THE PHONETIC PATTERNING OF SPONTANEOUS AMERICAN ENGLISH DISCOURSE

Steven Greenberg, Hannah Carvey, Leah Hitchcock and Shuangyu Chang

International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA

ABSTRACT

Statistical analysis of a manually annotated, 45-minute subset of the SWITCHBOARD corpus indicates that pronunciation variation observed in spontaneous American English discourse is highly structured at the level of the syllable, particularly when prosodic stress accent (i.e., syllable prominence) is taken into account. The pattern of segmental substitutions and deletions observed are largely associated with different constituents within the syllable (nuclei and codas, respectively); their frequency of occurrence is inversely proportional to stress-accent magnitude. The phonetic identity of vocalic nuclei is also related to stress accent, as is the probability of segmental deletion in the coda. Such data imply that "information" governs much of the phonetic patterning of spoken language characteristic of the real world.

1. INTRODUCTION

Virtually all of what is known about the phonetics of spoken language has been garnered from speech elicited in non-communicative contexts (e.g., read text, isolated words or nonsense syllables), typically recorded in a laboratory setting under highly controlled (and artificial) conditions. Such forms of spoken language largely conform to the "canonical" (i.e., standard) pronunciation found in a dictionary. The phonetic impact of such factors as prosody, emotion and meaning is largely unknown in such circumstances. To the extent that pronunciation is guided by "information" and other higher-level factors [12] the current state of phonetic knowledge is potentially unrepresentative of the real world (cf. [5]).

One means with which to remedy this gap in phonetic knowledge is to annotate large amounts of unscripted conversation and statistically analyze the corpus materials. Such analyses can serve to delineate phonetic patterns associated with spontaneous speech potentially capable of providing deep insight into the structure and organization of spoken language.

Towards this end a subset of the SWITCHBOARD corpus [3] has been manually annotated at the phonetic and prosodic levels, and this transcription material analyzed with respect to pronunciation variation (cf. [2][5][6][7][9][11][12][13][14]). The resulting analyses provide keen insight into the phonetic organization of spontaneous American English dialogues that may prove useful in the design of future-generation automatic speech recognition and text-to-speech applications [2][8]. In particular, the analyses strongly suggest that the syllable, rather than the phoneme, is the basic structural unit of spoken language, and that prosodic factors, such as stress accent [1], play an important role in the phonetic realization of spontaneous speech. In addition, such analyses suggest a non-arbitrary relation between sound (phonetic features) and symbol (words) that has important implications for the manner in which the lexicon is organized [10].

2. CORPUS MATERIALS

The SWITCHBOARD corpus [3] contains hundreds of brief (5-10 minute) telephone dialogues of a casual nature, spoken by native speakers of American English (representative of most major dialect regions). A subset of this material (45.43 minutes, consisting of 9,922 words, 13,446 syllables and 33,370 phonetic segments, comprising 674 utterances spoken by 581 different speakers) was hand-labeled (by students in Linguistics from the University of California, Berkeley, using Entropics Software to concurrently display the pressure waveform, spectrogram, word- and syllable-level transcripts) with respect to phonetic-segment identity and level of stress accent (associated with each syllable). The mean duration of the utterances transcribed was 4.76 seconds (ranging between 2 and 17 seconds, with ca. 60% of the material between 4 and 8 seconds in length), and the average number of words per utterance was 18.5 (range: 2 to 64 words). The average number of syllables per utterance was 23.25 (range: 5 to 81 syllables). Filled pauses (e.g., "um" and "uh") were excluded from analysis because of the high proportion of non-linguistic attributes associated with such forms.

Three transcribers phonetically labeled and segmented the material. The phonetic inventory used is a variant of Arpabet, originally applied to labeling the TIMIT corpus, but adapted to the exigencies of spontaneous material (cf. [6] for details of the transcription orthography). The interlabeler agreement was 74%. An analysis of the pattern of interlabeler disagreement for vocalic segments indicates that, in such instances, labelers typically disagreed only slightly, usually in terms of one level of height or front/back position. Rarely did transcribers disagree about whether a segment was a monophthong or diphthong.

