



The perceptual flow of phonetic information

Steven Greenberg¹ · Thomas U. Christiansen²

Published online: 31 January 2019
© The Psychonomic Society, Inc. 2019

Abstract

Over a long and distinguished career, Randy Diehl has elucidated the brain mechanisms underlying spoken language processing. The present study touches on two of Randy's central interests, phonetic features and Bayesian statistics. How does the brain go from sound to meaning? Traditional approaches to the study of speech intelligibility and word recognition are unlikely to provide a definitive answer. A finer-grained, Bayesian-inspired approach may help. In this study, listeners identified 11 Danish consonants spoken in a Consonant + Vowel + [l] environment. Each syllable was filtered so that only a portion of the original audio spectrum was presented. Three-quarter-octave bands of speech, centered at 750, 1,500, and 3,000 Hz, were presented individually and in combination. The conditional, posterior probabilities associated with decoding the phonetic-features Voicing, Manner, and Place of Articulation were computed from confusion matrices to delineate the perceptual flow of phonetic information processing. Analysis of the conditional probabilities associated with both correct and incorrect feature decoding suggest that Manner of articulation is linked to the decoding of Voicing (but not vice-versa), and that decoding of Place of articulation is associated with decoding of Manner of articulation (but not the converse). Such feature-decoding asymmetries may reflect processing strategies in which the decoding of lower-level features, such as Voicing and Manner, is leveraged to enhance the recognition of more complex linguistic elements (e.g., phonetic segments, syllables, and words), especially in adverse listening conditions. Such asymmetric feature decoding patterns are consistent with a hierarchical, perceptual flow model of phonetic processing.

Keywords Speech perception · Phonology · Bayesian modeling

Introduction

This study examines the perception of speech from two perspectives that have benefitted greatly from Randy Diehl's research over a long and distinguished career – articulatory-acoustic features and Bayesian statistics (e.g., Diehl & Lindblom, 2004; Redford & Diehl, 1999).

Consonants convey information critical for lexical identification (Bonatti, Peña, Nespor, & Mehler, 2005). While vowels often undergo significant changes in spectrum and duration without compromising intelligibility, this is rarely the case for

consonants (but see Kewley-Port, Burke, & Lee, 2007 and Lee and Kewley-Port, 2009 for an alternative perspective). When the acoustic properties of consonants are degraded (e.g., reverberation, background noise) or the listener's internal representation compromised (due to auditory pathology or cognitive impairment), speech comprehension may suffer.

The importance of consonants lies partly in their syntagmatic role in lexical representation, whether auditory or visual. They often occupy the skirts of syllables in either onset or coda position. Consonants are especially important at the onset of accented (i.e., linguistically stressed) syllables (Greenberg, Carvey, & Hitchcock, 2002). One likely reason is that sensory neurons respond most vigorously to the beginning of meaningful events, such as syllables or words (Greenberg & Ainsworth, 2004). If the onset excitation pattern strays too far from the norm, it may be ineffective in linking the sensory input to the brain's internal linguistic representation. In this sense, consonants serve as strategic linchpins for speech communication.

✉ Steven Greenberg
steven@siliconspeech.com

¹ Silicon Speech, Hidden Valley Lake, CA 95467, USA

² Oticon, Kongebakken 9, DK-2765 Smørum, Denmark

Despite their perceptual importance, comparatively little is known about the auditory or cognitive mechanisms that underlie consonant recognition. This study focuses on an intriguing property of consonant-confusion patterns, namely their asymmetry when analyzed as articulatory-acoustic phonetic features (APFs). This asymmetry offers clues as to how consonant segments are analyzed in the auditory system and beyond.

One of the most detailed and well known quantitative models of consonant recognition is the Articulation Index (AI), developed by Fletcher and others (ANSI, 1969; Allen, 2005; French & Steinberg, 1949; Fletcher, 1953 [1995]; Kryter, 1962). The AI is, at heart, a pattern-recognition model of auditory spectral processing. The speech signal's auditory representation is represented as a tonotopically organized array of frequency channels. Each channel's contribution depends (mostly) on two parameters – its signal-to-noise ratio (SNR) and its “frequency importance” (Bell, Dirks, & Trine, 1992; French & Steinberg, 1949; Fletcher, 1953 [1995]; Pavlovic, 1994; Pavlovic, 2006). The latter is a weighting function based on perceptual tests. The AI assumes speech recognition is mainly an SNR problem (to solve). At very low SNRs, the listener's ability to identify syllables and words is seriously compromised. The AI is essentially a distortion metric that quantifies the distance between a reference and the signal received by the listener. Any distortion (e.g., noise, reverberation, hearing impairment) that increases the distance between the two is likely to reduce the articulation score and intelligibility.

A key problem with the AI and kindred distortion metrics, such as the Speech Intelligibility Index (SII) (ANSI, 1997; Pavlovic, 2006), the Speech Transmission Index (STI) (Steeneken & Houtgast, 1980), and the Spectro-Temporal Modulation Index (STMI) (Elhilali, Chi, & Shamma, 2003), is the absence of a *principled* mechanism for how the brain goes from sound to meaning. Instead, these distortion-centric formulations assume speech comprehension is the end-product of a process by which a neural excitation pattern is linked to meaningful elements (e.g., phonemes, syllables, and words) via some form of template matching, the specifics of which are largely left to the reader's imagination.

Another drawback of distortion models is their omission of non-auditory information sources. Yet, it is well documented that visual speech cues can enhance speech intelligibility, especially in challenging listening conditions (Braidá, 1991; Cohen & Massaro, 1995; Grant & Braidá, 1991; Grant, Walden, & Seitz, 1998; Massaro, 1987). Linguistic (i.e., semantic and syntactic) context is also known to improve intelligibility (Greenberg & Christiansen, 2008). Such extraphonetic factors are not readily accommodated within distortion models. Indeed, linguistic context often plays a crucial role in speech comprehension (e.g., Boothroyd & Nitttrouer, 1988; Greenberg & Christiansen, 2008).

Jakobson, Fant, and Halle (1952) proposed a speech analysis model that differs from distortion models in several important ways. In their model, the speech signal is decomposed into acoustic primitives that reflect articulatory gestures associated with its production. These “distinctive features” are used to differentiate the sounds of a language's phonological inventory. By decomposing speech sounds into elementary features, the conundrum created by the AI's focus on distortion is potentially resolved. This is because a speech sound (a phonetic segment or “phone”) need not be processed as a monolithic entity, but rather as a cluster of articulatory-acoustic phonetic features. These could be recoded into consonantal or vocalic form, and then used to infer the words spoken (via a process analogous to a dictionary lookup or some other indexical operation). Distinctive-feature models view spoken language as a conglomeration of perceptual primitives from which higher-level, more abstract linguistic entities are derived. Because these primitives are rooted in vocal production, they may provide a principled basis for combining information that integrates acoustics, articulation, and perception into a coherent, unitary framework for decoding the speech signal.

