

Pronunciation Variation is Key to Understanding Spoken Language

Steven Greenberg

The Speech Institute
9 Sereno Circle, Oakland, CA 94619 USA

Abstract

Pronunciation variation defies the capability of a strictly alphabetic notation to fully characterize, as the basic parameters governing spoken language transcend the phoneme. The time scales required to accurately describe and model pronunciation are both shorter *and* longer than the phonetic segment, and are inherently wedded to the syllable. The specific properties of phonetic constituents are shaped by two inter-related parameters – prosodic prominence and their ordinal position within the syllable – which ultimately reflect the entropic potential associated with lexemic discriminability. The internal structure of the syllable can be parsimoniously portrayed in terms of articulatory features (e.g., voicing, place and manner of articulation, etc.) and used to characterize the microstructure of pronunciation variation. Parsing the utterance into syllabic units of variable prominence is essential for characterizing spoken language, and of utility for developing robust speech technology.

1 Introduction

Models of spoken language generally ignore pronunciation variation in favor of more abstract (and idealized) conceptualizations derived from laboratory and elicitation studies. Although such models may be useful in the classroom, they are of limited utility for developing speech technology because of the artificial nature of the materials involved. Language, as spoken in the “real world,” is fundamentally communicative in function. It is intended to convey desires, emotions and information that facilitate behavioral interaction. Within this context, the phonetic and prosodic variation observed may be as informative as the literal message itself. Such variation can be used as an interpretative framework with which to decode the speaker’s intent [5].

Despite the ubiquity of pronunciation variation in the spoken world, its linguistic significance is poorly understood. Utterances are generally represented as sequences of words, and words as strings of phones (or phonemes). The specific way in which the words are spoken is rarely an occasion for comment or concern.

The orthography used to represent words on the printed page also diminishes the importance of pronunciation variation. There is often but a single “correct” way to spell a word, and the amount of orthographic variation tolerated is extremely limited. And even publications that focus on pronunciation variation (e.g., pronouncing dictionaries) rarely list more than a few variants of a “canonical” pronunciation. In this sense, orthography is destiny, as it downplays the importance of variability in pronunciation.

Within the alphabetic tradition it is commonplace to represent a word as a mere sequence of discrete sounds (irrespective of the sound/symbol correspondence pattern). In the International Phonetic Alphabet (IPA) and other phonetic orthographies, each sound is associated with a distinctive symbol designed to retain its essential identity independent

of context. Although this orthographic convention has a singular advantage in terms of representational parsimony, it may also be a liability with respect to descriptive precision and accuracy if the phenomena under study conform poorly to the orthographic representation used.

It is the thesis of this paper that the sorts of pronunciation variation characteristic of spoken language defy the capability of a (strictly) alphabetic notation to adequately characterize, as the basic parameters governing pronunciation (and its variability) transcend the phone(me). The time scales required to accurately describe and model pronunciation are both shorter *and* longer than the phonetic segment, and are inherently wedded to the syllable.

2 The Phenomenological Unreality of the Phoneme

The phoneme is the phonological foundation of linguistic theory and practice. It is the primary means by which words are represented in scientific discourse, as well as in technology. For example, automatic speech recognition (ASR) systems generally represent words as strings of (context-dependent) phonemic elements, which are used to decode the acoustic properties of the speech signal. Much of current text-to-speech (TTS) technology also uses (context-dependent) phonemes as a basic unit to represent words.

Unfortunately, words are not spoken in terms of phones (or phonemes). In spontaneous dialogues (from the Switchboard corpus [6]) nearly a quarter of the phonemic elements (based on the canonical pronunciation) are “missing,” as the speaker has “deleted” such segments from the spoken representation (i.e., the speech signal) [7]. Another 30% of the phones are pronounced in a manner that differs substantially from the phonemic (i.e., canonical) representation [7]. If the dialogue material in Switchboard is representative of the genre, nearly half of the phonemic elements of spontaneous speech are not realized in the form classical representations of speech would imply. Perhaps this is why ASR systems have such difficulty decoding spoken language representative of the real world – many of the elements have gone missing or “undercover.”

3 The Syllable as an Island of Stability

In contrast to the ephemeral nature of the phone(me), stands the syllable. In Switchboard, less than 1% of all syllables are deleted [7]. Although the specific phonetic properties of a syllable may differ from the canonical, it retains its basic function within spoken discourse.

