r}$ for a given multi-band condition and the sum of $\mathrm{IT_r}$ from the contributing single bands when presented alone. For example, given the observed $\mathrm{IT_r}$ associated with Place for the $1500 + 3000$ Hz two-slit condition is 0.48, and that the observed information transmission for the single bands centered at 1500 and 3000 Hz is 0.07 and 0.06, respectively, the cross-spectral integration quotient $= 0.48/(0.07 + 0.06) = 3.69$.

The cross-spectral integration of phonetic information can further be quantified using the SIQ as shown in Fig. 4 and described in (3). We revisit these differing patterns of cross-spectral $\mathrm{IT_r}$ integration and SIQ in Section IV.

C. Comparison to Model Data

In this section, we analyze how well five stimulus-integration models (Flet1, Flet2, Pre-labeling, Post-labeling, and Fuzzy-logic, as described in [16]), account for the cross-spectral integration data presented above. By doing so, we show that the conventional methods for modeling cross-spectral integration fail to predict the pattern of phonetic-feature decoding errors observed in the current study. We discuss these models' deficiencies below as the first step in proposing a more robust framework for understanding the role played by cross-spectral integration in consonant recognition.

The Flet1 model refers to Harvey Fletcher's "Product of Errors" (PoE) formulation, which predicts that the probability of committing an error in condition AB is equal to the probability of erring in condition A, multiplied by the probability of erring in condition B (i.e., that the errors are independent of each other). In the present context, condition A refers to a condition in which pass-band A (i.e., Slit 1), is presented alone, while condition B refers to a condition where pass-band B (Slit 2) is presented alone, and condition AB refers to the condition where pass-bands A and B are presented concurrently. The Flet2 model is a variant of Flet1, in which error probabilities are determined separately for each consonant rather than being averaged across all consonants. The Flet1 and Flet2 models are described in more detail in [16].

In the Pre-labeling model [17], single-band data are represented in continuously valued form, which is combined prior (hence, "Pre-labeling") to category labels, i.e., consonant identity, being assigned. The statistical properties of the information are inferred from multi-dimensional scaling of single-band confusion matrices [49].

In the Post-labeling integration model [17], it is assumed the listener makes separate judgments, by means of labels for each data stream (i.e., slit), and subsequently combines these to arrive at a single judgment for the multi-stream signal (hence

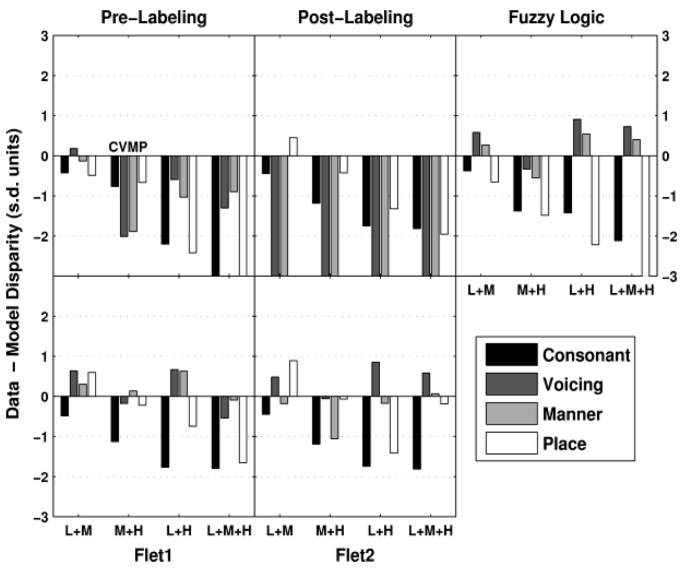

Fig. 5. Comparisons of consonant-recognition and feature-decoding performance predicted by the five integration models discussed in the text (Pre- and Post labeling, Fuzzy logic, Fletcher1 and Fletcher 2). Results are grouped on the x-axis corresponding to the four slit configurations 1) $\mathrm{L}+\mathrm{M}-[750\ \mathrm{Hz}+1500\ \mathrm{Hz}]$, 2) $\mathrm{L}+\mathrm{H}-[750\ \mathrm{Hz}+3000\ \mathrm{Hz}]$, 3) $\mathrm{M}+\mathrm{H}-[1500\ \mathrm{Hz}+3000\ \mathrm{Hz}]$, and 4) $\mathrm{L}+\mathrm{M}+\mathrm{H}-[750\ \mathrm{Hz}+1500\ \mathrm{Hz}+3000\ \mathrm{Hz}]$ and by phonetic feature (see legend). The y-axes show the disparity between the model prediction and the empirical data (expressed in standard-deviation units associated with the empirical data). For example, a datum of -1.2 means that the model under-predicts the empirical data by 1.2 standard deviations. Absolute values greater than 3 are truncated to ± 3.

