THE UNINVITED GUEST: INFORMATION’S ROLE IN GUIDING THE PRODUCTION OF SPONTANEOUS SPEECH

Steven Greenberg and Eric Fosler-Lussier
1nternational Computer Science Institute
1947 Center Street, Berkeley, CA 94704, USA
steveng,fosler@icsi.berkeley.edu

ABSTRACT

Statistical analysis of a large, phonetically transcribed corpus of spontaneous, American English dialogue suggests that an important factor governing the production of spoken language is the information associated with a given element (whether it be feature, phone, syllable, word or phrase), and that it is difficult to fully account for articulatory patterns (as inferred from the acoustic signal) purely on the basis of biomechanical factors. This entropic foundation of articulation is observed in terms of the probability of canonical pronunciation relative to the position within the syllable, as well as with respect to speaking rate and frequency of lexical occurrence. Such patterns of pronunciation variability imply that the phonetic realization (and hence production) of spoken language is highly dependent on the entropy associated with syllabic, lexical and phrasal contexts, and thus it is likely that the production of spontaneous speech is largely governed by mechanisms operating at an exceedingly high level of linguistic organization.

“....we speak to be heard in order to be understood”
Jakobson, Fant and Halle [14]

1 INTRODUCTION

There is an implicit assumption made by many that the phonetic segment (or “phone”) serves as a fundamental unit of linguistic organization in the production and perception of speech. The special status of the phone in linguistic theory is manifest in a variety of ways:

(1) The standard means of representing the speech stream is in terms of a sequence of phonetic segments. Descriptions of a language’s sub-lexical properties are typically based on an inventory of phonological elements in which these phones are arranged in certain patterns known as phonotactic constraints [3][17].

(2) Sociolinguistic accounts of pronunciation variation primarily describe such phenomena at the phonological (and occasionally phonetic) level [16].

(3) Cognitive models of speech perception and production typically posit a bank of lower tiers comprising the acoustics, the phonetic segment (and occasionally features) as well as a more abstract representation at the level of the phoneme [6] [18].

(4) It is common practice to use primitive stimuli, composed of simple combinations of consonants and vowels (e.g., [ba], [da], [ga], [i^{y}bi^y], [i^{y}di^y], etc.) for ascertaining the cognitive and neural mechanisms underlying the perception and production of speech [20].

(5) Automatic speech recognition systems use the phone as a basic building block. In an ASR system’s front-end individual acoustic frames are associated with specific phonetic segments, and sequences of these phones are subsequently linked with words (based on lexical models derived from phones) [21]. Even ASR systems incorporating multiple pronunciations typically use word models composed solely of phones [24].

Despite the long historical shadow cast by the phone in linguistic theory and practice it is becoming increasingly apparent that far more sophisticated and multifaceted models are required to account for the complexity and variety of spoken language. Thus, the ascendancy of auto-segmental [7] and “non-linear” [13] phonological models is an implicit acknowledgement that something other than the phone is required for the sub-lexical description of a language (cf. [1][2]). Sophisticated speech synthesis (by rule) systems are commonly based on models derived not from sequences of phones, but from diphones or demisyllables (or other suprasegmental unit) in order to capture “co-articulation” effects [22]. And even speech recognition systems often train on context-sensitive (but still) phone models incorporating information from surrounding phonetic environments [21].

Despite a growing sense of unease with the traditional phonological framework of language, it is not entirely clear what sort of structural unit(s) might serve as a worthy successor - syllable, word or something else. This uncertainty may reflect an inherent inability to capture the essence of language purely from within a structural perspective. Something else is needed in order to sculpt a fully formed theory of spoken language - an “uninvited guest,” otherwise known as “information.”