The syllable is not only important because of its articulatory stability, it also shapes the character of the phonetic constituents through two related mechanisms.

The first pertains to the ordinal position of a constituent relative to the syllabic nucleus. Consonant segments preceding the nucleus behave differently from those that follow. With reference to the canonical form, onset consonants tend to conform far more closely to their idealized representation

than coda constituents [10][11]. In some sense, coda segments are but pale shadows of their onset counterparts, and it does an injustice to this phonetic nuance to represent such constituents with precisely the same symbol irrespective of their position within the syllable.

The second mechanism pertains to the accentuation (or prominence) imparted to a syllable. Some syllables are heavily accented, while others are much less so (or not at all). Many phonetic properties of constituents within the syllable depend on the level of accentuation (see Section 6 for details).

The syllable's integrity is also crucial for intelligibility. Blurring the boundaries of contiguous syllables effectively destroys the capability of understanding spoken language [4] and has a significant impact of the speech signal's modulation spectrum (i.e., the low-frequency fluctuation pattern of energy as a function of time) [16]. There appears to be a close relationship between energy modulation, syllable duration and intelligibility [7].

4 Articulatory Feature Structure Within the Syllable

Although the syllable is most often described in terms of its segmental constituents, it is also possible to analyze its structure in terms of articulatory-feature dimensions. Such a fine-grained perspective provides additional insight into the nature of pronunciation variation at the level of the syllable.

4.1 Voicing

Voicing (i.e., quasi-periodic vibration of the glottis during phonation) serves as the articulatory foundation of the syllabic nucleus. It spreads outward from the core of the nucleus (the "peak") towards the onset and coda. This core is usually vocalic, but can also assume other sonorant forms (e.g., a syllabic nasal or liquid). Once voicing ceases its flow beyond the core, it does not resume *within* syllable. In this sense, voicing is controlled more at the syllabic than at the segmental level. Consistent with this observation is the complete devoicing of certain segments, particularly in coda (e.g., devoiced [z] in the Switchboard corpus) and word-final position.

Voicing serves three principal functions: (1) it builds the syllable's amplitude up through the core (i.e., peak), (2) provides a quasi-harmonic spectral structure for robustness in noise and reverberation, and (3) serves as the structural foundation for intonation and other pitch-related information [9].

4.2 Place of Articulation

Place-of-articulation (i.e., the locus of maximum constriction during production) is an inherently trans-segmental feature that binds the nucleus with the onset or coda. Acoustically, it is often associated with formant transitions connecting a consonant with a vowel. Although it is possible to deduce place of articulation solely from the consonantal release [21], a true sense of consonantal *identity* emerges only upon its conjunction with the nucleus.

Although there are many potential places of articulation in any given language (e.g., English has ca. 10), it is unusual for more than three or four to be associated with any single manner class [3]. In this sense, manner and place of articulation are inextricably linked.

Articulatory place cues serve to distinguish among words, particularly at onset [13], and constitute the most important feature set for lexical discriminability [18]. In the onset such fea-

tures are statistically distributed relatively evenly across the place dimension (from anterior to posterior loci of constriction) [10][11], a pattern consistent with their discriminative function.

The formant transitions binding the onset to the nucleus (and the nucleus to the coda) serve an important perceptual function – they serve to enhance the perceptual coherence of spectrally disparate elements as a single linguistic entity. They also serve as acoustic "scaffolding," guiding the ascending trajectory of energy towards the nucleic peak, as well as its descent to the syllable's rump. This "energy arc" provides the cohesive force required to meld phonetic constituents into a temporal compound, otherwise known as the syllable.

4.3 Manner of Articulation

In contrast to voicing and articulatory place, manner of articulation is inherently discrete in time [3]. It is rare for more than a single mode of production (associated with stops, nasals, fricatives, vowels and the like) to operate concurrently (although certain languages, such as French and Portuguese, may combine certain properties of different modes, as occurs in nasalized vowels). This temporally discrete characteristic of articulatory manner is the one closest to the concept of the phone, providing a principled method to segment the syllable into its constituent parts [3].