The distinctive-feature approach has been highly influential, both in theoretical discussions (e.g., Chomsky & Halle, 1968; Clements, 1985; Ladefoged, 1971; Trubetzkoy, 1969 [1939]) and in empirical studies of perception (e.g., Diehl & Lindblom, 2004; Kewley-Port, Pisoni, & Studdert-Kennedy, 1983; Miller & Nicely, 1955; Stevens, 2002; Sussman, McCaffrey, & Matthews, 1991) and neurophysiology (Cheung, Hamilton, Johnson, & Chang, 2016; Mesgarani, Cheung, Johnson, & Chang, 2014). Distinctive features have also been successfully deployed in automatic speech recognition (e.g., Chang, Wester, & Greenberg, 2005; Frankel, Wester, and King, 2007; Ghosh & Narayanan, 2011; Hasegawa-Johnson et al., 2005; Juneja, 2004; Kirckhoff, Finkard, & Sagerer, 2002; Livescu et al., 2007; Rasipuram & Magimai-Doss, 2016). The APFs used nowadays are more parsimonious than those of Jakobson, Fant, and Halle (1952) (see Rasipuram & Magimai-Doss, 2016 for a brief historical survey, and Diehl & Lindblom, 2004 for a comprehensive review), but retain the production-centric focus of the original framework.

Three feature classes are especially important for phonetic representations. These pertain to (a) laryngeal vibration (“Voicing”), (b) the manner through which air flows through (or is constricted in) the vocal tract (“Manner of articulation”), and (c) the locus of articulatory constriction (“Place of articulation”). An illustration of how these features distinguish the 11 Danish consonants used in the current study is shown in Table 1.

A signal advantage of APF models is their robustness. A word or phrase can usually be recognized successfully despite one or more features being decoded incorrectly. This

Table 1 The phonetic feature classes and their values associated with the 11 consonants used in the current study

Segment	Voicing	Manner	Place
[p]	–	Stop	Anterior
[t]	–	Stop	Central
[k]	–	Stop	Posterior
[b]	+	Stop	Anterior
[d]	+	Stop	Central
[g]	+	Stop	Posterior
[f]	–	Fricative	Anterior
[s]	–	Fricative	Central
[v]	+	Fricative	Anterior
[m]	+	Nasal	Anterior
[n]	+	Nasal	Central

perceptual resilience is the likely consequence of ancillary information (e.g., visual or semantic cues) compensating for feature-decoding errors. Such “side data” could be especially important in the context of linguistic variability (e.g., dialect, voice quality) or acoustic interference (e.g., background noise, reverberation).

Despite the benefits of distinctive-feature models, there are potential drawbacks. First, the features proposed in Jakobson, Fant, and Halle (1952) are grounded in *production* and *acoustics* rather than *perception*. Hence, the link between their articulatory foundation and perception is indirect. Jakobson, Fant, and Halle (1952)’s model implicitly entails some kind of “translation” into auditory form, perhaps via back-computing like the “motor theory” of speech perception (Lieberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967).

Second, Jakobson, Fant, and Halle (1952)’s features are treated as independent elements. The advantage of feature independence is its representational parsimony. Dozens of phonological elements (and by extension, thousands of words) can be reliably distinguished with just a handful of phonetic features (e.g., Chomsky & Halle, 1968; Diehl & Lindblom, 2004; Ladefoged, 1971; Ladefoged & Maddieson, 1996; Stevens, 2002). However, this assumption of independence may not be empirically justified.

The *perceptual* relationship between APFs and consonant recognition was originally investigated by Miller and Nicely (1955). They computed consonant-confusion matrices for 16 English consonants embedded in high- and low-pass filtered CV syllables presented in variable amounts of background noise (i.e., over a range of SNRs). The confusion patterns were analyzed in terms of distinctive APFs. Their study showed that the way phonetic information is distributed across the frequency spectrum differs among APFs. For example, the amount of information transmitted (IT) for “Place of articulation” increases almost linearly as the speech bandwidth broadens (Miller & Nicely, 1955, Fig. 3). In contrast, the

corresponding IT functions for “Voicing” and “Manner of articulation” (Miller & Nicely, 1955, Fig. 4) asymptote at relatively narrow bandwidths (ca. 1 octave). Moreover, Voicing information can be reliably decoded from most any region of the frequency spectrum for bandwidths of an octave or more. Manner of articulation information is spectrally more constrained than Voicing, but requires more of the spectrum to successfully decode. Place of articulation is the most fragile and brittle of the features, requiring much of the speech spectrum (0.5–3.5 kHz) to reliably distinguish anterior, central, and posterior loci of vocal-tract constriction.

Miller and Nicely (1955)’s study is important because it focused on three aspects of speech processing that had previously been ignored: (1) decomposition of consonants into structural primitives (i.e., APFs), (2) detailed error analyses derived from confusion matrices, and (3) the use of an information-theoretic analysis based on (1) and (2) that provides insight into the perceptual processes underlying consonant recognition. Their principal finding – that Place-of-articulation decoding is considerably more vulnerable to background noise than Manner and Voicing – is not called into question by the assumption of feature independence. The perceptual fragility of the Place feature is likely a consequence of its information being broadly distributed across the speech spectrum.

Christiansen and Greenberg (2012) extended Miller and Nicely (1955)’s study, with a focus on robustness and cross-spectral integration of phonetic information. Using an information-transmission metric, they concluded that consonant recognition requires accurate decoding of Place of articulation, and relies on nonlinear, cross-spectral synergy (in contrast to Voicing and Manner which do not exhibit such nonlinear integration across the speech spectrum).

In the current study, we used the same consonant-recognition data reported in Christiansen and Greenberg (2012) to examine the perceptual flow of APF processing. Key to the recognition of consonants is the concept of decoding. Decoding refers to “partial recognition,” whereby the part correctly recognized refers to a structured constituent (or element) of the “whole.” In the present study, consonants represent the whole, and APFs the elements. Decomposition of the consonant into primitive elements, when embedded within the appropriate structure, provides a powerful means by which to examine phonetic processing in greater detail than conventional segment-based analyses afford.

Methods and procedures

Design

This study is part of a larger project investigating how consonants are processed and recognized by human listeners. The

stimuli and experimental methods are the same as in Christiansen and Greenberg (2012). Its focus is on the conditional probabilities associated with phonetic-feature decoding derived from consonant-recognition confusion matrices.

Stimuli

Stimuli were Danish syllables recorded in a sound-insulated chamber (Chan et al., 1995). Each presentation was a concatenation of a short, unfiltered carrier phrase “På pladsen mellem hytten og...” (English translation: "On the square between the cottage and...") followed by a test syllable, which contained one of 11 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 [l] (e.g., [til] [tal], [tul]). Recordings of a female talker and a male talker, each enunciating the carrier sentence and the test syllables, were used.

