Power spectral entropy as an information-theoretic correlate of manner of articulation in American English

Fernando Llanos[a)]
School of Languages and Cultures, Purdue University, 640 Oval Drive, West Lafayette, Indiana 47907, USA
fllanos@utexas.edu

Joshua M. Alexander
Department of Speech, Language, and Hearing Sciences, Purdue University, 715 Clinic Drive, West Lafayette, Indiana 47907, USA
alexan14@purdue.edu

Christian E. Stilp
Department of Psychological and Brain Sciences, University of Louisville, 308 Life Sciences Building, Louisville, Kentucky 40292, USA
christian.stilp@louisville.edu

Keith R. Kluender
Department of Speech, Language, and Hearing Sciences, Purdue University, 715 Clinic Drive, West Lafayette, Indiana 47907, USA
kkluender@purdue.edu

Abstract: While all languages differentiate speech sounds by manner of articulation, none of the acoustic correlates proposed to date seem to account for how these contrasts are encoded in the speech signal. The present study describes power spectral entropy (PSE), which quantifies the amount of potential information conveyed in the power spectrum of a given sound. Results of acoustic analyses of speech samples extracted from the Texas Instruments–Massachusetts Institute of Technology database reveal a statistically significant correspondence between PSE and American English major classes of manner of articulation. Thus, PSE accurately captures an acoustic correlate of manner of articulation in American English.

[AL]
Date Received: June 19, 2016 **Date Accepted:** October 31, 2016

1. Introduction

In speech science, manner of articulation typically refers to the type of aerodynamic obstruction imposed in the vocal tract to the airstream initiated at the trachea during the act of speaking (Catford, 1977; Ladefoged and Johnson, 2014; Hewlett and Beck, 2013). The ability to produce and distinguish speech sounds by their manner of articulation plays an active role in everyday speech communication, and all known languages in the world contain speech contrasts based on this feature (Ladefoged and Maddieson, 1998). Although all speech sounds can be grouped in several major classes with respect to their manner of articulation (e.g., in American English: vowels, approximants, nasals, fricatives, affricates, and stops), a definitive acoustic correlate of this feature has been very elusive. As a consequence, manner of articulation is typically described as a combination of multiple acoustic attributes, such as segmental duration, signal amplitude, or the frequency of specific acoustic resonances (i.e., formants). In isolation, none of these acoustic attributes seems to capture the difference between major classes of manner of articulation. For instance, although intensity tends to increase for sonorant sounds, nasals (which are sonorant) are not always produced with higher intensity than fricatives (which are not sonorant) (e.g., Fletcher, 1953). Despite this, listeners seem to be able to rely on qualitatively different acoustic properties to decode manner of articulation depending on the phonetic quality of the speech contrast.

[a)]Also at Communication Sciences and Disorders, University of Texas at Austin, 2504A Whitis Avenue, Austin, Texas 78712, USA. Author to whom correspondence should be addressed.

In the present study we introduce an information-theoretic characterization of manner that does not focus on the identification of local acoustic properties in the spectrum but on the overall distribution of spectral power. Specifically, it is hypothesized that speech production leaves a manner-specific trace in the distribution of spectral power along the decibel range that can be appropriately quantified by the Shannon entropy formula (Shannon, 1949).

The formula for Shannon entropy [Eq. (1)] estimates the average number of bits of information required to efficiently code the outcome of a source of potential events $S = \{s_i, ..., s_m\}$ based on their probability of occurrence $P = \{p(s_1), ..., p(s_m)\}$. Typically, the higher the entropy, the more bits are required on average to code the event. Also, when entropy is high, the outcome of the source is more uncertain because is not biased toward any particular event. Thus, Shannon entropy maximizes when the outcome is uniformly distributed, or dispersed, across all the potential events (i.e., all events are equally likely and thus equally informative)

$$H(P) = -\sum_{i=1}^{m} p(s_i)\ \log_2\{p(s_i)\}. \tag{1}$$

We quantified the entropy of the distribution of power along the decibel range, as it was provided by the discrete Fourier transform (DFT) of different English sounds. Average entropy across sounds sharing the same manner of articulation was then used to index the amount of information, in bits, of the corresponding class. This approach was motivated by the way in which spectral power seems to be distributed along the decibel range as a function of manner of articulation.