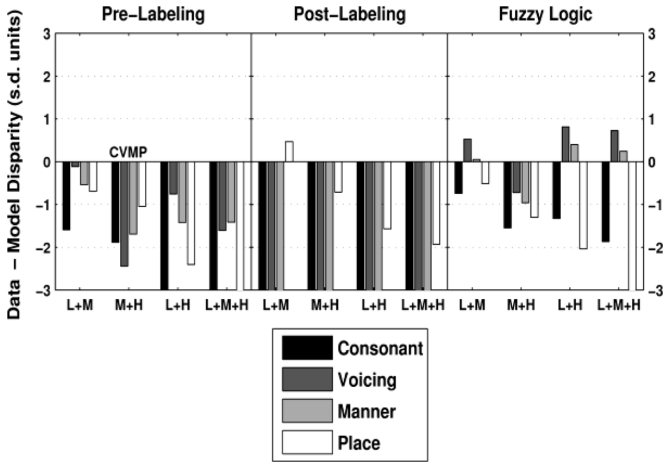

Fig. 6. Comparisons of relative information transmitted predicted by the three of integration models (Pre- and Post labeling, Fuzzy logic). The Flet1 and Flet2 models are excluded because they are generally not used to predict $\mathrm{IT_r}$. Other parameters and designations are the same as in Fig. 5.

"Post-labeling"). For example, in the AB condition, presentation of stimulus S generates a pair of labels (LA, LB) associated with bands A and B, respectively. These labels correspond to the responses that would be given in single-band conditions. In order to arrive at multi-band judgments, labels are combined using maximum likelihood [17].

In the Fuzzy-logic model [18], the response to each stimulus is determined in a probabilistic way by the "feature value" of that stimulus for each of the possible responses. In single-band presentations, the feature value is estimated as a conditional probability. In multi-band conditions, the feature value is assumed to be proportional to the product of the feature values for the corresponding single-band conditions [50].

Fig. 5 shows the consonant-recognition scores predicted for each of the five stimulus-integration models. None of the models predict consonant recognition with great precision (i.e., within one standard deviation of the average measures). The Flet1 and Flet2 models underestimate consonant-recognition performance by slightly less than two standard deviations in Conditions $\mathrm{M}+\mathrm{H}$ and $\mathrm{L}+\mathrm{M}+\mathrm{H}$. The decoding of Voicing is seriously underestimated by both the Pre- and Post-labeling models for most conditions, while the Fuzzy-logic, Flet1, and Flet2 Models predict the integration of Voicing information within one standard deviation of the observed data.

For Manner, the Pre- and Post-labeling models significantly underestimate the consonant recognition scores for the three-slit and most of the two-slit conditions, the exception being Condition $\mathrm{L}+\mathrm{M}$. In contrast, the Fuzzy-logic model generally overestimates the precision of Manner decoding, the exception being, once again, for Condition $\mathrm{L}+\mathrm{M}$. Both the Flet1 and Flet2 models predict Manner decoding based on cross-spectral integration within one standard deviation of the observed performance.

Place decoding based on spectral integration is underestimated for all conditions by the Pre- and-, Post-labeling models, as well as by the Fuzzy-logic model, except for condition $\mathrm{L}+\mathrm{M}$ by the Post-labeling model. In about half the conditions, the deviation is substantial (i.e., more than 1 standard deviation); in these, consonant recognition performance is underestimated. In contrast, both the Flet1 and Flet2 models predict Place decoding within one standard deviation for most conditions, the two exceptions being Flet1 for Condition $\mathrm{L}+\mathrm{M}+\mathrm{H}$ and Flet 2 for Condition $\mathrm{M}+\mathrm{H}$.

Although the Flet1 and Flet2 are the most accurate of the five in estimating consonant-recognition scores across slit conditions, their ability to predict the details of PF decoding and consonant recognition is imperfect. For example, the Flet1 model underestimates consonant-recognition scores in Conditions $\mathrm{M}+\mathrm{H}$ and $\mathrm{L}+\mathrm{M}+\mathrm{H}$ by approximately the same amount in standard deviation units. In Condition M+H, the decoding of Voicing is overestimated, but not for Condition $\mathrm{L}+\mathrm{M}+\mathrm{H}$. Such inconsistencies suggest that the specific relation (reflected in the correlation) between consonant recognition and PF decoding is inadequately modeled.

Fig. 6 shows the predictions of $\mathrm{IT_r}$ made by the Pre-labeling-, Post-labeling- and Fuzzy-logic models. Flet1 and Flet2 models only predict correct identification, (i.e., elements in the confusion-matrix diagonal). Calculating the amount of information transmitted would require predictions of the off-diagonal elements. Consonant and PF $\mathrm{IT_r}$ are underestimated by the Pre- and Post-labeling models for virtually all stimuli in Condition L+M. This underestimation of $\mathrm{IT_r}$ is more than one standard deviation for consonants. The Fuzzy-logic model is better at predicting Consonant recognition. However, it underestimates $\mathrm{IT_r}$ by more than one standard deviation for Conditions $\mathrm{L}+\mathrm{H}, \mathrm{M}+\mathrm{H}$ and $\mathrm{L}+\mathrm{M}+\mathrm{H}$.