It is rare for two segments of the same manner class to occupy contiguous positions within the syllable [2][10]. Why should this be so? One possibility pertains to the energy arc. Each mode of articulation is associated with an overall level of energy, vowels being the most intense, followed by liquids, semi-vowels and nasals (these are the classic sonorants). Of lesser intensity are the manner classes, such as stops, fricatives and affricates, in which voicing is optional. Clearly, there is a gradient of energy associated with different manner classes that may underlie the phonotactic constraints observed within the syllable (a.k.a. the "sonority hierarchy" e.g., [22]). Might not the sequence of articulatory manner modes observed within the syllable provide the biomechanical means with which to shape the energy trajectory into an arc synchronized to the syllable?

4.4 Interaction Among Articulatory Features

Traditionally, the articulatory features associated with voicing, place and manner of articulation, have been treated as independent dimensions [17]. Examination of the Switchboard corpus suggests otherwise. Voicing appears to be linked to both articulatory manner and place features. The patterns of deviation from canonical pronunciation indicate that voicing rarely deviates from the canonical unless manner and/or place features do as well (but not vice versa) [8]. Clearly, voicing, place and manner are under some form of central nervous control that is probably organized at the level of the syllable (rather than at the level of the phone).

5 Phonetic Segmentation Within the Syllable

The energy arc, in tandem, with manner-of-articulation classifiers, provides a principled means with which to phonetically segment the syllable into its underlying phonetic constituents. In theory, it should be possible to estimate the number of phones within a syllable, as well as specify their identity through a hierarchical network of classifiers trained to recognize a variety of articulatory dimensions (e.g., [3]). Unfortunately, this objective is unlikely to be met *exclusively* through the use of articulatory-feature classifiers, irrespec-

tive of their nominal efficacy. An important ingredient is missing, one that is required to interpret the acoustic signal within the appropriate linguistic framework.

6 The Importance of Syllabic Prominence

The missing ingredient is syllable prominence, associated with the prosodic concept of “accent.” Syllables vary in perceived prominence across an utterance in a manner roughly proportional to the information contained within. Certain languages, such as Japanese, nominally instill prominence largely through pitch variation (“pitch accent”)[1], while others (such as English and German) use a broad constellation of acoustic properties (e.g., duration, energy, vowel quality, and pitch) to emphasize syllables (“stress accent”)[1]. Irrespective of its acoustic basis, prominence strongly impacts the phonetic realization of constituents within the syllable [10][11].

In English, the vowel system varies, depending on the stress-accent magnitude of the syllable [14][15]. In heavily accented syllables, the full spectrum of vocalic identity is employed, while in their unaccented counterparts the vowel space is largely confined to the high-front and high-central regions of the articulatory space [14][15]. The low and back vowels are hardly present in unaccented syllables (which account for 40% of the Switchboard corpus) [10]. Not only identity, but also duration varies as a function of prominence; vowels in accented syllables are, on average, twice as long as their unaccented counterparts.

The onset and coda constituents also vary as a function of syllabic prominence, but in different ways. The primary impact of accent on onsets is via duration [12]. In heavily accented syllables such segments are 40-60% longer, on average, than their unaccented counterparts [12]. With respect to segmental identity, there is relatively little impact of prominence on onsets (except for a somewhat higher deletion probability for segments like [ð], which occur in demonstratives, pronouns and articles such as “them,” “these,” “that” and “the”) [10].

Prominence affects the coda in a somewhat different way. The overwhelming majority of coda segments are coronals [11]. The probability of coda deletion is considerably higher for such segments in unaccented syllables relative to their accented counterparts [10]. However, non-coronal segmental identity is largely unaffected by prominence [10][11]. Nor are the durational properties of codas significantly affected significantly by accent [12].

Such patterns imply that models of pronunciation variation would profit from deeper insight into the manner in which syllable structure and prominence interact [5].

7 The Importance of “Information” for Phonetic Patterning Within and Across Syllables

The patterns of pronunciation described in Sections 4 and 6 suggest that “information” is the single most important parameter controlling the phonetic realization of spoken language, particularly within the context of spontaneous interactions. It would be difficult to interpret the acoustic signal with precision in the absence of such knowledge. Prominence provides one metric with which to gauge the amount of information associated with a given syllable, and thereby provides an entropy-based framework for decoding the acoustic signal into units tied to the word, syllable, segment and articulatory feature (and may provide a principled basis for “underspecification” in phonological theory).