The carrier phrase was used 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 sampling rate was 20 kHz. The signals were subsequently up-sampled to 44.1 kHz for stimulus presentation. The acoustic-frequency spectrum was partitioned into three spectrally delimited bands (slits). The lowest slit was centered at 750 Hz, the middle slit at 1,500 Hz, and the highest at 3,000 Hz. All slit combinations were tested, yielding a total of 3 (single slits) + 3 (two-slit combinations) + 1 (all three-slits) = 7 slit configurations. The band-pass filters had a 3-dB bandwidth of 3/4-octave and nominal slopes of 120 dB/octave outside the pass band.

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 slits presented concurrently was close to, but not quite 100% to avoid ceiling effects, (2) recognition accuracy with two slits was 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 was clearly above chance level in order to avoid floor effects, and (4) consonant-recognition accuracy was roughly comparable across single-slit conditions.

The seven conditions are listed, along with 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).

Miller and Nicely (1955) used 16 English consonants [p t k f θ s b d g v δ z m n], of which [θ δ z] were excluded in the current study because [θ z] lack Danish counterparts and [δ] does not occur syllable-initially in Danish. The remaining 11 consonants bear similarity to their English counterparts (Basbøll, 2005; Grønnum, 1998).

Procedure and subjects

The data associated with the seven slit configurations were collected as part of a larger study encompassing 83 slit combinations, where the presentations were divided into nine sessions, each lasting less than 2 h, during which subjects could take short breaks. The data 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. The 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.

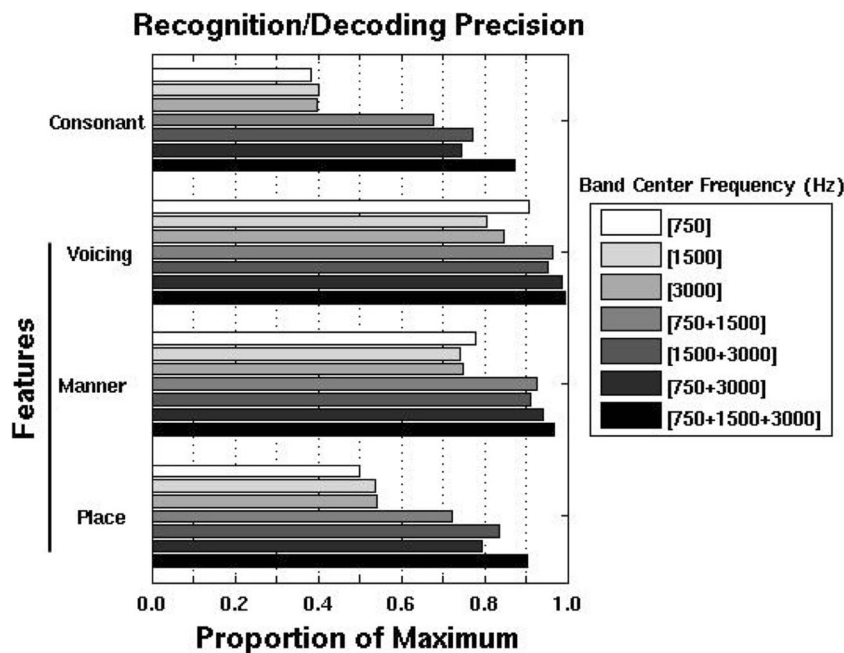
The subject was seated in a double-walled sound booth. His/her task was to identify the initial consonant of the test signal by mouse-selecting it from the 11 consonant alternatives displayed on a computer display. No response feedback was provided. Six native speakers of Danish (three males, three females)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

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.

Analysis

Consonant recognition is usually studied from the perspective of whether the speech sound is correctly identified or not. This metric gauges how well the speech signal is processed, often in listening conditions involving background noise, reverberation, or competing talkers. As useful as this metric may be, it is limited in its ability to delineate the specific mechanisms underlying consonant recognition because there are various

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. Of the consonants, 99.0% were correctly recognized in the absence of band-pass filtering (i.e., the original, unprocessed signals). Reprinted from Christiansen and Greenberg (2012) with permission



ways in which a consonant may be incorrectly recognized. Error pattern analysis at the phonetic-feature level provides insight into the underlying decoding process that goes well beyond what's possible when measuring consonant-recognition alone. Because decoding and recognition accuracy were comparable across the single-slit and two-slit conditions (Fig. 1), the data illustrated in the current study are averages across the three data sets associated with each slit condition (for simplicity of exposition). Collapsing the data in this way does not alter any of the conclusions drawn.

In what follows, we describe the steps used in the conditional probability analysis, as well as its logic.

When a consonant is identified accurately, its constituent features are also decoded correctly. But what happens when a consonant is not identified correctly? Are the errors entirely random? Or is there a systematic structure to the consonant confusions? If the latter, what do the error patterns reveal about the perceptual mechanisms responsible for consonant recognition?

When a consonant is misrecognized, it is unusual for all constituent features to be decoded inaccurately. There is structure to feature-decoding errors, one that can shed light on how consonants (and by extension, words and phrases) are processed and perceived.

To understand how phonetic-feature-error analysis can shed light on consonant recognition, let's first examine the 11 Danish consonants used in this study (Table 1). These consonants are decomposed into the articulatory-acoustic feature classes of Voicing, Manner, and Place. For example, Voicing distinguishes the voiced stop consonants, [b], [d],

[g] [from their unvoiced counterparts [p], [t], [k]].¹ Manner distinguishes [b] (a stop) from [m] (a nasal), and [v] (a fricative). Place distinguishes among the bilabial, anterior constriction of [p] from [t] (medial constriction), and [k] (posterior constriction). Two of the feature classes (Manner and Place) can assume one of three values, while Voicing is a binary feature.

Two hypothetical examples of a consonant confusion are shown in Table 2. In Table 2, a single decoding error, confined to the Place feature, is illustrated. In this instance, the segment [d] is mistaken for the segment [g]. In Table 3, a decoding error occurs for two features, Manner and Place, resulting in a [g] being misidentified as [n].

Key to the current analyses is the consonant-confusion matrix, which provides a method for analyzing error patterns in terms of articulatory-acoustic features. Table 4 shows an example of one such consonant-confusion matrix where row values refer to the stimulus presented, while column values denote listener responses. The number of correct responses is indicated in bold along the diagonal. The corresponding APF confusion matrices for Voicing, Manner, and Place are shown in Tables 5, 6 and 7.

¹ 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 conventional term, "Voicing," is retained in the present analysis and discussion.