The models' ability to predict consonant-recognition performance is generally mediocre (Fig. 5), as discussed above. Their ability to predict *specific patterns of consonant confusion* is even

worse. This is reflected in the systematic under-prediction of PF decoding (Figs. 5 and 6). None of the models examined predict the *specific* error patterns well, either quantitatively or even qualitatively, which underlines the fact that a principled understanding of how speech (consonants) is decoded is currently lacking. This is where the current study may contribute to better understanding the mechanisms underlying consonant recognition.

D. Measuring Redundancy

Table III shows a SyR matrix associated with consonant recognition and PF decoding for single- and two-slit signals. SyR is a measure of response similarity formulated in terms of consonant-confusion patterns. The more similar the slit-response patterns, the greater the redundancy between them. In a sense SyR is for IT what PoE is for recognition accuracy; only PoE presumes band-independence whereas SyR does not.

For example, the Consonant SyR of single-slit signals (shaded cells) is approximately 0.40. SyR of this magnitude is associated with an intermediate degree of overlap in the response patterns; they are not randomly distributed among the 11 consonants. Were they, the SyR would be ca. 0.02 (as computed by a Monte-Carlo simulation) even if consonant-recognition accuracy was comparable across slits (ca. 40%). This is one reason why recognition accuracy alone cannot provide the sort of performance details required to understand how listeners recognize consonants.

When SyR is computed for the individual PFs, an interesting pattern emerges. The SyR computed for Voicing and Manner across *single-slit* stimuli is somewhat lower than for Consonants (0.21 to 0.30), but roughly comparable to each other. In contrast, the SyR for single-slit Place information is much lower (0.035 to 0.055)—so low that it approaches orthogonality. This means there is virtually no overlap in the Place errors across the spectral bands. These are precisely the sort of response patterns deemed optimal for combining in machine learning [51] and account, at least partially, for why Place information is integrated across the frequency spectrum so differently than Manner and Voicing. Moreover, a very low SyR implies virtual independence of information in each band, a condition consistent with the product of errors formulation of the Flet1 and Flet2 models. The *low* SyR associated with Place is also consistent with this feature's *high* SIQ. Adding sources of quasi-orthogonal data streams results in a disproportionate gain in information. We return to these key points in Section IV.

SyR can also be computed for multi-slit conditions. Calculating SyR for conditions where two slits have the same center frequency, may appear counter-intuitive; SyR simply computes the amount of response overlap between conditions irrespective of their spectral content and is hence "blind" to overlap in frequency content.

There are three cases that do not contain slits with identical center frequencies—(1) 750 Hz+1500 Hz versus 3000 Hz, (2) 750 Hz+3000 Hz versus 1500 Hz), and (3) 1500 Hz+3000 Hz versus 750 Hz. For Consonants, the SyR associated with these multi-slit conditions is only slightly higher (0.45 to 0.52) than for the single-slit conditions (0.39–0.42). In contrast, the SyR for the corresponding multi-slit conditions associated with PFs is higher than the single-slit conditions (Voicing: 0.33 to 0.55 versus 0.24 to 0.30; Manner: 0.33 to 0.42 versus 0.21 to 0.25; Place: 0.07 to 0.11 versus 0.04 to 0.06). Such comparisons imply that the amount of information gained by adding a third slit is rather less than adding a second slit to what was originally a single-slit signal; this is the case for both PF decoding and consonant recognition. Such leveling off of information growth is reflected in the SyR for non-overlapping, multi-slit conditions.

When SyR is computed for two conditions with slits possessing identical center frequencies, there is a general tendency for the associated indices to increase substantially relative to other conditions. This is hardly surprising. However, there is

TABLE III
SyR

Slit Center Frequency in Hz	**Consonants**	750	1500	3000	750 1500	750 3000	1500 3000
	750	1	0.408	0.417	0.527	0.522	0.517
	1500		1	0.389	0.492	0.486	0.498
	3000			1	0.447	0.491	0.483
	750 +1500				1	0.613	0.583
	750 +3000					1	0.681
	1500 +3000						1

Slit Center Frequency in Hz	**Voicing**	750	1500	3000	750 1500	750 3000	1500 3000
	750	1	0.235	0.303	0.558	0.558	0.546
	1500		1	0.263	0.363	0.331	0.333
	3000			1	0.425	0.469	0.459
	750 +1500				1	0.768	0.664
	750 +3000					1	0.754
	1500 +3000						1

Slit Center Frequency in Hz	**Manner**	750	1500	3000	750 1500	750 3000	1500 3000
	750	1	0.248	0.214	0.415	0.414	0.422
	1500		1	0.256	0.369	0.373	0.386
	3000			1	0.334	0.362	0.361
	750 +1500				1	0.683	0.613
	750 +3000					1	0.630
	1500 +3000						1

Slit Center Frequency in Hz	**Place**	750	1500	3000	750 1500	750 3000	1500 3000
	750	1	0.056	0.036	0.137	0.131	0.108
	1500		1	0.035	0.120	0.093	0.103
	3000			1	0.068	0.110	0.111
	750 +1500				1	0.225	0.249
	750 +3000					1	0.415
	1500 +3000						1