2 THE UNCONVENTIONAL NATURE OF THE PHONE

The long reach of information’s hand is manifest in the statistical properties of spoken language. The current discussion focuses on a corpus of spontaneous telephone dialogues spoken by hundreds of different speakers of American English (“Switchboard” [5]) that has been used for evaluation of automatic speech recognition performance over the past seven years. Four hours of this material has been labeled at the phonetic segment level by highly trained individuals with an academic background in phonetics [8][11]. This Switchboard transcription corpus forms the empirical basis of the current analyses and discussion.

Contained within the four hours of transcribed material are many instances of non-canonical phonetic phenomena. Among the most common are:

(1) Spurious frication (particularly among stop segments)

(2) Devoicing of normally voiced segments, such as [z]

(3) Acoustic cue trading (e.g., amplitude modulation of the waveform in place of a rapid F_2 and F_3 transition in liquid segments, such as [l] and [r])

One means by which to quantify the properties of this phonetic “bestiary” is to ascertain the number of different phonetic realizations associated with each of the most frequently occurring words. Table 1 illustrates the diversity of pronunciation for the 50 most common words (accounting for over 50% of the lexical tokens in the corpus), most of which are phonetically realized in dozens of different ways. The most popular pronunciation generally accounts for only 10-25% of the variants. Clearly, there is little in the way of lexical constancy at the level of the phone.

3 SOME MEASURE OF STABILITY IN THE SYLLABLE

In contrast to the phone, the syllable exhibits some measure of stability with respect to pronunciation variation. Whereas nearly a quarter (22%) of the phonetic segments associated with a word (in terms of its dictionary form) are omitted (on average) in production, only ca. 1% of syllables fail to be uttered [9].

There is a consistent pattern of pronunciation variation in the Switchboard corpus when viewed from the perspective of the syllable. The onset of a syllable is generally realized in canonical (i.e. dictionary) form ca. 85-90% of the time (Table 2), in contrast to both the nucleus and coda, which are canonically realized in only ca. 60-65% of the instances. In coda position

	Word	N	#Pr	Most Common Pronunciation	% Total		Word	N	#Pr	Most Common Pronunciation	% Total
1	I	649	**53**	ay	**53**	26	not	92	**24**	m aa q	**24**
2	and	521	**87**	ae n	**16**	27	think	92	**23**	th ih ng kcl k	**32**
3	the	475	**76**	dh ax	**27**	28	for	87	**19**	f er	**46**
4	you	406	**68**	y ix	**20**	29	well	84	**49**	w eh l	**23**
5	that	328	**117**	dh ae	**11**	30	what	82	**40**	w ah dx	**14**
6	a	319	**28**	ax	**64**	31	about	77	**46**	ax bcl b aw	**12**
7	to	288	**66**	tcl t uw	**14**	32	all	74	**27**	ao l	**24**
8	know	249	**34**	n ow	**56**	33	that's	74	**19**	dh eh s	**16**
9	of	242	**44**	ax v	**21**	34	oh	74	**17**	ow	**61**
10	if	240	**49**	ih	**22**	35	really	71	**25**	r ih l iy	**45**
11	yeah	203	**48**	y ae	**43**	36	one	69	**8**	w ah n	**78**
12	in	178	**22**	h n	**45**	37	are	68	**19**	er	**42**
13	they	152	**28**	dh ey	**60**	38	I'm	67	**9**	q aa m	**26**
14	do	131	**30**	dcl d uw	**54**	39	right	61	**21**	r ay	**28**
15	so	130	**14**	s ow	**74**	40	uh	60	**16**	ah	**41**
16	but	123	**45**	bcl b ah tcl t	**12**	41	them	60	**18**	ax m	**23**
17	is	120	**24**	ih z	**50**	42	at	59	**36**	ae dx	**8**
18	like	119	**19**	l ay kcl k	**46**	43	there	58	**28**	dh eh r	**22**
19	have	116	**22**	hh ae v	**54**	44	my	58	**9**	m ay	**66**
20	was	111	**24**	w ah z	**23**	45	mean	56	**10**	m iy n	**58**
21	we	108	**13**	w iy	**83**	46	don't	56	**21**	dx ow	**14**
22	it's	101	**14**	ih tcl s	**20**	47	no	55	**8**	n ow	**77**
23	just	101	**34**	jh ix s	**17**	48	with	55	**20**	w ih th	**35**
24	on	98	**18**	aa n	**49**	49	if	55	**18**	ih f	**41**
25	or	94	**23**	er	**36**	50	when	54	**18**	w eh n	**31**