8 Segmentation Across Syllables

This information-centric approach focuses on the relation among syllables over the course of an utterance. Some syllables are more prominent than others. Typically, heavily accented syllables are surrounded by less prominent ones [7]. Often, the phonetic demarcation between syllables reflects the associated prominence pattern [10], particularly if they are bound to the same word or lexemic compound.

The concept of “juncture” is particularly useful in this context. The juncture separating syllables often provides important clues as to their prosodic and entropic relationship [10]. In English, contiguous syllables within a word are often separated by a “flap” (as in the words “winner” and “rider”), which is generally treated as an ambisyllabic segment. However, close inspection of such “pure” junctures makes it clear that a flap in this context is not really a segment, but rather serves to separate an accented syllable from an unaccented one – it is simply a juncture that lies betwixt and between two syllables of unequal prominence [10]. Other forms of juncture serve to separate syllables of more equal prosodic status.

Clear demarcation between syllables is essential for spoken language, and provides a means by which to understand pronunciation variation within a broader context conducive to advancements in speech technology.

9 The Modulation Spectrum and the Energy Arc

The low-frequency modulation spectrum is the acoustic manifestation of syllables in spoken discourse [7]; its integrity is essential for intelligibility. But what underlies the modulation spectrum? And why is it so important for speech understanding?

The human cortex appears to require that information be packaged in syllable-length chunks for proper decoding [9]. The auditory, visual and motor systems all operate at rates consistent with that of spoken language [9]. In order for a syllable to carry a significant amount of information it is essential that both the spectrum and time be divisible into smaller units that are typically associated with segments and articulatory features [9]. But these syllabic constituents must conform to the energy arc in order for the information they contain to be readily digested by the auditory system as well as the brain. Many phonetic properties of spoken language can be understood within this information-centric perspective, and the constraints imposed by this framework can be used for developing realistic pronunciation models for ASR and TTS applications.

10 Pronunciation Models and Speech Technology

Much of current automatic-speech-recognition and text-to-speech technology treat the phone(me) as a principal unit of spoken language, despite ample evidence to the contrary. The phone is ill-suited for characterizing pronunciation variation because of its egalitarian approach to entropic distribution. Phonetic and prosodic analysis of spontaneous discourse demonstrate that not all phones are created equal, and that a representational framework adapted to the exigencies of spoken language would enhance the capability of capturing the full range of pronunciation variability observed in the real world.

Such a framework would give precedence to the syllable over the phone(me). Moreover, it would place great emphasis on segmentation, identifying individual syllables through

temporal demarcation of the vocalic nucleus [3][20] as well as delineating the phonetic boundaries that pertain within and across syllables [3][19]. For example, knowledge of a segment's position within the syllable greatly enhances machine-based phonetic and articulatory-feature classification, especially in noisy backgrounds [2], consistent with the notion that phonetic properties vary systematically as a function of syllable structure. Multiple-time-scale windows of analysis could thus be utilized to selectively extract information pertaining to voicing (ca. 20-40 ms), manner (ca. 60-160 ms) and place (ca. 120-160 ms). Each syllable would be evaluated relative to its neighbors with respect to prominence, and this information used to interpret the phonetic constituents contained within. ASR applications, in particular, could exploit structural constraints of the sort described in this paper, as a means of pruning phonetic constituents so as to more faithfully decode the speech signal and thus lessen a system's dependence on training material. TTS applications could utilize these same structural constraints for developing more robust concatenative models, as well as combining a statistical approach with more flexible methods for generating speech capable of simulating a wide range of voice qualities and speaking styles.

11 Understanding Pronunciation Variation

Pronunciation variation is crucial for understanding spoken language. Answering such questions as:

- (1) What are the basic units of speech if not the phone?
- (2) How do the phonetic and prosodic properties of the speech signal interact?
- (3) What is the relation between the information contained in the speech signal and its phonetic/prosodic properties?
- (4) How does the presence of visual speech cues affect pronunciation in spontaneous interactions?
- (5) How can phonetic and prosodic knowledge be used effectively to enhance the quality of speech technology?
- (6) Can speech technology applications be used as a "proving ground" for testing the validity of pronunciation models?

may serve to place the science and technology of spoken language on empirically firmer ground, and thereby foster a golden age for speech-related research.