The SyR for the one- and two-slit combinations for consonants and PFs. The three-slit data are omitted for illustrative clarity. Because the table is symmetric along the diagonal, only the upper portion of the Syr matrix is shown. SyR computed for conditions without identical slits are shaded. A light shading indicates single-slit conditions; darker shading indicates multi-slit conditions.

one pattern of particular interest—it is the SyR associated with the 750 Hz + 3000 Hz versus 1500 Hz + 3000 Hz signals, which is generally the highest (or almost so in the case of Voicing and Manner) of any SyR computed. This result implies that the 3000-Hz band may be slightly more redundant than the lower-frequency slits. Analogously, for the 750 Hz + 1500 Hz versus 750 + 1500 Hz conditions, the SyR is high for Voicing and Manner, suggesting that the lowest-frequency channel may carry somewhat more redundant information for these features. The 1500-Hz slit appears to be the least redundant channel except for Place, where the SyR is marginally higher than for the 750-Hz slit.

In summary, the SyR in Table III are consistent with MN55's finding that the amount of Place IT_r increases as the bandwidth of the masking noise decreases. This decorrelation of phonetic information across the frequency spectrum is precisely what one would expect for a feature (i.e., Place) whose decoding depends on cues broadly distributed across the frequency spectrum and where no single region contains sufficient information for highly accurate feature decoding to distinguish among consonants.

IV. Discussion

MN55 is one of the most frequently cited papers in speech research. Its merits rest largely on elegantly melding two perspectives—distinctive-feature theory [25] and information theory [52]—into a single framework for understanding how human listeners distinguish speech sounds. By focusing on consonants, MN55 cut to the core of what is important for decoding spoken language, given their central role in recognizing words and higher-level linguistic units. MN55 made two specific contributions: 1) it was the first formal application of information theory to the study of speech perception; and 2) it showed that Place information is considerably more vulnerable to background noise than other PFs. However, the issue of precisely how information is distributed and integrated across the spectrum was left largely unexplored given the difficulty of doing so using only high- and low-pass filtered signals.

By using narrowband, filtered speech, it is possible to ascertain how band-limited phonetic information is combined in a way that is rather independent of its spectral distribution. This is because the relative *gain* in information transmitted is measured irrespective of its magnitude in any single band. The SIQs associated with Place decoding are much higher than those associated with Manner and Voicing, implying that Place information combines more effectively across channels than Voicing and Manner—this despite, and potentially because of, the fact that the amount of Voicing and Manner information in single slits is much higher than Place. Such disparities illustrate the utility of distinguishing between the *spectral distribution* and *perceptual integration* of phonetic information.

For Place, perceptually combining information from separate frequency channels results in a dramatic gain in decoding, much greater than would be predicted from just the distribution of information in the individual channels (cf. Fig. 4). This synergistic integration may be why background noise has such a devastating effect on Place decoding; it interferes with the listener's ability to combine information across spectral channels. Voicing and Manner information is distributed quite differently. Voicing can be decoded accurately from many different parts of the spectrum, and combining such information from two regions results in a relatively linear gain in decoding. Adding a third slit produces little, if any, improvement. A similar integration pattern is observed for Manner.

One important clue to understanding this differential pattern of integration is provided by the SyR metric. Physically similar subsets of a pattern, when combined, offer little gain in recognition compared to highly dissimilar subsets—we refer to this as the Principle of Complementarity (PoC). SyR is essentially a behavioral measure of pattern similarity based on consonant-confusion patterns. The SyR associated with Voicing and Manner is comparatively high; there is relatively small potential for decoding benefit from combining patterns or fragments. The sub-patterns would need to differ far more for their combination to produce a significant gain in decoding. PoC is used in machine-learning to optimize recognition performance by selecting only those data streams whose information complements others (e.g., [51]). In the current study, Place is associated with the most complementary spectral fragments. Its SyR is so low (ca. 0.05) that information in single slits is virtually independent of others. For this reason, the potential for information gain is high when such fragments are combined. However, combining quasi-orthogonal fragments is fragile. Disrupting any significant part of the pattern jeopardizes the recognition of the whole given the interconnected nature of the features and consonant recognition. In everyday speech communication, high-level, semantic, and contextual factors, as well as visual cues, may compensate for this decoding fragility.

The SIQ and SyR are related metrics. Although they measure different properties of feature-decoding performance, they are highly inversely correlated. A low degree of SyR is generally associated with a high SIQ, and vice versa. In our view, information redundancy/complementarity and cross-spectral integration are crucial for robust speech recognition by human listeners. The principle of inverse effectiveness, used to model cross-modal sensory integration (e.g., [48]), yields similar results to the PoC in many circumstances.

