INSIGHTS INTO SPOKEN LANGUAGE GLEANED FROM PHONETIC TRANSCRIPTION OF THE SWITCHBOARD CORPUS

Steven Greenberg, Joy Hollenback, Dan Ellis

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

ABSTRACT

Models of speech recognition (by both human and machine) have traditionally assumed the phoneme to serve as the fundamental unit of phonetic and phonological analysis. However, phoneme-centric models have failed to provide a convincing theoretical account of the process by which the brain extracts meaning from the speech signal and have fared poorly in automatic recognition of natural, informal speech (e.g., the Switchboard corpus).

Over the past five months the Switchboard Transcription Project has phonetically transcribed a portion of the Switchboard corpus in an effort to better understand the failure of phoneme-centric models for machine recognition of speech, as well as to provide a database through which to improve the performance of recognition systems focused on conversational dialogs.

Transcription of spoken dialogs illustrates the pitfalls of a phoneme-based system. Many words are articulated in such a fashion as to either omit or significantly transform the phonetic properties of phonemic constituents, thus resulting in wide variation of word pronunciations. Often, only the barest hint of a segment is realized phonetically, in spite of good intelligibility.

Despite this large variability in phonetic realization of words, the temporal properties of speech segments, both phones and syllables, appear to conform to regular patterns. This temporal regularity suggests that much of the linguistic information in speech may be signaled through variations in amplitude, pitch and the coarse spectrum, and that such patterns may be useful in the design of future-generation speech recognition systems.

1. INTRODUCTION

Models of speech perception and recognition focus on the phone(me) as the basic representational unit from which lexical units are ultimately derived. Although this representational model often provides an adequate (if not completely comprehensive) descriptive basis for the acoustics of carefully articulated and read speech, it fails to capture many of the spectro-temporal properties of spontaneous speech typical of informal spoken dialog.

The Switchboard corpus provides an excellent test-bed with which to compare the phonetic properties of spontaneous speech with those characteristic of more formal speaking situations as encapsulated by the TIMIT, ATIS and Wall Street Journal corpora. For the latter three corpora, performance by automatic speech recognition systems typically range between 85 and 98% correct. In contrast, material from the Switchboard corpus is typically recognized with only 40-60% accuracy.

In the Switchboard corpus two individuals discuss a specific topic, such as summer vacations, professional dress codes, the international political situation, credit cards and so on for several minutes over the telephone. The dialog contains a significant proportion of "filled pauses" (e.g., "um," "uh-huh", etc.), "misarticulations" (e.g., transpositions of specific phonetic segments), phonetic and lexical deletions ("University of Nebraska" being pronounced [yuw nix ver six n dix bclbrae skclkaeq], where the "of" is entirely deleted, and the final syllable of "University" [dix] delayed till after the initiation of the nasal consonant in "Nebraska").

Often, only the vaguest hint of the "appropriate" spectral cues are present in the spectrographic representation. Typically, formant transitions usually associated with specific segments (such as liquids or nasals) are either entirely missing or differ appreciably from the patterns observed in more formally articulated speech. Such deviations from the "canonical" phonetic representation pose a significant challenge to current models of speech recognition.

2. SWITCHBOARD TRANSCRIPTION PROJECT

In order to more fully characterize the phonetics of spontaneous speech, seventy-two minutes of the Switchboard corpus (comprising portions of 618 conversations from 750 speakers, representing both genders, and spanning a wide range of adult ages and dialectal patterns from American English) were phonetically transcribed by a group of eight Linguistics students (7 undergraduates and 1 graduate student) all of whom had received previous training in phonetic transcription and general phonetics/phonology at the University of California, Berkeley. The transcribers were closely supervised by both the senior author and Professor John Ohala in order to insure as accurate and as uniform a transcription of the materials as possible. Specific transcription issues were discussed at weekly project meetings, using a 60" BARCO projection screen for computer display and audio feedback.

The phonetic transcriptions were encoded with a variant of the Arpabet transcription system used for the TIMIT corpus. This transcription system was augmented with a set of diacritics representing such phonetic properties as glottalization ("creaky voice"), nasalization (typically applied to vocalic segments), frication, aspiration, de-voicing, unusual voicing, and velarization. In addition, transitional elements between adjacent vocalic or glide-like segments were explicitly marked.