This distributional pattern is illustrated in Fig. 1, which contrasts the distribution of spectral power of English sounds produced with a different manner: stop [b], fricative [ʃ], approximant [ɹ], and vowel [ae]. In Fig. 1, the distribution of spectral power along the decibel range becomes more uniform (higher entropy) as the overall degree of spectral prominence gradually increases from stop [b] to vowel [ae]. Here, spectral prominence refers to degree of resistance, or loss, of the acoustic system defined by the vocal tract configuration characteristic of the sound class (Clements, 2009). Sounds with a low degree of resistance (e.g., vowels) are characterized by a slow decay of formant oscillation, which is manifested in the spectrum as a reduction in formant bandwidth and an increment in spectral kurtosis (i.e., more sharply peaked formants). On the contrary, sounds with a high degree of resistance (e.g., stop consonants) are characterized by a faster decay of formant oscillation that increases formant bandwidth and leads to a flatter spectrum and a lower spectral kurtosis. This relationship between manner of articulation, spectral prominence, and the distribution of power along the decibel range motivated the usage of the Shannon entropy formula that is described in Sec. 2.

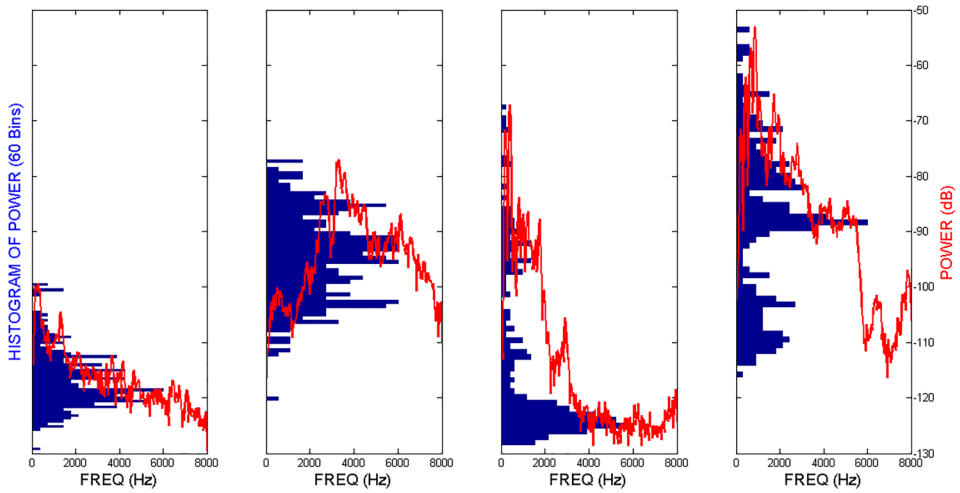

Fig. 1. (Color online) Relationship between spectral power distribution and manner of articulation. DFT spectrums of four 25-ms windows centered at the mid-time point of four corresponding English sounds gradually increasing in spectral prominence, from the leftmost to the rightmost panel: [b] (stop), [ʃ] (fricative), [ɹ] (approximant), and [ae] (vowel). A re-scaled 60-bin histogram of spectral power computed in decibels is projected along the *y* axis of each panel. Histograms are used to visualize the distribution of power in the DFT spectrum, which becomes gradually more uniform (or less peaky) from the leftmost to the rightmost panel, thus increasing the entropy of the corresponding distribution.

2. Methods

2.1 Power spectral entropy (PSE)

PSE [Eq. (2)] was estimated as follows. Let $S = \{dB_1, dB_m\}$ be a partition of the range of decibels of a DFT spectrum with N points of frequency resolution, where dB_i is the number of frequencies with a decibel value falling within the interval $I_i = (dB_i - \varepsilon, dB_i + \varepsilon)$, for an arbitrary ε. Next, S was converted into a probability distribution function of power $P = \{p(dB_1), ..., p(dB_m)\}$, in which $p(dB_i) = dB_i/N$ denotes the probability of having a frequency with a decibel value falling within the interval I_i. PSE was then computed by submitting P to the Shannon information formula [Eq. (1)]. Following this method, PSE maximizes when power is uniformly distributed across all the intervals $I = \{I_1, ..., I_m\}$

$$H(P) = -\sum_{i=1}^{m} p(dB_i)\ \log_2\{p(dB_i)\}. \quad (2)$$

Equation (2) was inspired by previous applications of information theory to the analysis of speech production and processing (Lufti, 1992; Rallapalli and Alexander, 2015; Stilp and Kluender, 2010). However, PSE was designed to quantify the entropy of the distribution of spectral power in the decibel—not the frequency—domain. From this perspective, speech sounds in which spectral power tends to concentrate around the same decibel values in the spectrum (i.e., sounds that are spectrally less prominent, such as stop and fricative consonants) are expected to have a lower degree of PSE. This means that the encoding of power in these sounds requires a lower number of bits of information because its outcome in decibels is more predictable and thus less informative. For example, white noise has a relatively flat spectrum, but a concentration of spectral power over a narrow range, resulting in very low entropy.