MN55 has been criticized in three principal ways. First, their application of PFs is controversial. Apart from articulatory-acoustic-phonetic studies of the speech signal (e.g., [25]), there is little independent evidence for their objective, (i.e., neurological) existence. Despite this, the PF-based framework has been productive as a working hypothesis for speech research over the intervening period (e.g., [41]–[43] and [53], [54]). Another criticism is that MN55 used an unconventional feature representation. Particularly controversial is how they dealt with Manner. Because MN55 were following the binary classification scheme of [25], Manner was divided into the binary features of Nasality and Affrication (where the class "Stop" is designated by a nil value for both explicitly marked classes). Since the 1970s, Manner has generally been represented as an n-ary feature, where n ranges between 3 and 5 [55]. This change has simplified the modeling of consonant features, which are generally reduced to three—Voicing (binary), Manner (3 to 5 classes) and Place (usually 3 classes). However,

the use of an obsolete feature structure does not invalidate MN55's analyses, it only requires additional interpretation. A third criticism is that the features are assumed to be statistically independent. In fact, MN55's own data show that PFs are not entirely independent of each other. MN55 were aware of this but argued that their general conclusions—Place is far more vulnerable to background interference and its information more broadly distributed across the frequency spectrum—do not depend on absolute feature independence, a conclusion with which we concur.

Some (e.g., Allen and associates [28], [29]) have argued that using *a priori* perceptual features, based on labels originating from speech production, is fundamentally wrong. In their view, perceptual features should be derived only through a "data-driven" approach in which patterns of consonant confusions are used to determine the identity of the underlying features. Although appealing in principle, this approach is subject to its own set of limitations. In the case of [29], the consonant-confusion patterns reflect the interaction between the spectral properties of the masking noise and the consonants (i.e., energetic masking) rather than some more fundamental, central, linguistic representation that generalizes to a large variety of contexts and applications. As shown in the current study, PFs do provide a means of identifying some important characteristics of the spectral integration process for consonants. This, in turn, may be important for understanding the fundamental linguistic representation of speech. These observations do not depend on the precise perceptual status of PFs nor whether they are entirely independent of each other. Rather, PFs can be used to specify a formal structure that is effective in distinguishing among consonants.

Although the current study does not allow us to specify precisely how phonetic-feature information is combined across the frequency spectrum, it shows that the spectral integration process plays an important role for recognition of linguistic entities. In the current study, these entities have been consonants, but the framework could be applied to other levels of linguistic analysis utilizing different representational units than phonetic features.

V. Summary and Conclusion

In this study, an information-theoretic based analysis was used to characterize the spectral integration of phonetic information in Danish consonants.

Two information-theoretic metrics were used. One, the SIQ, quantifies the way in which information associated with consonants and distinctive PFs combines across the acoustic-frequency spectrum. The second, SyR, quantifies response similarity across separate spectral regions. These two measures are closely related, reflecting the potential gain in pattern decoding/recognition when two or more data streams are combined.

Consonant decomposition into PFs provides a principled and structured representation of consonants for analyzing patterns of perceptual confusions essential for quantifying how well the decoding and recognition processes are performed in and across different spectral regions. From such analyses we conclude that Voicing and Manner are decoded very differently from Place. Manner and Voicing error patterns are largely similar, and their SyR indices are relatively high (> 0.40), indicating that Voicing and Manner information is rather redundantly *distributed* across the frequency spectrum. This may be why these features are relatively resilient to background noise and acoustic interference—*comparable* phonetic information can be decoded from *different* regions of the spectrum. For Voicing and Manner, distortion/interference has relatively little impact on decoding. In contrast, Place confusion patterns differ significantly across spectral regions; their SyR indices are low ($\sim$)0.05, which means there is virtually no redundancy in Place information distributed across frequency channels. Such low cross-spectral redundancy is why combining information from different spectral regions improves feature decoding and by implication, consonant recognition. The consequence of this highly distributed encoding is a potential vulnerability of feature decoding and pattern recognition dependent on such decoding.

The way in which information combines across the spectrum differs among PFs. For Manner and Voicing, the integration is approximately linear when two slits are combined. The amount of information contained in two slits is roughly double that of one. In contrast, Place information combines across frequency channels very differently. Two slits often provide four or more times the information contained in one. Information integration is expansive and continues to be so when a third slit is added. In contrast, integration of Voicing and Manner information often exhibits a compressive function with the addition of a third slit—there is little or no information gain. However, Consonant recognition depends on correctly decoding all three PFs. When the information-integration functions of the features are combined for the two- and three-slit conditions, the composite integration functions appear to be approximately linear, similar to the IT growth function associated with consonant recognition. In this sense, consonant recognition may only appear to be linear. In actuality, consonant recognition may be the composite of three separate cross-spectral-integration functions none of which is strictly linear over its full dynamic range.

One interpretation of the data and analyses reported in this study is that the PFs obscure and consequently complicate interpretation of the perceptual data. In this perspective, the results could serve as an argument against using PFs as a means of understanding the speech decoding process.

In our view, this argument is contradicted by the fact that the confusion patterns observed are clearly accounted for in terms of PFs, particularly place of articulation. Without some principled structural framework with which to relate consonants to each other the error patterns seem arbitrary and without a coherent perceptual basis.