Table 2 A hypothetical phonetic-feature analysis, in which a consonant confusion (mistaking a [g] for the segment [d]) is illustrated

	Segment	VOICING	MANNER	PLACE	
	[p]	–	Stop	Anterior	
	[t]	–	Stop	Central	
	[k]	–	Stop	Posterior	
	[b]	+	Stop	Anterior	
Stimulus	[d]	+	Stop	<u>Central</u>	Place
Response	[g]	+	Stop	<u>Posterior</u>	Error
	[f]	–	Fricative	Anterior	
	[s]	–	Fricative	Central	
	[v]	+	Fricative	Anterior	
	[m]	+	Nasal	Anterior	
	[n]	+	Nasal	Central	

In this example, Voicing and Manner are correctly decoded; only Place is incorrect. The incorrect feature is shown in bold, underlined text

The APF-confusion matrix is derived by first grouping the consonant-by-APF associations (see Table 1). In Table 6, the groups are the Manner classes “stop” ([p], [t], [k], [b], [d], [g]), “fricative” ([s], [f], [v]), and “nasal” ([m], [n]). 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 6). 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,

Table 3 A second hypothetical phonetic-feature analysis, in which the error (mistaking a [g] for the segment [n]) involves two different phonetic features, Manner and Place

	Segment	VOICING	MANNER	PLACE	
	[p]	–	Stop	Anterior	
	[t]	–	Stop	Central	
	[k]	–	Stop	Posterior	
	[b]	+	Stop	Anterior	
	[d]	+	Stop	Central	
Stimulus	[g]	+	Stop	Posterior	PLACE
	[f]	–	Fricative	Anterior	
	[s]	–	Fricative	Central	Errors
	[v]	+	Fricative	Anterior	
	[m]	+	Nasal	Anterior	
Response	[n]	+	Nasal	Central	MANNER

Only the VOICING feature is decoded correctly. The incorrect features are shown in bold, underlined text

as shown in Table 4 – 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%.

Results

The consonant-recognition data were analyzed in terms of accuracy (i.e., percent correct) at both the segment (i.e., consonant) and phonetic-feature levels. We first examined the relationship between consonant recognition and decoding of the three phonetic-feature classes, Voicing, Manner, and Place of articulation. If Voicing, Manner, and Place are truly independent perceptual elements, decoding of each feature should be uncorrelated with consonant recognition accuracy. However, as Table 4, and subsequent perceptual confusion data show, the relationship between consonant-recognition and feature-decoding errors is anything but random. Voicing is the feature least correlated with consonant recognition (Fig. 2). It is frequently decoded correctly even when the associated consonant is misrecognized. This relatively loose connection between voicing and consonant recognition is reflected in their relatively low correlation ($r^2=0.76$). Manner of articulation is more closely linked to consonant recognition ($r^2=0.94$). However, this feature, like Voicing, is often decoded accurately despite the consonant being identified incorrectly. Unlike Voicing and Manner, there is a very tight connection between consonant recognition and Place-of-articulation decoding ($r^2=0.99$). If the Place feature is decoded correctly, the consonant is virtually always identified correctly as well. In this sense, consonant recognition and Place decoding are flip sides of the same perceptual coin.

Let’s delve a little deeper into the feature-decoding patterns to gain further insight into how each phonetic feature is decoded, as well as its likely function in recognizing consonants. We focus first on the probability that a given feature is decoded correctly (or not), conditioned on a different feature being decoded successfully (or not).

If all features are decoded independently, feature-decoding errors should be roughly symmetric. For example, decoding of Manner should not be impacted if Voicing is correctly decoded (or not), and vice versa. Place decoding should not be affected if Manner or Voicing are decoded correctly (or not), and so on. In the analyses that follow, we examine the impact of decoding feature X (e.g., Place), given the decoding accuracy of feature Y (e.g., Manner) for each phonetic-feature pair (Voice–Manner, Voice–Place, Manner–Place, Manner–Voice, Place–Voice, Place–Manner).

An example of a conditional-probability analysis is illustrated in Figure 3 for instances where the base feature is correctly or incorrectly decoded. In the example, the base (i.e., reference) feature is Voicing and the conditional feature is Manner. The same procedure is also performed for the

Table 4 Example confusion matrices for the 1,500-Hz slit condition summed across the six test subjects

		R e s p o n s e											
		p	t	k	b	d	g	s	f	v	n	m	Total
S t i m u l u s	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

Row values pertain to the stimulus presented, while column values are the listener responses. Table 4 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. Tables 5, 6 and 7 show the confusion matrices derived from Table 4 for Voicing, Manner, and Place, respectively, as described in the *Analysis* section. The sum of all elements from the left-most sub-matrix in Table 4 yields the number of correctly identified stop consonants (179), which in turn is the upper left-hand element in Table 6. Similarly, the 23 stop consonants confused for fricatives in the right-most box of Table 4 correspond to the upper middle element of Table 6. Reprinted from Christiansen and Greenberg (2012) with permission

Consonant-recognition accuracy: $159/396 = 40.2\%$

mirror-image condition in which Manner serves as the *Base* feature and Voicing as the *conditional* feature. If the features Voicing and Manner are processed independently, decoding accuracy should be comparable across conditions. However, they are not. Instead, we see that Manner is decoded more poorly (55%) when Voicing is incorrectly decoded, but that Voicing is decoded more accurately (than Manner) when the latter feature is decoded incorrectly (80%). It is this form of asymmetry in APF-decoding accuracy (i.e., 55% vs. 80%) that is of primary interest in our study.

In Table 8, the values of interest are the percent correct for Manner given incorrect decoding of Voicing (55%) relative to the percent correct for Voicing, given incorrect decoding of Manner (80%). This ratio (55/80) yields a “conditional feature accuracy ratio” (CFAR) of 0.69, well below the neutral (i.e., hypothetical independence) value of 1.0. Analogous computation for Voicing, relative to Place decoding, yields comparable results (69% correct decoding for Place when Voicing is incorrect; 96% correct decoding of Voicing, when Place is incorrect). The CFAR for this condition is $69/96 = 0.72$.

Table 5 Derived confusion matrix for Voicing

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

Decoding precision: $319/396 = 80.6\%$

Table 6 Derived confusion matrix for Manner

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

Decoding precision: $293/396 = 74.0\%$

Table 7 Derived confusion matrix for Place

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

Decoding precision: 213/396 = 53.8%

This computation can be repeated for the six Feature X / Feature Y conditions to determine which features are decoded accurately without other features being decoding correctly, as illustrated in Table 9.

In this example, there is an asymmetry between the accuracy of feature decoding for Manner and Voicing. When Manner is decoded correctly (98%), Voicing is correctly decoded almost as well (80%). In contrast, when Voicing is decoded correctly (94%), Manner is not decoded nearly as well (55%) for this condition.