PSE increases in proportion to the number of intervals needed to cover the range of spectral decibels characteristic of the sound. Since the width of this range may vary according to factors (e.g., signal amplitude or the vocal tract physiology of the talker) that are not directly related to the distribution of spectral power within that range, the constant ε that determines the width of each interval $I_i = (dB_i - \varepsilon, dB_i + \varepsilon)$ was systematically modified for each DFT spectrum to achieve a total of 60 intervals or bins. This adjustment allowed us to minimize changes in PSE due to factors extrinsic to the distribution of spectral power that is characteristic of each sound. This number of bins approximated quite well the average range of decibels covered by the spectrum of all sounds included in the Texas Instruments–Massachusetts Institute of Technology (TIMIT) database, which was approximately 60 dB. Preliminary analyses using a different number of intervals (e.g., 30 and 15 intervals) revealed no substantial differences in terms of relative PSE across sounds.

2.2 Speech samples

Speech sounds were extracted from the TIMIT database. It contains a total of 6300 sentences recorded from 630 speakers from eight major dialects of American English (New England, Northern, North Midland, South Midland, Southern, New York City, Western, and Army Brat). Recordings were digitalized at 16 kHz. Text materials prompted to speakers consisted of two dialect sentences, 450 phonetically compact sentences, and 1890 phonetically diverse sentences. Each speaker read the two dialect sentences, five phonetically compact sentences, and three phonetically diverse sentences (Garofolo *et al.*, 1993; Zue *et al.*, 1990). The sets of sentences read by each talker was designed to maximize the number of different sentences read across talkers. All sentences were phonetically labeled by expert phoneticians.

Table 1 shows the 58 speech sounds that were included in the study, arranged into the six major natural classes proposed in the documentation of the TIMIT database. These classes were vowel, approximant, nasal, fricative, affricate, and stop. This selection of classes provides a feasible ranking of manner for American English, with natural classes gradually decreasing in terms of the average expected degree of spectral prominence (cf. Clements, 2009). Also, these classes are common in broad classifications of manner proposed for American English (e.g., Ladefoged and Johnson, 2014; Catford, 1977).

Following the phonetic criterion of the TIMIT database (cf. Garofolo *et al.*, 1993), English flaps were grouped with either stops ([ɾ], as in *dirty*) or nasals ([n̆], as in *winner*) depending on whether they presented or not a nasal closure. Similarly, glottal fricatives [h] (as in *hey*) and [ɦ] (as in *ahead*) were placed into the approximant class because of their type of oral constriction, which is wide enough as to avoid the noise

Table 1. Speech sounds included in the analysis.

		Approximant					
Vowel		Glide	Liquid	Nasal	Fricative	Affricate	Stop
i	ə	w	l	m	s	tʃ	p
ɪ	o	j	ɹ	n	z	ʤ	t
ɛ	u	ɦ	ɺ	ŋ	f		k
æ	ɨ	h		ɳ	ʃ		b
ɑ	aɪ			ɳ̩	ʒ		d
ɔ	aʊ			m̩	*θ*		g
ʌ	ə̹			ɾ̃	ð		ɾ
ʊ	e				v		ʔ
ᵊ	oɪ						
ᵻ	ụ						

characteristic of fricatives and thus generate approximant-like spectral resonances. For similar reasons, no major distinctions were made between lateral and central approximants, which tend to exhibit a similar degree of spectral prominence (Clements, 2009). Analyses of PSE covered a total of 189,354 speech tokens.

2.3 Acoustic measurements

Every speech token was split into a series of 20-ms consecutive time frames, with 50% overlap at the 16-kHz sampling rate. This particular selection of frame length and overlap was chosen so that spectral changes over time could be adequately defined for most speech sounds. For example, vowel formants may systematically vary over the course of time (Hillenbrand and Nearey, 1999). Similarly, both stop and affricate consonants tend to exhibit rapid spectral changes before and after the consonant release that may not be captured by the standard long-term Fourier analysis. Therefore, the use of a short-term DFT analysis allowed us to incorporate part of the variation of PSE across frames into the analysis.

PSE was estimated from the DFT spectrum of each time frame as specified by Eq. (2). The frequency resolution of the DFT was set to the next power of 2 (512 points) after the frame length (320 points). Power spectral entropy for each token was computed as the average entropy across all time frames within the token.