We believe that the SIQ and SyR analyses call into question some of the assumptions underlying the AI, SII, and STI, and imply that these theoretical perspectives are not entirely in accord with how human listeners decode and recognize consonants. By using PFs as well as the SIQ and SyR metrics, it may be possible to more accurately model human speech processing and use this knowledge for improving a variety of technologies, including automatic speech recognition and speech enhancement.

APPENDIX

A. Appendix

The following example illustrates how information transmitted is calculated based on the confusion matrix given in Table II(b) (voicing). The derived probabilities of the four cells in the confusion matrices are $p_{11} = 201/396 = 0.5076$, $p_{12} = 15/396 = 0.03788$, $p_{21} = 62/396 = 0.1566$, $p_{22} = 118/396 = 0.2980$ respectively. Applying (1) the number of bits transmitted is: $p_{11}\log_2((p_{11}+p_{12})(p_{11}+p_{21})/p_{11}) - p_{12}\log_2((p_{12}+p_{11})(p_{12}+p_{22})/p_{12}) - p_{21}\log_2((p_{21}+p_{22})(p_{11}+p_{21})/p_{21}) - p_{22}\log_2((p_{21}+p_{22})(p_{12}+p_{22})/p_{22})) = 0.2999$.

ACKNOWLEDGMENT

The authors would like to thank Prof. T. Dau and the four anonymous reviewers for extremely helpful comments and suggestions on earlier versions of this paper. The authors would also like to thank Prof. L. Atlas for advice concerning the use of the "Modulation Toolbox" developed in his laboratory at the University of Washington, Seattle, Prof. T. Høholt, of the Technical University of Denmark, for valuable discussion concerning the information-theoretic analyses used in our study, and Prof. L. Braida of MIT for providing predictions for the pre-labeling and post-labeling models.