The data used to illustrate the conditional probability calculation and the CFAR up to this point are from one condition out of the many investigated. Let’s now examine the conditional probability patterns for the *entire* data set in two different ways. In Table 9, the proportion of features (or consonants) correctly decoded (or recognized) are shown for two different scenarios. In the first (to the left of the grey vertical bar), the conditional feature,

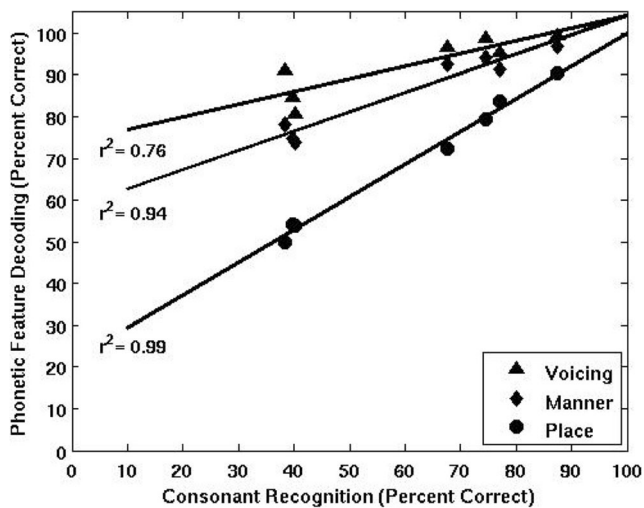


Fig. 2 Voicing, Manner, and Place decoding accuracy as a function of consonant-recognition accuracy for the same conditions and listeners as shown in Fig. 1. With each phonetic feature, a best-fit linear regression and a correlation coefficient (r^2) are shown. Not all phonetic features are decoded equally well. Manner and Voicing are often correct, even when the consonant is incorrectly recognized. Place of articulation is the only feature (of the three) for which decoding is (almost) perfectly correlated with consonant recognition. Reprinted from Christiansen and Greenberg (2012) with permission

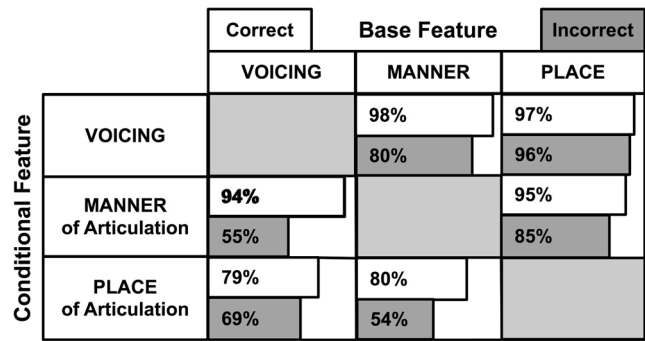


Fig. 3 Diagram showing the accuracy of the different base and conditional feature-decoding combinations. The open (white) rectangles designate conditions where the base feature is correctly decoded. The gray boxes are associated with conditions where the base feature is incorrectly decoded. Note that the decoding of Voicing is mostly correct when Manner (80%) and Place (96%) are incorrectly decoded. In contrast, Place is not well decoded when Voicing (69%) or Manner (54%) are incorrectly decoded. Manner decoding is severely impacted when Voicing (55%) is inaccurately decoded, but isn’t so much when Place (85%) is incorrectly decoded

Place, Manner, or Voicing, is correctly decoded. In the second, these same features are decoded inaccurately.

Place is closely associated with Consonant recognition, while Voicing and Manner are not nearly as much. Moreover, there appears to be a hierarchical relationship between Place and Manner. Although decoding of the Manner feature is not so tightly linked to the decoding of Place, the latter feature’s decoding is considerably more accurate when the former is correctly decoded.

The Conditional Feature Accuracy Ratio exhibits an asymmetric pattern across many stimulus conditions (e.g. Voicing/Manner values differ greatly from Manner/Voicing values for many of the conditions). In slit conditions spanning a broad frequency range, cross-spectral integration may account for the CFARs close to 1 (Table 10).

Table 8 An illustration of how conditional probabilities are computed in the analysis of feature decoding

Feature	Class	Accuracy	Accuracy	Math	CFAR	
Base	X	VOICING	Correct	Incorrect		
Conditional	Y	Manner	Correct	Correct		
			94%	55%	0.55	
Base	Y	Manner	Correct	Incorrect	×	=
Conditional	X	VOICING	Correct	Correct		
			98%	80%	0.8	0.69

In this example, there is an asymmetry between the accuracy of feature decoding for Manner and Voicing. When Manner is decoded correctly (98%), Voicing is decoded correctly almost as well (80%). In contrast, when Voicing is decoded correctly (94%), Manner may not be decoded nearly as accurately (55%). The conditional feature accuracy (CFAR) is the proportion of feature Y’s conditional accuracy (given incorrect decoding of feature X) [0.55] divided by feature X’s conditional accuracy (given incorrect decoding of feature Y) [0.80]

Table 9 The proportion of phonetic features (and consonants) correctly decoded (or recognized), depending on whether the conditional element is correctly or incorrectly decoded

Reference Element		Conditional element					
		Correct			Incorrect		
Element		PLACE	MANNER	VOICING	PLACE	MANNER	VOICING
Slits	Consonant	0.965	0.901	0.878	0	0	0
	PLACE	–	0.907	0.904	–	0.811	0.875
3	MANNER	0.971	–	0.969	0.937	–	0.813
	VOICING	0.993	0.994	–	0.991	0.959	–
	Consonant	0.915	0.759	0.713	0	0	0
	PLACE	–	0.773	0.754	–	0.526	0.638
2	MANNER	0.931	–	0.919	0.815	–	0.489
	VOICING	0.966	0.979	–	0.943	0.793	–
	Consonant	0.749	0.521	0.462	0	0	0
	PLACE	–	0.557	0.539	–	0.431	0.448
1	MANNER	0.800	–	0.813	0.707	–	0.425
	VOICING	0.875	0.918	–	0.829	0.655	–

The base elements are listed vertically. For example, when MANNER is decoded correctly, Place is decoded correctly with a probability of 0.971 in the three-slit condition

Decoding of Manner appears to be associated with accurate decoding of Voicing. This is reflected in the relatively low CFAR for the Voicing/Manner conditions. The disparity between Voicing/Manner and Manner/Voicing CFARs suggests that Voicing may be the more fundamental feature, in that the CFAR is usually greater for the Manner/Voicing condition relative to the Voicing/Manner condition. In this sense, Manner and Voicing may be kindred (i.e., linked) features with respect to their conditional behavior (e.g., Table 10).

The decoding of Place information may be tied to accurate decoding of Manner and Voicing, particularly when the speech spectrum is only partially available, as is the case for

the one-slit and 2-slit conditions. This is because Place is not decoded nearly as accurately when Manner and Voicing are incorrectly decoded. The absence of such a pattern for the three-slit condition may be a consequence of the broad distribution of spectral information associated with Place (consistent with Miller & Nicely (1955)’s study). A comparison of the Manner/Place CFARs relative to Place/Manner CFARs suggests that decoding of Place is facilitated when Manner is correctly decoded. This result is also consistent with machine-learning experiments reported by Chang, Wester and Greenberg (2005), in which Place-feature classification benefitted appreciably from Manner-dependent training.