3. Analysis and results

Values of PSE (bits of information) for each token were submitted to a mixed-effects linear regression model with manner class as the predictor variable (coded categorically with six levels for the six phonetic manner classes) and entropy measures as the dependent variable. The model included a random effect of talker, with random intercepts for different talkers in the TIMIT database. Degrees of freedom and *p*-values were estimated using the Satterthwaite approximation (lmerTest package in R). However, this model structure only tests for differences between each manner class and one class designated as the default level (here, stops). While large differences in entropy were expected to be statistically significant, differences between manner classes with neighboring values of entropy were of primary interest. These pairwise contrasts were conducted using the multcomp package in R. Mean entropy measures significantly differed between stops and affricates (Wald test: $Z = 1391.37$), affricates and fricatives ($Z = 17.91$), fricatives and nasals ($Z = 22.47$), nasals and approximants ($Z = 8.95$), and approximants and vowels ($Z = 70.88$; all $p < 2 \times 10^{-16}$, Bonferroni-corrected for multiple comparisons). Therefore, manner classes were ranked as follows, from higher to lower mean PSE: vowel > approximant > nasal > fricative > affricate > stop. Table 2 provides the mean number of bits, standard deviation, and number of tokens for each major class.

In addition to the analysis of PSE for each major class, PSE was also estimated for each individual speech sound included in the analysis (Table 1). This complementary analysis allowed a more detailed exploration of PSE within each major class. For instance, some finer classifications of manner in English distinguish between lateral approximants (e.g., [l] in *lateral*) and central approximants (e.g., [ɹ] in *raw*, and [j] in *yes*) (Ladefoged and Johnson, 2014). Table 3 shows all the speech sounds included in Table 1 ranked by PSE.

Table 2. Descriptive statistics of power spectral entropy (in bits of information) for each class of speech sounds. Number of tokens refers to how many speech sounds from each class were included in the analyses.

Natural Class	Mean	Standard Deviation	Number of Tokens
Vowel	5.32	0.11	78070
Approximant	5.23	0.14	27742
Nasal	5.22	0.15	18748
Fricative	5.19	0.14	2840
Affricate	5.13	0.11	2662
Stop	5.08	0.17	33731

4. Discussion

This study revealed the existence of a significant statistical correspondence between power spectral entropy (PSE) and the six major classes of manner of articulation in American English. This correspondence suggests that PSE could be a suitable correlate of manner of articulation. In particular, PSE gradually decreased from vowels to stop consonants according to the following ranking: vowel > approximant > nasal > fricative > affricate > stop. This ranking can be interpreted as follows. As the degree of spectral prominence gradually increased from stop consonants to vowels the distribution of power in the spectrum became more uniformly distributed across the decibel range, thus raising the overall level of PSE of the corresponding class. From an information-theoretic perspective, vowels (highest entropy) could be considered more informative than consonants, in that the efficient encoding of their power spectral density would require more bits of information than for consonants.

The ranking of major classes predicted reasonably well the pattern of phonotactic variation of speech sounds within the syllable. According to this pattern, the position of nucleus is typically occupied by vowels, followed by approximants and nasals. On the contrary, the margins of the syllable tend to be occupied by stops, affricates and fricatives (for a detailed review of this topic see Blevins, 2003; Kawasaki-Fukumory and Ohala, 1997; Parker, 2008). Our results suggest that the structure of the syllable in English could be broadly outlined as a fluctuation of information from the nucleus (higher entropy) to the syllable boundaries (lower entropy). This is illustrated in Fig. 2, which shows the fluctuation of PSE across different syllables within the same English sentence.

Interestingly, the difference between the nasal and approximant class in terms of PSE was considerably smaller than the one between any other two classes. This suggests that the degree of spectral prominence in nasals and approximants might be quite similar. In fact, in the ranking of speech sounds (Table 3), some nasals (nasal flap [ɾ̃] and syllabic nasals [n̩] and [m̩]) ranked higher than many approximants, although these reversals in individual speech sounds did not have a strong impact on the ranking of major classes.

Two other interesting reversals in our data were (1) back vowels ([u], [ɔ], [o]) and labial consonants ([w], [v], [f]), which ranked lower than their phonetic peers in the

Table 3. Ranking of speech sounds by power spectral entropy.