Two individuals (distinct from those involved with the phonetic labeling) marked the same material with respect to stress accent. Three levels of stress were distinguished – (1) fully accented ("heavy"), (2) completely unaccented ("no accent") and (3) an intermediate level of accent ("light"). The transcribers were instructed to label each syllabic nucleus on the basis of its perceptually based prominence rather than using knowledge of a word's canonical stress pattern derived from a dictionary. All of the material was labeled by both transcribers and the accent labels averaged. In the vast majority of instances the transcribers agreed as to a syllable's stress-accent level – interlabeler agreement was 85% for unaccented nuclei, 78% for fully accented nuclei (and 95% for any level of accent, where both transcribers ascribed some measure of stress to the syllable). In those instances where the transcribers were not in complete accord, the difference in their labeling was usually a half (rather than a whole) step of accent. Moreover, disagreement was typically associated with circumstances where there was a genuine ambiguity in accent level (as determined by an independent observer).

3. THE SYLLABLE'S ROLE IN PRONUNCIATION

Many phonetic properties associated with the SWITCHBOARD corpus are consistent with a syllabic organization of spoken language. For example, the syllable is a remarkably stable unit – only 1% of *canonical* syllables are phonetically unrealized (i.e., deleted) [7], in contrast to the large proportion (22%) of (canonical) phonemes that are unarticulated (or significantly reduced) in the corpus. Moreover, many patterns of pronunciation variation, which appear chaotic and unsystematic at the phonemic level, gain structure and coherence when analyzed in terms of the syllable (*ibid)*.

Figure 1 (upper left panel) illustrates the general pattern associated with phonetic variation (relative to the canonical pronunciation) as a function of the segment's position within the syllable (i.e., onset, nucleus, coda) and stress-accent level (heavy, light, none). *Onsets* of *stressed* syllables are likely to be pronounced canonically, consistent with models of spoken language focusing on the importance of onsets for lexical access [4][16]. The nuclei and codas are far less likely to be canonically realized, and the likelihood of deviation from the canonical rises dramatically as the magnitude of stress-accent diminishes.

Further insight is gained when the pronunciation patterns are partitioned according to the *type* of deviation observed (i.e., substitutions, deletions and insertions). Most *substitution* forms of deviations (lower left panel) are found in the *nucleus* and are inherently vocalic in nature. Substitutions are rarely encountered in either the onset or the coda. Segmental deletion, on the other hand, is rarely observed in either the nucleus or onset, but commonly occurs in the coda (upper right panel). Insertions occur infrequently, and are concentrated in the onset (lower right panel). The absence of stress accent dramatically increases the probability that a nucleus or coda constituent will deviate from the canonical pronunciation. However, accent's impact is highly selective. Its influence is most apparent for substitutions in the nucleus and deletions in the coda. The probability of insertion, as well as coda substitution, is generally unaffected by accent level.

4. STRESS-ACCENT'S IMPACT ON VOWELS

Much of prosody's impact on phonetic identity is found in the vocalic nucleus. Figure 2 illustrates the dramatic changes imposed by stress accent on the phonetic composition and structure of the vocalic system. In heavily accented syllables there is a relatively even distribution of vocalic segments across the articulatory space, particularly with respect to front vowels. Back vowels are mainly represented in terms of the diphthongs [ow] and [uw]. The articulatory distribution of vowels differs markedly in unaccented syllables. Within this context the overwhelming majority of segments lie in the high-front ([ih], [iy]) and high-central ([ax]) regions of the articulatory space. Moreover, the proportion of low- and mid-height vowels is considerably lower than observed in accented syllables. Among unaccented syllables there is a decided skew in the distribution towards high vowels for both canonical and non-canonical forms (cf. Figure 4 in [9]). Changes in vowel height are heavily skewed towards raising in