Table 10 The Conditional Feature Accuracy Ratio (CFAR) computed for six phonetic-feature pairs across seven slit conditions

Stimulus Condition			Phonetic Features					Base Feature Conditional Feature
Slit Center Frequency (Hz)			VOICING MANNER	MANNER VOICING	MANNER PLACE	PLACE MANNER	VOICING PLACE	
750			0.47	0.78	0.98	0.96	0.76	0.96
		15000	0.51	0.63	0.58	0.95	0.87	0.95
		3000	0.57	0.74	0.81	0.93	0.79	0.93
750	1500		0.43	0.74	0.72	0.99	0.92	0.99
	1500	3000	0.56	0.78	0.58	0.94	0.78	0.94
750		3000	1.06	1.01	0.76	1.01	1.14	1.01
750	1500	3000	0.84	0.97	0.88	1.00	0.97	1.00

Note that the conditional features, Voicing and Manner, are somewhat insensitive to whether the base Place feature is correctly decoded or not. The CFAR is asymmetric across many of the phonetic feature pair comparisons. Manner and Place show a greater statistical reliance on accurate decoding of the base feature than Voicing

Discussion

Historically, articulatory-acoustic features have been treated as independent, descriptive elements within an analytical framework for distinguishing the phonological units of a language (e.g., Chomsky & Halle, 1968; Jakobson, Fant, & Halle, 1963). But even perceptual scientists (e.g., Miller & Nicely, 1955) often treat APFs as independent. Feature independence has recently been called into question by Christiansen and Greenberg (2012), as well as by this study. For if APFs are truly independent elements, there should be little or no asymmetry in decoding accuracy. However, as Tables 8, 9 and 10 and Fig. 3 illustrate, decoding asymmetries are pervasive. Decoding of higher-level features appears to be linked to successful decoding of lower-level features.

To understand why, we present a hierarchical model of phonetic-feature processing, and discuss how such an approach might account for the feature-decoding patterns shown in Tables 9 and 10 and Figs. 2 and 3. In our view, the APFs of Voicing, Manner and Place perform complementary roles in the recognition of consonant segments (and by extension, non-consonants, such as vowels and semi-vowels). But before discussing the model, let's examine each phonetic feature by itself to better understand its perceptual relationship to the other APFs.

Voicing is the feature most accurately decoded across stimulus conditions (Fig. 1). One reason is its binary character (i.e., +Voice, -Voice), meaning that its decoding "floor" (i.e., chance level) is 0.5. In contrast, the decoding floor for Manner and Place is 0.333, as each feature may assume one of three values.

However, there may be other reasons why Voicing's decoding is superior. It is the most resilient of the APFs, perceptually robust even in the face of background noise or distortion (e.g., narrow spectral slits, like those used in the current study, see Fig. 1). Moreover, voicing cues are broadly distributed across the speech spectrum. And yet, reliable decoding only requires a narrow portion of the spectrum (i.e., a single slit) (Fig. 1). Additional slits improve decoding only slightly (Fig. 1). Such a pattern suggests that Voicing information can be reliably extracted from most parts of the frequency spectrum, and that it is not tied to any given region. This means that the "ideal" listener need not analyze the entire spectrum to determine whether a segment is Voiced or not. Such spectral flexibility likely facilitates the decoding process.

Voicing is also the APF least necessary to infer a phonetic segment's "underlying" (i.e., phonemic) identity. This is because a consonantal segment's Voicing "sign" (from +Voice to -Voice, or vice versa) is mutable in that linguistic context can disambiguate the phonetic fuzziness of spoken discourse. A consonant's Voicing quality, especially in coda position, reflects a syllable's prosody (e.g., stress-accent) not just its phonological identity (Greenberg, Carvey, & Hitchcock, 2001),

and may enhance intelligibility in challenging listening conditions (e.g., Divenyi, 2004).

Manner of articulation is often temporally aligned with the phonetic segment. Contiguous segments within a syllable usually differ in their Manner affiliation, an observation consistent with this feature being closely associated with the phone and its underlying representation (i.e., the phoneme). Unlike Voicing, Manner cues are distributed over approximately an octave range of the spectrum (Miller & Nicely, 1955). This may be why Manner features are decoded reasonably well (if imperfectly) in single-slit stimuli. Adding a second slit improves decoding only slightly (to near-ceiling performance). A third slit benefits Manner decoding hardly at all (Fig. 1).

Place of articulation differs from Voicing and Manner in several ways. Its decoding cuts across a broad frequency range that encompasses much of the speech spectrum. This is reflected in the following association between spectral peaks and Place of articulation. Anterior vocal tract constrictions ("Front") are associated with a low-frequency (<1,000 Hz) peak, while posterior occlusions ("Back") are linked to a high-frequency (> 2,500 Hz) peak. Intermediate constrictions ("Central") have their energy peak between 1,200 and 2,400 Hz. These loci roughly correspond to the lowest three vocal-tract resonances or formants (F1 < 800 Hz, F2 1,000–2,400 Hz, F3 2,500–3,200 Hz).

Place decoding is relatively poor for single-slit stimuli, irrespective of their spectral locus. The addition of a second slit significantly improves decoding accuracy. However, near-perfect decoding requires all three slits, a pattern consistent with Miller and Nicely (1955)'s finding that Place information is distributed over a broad swath of the speech spectrum. A potential "downside" of Place's broad spectral distribution is its susceptibility to acoustic interference, distortion or hearing impairment. Place of articulation cues are also widely distributed over time, so much so that information in the preceding and/or following vowel is often used to recognize consonants (e.g., Cole et al., 1996; Kewley-Port, Burke, & Lee, 2007; Lee & Kewley-Port, 2009).

Place of articulation also differs from other APFs in its close connection to visual speech cues (Grant, Walden, & Seitz, 1998; McGurk and MacDonald, 1976). These cues pertain to the movement of the lips, tongue, and jaw, and are highly dynamic, closely paralleling the formant patterns associated with Place of articulation.

Let's now consider how APFs may interact during the process of consonant recognition. Phonetic features can be likened to (visual) perceptual primitives (Marr, 1982). In a hierarchical processing model, the feature classes, Voicing, Manner, and Place carry different (often adjustable) weights, and perform complementary roles in the recognition of consonant segments. Information critical for decoding Voicing, Manner, and Place of articulation is differentially distributed across time and frequency. Such differences are key to how each feature interacts with the others during consonant recognition.

Voicing is a temporally fine-grained feature in which the critical interval is 30–40 ms (Lisker & Abramson, 1964). The distinction between voiced and unvoiced stop consonants may involve voice-onset time differences as brief as 20–30 ms (Abramson & Lisker, 1970; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967). This acute time sensitivity, in tandem with spectral flexibility, allows for Voicing decisions to be made quickly and reliably. Voicing is a sub-phonemic feature associated with the speech signal's spectral and temporal microstructure. Its decoding time is likely to be very fast, about 40 ms, about half a consonant's duration.

Manner and Place information unfold over much longer time intervals. Manner's information is usually distributed over the entire segment (80–100 ms for consonants) and spectrally localized. Such localization likely requires more time to analyze than Voicing.

Place of articulation decoding relies on a broadband, spectral-shape analysis over a comparatively long interval of time spanning two or more segments.