Acknowledgements

I thank Shawn Chang, Björn Granström and Marion Jäger for their comments on a previous version of this paper.

References

- [1] Beckman M. *Stress and Non-Stress Accent*. Dordrecht: Fortis, 1986.
- [2] Chang, S. *A Syllable, Articulatory-Feature, and Stress-Accent Model of Speech*. Ph.D. Thesis, University of California, Berkeley. ICSI Technical Report 2002-007; www.icsi.berkeley.edu/techreports/2002.html, 2002.
- [3] Chang S., Greenberg S., Wester, M. "An elitist approach to articulatory-acoustic feature classification," Proc. 7th Eur. Conf. Speech Comm. Tech. (Eurospeech-2001), pp. 1725-1728, 2001.
- [4] Drullman R., Festen J.M., Plomp R. "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. 95: 1053-1064, 1994.
- [5] Firth, J.R. "Sounds and prosodies," Trans. Philological

- Soc., 1948, pp. 127-152 [reprinted in J.R. Firth, *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press, pp. 121-138].
- [6] Godfrey, J.J., Holliman, E.C., McDaniel, J., "SWITCHBOARD: Telephone speech corpus for research and development," Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc. (ICASSP-92), pp. 517-520, 1992.
- [7] Greenberg, S. "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication* 29: 159-176, 1999.
- [8] Greenberg, S. "Understanding spoken language using statistical and computational methods," Presentation at Workshop on Patterns of Speech Sounds in Unscripted Communication - Production, Perception, Phonology, Akademie Sankelmark, 2000.
- [9] Greenberg, S., Ainsworth, W.A. "Speech processing in the auditory system: An overview," In Greenberg, S., Ainsworth, W.A. (eds.) *Speech Processing in the Auditory System*. New York: Springer-Verlag, 2003.
- [10] Greenberg, S., Carvey, H., Hitchcock, L., Chang, S. "Beyond the phoneme - A juncture-accent model for spoken language," Proc. Second International Human Language Technology Conference, pp. 36-43.
- [11] Greenberg, S., Carvey, H., Hitchcock, L., Chang, S. "The phonetic patterning of spontaneous American English discourse," Proc. ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, 2003.
- [12] Greenberg S., Carvey H., Hitchcock L., Chang, S. "Temporal properties of spontaneous speech - A syllable-centric perspective," *J. Phon.*, in press.
- [13] Greenberg, S., Chang, S. "Linguistic dissection of switchboard-corpus automatic speech recognition systems," Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium, pp. 195-202, 2000.
- [14] Greenberg, S., Chang, S., Hitchcock, L. "The relation between stress accent and vocalic identity in spontaneous American English discourse," Proc. ISCA Workshop on Prosody in Speech Recognition and Understanding, pp. 51-56, 2001.
- [15] Hitchcock, L., Greenberg, S. "Vowel height is intimately associated with stress accent in spontaneous American English discourse," Proc. 7th Eur. Conf. Speech Comm. Tech. (Eurospeech-2001), pp. 79-82, 2001.
- [16] Houtgast T., Steeneken H. "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *J. Acoust. Soc. Am.* 77: 1069-1077, 1985.
- [17] Jakobson R., Fant G., Halle M. *Preliminaries to Speech Analysis*. Cambridge, MA: MIT Press, 1963.
- [18] Rabinowitz W.M., Eddington D.K., Delhorne L.A., Cuneo P.A. "Relations among different measure of speech reception in subjects using a cochlear implant," *J. Acoust. Soc. Am.* 92: 1869-1881, 1992.
- [19] Shastri L., Chang S., Greenberg S. "Syllable detection and segmentation using temporal flow neural networks," Proc. 14th Int. Cong. Phon. Sci., pp. 1721-1724, 1999.
- [20] Sirigos J., Fakotakis N., Kokkinakis G. "A hybrid syllable recognition system based on vowel spotting," *Speech Comm.* 38: 427-440, 2002.
- [21] Stevens K.N., Blumstein S.E. "Invariant cues for place of articulation in stop consonants," *J. Acoust. Soc. Am.* 64: 1358-1368, 1978.
- [22] Zec D. "Sonority constraints on syllable structure," *Phonology* 12: 85-129, 1995.