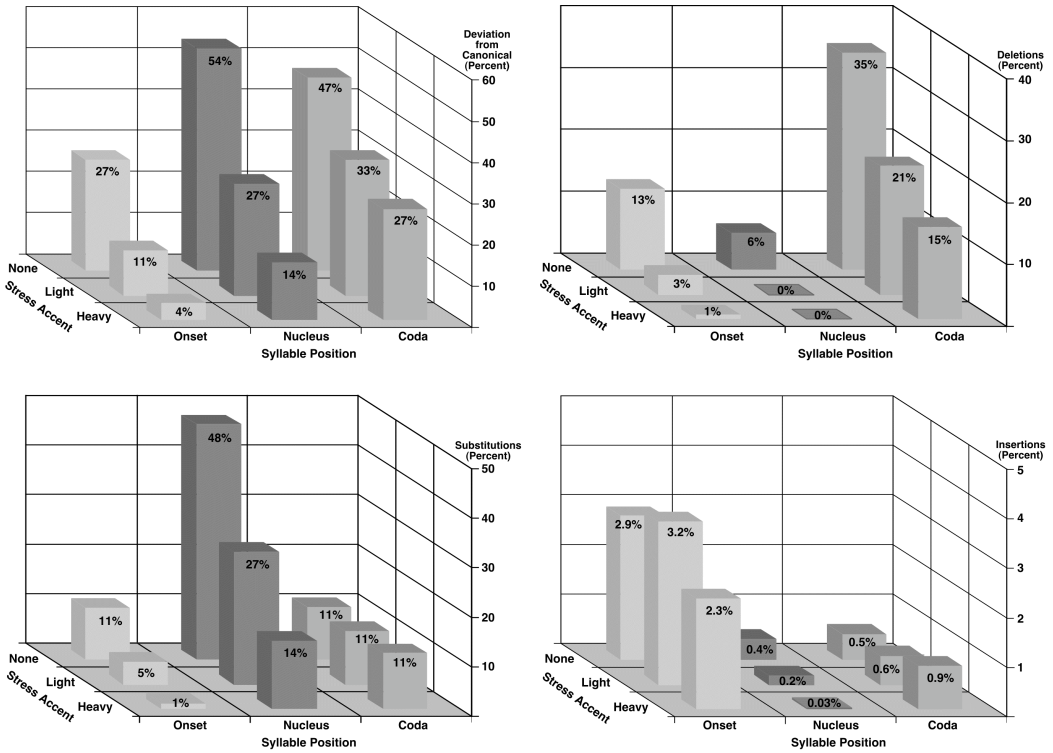

Figure 1 The impact of stress accent on pronunciation variation in the Switchboard corpus, partitioned by syllable position and the type of pronunciation deviation from the canonical form. The height of the bars indicates the percent of segments associated with onset, nucleus and coda components that deviate from the canonical phonetic realization. The magnitude of the deviation is also shown in terms of percentage figures for each bar. Note that the magnitude scale differs for each panel. The sum of the "Deletions," (upper right panel) "Substitutions" (lower left) and "Insertions" (lower right) equals the total "Deviation from Canonical" shown in the upper left panel. Canonical onsets = 10,241, nuclei = 12,185, codas = 7,965. Adapted from [9].