Given that decoding of Voicing, Manner, and Place of articulation occur over different time scales and spectral regions,

some high-level mechanism likely monitors and guides their interaction during the process of consonant recognition.

Such differences in feature information distribution provide a theoretical framework for how Voicing, Manner, and Place interact in consonant recognition. Voicing requires a relatively brief amount of time to reliably decode, usually less than half a segment's duration. For this reason, a listener is likely to have made a Voicing decision in advance of decoding Manner of articulation, which requires most of the segment to accomplish. Place of articulation requires a longer time span to process, as critical information frequently spans two or even three segments. Under challenging listening conditions, speech may require even longer to decode (e.g., Greenberg & Christiansen, 2008).

The perceptual flow model (Fig. 4) posits that a lower-level feature, such as Voicing, is decoded in advance of higher-level features, which integrate information from below. This supposition is consistent with our finding that a lower-level feature (e.g., Voicing) is rarely impacted by the decoding of a higher-level feature (e.g., Manner), but not vice versa. By the time an intermediate-level feature (i.e., Manner) or high-level feature (i.e., Place) is processed, the lower-level feature has already been decoded.

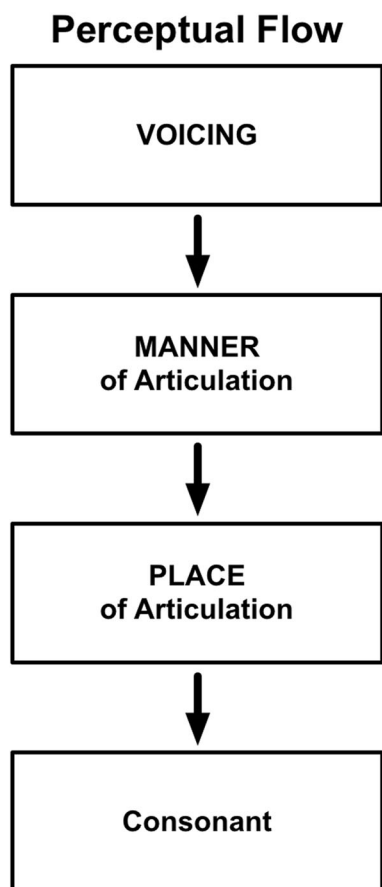


Fig. 4 A schematic illustration of the phonetic flow model of phonetic processing. The coarsest, lowest level of analysis is Voicing, with a postulated processing time of 40 ms. The flow of phonetic information proceeds to the finer-grained, higher-level features of Manner and Place

Phonetic Information

VOICING is highly resilient to acoustic interference, and can be accurately decoded from several different parts of the frequency spectrum. It reflects the harmonic structure of the syllable. It not only distinguishes voiced from unvoiced segments, but also conveys prosodic information related to syllable prominence and intonation.

MANNER is a feature that demarcates boundaries between segments. Its spectral features are finer grained than VOICING, but coarser than PLACE. MANNER reflects a syllable's energy pattern at a segmental time scale, and forms the analytical substrate for PLACE analysis.

PLACE decoding requires analysis of much of a segment's frequency spectrum, and is the result of cross-spectral and cross-temporal synergy. Its information can combine with visual speech cues for robust processing in acoustically challenging conditions. Consonant recognition is virtually isomorphic with PLACE decoding.

All phonetic features of the consonant segment have to be decoded correctly to accurately identify a consonant in this Maximum Entropy, low context recognition task.

of articulation. The (estimated) interval associated with each analysis is shown in the middle column. A functional description of each level's role is shown at right

Support for such a hierarchical, perceptual-flow model also comes from studies of audio-visual speech processing. Van Wassenhove, Grant, and Poeppel (2005) have shown that the inclusion of visual speech cues may accelerate speech processing by approximately 100 ms. With the integration of visual, speech-reading information, the brain would have access to key information about Place of articulation well in advance of it being available through the auditory modality alone. This bi-sensory input could also facilitate integration with Manner and Voicing cues (that are usually not accessible in speech-reading cues) to provide a more robust representation of speech. The perceptual flow model is also consistent with a study by Grant and Greenberg (2001), in which the intelligibility of sparse spectral sentences was shown to be relatively insensitive to desynchronization of the audio and visual channels when the video channel preceded the audio by as much as 200 ms (but not vice versa).

Although Place of articulation is the key phonetic feature for consonant recognition, Manner and Voicing also play important roles. Voicing is a hybrid feature in that it serves not only to distinguish voiced from unvoiced consonants, but also provides a perceptual foundation for prosodic (syllabic) prominence and lexical stress in speech communication. Manner of articulation is the feature most temporally aligned with the phonetic segment, and as such provides important information germane to phonemic representations. It is the intricate interplay of phonetic-feature decoding that provides the perceptual foundation for the (usually) seamless recognition of segments and higher-level elements (e.g., syllables, words, and phrases) in spoken discourse.

Conclusions

Consonant-confusion patterns were analyzed to gain insight into phonetic-feature decoding involved in consonant recognition. Asymmetries in decoding articulatory-acoustic phonetic features are consistent with a perceptual flow model in which the decoding of Manner of articulation derives substantial benefit from the decoding of Voicing (but not vice-versa), and decoding of Place of articulation benefits from decoding of Manner (but not the converse). Such a hierarchical pattern of feature decoding may facilitate the integration of fine-grained phonetic detail into more global linguistic representations associated with prosody and other supra-segmental elements important for understanding speech, especially in adverse acoustic environments.

Acknowledgements This research was funded by the Carlsberg Foundation, Technical University of Denmark, and the United States Air Force Office of Scientific Research. The authors thank Torsten Dau for helpful suggestions and comments on various aspects of this research, as well as Andy Lotto and an anonymous reviewer for helpful suggestions on improving the original draft of this paper.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