REFERENCES

[1] N. R. French and J. C. Steinberg, "Factors governing the intelligibility of speech sounds," *J. Acoust. Soc. Amer.*, vol. 19, pp. 90–119, 1947.
[2] *Methods for the Calculation of the Articulation Index*, ANSI Standard S3.5-1969.
[3] *Methods for the Calculation of the Speech Intelligibility Index*, ANSI Standard S3.5-1997.
[4] H. J. Steeneken and T. Houtgast, "A physical method for measuring speech-transmission quality," *J. Acoust. Soc. Amer.*, vol. 67, pp. 318–326, 1980.
[5] H. Fletcher, "Speech and Hearing in Communication," in *The ASA Edition*, J. B. Allen, Ed. New York: Acoustical Society of America, 1995, pp. 278–302, 318–414.
[6] "IEC 60268-16: Sound System Equipment—Part 16," in Objective Rating of Speech Intelligibility by Speech Transmission Index, 3rd ed. International Electrotechnical Commission, Geneva, Switzerland, 2003.
[7] K. D. Kryter, "Speech bandwidth compression through spectrum selection," *J. Acoust. Soc. Amer.*, vol. 32, pp. 547–556, 1960.
[8] R. P. Lippmann, "Accurate consonant perception without mid-frequency speech energy," *IEEE Trans. Speech Audio Process.*, vol. 4, no. 1, pp. 66–69, Jan. 1996.
[9] R. M. Warren, K. R. Riener, J. A. Bashford, Jr., and B. S. Brubaker, "Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits," *Percept. Psychophys.*, vol. 57, pp. 175–182, 1995.
[10] S. Greenberg, T. Arai, and R. Silipo, "Speech intelligibility derived from exceedingly sparse spectral information," in *Proc. 5th Int. Conf. Spoken Lang. Process.*, 1998, pp. 74–77.
[11] H. Müsch and S. Buus, "Using statistical decision theory to predict speech intelligibility. II. Measurement and prediction of consonant-discrimination performance," *J. Acoust. Soc. Amer.*, vol. 109, pp. 2910–2920, 2001.
[12] S. Greenberg and T. Arai, "What are the essential cues for understanding spoken language?," *IEICE Trans. Inf. Syst.*, vol. E87, pp. 1059–1070, 2004.
[13] F. Li and J. B. Allen, "Additivity law of frequency integration for consonant identification in white noise," *J. Acoust. Soc. Amer.*, vol. 126, pp. 347–353, 2009.
[14] H. J. M. Steeneken and T. Houtgast, "Mutual dependence of the octave-band weights in predicting speech intelligibility," *Speech Commun.*, vol. 28, pp. 109–123, 1999.
[15] L. A. Chistovich and V. V. Lublinskya, "The 'center of gravity' effect in vowel spectra and critical distance between formants: Psychoacoustical study of the perception of vowel-like stimuli," *Hear. Res.*, vol. 3, pp. 185–195, 1979.
[16] D. Ronan, A. K. Dix, P. Shah, and L. D. Braida, "Integration across frequency bands for consonant identification," *J. Acoust. Soc. Amer.*, vol. 116, pp. 1749–1762, 2004.
[17] L. D. Braida, "Crossmodal integration in the identification of consonant segments," *Q. J. Exp. Psychol.*, vol. 43, pp. 647–677, 1991.
[18] D. W. Massaro, *Speech Perception by Ear and Eye: A Paradigm for Psychological Inquiry*. Hinsdale, NJ: Lawrence Erlbaum, 1987, pp. 61–66, 113–127, 209–217, 255–268.
[19] K. W. Grant and L. D. Braida, "Evaluating the articulation index for auditory-visual input," *J. Acoust. Soc. Amer.*, vol. 89, pp. 2952–2960, 1991.
[20] I. Pollack, "Effect of high pass and low pass filtering on the intelligibility of speech in noise," *J. Acoust. Soc. Amer.*, vol. 20, pp. 259–266, 1948.
[21] T. S. Bell, D. D. Dirks, and T. D. Trine, "Frequency-importance functions for words in high- and low-context sentences," *J. Speech Hear Res.*, vol. 35, no. 4, pp. 950–959, 1992.
[22] G. A. Studebaker, C. Gilmore, and R. Sherbecoe, "Performance-intensity functions at absolute and masked thresholds," *J. Acoust. Soc. Amer.*, vol. 93, pp. 3418–3421, 1993.
[23] C. V. Pavlovic, "Band importance functions for audiological applications," *Ear Hear.*, vol. 15, pp. 100–104, 1994.
[24] G. A. Miller and P. Nicely, "An analysis of perceptual confusions among some English consonants," *J. Acoust. Soc. Amer.*, vol. 27, pp. 338–352, 1955.
[25] R. Jakobson, G. Fant, and M. Halle, *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press, 1963, [Originally published in 1952 as a research monograph by the MIT Research Laboratory of Electronics].
[26] *Loudness*, M. Florentine, A. N. Popper, and R. R. Fay, Eds. New York: Springer, 2010, ch. 10.
[27] E. M. Relkin and J. R. Doucet, "Is loudness simply proportional to the auditory nerve spike count?," *J. Acoust. Soc. Amer.*, vol. 101, pp. 2735–2740, 1997.
[28] J. B. Allen, "Consonant recognition and the articulation index," *J. Acoust. Soc. Amer.*, vol. 117, pp. 2212–2223, 2005.
[29] S. A. Phatak and J. B. Allen, "Consonant and vowel confusions in speech-weighted noise," *J. Acoust. Soc. Amer.*, vol. 121, pp. 2312–2326, 2007.
[30] T. B. Schalk and M. B. Sachs, "Nonlinearities in auditory-nerve fiber responses to bandlimited noise," *J. Acoust Soc. Amer.*, vol. 67, pp. 903–913, 1980.
[31] J. A. Costalupes, E. D. Young, and D. J. Gibson, "Effects of continuous noise backgrounds on rate response of auditory nerve fibers in cat," *J. Neurophysiol.*, vol. 51, no. 6, pp. 1326–1344, 1984.
[32] W. S. Rhode and S. R. Greenberg, "Lateral suppression and inhibition in the cochlear nucleus of the cat," *J. Neurophysiol.*, vol. 71, pp. 493–519, 1994.
[33] M. C. Liberman and M. C. Brown, "Physiology and anatomy of single olivocochlear neurons in the cat," *Hear. Res.*, vol. 24, pp. 17–36, 1986.
[34] J. A. Bashford, R. M. Warren, and P. W. Lenz, "Enhancing intelligibility of narrowband speech with out-of-band noise: Evidence for lateral suppression at high-normal intensity," *J. Acoust. Soc. Amer.*, vol. 117, pp. 365–369, 2005.
[35] A. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1991.
[36] *Speech Separation by Humans and Machines*, P. Divenyi, Ed. New York: Springer, 2004, pp. 1–53, 135–154, 221–294, 305–314.
[37] D. Wang and G. Brown, *Computational Auditory Scene Analysis*. New York: Wiley, 2006.
[38] S. E. Blumstein and K. N. Stevens, "Perceptual invariance and onset spectra for stop consonants in different vowel environments," *J. Acoust. Soc. Amer.*, vol. 67, pp. 648–662, 1980.
[39] D. Kewley-Port, D. B. Pisoni, and M. Studdert-Kennedy, "Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants," *J. Acoust. Soc. Amer.*, vol. 73, pp. 1779–1793, 1983.
[40] H. M. Sussman, H. A. McCaffrey, and S. A. Matthews, "An investigation of locus equations as a source of relational invariance for stop consonant place categorization," *J. Acoust. Soc. Amer.*, vol. 90, pp. 1309–1325, 1991.
[41] J. C. Ziegler, C. Pech-Georgel, F. George, F. X. Alario, and C. Lorenzi, "Deficits in speech perception predict language learning impairment," in *Proc. Nat. Acad. Sci. USA*, 2005, vol. 102, pp. 14110–14115.