Table 1 Pronunciation variability for the 50 most frequently occurring words in the phonetically segmented portion of the Switchboard transcription corpus. "N" is the number of instances each word appears in this 72-minute portion of the corpus. "#Pr." is the number of distinct phonetic expressions for each word. "%Tot" is the percentage of the total number of pronunciations accounted for by the single most common variant. The phonetic representation is derived from a variant of the Arpabet orthography. Further details concerning both the pronunciation data and the transcription orthography may be found in [9], from which this table is adapted.

non-canonical pronunciation usually assumes the form of segmental deletion, while pronunciation variation in the nucleus usually entails phonetic substitution. Thus, there appears to be an overall patterning of pronunciation based on syllabic units that is far more consistent than is observed at the phonetic segment level, consistent with the premise that the syllable is far more grounded in articulatory dynamics than the phone.

4 THE INTRUSION OF ENTROPY

As important as the syllable may be for understanding pronunciation variation (and hence the dynamics of production) there is at least one other factor of importance. This factor is manifest in the pronunciation profile associated with onset, nucleus and coda that is affected by speaking style. Table 2 compares the pronunciation pattern of Switchboard with that of another corpus (TIMIT)[26], consisting of individually read sentences. In this rather more formal material there is an even greater tendency for onsets to assume canonical form than is observed in Switchboard. Moreover, the probability of the coda portion being pronounced canonically is nearly as high as for onsets - only the pronunciation pattern of the nucleus is the same (in terms of canonicality) across the corpora. There appears to be something about the information associated with an utterance that affects very basic patterns of pronunciation.

5 WORD FREQUENCY AND SPEAKING RATE

It has been known for many years that speaking rate can have a dramatic impact on pronunciation. Reduction of vocalic nuclei and deletion of (mostly) coda segments are commonly associated with very fast rates of speaking. Figure 1 illustrates this relation between speaking rate and deviation from canonical pronunciation. As speaking rate increases, so does the probability of non-canonical production. This relation between speaking rate and canonical pronunciation is not, in and of itself, surprising, since "something has to give" with respect to packing the same linguistic message into a smaller interval of time.

What is of interest is the effect that word frequency has on this relation. For the 100 most common words there is marked increase in non-canonical pronunciation with faster rates of speaking (as measured in terms of syllables/sec). For less common words (which typically carry more information as a consequence of their atypicality) the effect of speaking rate is much less marked, suggesting that words of high information content are more likely to be pronounced in canonical form regardless of speaking style and context.

This effect of word frequency suggests that such concepts as "economy of effort" and "ease of articulation," often invoked to explain pronunciation patterns associated with fast, casual speech [19], may not serve as governing factors controlling the manner in which speech is produced under such conditions.

6 ENTROPY AS A POTENTIALLY CONTROLLING FACTOR IN SPEECH PRODUCTION

Why should the pronunciation of words differ so dramatically as a function of their frequency of occurrence? Many years ago Zipf observed that the length of a word (in orthographic form) tends to be inversely correlated with its frequency of occurrence (in printed matter)[25]. Hence, long words tend to be rarely encountered in print, while common words usually consist of short alphabetic sequences. Zipf's law can be shown mathematically to conform to elementary principles of information theory, and that to a first approximation, word frequency is also inversely related to the amount of entropy associated with a lexical element. The relationship between word frequency and length is likely to pertain to spoken forms of language as well [9].