Rank	Sound	Entropy	Rank	Sound	Entropy	Rank	Sound	Entropy	Rank	Sound	Entropy
1	æ	5.376	14	ʊ	5.308	27	ŋ̩	5.236	40	tʃ	5.158
2	aɪ	5.371	15	oɪ	5.307	28	m̩	5.233	41	ʤ	5.125
3	aʊ	5.361	16	o	5.301	29	l	5.228	42	w	5.123
4	ɛ	5.358	17	j	5.300	30	m	5.226	43	k	5.110
5	e	5.353	18	ɦ	5.282	31	ŋ	5.225	44	ð	5.100
6	ʌ	5.350	19	ɝ	5.281	32	n	5.219	45	v	5.089
7	ɪ	5.336	20	h	5.279	33	ɾ	5.217	46	g	5.082
8	ɑ	5.333	21	ɚ	5.263	34	n̩	5.216	47	f	5.081
9	i	5.329	22	u	5.261	35	ʒ	5.214	48	θ	5.080
10	ɨ	5.323	23	s	5.259	36	ə̹	5.210	49	t	5.080
11	ʉ	5.318	24	ɹ	5.259	37	ʃ	5.206	50	d	5.011
12	ɾ̃	5.317	25	ɔ	5.254	38	l̩	5.197	51	p	5.000
13	ə	5.309	26	z	5.250	39	ʔ	5.193	52	b	4.887

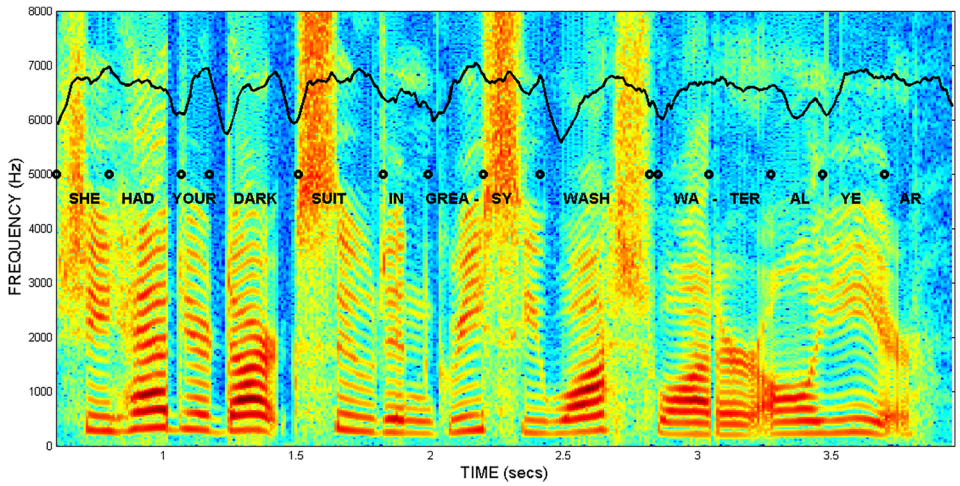

Fig. 2. (Color online) Information-theoretic encoding of syllable phonotactics. Spectrogram of the English utterance: *she had your dark suit in greasy wash water all year*, extracted from the TIMIT database, with syllable boundaries (dark circles) and fluctuations of power spectral entropy over time (black curve).

vowel and fricative class, and (2) grooved sibilants ([s], [z]), which ranked higher than their phonetic peers in the fricative class. Again, none of these reversals seemed to have a strong impact in the ranking of major classes. However, they may inform about specific articulatory gestures that could contribute to amount of variation in PSE within each natural class. For example, in the case of back vowels and labial consonants, the decay of PSE could be related to the gesture of labialization involved in the articulation of these sounds (back vowels tend to be labialized, or rounded, in English; Diehl, 2008; Flemming, 2004). With respect to the grooved sibilants ([s], [z]), their greater amount of PSE could be a consequence of how the back of the tongue is curved in the production of these consonants (Fletcher and Newman, 1991), creating a greater aerodynamic channel that could be responsible for the increment of entropy. Similarly, the ranking of individual sounds suggests that voiceless sounds (e.g., [p], [t], [k]) tend to rank higher in entropy than their voiced counterparts (e.g., [b], [d], [g]). The real contribution of these articulatory gestures related to the degree of vocal stricture is difficult to estimate from the results of the present study, which are informed by acoustic analysis. Similarly, although each of the English central approximants [j] and [ɹ] ranked significantly higher in PSE than the English lateral approximant [l] ($p < 0.05$), it is difficult to estimate the exact contribution of the articulatory gestures involved in the production of this contrast.