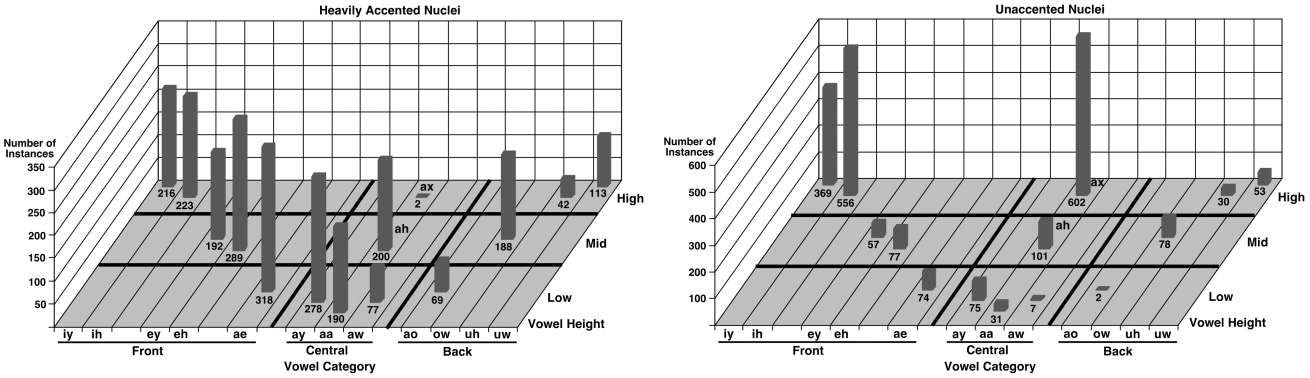

Figure 2 The impact of stress accent ("Heavy" and "None") on the number of instances of each vocalic segment type in the corpus. The vowels are partitioned into their articulatory configuration in terms of horizontal tongue position ("Front," "Central" and "Back") as well as tongue height ("High," "Mid" and "Low"). Note the concentration of vocalic instances among the "Front" and "Central" vowels associated with "Heavy" accent and the association of high-front and high-central vowels with unaccented syllables. The data shown pertain solely to canonical forms realized as such in the corpus. The skew in the distributions would be even greater if non-canonical forms were included. Adapted from [9].

unaccented syllables (cf. Figure 5 in [9]). Overall, there is a tendency for lax, high vowels to occur primarily in unaccented syllables and for low vowels to be present in accented forms, as illustrated in Figure 3 (cf. [9][14]).

5. THE STABILITY OF ONSET SEGMENTS

Onsets are usually pronounced canonically, particularly in accented syllables (cf. Figure 1). Only in unaccented syllables is there a significant tendency for a certain proportion of onsets to be non-canonically pronounced. Most of these deviations are in the form of segmental deletions and are associated with common pronominals, such as "them," "they," "him" and "her," the definite article "the," and the demonstratives "these" and "those." The other segments that tend to be non-canonically realized in onset position are the centrally articulated stops and nasals ([t], [d], [n]), as well as the liquids ([l], [r]); however these segments deviate from the canonical principally in unaccented syllables.

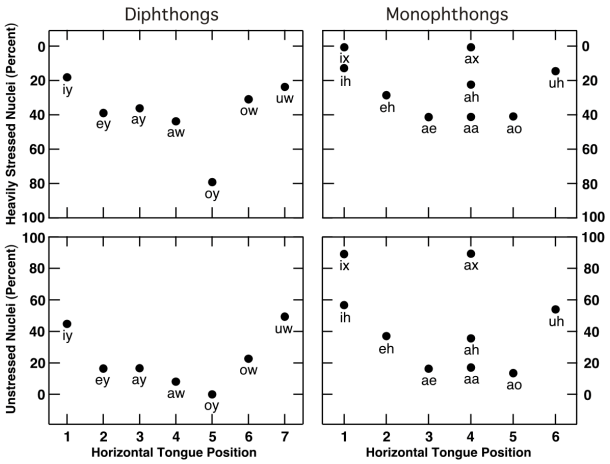

Figure 3 Spatial representation of the mean proportion of nuclei associated with syllables that are heavily stressed or completed unstressed as a function of vocalic identity. Vowels are segregated into diphthongs and monophthongs for illustrative clarity. Note that the polarization of the y-axis scale for the unaccented syllables is the reverse of that associated with the heavily accented syllables (performed in order to highlight the spatial organization of the data). The x-axis refers to the hypothetical position of the tongue in the horizontal place and is intended purely for illustrative purposes. From [14].

6. THE EPHEMERAL NATURE OF THE CODA

The coda is far less likely to be canonically pronounced than the onset (cf. Figure 1). Most of the deviations observed are in the form of segmental deletions; their frequency is extremely sensitive to stress accent magnitude (*ibid*).