	Switchboard	**Switchboard**	**TIMIT**	**TIMIT**
Syllable Constituent	All Instances	Percent Canonical	All Instances	Percent Canonical
Onset (total)	39,214	84.7	57,868	90.0
Simple [C]	32,851	84.4	42,992	88.9
Complex [CC[C]]	6,363	89.4	14,876	93.3
Nucleus	48,993	65.3	62,118	62.2
with/without onset	35,979 / 13,104	69.6 / 53.4	50,166 / 11,952	64.7 / 51.8
with/without coda	26,258 / 15,101	64.4 / 66.4	32,598 / 29,520	58.2 / 66.6
Coda (total)	32,512	63.4	40,095	81.0
Simple [C]	20,282	64.7	25,732	81.3
Complex [CC[C]]	12,230	61.2	14,363	80.5

Table 2. The frequency with which the phonetic pronunciation corresponds to the lexicon's canonical pronunciation, as a function of syllabic constituent for the phonetically transcribed portion of the Switchboard corpus as well as the for TIMIT corpus of read sentential material. For both corpora the onsets of syllables tend to be phonetically realized as the canonical form most of the time. There is a slightly greater probability of canonical pronunciation for onsets containing two or more consonants. The vocalic nuclei are realized in canonical form far less frequently than syllabic onsets. When the syllable lacks an onset constituent the probability of canonical realization for the nucleus is significantly reduced. However, the absence of a coda element has relatively little impact on the phonetic realization of the nucleus. The primary difference in the pattern of canonical realization between read and spontaneous speech appears localized to the coda constituent. In TIMIT the coda is canonically realized nearly as often as the onset. In contrast, the coda is canonically realized significantly less frequently in Switchboard. Adapted from [4][9].

Because the brain appears to process words of high frequency more quickly than their infrequent counterparts [12] it is likely that a primary factor controlling the specific pattern of pronunciation involves the synchronization of encoding (i.e., speaking) and decoding (i.e., listening) during the course of verbal communication.

Such considerations suggest that speakers may dynamically adjust their articulation in order to provide an acoustic signal that is fine-tailored for the listener(s) in terms of the acoustic background, familiarity with the discussion topic, language background, shared values and experience, etc. Over the course of a lifetime a speaker develops a large repertoire of internal models associating specific patterns of pronunciation with the communication tasks at hand and uses these to craft a variety of speaking styles for the myriad of verbal situations encountered in everyday life.

7 A SPECTRO-TEMPORAL REPRESENTATION INDEPENDENT OF SPEAKING RATE AND STYLE

How do listeners recognize the words of an utterance under such a wide range of speaking (and listening) conditions? It appears that intelligibility does not necessarily require a detailed spectral representation of the acoustic signal as long as a representative sampling of the modulation patterns distributed across the full span of the spectrum is preserved. Thus, four narrow (1/3-octave) channels (accounting for less than 20% of the full spectrum) are sufficient to insure ca. 90% intelligibility when distributed in such a manner as to cover the full bandwidth of the original signal [10].

The sufficiency of such a sparse spectral profile (under optimum listening conditions) suggests that the important information contained in the acoustic signal may be likened to the location of peaks and valleys in a mountainous terrain,