Another aspect that would be worthwhile to explore in future research is whether the general ranking of manner classes reported for American English can be extrapolated to other languages. Although manner of articulation is sometimes described in terms of cross-linguistically invariant phonological features (e.g., [continuant], [syllabic]; Clements, 1985), the acoustic relationship between phonetic and phonological features is rarely one-to-one, and languages may rely on different phonetic properties to encode the same phonological contrast. Furthermore, although the American English speech-sound inventory is phonetically quite rich, it lacks several speech sounds and manners of articulation that could not be examined in the present study, such as trills (e.g., [r]) or retroflex consonants (e.g., [ɖ]). Similarly, the number and type of speech sounds included in each major class of manner may vary across languages and authors, and these are factors that might give rise to several types of cross-linguistic differences in terms of PSE.

References and links

Blevins, J. (**2003**). "The independent nature of phonotactic constraints: An alternative to syllable-based approaches," in *The Syllable in Optimality Theory*, edited by C. Fery and R. van de Vijver (Cambridge University Press, Cambridge, UK), pp. 375–403.

Catford, J. C. (**1977**). *Fundamental Problems in Phonetics* (Midland Books, Bloomington, IN).

Clements, G. N. (**1985**). "The geometry of phonological features," Phonology **2**(1), 225–252.

Clements, G. N. (**2009**). "Does sonority have a phonetic basis?," in *Contemporary Views on Architecture and Representations in Phonology*, edited by E. Raimy and C. E. Cairns (MIT Press, Cambridge, MA), pp. 165–176.

Diehl, R. L. (**2008**). "Acoustic and auditory phonetics: The adaptive design of speech sound systems," Philos. Trans. R. Soc. London B **363**(1493), 965–978.

Flemming, E. (**2004**). "Contrast and perceptual distinctiveness," in *Phonetically-based Phonology*, edited by B. Hayes, R. Kirchner, and D. Steriade (Cambridge University Press, Cambridge, UK), pp. 232–276.
Fletcher, H. (**1953**). *Speech and Hearing in Communication* (D. Van Nostrand Company, New York), pp. 84–86.
Fletcher, S. G., and Newman, D. G. (**1991**). "[s] and [ʃ] as a function of linguapalatal contact place and sibilant groove width," J. Acoust. Soc. Am. **89**(2), 850–858.
Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (**1993**). "DARPA TIMIT acoustic-phonetic continous speech corpus [CD-ROM]," NIST speech disc 1-1.1, NASA STI/Recon Technical Report 93.
Hewlett, N., and Beck, J. M. (**2013**). *An Introduction to the Science of Phonetics* (Routledge, London, UK).
Hillenbrand, J. M., and Nearey, T. M. (**1999**). "Identification of resynthesized/hVd/utterances: Effects of formant contour," J. Acoust. Soc. Am. **105**(6), 3509–3523.
Kawasaki-Fukumori, J., and Ohala, J. (**1997**). "Alternatives to the sonority hierarchy for explaining segmental sequential constraints," in *Language and its Ecology: Essays in Memory of Einar Haugen*, edited by S. Eliasson and E. H. Jahr (Mouton De Gruyter, Berlin, Germany), Vol. 100, pp. 343–366.
Ladefoged, P., and Johnson, K. (**2014**). *A Course in Phonetics* (Nelson Education, Scarborough, ON, Canada), pp. 15–17.
Ladefoged, P., and Maddieson, I. (**1998**). *The Sounds of the World's Languages* (Blackwell, Malden, MA).
Lutfi, R. A. (**1992**). "Informational processing of complex sound. III: Interference," J. Acoust. Soc. Am. **91**(6), 3391–3401.
Parker, S. (**2008**). "Sound level protrusions as physical correlates of sonority," J. Phonet. **36**(1), 55–90.
Rallapalli, V. H., and Alexander, J. M. (**2015**). "Neural-scaled entropy predicts the effects of nonlinear frequency compression on speech perception," J. Acoust. Soc. Am. **138**(5), 3061–3072.
Shannon, C. E. (**1949**). "Communication theory of secrecy systems," Bell Syst. Tech. J. **28**(4), 656–715.
Stilp, C. E., and Kluender, K. R. (**2010**). "Cochlea-scaled entropy, not consonants, vowels, or time, best predicts speech intelligibility," Proc. Natl. Acad. Sci. **107**(27), 12387–12392.
Zue, V., Seneff, S., and Glass, J. (**1990**). "Speech database development at MIT: TIMIT and beyond," Speech Commun. **9**(4), 351–356.