Coda deletions are of a highly selective nature. Very few of the segments with an anterior or posterior articulatory constriction are deleted (the exceptions are segments articulated as flaps). In contrast to the anterior and posterior codas, the centrally articulated segments, particularly [t], [d] and [n], are extremely likely to be non-canonically realized, even in heavily accented syllables (the level of accent exerts a significant impact on the probability of non-canonical pronunciation). In many contexts the *default pronunciation of such segments is non-canonical* (usually in the form of segmental deletion or junctural [flap] substitution).

7. ENTROPY'S ROLE IN PHONETIC IDENTITY

What distinguishes centrally articulated codas from their more forward (and backward) counterparts (besides place of articulation)? In contrast to onsets, where in terms of place of articulation there is a *relatively* even numerical distribution among anterior, central and posterior segments (cf. Table 1), the codas manifest a decided skew towards the central phones. Fully 72% of (canonical) coda segments are centrally articulated (*ibid)*, a bias exaggerated even further in unaccented syllables (*ibid).* In other words, the default place of articulation in coda segments is central (i.e., coronal). Anterior and posterior segments are relatively rare, and in this sense are more "informative" (when they occur) with respect to lexical and syllabic differentiation (a similar pattern is characteristic of German spontaneous material in the Kiel Corpus [Marion Jaeger, personal communication]).

A separate property distinguishing coronal codas from their more forward and backward counterparts is the locus of peak spectral energy. The locus of spectral energy maxima among coronal non-continuant consonants ranges between 1500 and 2500 Hz, in the mid-range between the anterior (typically below 1200 Hz) and posterior (generally above 2800 Hz) segments [15]. This locus-frequency area for coronal consonants is similar to the (perceptual) second formant region associated with front and central vowels (cf. Figure 2) which form the overwhelming majority of nucleic segments in spontaneous speech. Thus, there appears to be a direct relationship between the vocalic identity of the syllabic nucleus and the following coda segment. Approxi-

		Syllable Onset								Syllable Coda							
	Stress	**Heavy**		**Light**		**None**		**Total**		**Heavy**		**Light**		**None**		**Total**	
Segments	***Parameter***	***Can***	***Tran***	***Can***	***Tran***	***Can***	***Tran***	***Can***	***Tran***	***Can***	***Tran***	***Can***	***Tran***	***Can***	***Tran***	***Can***	***Tran***
Anterior Constriction	**Number**	857	851	1301	1244	1491	1212	3649	3307	261	235	357	331	372	264	990	830
[p, b, m, f, th, dh, y]	**Percent**	**37.8**	**36.7**	**43.2**	**40.8**	**45.7**	**41.2**	**42.7**	**39.8**	**15.3**	**17.9**	**13.9**	**17.8**	**13.3**	**15.4**	**14.0**	**17.0**
Central Constriction	**Number**	818	831	962	965	1110	1017	2890	2813	1154	766	1828	1121	2127	1186	5109	3073
[t, d, dx, n, nx, s, z]	**Percent**	**36.1**	**35.9**	**31.9**	**31.6**	**34.0**	**34.6**	**33.8**	**33.9**	**67.7**	**58.4**	**71.2**	**60.1**	**76.3**	**69.2**	**72.4**	**62.8**
Posterior Constriction	**Number**	590	635	752	840	665	713	2007	2188	289	310	383	412	289	265	961	987
[k,g,ng, sh, zh,ch,jh,w,q]	**Percent**	**26.0**	**27.4**	**24.9**	**27.6**	**20.4**	**24.2**	**23.5**	**26.3**	**17.0**	**23.6**	**14.9**	**22.1**	**10.4**	**15.5**	**13.6**	**20.2**
All Consonants	**Number**	2265	2317	3015	3049	3266	2942	8546	8308	1704	1311	2568	1864	2788	1715	7060	4890