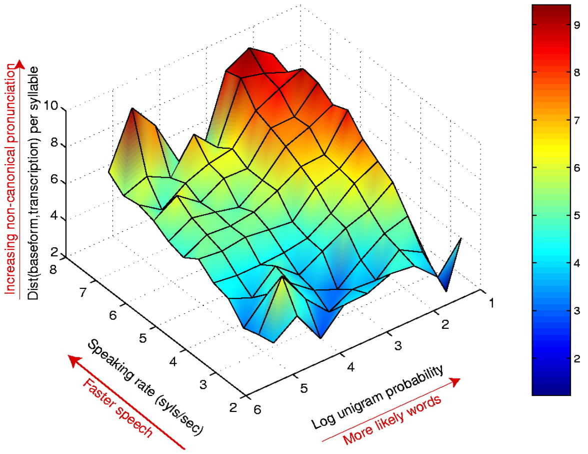
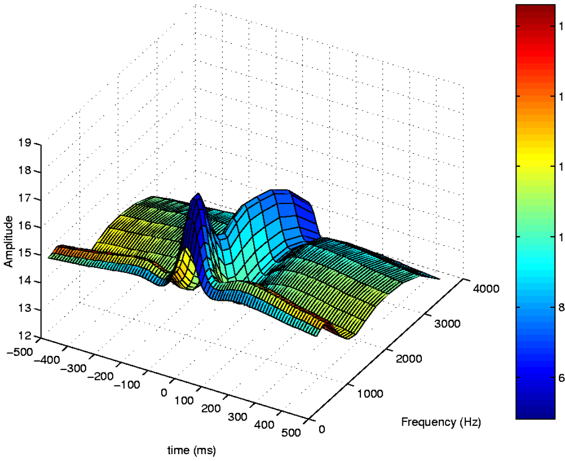


Figure 1 Magnitude of the deviation from canonical (dictionary) pronunciation of individual words in the phonetically transcribed portion of the Switchboard corpus as a function of speaking rate (syllables/sec) and frequency of occurrence. Note that the effect of speaking rate on pronunciation is far greater for high-frequency words than for less common lexical items. Adapted from [4].



igure 2 Spectro-Temporal Profile (STeP) associated with the feature [+vocalic]. The STeP was computed from thousands of instances of vocalic segments (across dozens of speakers) contained in the OGI Stories corpus. The common energy pattern across frequency and time (“0” is the segment’s center) is shown. Virtually the entire STeP exhibits very low variance (blue and green in the color version of the figure).

where the coordinates of the terrain are laid out in units of spectral (i.e., tonotopically organized) frequency along one axis and as a function of time along the other dimension (with the height of the peaks and depth of the valleys correlated with the magnitude of energy at any given place. An illustration of this sort of representation, a Spectro-Temporal Profile (STeP), is shown for the feature [+vocalic] in Figure 2.

In those situations where the entropy of an utterance is high it is likely that a detailed "image" of the spectro-temporal profile would be required for accurate decoding. Under such conditions, in order to be effective, the speaker will have to enunciate with sufficient precision (i.e., speak more "clearly") in order to insure that most of the terrain's details are "visible" to the listener.

When the entropy is low (as would occur for highly predictable words and phrases in casual conversation) a much coarser picture of the STeP is sufficient to recognize the patterns associated with the signal. Such coarse-grained perspectives (such as the one shown in Figure 2) provide for rapid analysis of familiar patterns.

Speaking fluently depends on the ability to modulate the granularity of the spectro-temporal profile in a seamless fashion, dynamically adjusting this level of detail on a moment-to-moment basis in order to maintain a steady stream of information that is readily understood and quickly assimilated by the listener.

Currently, most models of speech production leave the uninvited guest out in the cold, to languish outside for lack of a proper host. Language has been likened to a generator sufficiently powerful to run an elevator, but that normally services only a doorbell [23]. In some sense this analogy is off the mark with respect to pronunciation variation since the brain mechanisms required to drive the production of speech in all of its intricacy and subtlety possess a sophistication that far outstrips the capability of even the most powerful machines currently in existence. Students of speech production would be wise to call the uninvited guest in from the cold and to offer a place at the head of the table.

8 ACKNOWLEDGEMENTS

The research described in this paper is supported by grants from the U.S. Department of Defense and the National Science Foundation. We thank Shawn Chang for computing the STeP illustrated in Figure 2.

9 REFERENCES

[1] Browman, C. P. and Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology*, 6, 201-251.

[2] Browman, C. P. and Goldstein, L. (1990) Tiers in articulatory phonology, with some implications for casual speech. In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, J. Kingston and M. Beckman (Eds.). Cambridge: Cambridge University Press, pp. 341-376.