- Abramson, A., & Lisker, L. (1970). Discriminability along the voicing continuum: Cross language tests. *Proceedings of the 6th International Congress of Phonetic Sciences* (pp. 569–573).
- Allen, J. B. (2005). Consonant recognition and the articulation index. *The Journal of the Acoustical Society of America*, *117*, 2212–2223.
- ANSI (1969). Methods for the calculation of the articulation index. ANSI Standard S3.5-1969.
- ANSI (1997). Methods for the calculation of the speech intelligibility index. ANSI Standard S3.5-1997.
- Basbøll, Hans (2005). *The phonology of Danish*. Oxford: Oxford University Press.
- Bell, T. S., Dirks, D. D., & Trine, T. D. (1992). Frequency-importance functions for words in high- and low-context sentences. *Journal of Speech and Hearing Research*, *35*, 950–959.
- Bonatti, L., Peña, M., Nespore, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science*, *16*, 451–459.
- Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America*, *84*, 101–114.
- Braida, L. D. (1991). Crossmodal integration in the identification of consonant segments. *Quarterly Journal of Experimental Psychology*, *43*, 647–677.
- Chan, D., Fourcin, A., Gibbon, D., Granström, B., Huckvale, M., Kokkinakis, G., ..., Zeiliger, J. (1995). EUROM—A spoken language resource for the EU, in *Proceedings of the 6th European Conference on Speech Communication and Technology (Eurospeech'95)*, pp. 867–870.
- Chang, S., Wester, M., & Greenberg, S. (2005). An elitist approach to automatic articulatory-acoustic feature classification for phonetic characterization of spoken language. *Speech Communication*, *47*, 290–311.
- Cheung, C., Hamilton, L. S., Johnson, K., & Chang, E. F. (2016). The auditory representation of speech sounds in the human motor cortex. *eLife*, *5*, e12577.
- Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.
- Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, *2*, 225–252.
- Cohen M. M., & Massaro, D. W. (1995). Perceiving visual and auditory information in consonant-vowel and vowel syllables, In C. Sorin, J. Mariani, H. Meloni, & J. Schoentgen, (Eds.), *Levels in speech communication: Relations and interactions* (pp. 25–37). Amsterdam: Elsevier.
- Cole, R., Yan, Y., Mak, B., Fany, M., & Bailey, T. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* (pp. 853–856).
- Diehl, R., & Lindblom, B. (2004). In S. Greenberg, W. Ainsworth, A. Popper, & R. Fay (Eds.), *Speech processing in the auditory system* (pp.101–162). New York: Springer.
- Divenyi, P. (Ed.) (2004) *Speech separation by humans and machines*. Boston: Kluwer.
- Elhilali, M., Chi, T., & Shamma, S. (2003). A spectro-temporal modulation index (STMI) for assessment of speech intelligibility. *Speech Communication*, *41*, 331–348.

- Fletcher, H. (1953). *Speech and hearing in communication*. New York: Van Nostrand. Reprinted by the Acoustical Society of America, with a forward by J. Allen (1995).
- Frankel, J., Wester, M., & King, S. (2007). Articulatory feature recognition using dynamic Bayesian networks. *Computer Speech and Language, 21*, 620–640.
- French, N. R., & Steinberg, J. C. (1949). Factors governing the intelligibility of speech sounds. *The Journal of the Acoustical Society of America, 19*, 90–119.
- Ghosh, P. K., & Narayanan, S. (2011). Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. *The Journal of the Acoustical Society of America, 130*, EL251–257.
- Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America, 103*, 2677–2690.
- Grant, K. W., & Braida, L. D. (1991). Evaluating the articulation index for auditory-visual input. *The Journal of the Acoustical Society of America, 89*, 2952–2960.
- Greenberg, S., & Ainsworth, W. A. (2004). Speech processing in the auditory system: An Overview. In S. Greenberg, W. A. Ainsworth, A. R. Popper, & R. R. Fay (Eds.), *Speech processing in the auditory system* (pp. 1–62). New York: Springer.
- Greenberg, S., Carvey, H., & Hitchcock, L. (2002). The relation between stress accent and pronunciation variation in spontaneous American English discourse. *Proceedings of the ISCA Workshop on Prosody and Speech Processing*.
- Greenberg S., & Christiansen, T. U. (2008). Linguistic scene analysis and the importance of synergy, in T. Dau, J. M. Buchholz, J. M. Harte, T. U. Christiansen (Eds.), *Auditory signal processing in hearing impaired listeners*. Elsinore, Denmark: Danavox, (pp. 351–364).
- Grønnum, N. (1998). Illustrations of the IPA: Danish. *Journal of the International Phonetics Association, 28*, 99–105.
- Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S.,... Wang, T. (2005). Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop. *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol.1.
- Jakobson, R. Fant, G., & Halle, M. (1963). *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press. [Originally published in 1952 as a research monograph by the MIT Research Laboratory of Electronics].
- Juneja, A. (2004). *Speech recognition based on phonetic features and acoustic landmarks*. Ph.D. thesis, University of Maryland.
- Kewley-Port, D. Pisoni, D. B., & Studdert-Kennedy, M. (1983). Perception of static and dynamic acoustic cues to place of articulation in initial stop consonants. *The Journal of the Acoustical Society of America, 73*, 1779–1793.
- Kewley-Port, D., Burkley, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 122*, 2365–2375.
- Kryter, K. D. (1962). Methods for the calculation and use of the articulation index. *The Journal of the Acoustical Society of America, 34*, 1689–1697.
- Ladefoged, P. (1971). *Preliminaries to linguistic phonetics*. Chicago: University of Chicago Press.
- Ladefoged, P., & Maddieson, I. (1996). *The sounds of the world's languages*. Oxford: Blackwell.
- Lee, J. H., & Kewley-Port, D. (2009). Intelligibility of interrupted sentences at subsegmental levels in young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 125*, 1153–1163.
- Lieberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review, 74*, 431–461.
- Lisker L., & Abramson A. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*, 384–442.
- Livescu, K., Çetin, Ö., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N.,... Saenko, K. (2007). Articulatory feature-based methods for acoustic and audio-visual speech recognition: 2006 JHU summer workshop final report. *Proceedings of IEEE International Conference on Acoustic, Speech, and Signal Processing* (pp. 621–624).
- Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York: Freeman.
- McGurk H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*, 746–748.
- Massaro, D. W. (1987). *Speech perception by ear & eye: A paradigm for psychological inquiry*. Hinsdale, NJ: Lawrence Erlbaum.
- Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science, 343*, 1006–1010.
- Miller G. A., & Nicely, P. (1955). An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America, 27*, 338–352.
- Pavlovic, C. V. (1994). Band importance functions for audiological applications. *Ear and Hearing, 15*, 100–104.
- Pavlovic, C. V. (2006). The speech intelligibility index standard and its relationship to the articulation index and the speech transmission index. *The Journal of the Acoustical Society of America, 119*, 3326.
- Rasipurama, R., & Magimai-Doss, M. (2016). Articulatory feature based continuous speech recognition using probabilistic lexical modeling. *Computer Speech & Language, 36*, 233–259.
- Redford M., & Diehl, R. L. (1999). The relative perceptual distinctiveness of initial and final consonants in CVC syllables. *The Journal of the Acoustical Society of America, 106*, 1555–1565.
- Steeneken, H. J., & Houtgast, T. (1980). A physical method for measuring speech-transmission quality. *The Journal of the Acoustical Society of America, 67*, 318–326.
- Stevens, K. N. (2002). Toward a model for lexical access based on acoustic landmarks and distinctive features. *The Journal of the Acoustical Society of America, 111*, 1872–1891.
- Sussman, H. M., McCaffrey, H. A., & Matthews, S. A. (1991). An investigation of locus equations as a source of relational invariance for stop consonant place categorization. *The Journal of the Acoustical Society of America, 90*, 1309–1325.
- Trubetzkoy, N. (1969). *Principles of phonology*. Berkeley: University of California Press. Originally published in 1939 as *Grundzüge der Phonologie*. Travaux du Cercle Linguistique de Prague, 7. Prague.
- van Wassenhove, V., Grant, K. W., & Poeppel, D. (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences, 102*, 1181–1186.