Table 1 The impact of stress accent (heavy, light, none), syllable position (onset, coda) and place of articulation (anterior, central, posterior) on the likelihood of canonical pronunciation (canonical vs. transcribed). In instances where the number of canonical and transcribed instances are similar, the pronunciation is generally realized as canonical. Percentages pertain to the proportion of segments associated with a specific place of articulation (anterior, central, posterior. The place "chameleons" ([h], [l], [r]) are excluded from analysis. Abbreviations: Can – Canonical; Tran – Transcribed. Percentages and parameter labels are indicated in **bold** characters.

mately half of the non-continuant (i.e., [t], [d], [n]) segments in SWITCHBOARD are phonetically unrealized (i.e., deleted) [10]. Such articulatory deletions are unlikely to be *perceived* as segmental omission given the *implication* of coronal articulation latent in the formant patterns associated with the nucleus and thus this context is perceptually more "forgiving" of coda deletion (and reduction) than its non-coronal counterparts. Coronal codas of low-entropy (i.e., unaccented) syllables are far more likely to be reduced or entirely deleted than codas associated with anterior or posterior place of articulation (cf. Table 1 in [10]). In this sense, information appears to serve as a controlling parameter governing the fine phonetic detail of pronunciation.

8. ACKNOWLEDGEMENTS

The research described was supported by the U.S. Department of Defense and the National Science Foundation. We thank Candace Cardinal, Rachel Coulston, Jeff Good and Colleen Richey for manually annotating the SWITCHBOARD corpus. Shuangyu Chang is now affiliated with BISC, EECS Dept., UC-Berkeley.

9. 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 Recognition*, Ph.D. Thesis, University of California, Berkeley. ICSI Technical Report TR-02-007 (available at www.icsi.berkeley.edu/techreports).

[3] Godfrey, J.J., Holliman, E.C., and McDaniel, J. "SWITCHBOARD: Telephone speech corpus for research and development," *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc.*, pp. 517-520, 1992.

[4] Gow, D., Melvold, J. and Manual, S. "How word onsets drive lexical access and segmentation: Evidence from acoustics, phonology and processing," *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Phila., 1996.

[5] Greenberg, S. "On the origins of speech intelligibility in the real world," *Proc. ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mouson, 1997.

[6] Greenberg, S. "The Switchboard Transcription Project," in *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD, 1997.

[7] Greenberg, S. "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, 29, pp. 159-176, 1999.

[8] Greenberg, S. "From here to utility – Melding phonetic insight with speech technology," in *Integrating Phonetic Knowledge with Speech Technology*, W. Barry and W. Domelen (eds.). Dordrecht: Kluwer, in press.

[9] Greenberg, S., Carvey, H.M. and Hitchcock, L. "The relation of stress accent to pronunciation variation in spontaneous American English discourse," *Proc. Int. Conf. Speech Prosody*, Aix-en-Provence, 2002.

[10] Greenberg, S., Carvey, H.M., Hitchcock, L. and Chang, S. "Beyond the phoneme – A juncture-accent model for spoken language," *Proc. Human Lang. Tech. Conf.*, San Diego, 2002.

[11] Greenberg, S., Chang, S. and 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.

[12] Greenberg, S. and Fosler-Lussier, E. "The uninvited guest: Information's role in guiding the production of spontaneous speech," *Proc. Crest Workshop on Models of Speech Production: Motor Planning and Articulatory Modelling*, 2000.

[13] Greenberg, S., Hollenback, J. and Ellis, D. "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," *Proc. Int. Conf. Spoken Lang. Process. (ICSLP)*, Phila., pp. S24-S27, 1996.

[14] Hitchcock, L. and Greenberg, S. "Vowel height is intimately associated with stress accent in spontaneous American English discourse," *Proc. 7th Int. Conf. Speech Tech. Comm. (Eurospeech)*, pp. 79-82, 2001.

[15] Liberman, A. M., Cooper, F. S., Shankweiler, D. P., Studdert-Kennedy, M. "Perception of the speech code," *Psych. Rev.*, 74, pp. 431-461, 1967.

[16] Marslen-Wilson, W.D. and Zwitserlood, P. "Accessing spoken words: The importance of word onsets," *J. Exp. Psych: Human Percept. Perform.*, 15, pp. 576-585, 1989.