[3] Comrie, B. (1990) *The World's Major Languages*. Oxford: Oxford University Press.

[4] Fosler-Lussier, E., Greenberg, S. and Morgan, N. (1999) Incorporating contextual phonetics into automatic speech recognition. *Proc. Int. Cong. Phon. Sci.*, San Francisco, pp. 611-614.

[5] Godfrey, J. J., Holliman, E. C. and McDaniel, J. (1992) SWITCHBOARD: Telephone speech corpus for research and development, *Proc. IEEE ICASSP*, 1, pp. 517-520.

[6] Goldinger, S. D., Pisoni, D. B. and Luce, P. (1996) Speech perception and spoken word recognition: research and theory, in *Principles of Experimental Phonetics*, N. Lass (ed.). St. Louis: Mosby, pp. 277-327.

[7] Goldsmith, J. (1990) *Autosegmental and Metrical Phonology.* Oxford: Blackwell.

[8] Greenberg, S. (1997) 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.

[9] Greenberg, S. (1999) Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29, 159-176.

[10] Greenberg, S., Arai, T. and Silipo, R. (1998) Speech intelligibility derived from exceedingly sparse spectral information, *Proc. Int. Conf. Spoken Lang. Process.*, Sydney, pp. 74-77.

[11] Greenberg, S., Hollenback, J. and Ellis, D. (1996) Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus, *Proc. Int. Conf. Spoken Lang. Process*., Philadelphia, S24-27.

[12] Howes, D. (1967) Equilibrium theory of word frequency distributions. *Psychnom. Bull.* 1, 18.

[13] Hulst, H. van der and Smith, N. (eds.) (1985) *Advances in Nonlinear Phonology.* Dordrecht: Foris Publications.

[14] Jakobson R, Fant G, Halle M (1963) *Preliminaries to Speech Analysis*. Cambridge: MIT Press.

[15] Kent, R. D., Adams, S. G. and Turner, G. S. (1996) Models of speech production, in *Principles of Experimental Phonetics*, N. J. Lass (Ed.). Mosby, St. Louis, pp. 3-45.

[16] Labov, W. (1972) *Sociolinguistic Patterns.* Philadelphia: University of Pennsylvania Press.

[17] Ladefoged, P. and Maddieson, I. (1995) *The Sounds of the World's Languages*. Oxford: Blackwell.

[18] Levelt, W. (1989) *Speaking*. Cambridge: MIT Press.

[19] Lindblom, B. (1990) Explaining phonetic variation: A sketch of the H and H theory," in *Speech Production and Speech Modeling*, W. Hardcastle and A. Marchal (eds.). Dordrecht: Kluwer, pp. 403-439.

[20] Löfqvist, A. and Gracco, V. (1997) Lip and jaw kinematics in bilabial stop consonant production. *J. Speech, Lang. Hear. Res.*, 40, 877-893.

[21] Rabiner, L. and Huang, B.-H. (1993) *Fundamentals of Speech Recognition.* Englewood Cliffs: Prentice Hall.

[22] Santen J. van, Sproat R.W., Olive J., and Hirschberg J. (eds.) (1996) *Progress in Speech Synthesis*. Springer Verlag, New York.

[23] Sapir, E. (1921) *Language.* New York: Harcourt, Brace.

[24] Weintraub, M., Fosler, E., Galles, C., Kao, Y.-H., Khudanpur, S., Saraclar, M. and Wegmann, S. (1997) WS96 Project Report: Automatic Learning Of Word Pronunciation from Data, in *Research Report #24, Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.

[25] Zipf, G. K. (1945) The meaning-frequency relationship of words. *J. Gen. Psych.*, 33, 251-256.

[26] Zue, V.W. and Seneff, S. (1996) Transcription and alignment of the TIMIT database, in *Recent Research Towards Advanced Man-Machine Interface Through Spoken Language*, H. Fujisaki (ed.). Amsterdam: Elsevier, pp. 515-525.