[42] C. Füllgrabe, F. Berthommier, and C. Lorenzi, "Masking release for consonant features in temporally fluctuating background noise," *Hear. Res.*, vol. 211, pp. 74–84, 2006.
[43] A. Lovitt and J. B. Allen, "50 years late: Repeating Miller-Nicely 1955," in *Proc. Interspeech' 06 and 9th Int. Conf. Spoken Lang. Process.*, 2006, vol. 1–5, pp. 2154–2157.
[44] D. Chan, A. Fourcin, D. Gibbon, B. Granström, M. Huckvale, G. Kokkinakis, K. Kvale, L. Lamel, B. Lindberg, A. Moreno, J. Mouropoulos, F. Senia, I. Trancoso, C. Veld, and J. Zeiliger, "EUROM—A spoken language resource for the EU," in *Proc. 6th Eur. Conf. Speech Commun. Tech (Eurospeech'95)*, 1995, pp. 867–870.
[45] N. Grønnum, "Illustrations of the IPA: Danish," *J. Int. Phon. Assoc.*, vol. 28, pp. 99–105, 1998.
[46] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques*. San Mateo, CA: Morgan Kaufmann, 2005, pp. 290–292.
[47] S. Greenberg and T. U. Christiansen, "Linguistic scene analysis and the importance of synergy," in *Proc. 1st Int. Symp. Auditory Audiol. Res.*, Elsinore, Denmark, 2008, pp. 351–364.
[48] N. P. Holmes, "The law of inverse effectiveness in neurons and behavior: Multisensory integration versus normal variability," *Neuropsychologia*, vol. 45, pp. 3340–3345, 2007.
[49] N. A. Macmillan, R. F. Goldberg, and L. D. Braida, "Resolution for speech sounds—Basic sensitivity and context memory on vowel and consonant continua," *J. Acoust. Soc., Amer.*, vol. 84, pp. 1262–1280, 1988.
[50] M. M. Cohen and D. W. Massaro, "Perceiving visual and auditory information in consonant-vowel and vowel syllables," in *Levels in Speech Communication: Relations and Interactions*, C. Sorin, J. Mariani, H. Meloni, and J. Schoentgen, Eds. Amsterdam, The Netherlands: Elsevier, 1995, pp. 25–37.
[51] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.
[52] C. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379–423, 623–656, 1948, pp. 623–656.
[53] K. W. Grant, B. E. Walden, and P. F. Seitz, "Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration," *J. Acoust. Soc., Amer.*, vol. 103, pp. 2677–2690, 1998.
[54] R. Shannon, F. G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, 1995.
[55] P. Ladefoged, *Preliminaries to Linguistic Phonetics*. Chicago, IL: Univ. of Chicago Press, 1971.

Thomas U. Christiansen received the M.Sc. degree in computer science as a major and linguistics as minor from the University of Copenhagen, Copenhagen, Denmark, in 1999 and the Ph.D. degree in electrical engineering from the Technical University of Denmark, Lyngby, in 2004.

From 2004 to 2009, he was a Post-doc and an Assistant Professor with the Centre for Applied Hearing Research at the Technical University of Denmark. His major topics of interests include models of the normal and impaired auditory periphery and its relation to phonetics and speech perception. Starting in 2010, he has been currently employed as an Associate Professor with the Centre for Applied Hearing Research, Technical University of Denmark, working with speech perception, linguistic-based speech enhancement, and phonetics.

Dr. Christiansen is a board member of the Danish Acoustical Society, and is a chair of the committee for psychoacoustics for the Danish Acoustical Society.

Steven Greenberg (M'02–SM'07) received the A.B. degree in linguistics from the University of Pennsylvania, Philadelphia, in 1974 and the Ph.D. degree in linguistics from the University of California, Los Angeles, in 1980. His doctoral research on temporal coding of pitch and vowel quality was performed in the Brain Research Institute within the School of Medicine.

From 1980 to 1982, he was a Postdoctoral Fellow in the Communicative Disorders Department, Northwestern University, working on the psychoacoustics of complex sounds. From 1982 to 1983, he was a Postdoctoral Fellow in the Department of Neurophysiology, University of Wisconsin Medical School, Madison, researching the auditory physiology associated with speech and music processing. From 1983 to 1987, he was Research Assistant Professor and from 1987 to 1991 served as Research Associate Professor within that same department. From 1991 to 1995, he was an Associate Professor in the Department of Linguistics, University of California, Berkeley. From 1995 to 2002, he was a Senior Scientist and Faculty Affiliate at the International Computer Science Institute, Berkeley. Since 2002, he has been with Silicon Speech, a speech technology and research company based in Kelseyville, CA. He has edited several books on auditory speech processing including *Listening to Speech: An Auditory Perspective* (Lawrence Erlbaum, 2006), *Speech Processing in the Auditory System* (Springer, 2004), *Computational Models of Auditory Function* (Ios Pr, 20010 , and *Dynamics of Speech Production and Perception* (IOS, 2006). He has served on the editorial board of *Speech Communication* (1996–2003) and the *Journal of Audio, Speech, and Music Processing* (2006-present).