For each short span of speech, the time-domain waveform and its time-aligned wideband spectrographic representation were displayed on a color SparcStation using Entropics waves+ software. Below the spectrographic display was a forced-Viterbi (time) aligned and labeled transcription providing the transcribers with an "initial guess" as to the identity and temporal boundaries of each phonetic segment. Below this phonetic transcription was displayed the (time-aligned) word transcription associated with the speech signal. Transcribers used the initial Viterbi-aligned transcriptions only as a starting point, and typically modified (or moved), added and/or deleted segments and associated segmental boundaries, in accordance with their phonetic training.

3. □PHONE FREQUENCIES

The phonetic transcription of Switchboard provides an opportunity to examine the frequency of occurrence of phonetic segments in informal speech. Table 1 lists the phonetic elements (stripped of their diacritical modification) for the transcribed portion of the corpus. The general patterns illustrated conform to previous accounts [e.g., 1]. Perhaps the most interesting observation is the relatively high frequency of the glottal stop [q] in spontaneous speech. Its frequency (ca. 1.5%) is in the midrange of occurrence for phonetic elements. The glottal stop has begun to function in place of many syllable-final (usually voiceless) stops and at the beginning of many syllable-initial vocalic segments. In this sense it often serves as an element to demarcate the beginning and ending of syllables.

	1 -14		15 - 28		29 - 42		43 - 56	
1	n	.0540	dh	.0243	v	.0145	th	.0076
2	s	.0439	ae	.0239	bcl	.0144	nx	.0069
3	ih	.0430	ow	.0232	ey	.0142	aw	.0066
4	ax	.0420	ay	.0231	uh	.0142	el	.0048
5	iy	.0359	l	.0219	q	.0136	sh	.0047
6	tcl	.0356	w	.0219	y	.0134	ux	.0047
7	ix	.0354	dcl	.0218	dx	.0132	jh	.0046
8	t	.0345	z	.0190	f	.0131	ch	.0044
9	eh	.0309	d	.0189	ao	.0130	en	.0040
10	r	.0299	er	.0181	uw	.0116	axr	.0034
11	ah	.0295	aa	.0177	hh	.0102	oy	.0011
12	kcl	.0266	pcl	.0163	g	.0094	zh	.0010
13	k	.0261	p	.0158	ng	.0093	em	.0008
14	m	.0251	b	.0149	gcl	.0078	eng	.0001

Table 1: The frequency of occurrence, in descending order, of each of the 56 phones used to transcribe the Switchboard corpus.

4. DURATIONAL PROPERTIES OF PHONETIC ELEMENTS

Phonetic transcription also provides a means to analyze the durational properties of the phonetic elements in the Switchboard corpus, based on temporal boundaries associated with nearly 23,000 segments. These data are illustrated in Table 2, on the following page.

The durational patterns revealed by these data are rather interesting. The median duration for most phonetic classes is 60-100 ms. Diphthongs (except [uw] and [iy]) tend to be slightly longer (generally 120-150 ms) while the flaps and glottal stop tend to be shorter (25-25 ms). The short duration of the latter classes is consistent with their association with syllabic boundaries. Diphthongs are generally located within the nucleus of a syllable and possess certain properties similar to vocalic segments followed by glides. Their relatively long duration is therefore not surprising.

Within the Arpabet transcription system stop consonants are partitioned into closure (e.g. [pcl]) and release (e.g., [p]) components. When their durations are combined (in this analysis the combination is performed on a group, rather than the more desirable individual basis) the median durations fall within the 60-100 ms range typical of most other phonetic classes.

5. SYLLABLE DURATIONS

A durational analysis was also performed for syllables, which, on average, contain ca. 2.5 phonetic segments per unit. The median duration for all syllables in the transcribed portion of the Switchboard corpus is 167 ms, with the 20% and 80% percentiles corresponding to 107.5 and 260 ms, respectively. There is a slight asymmetry favoring longer intervals in the distribution of durations (on a linear axis) that is reflected in the mean (190 ms) being ca. 33 ms greater than the median duration. When syllable durations are plotted on a $\log_2$ scale, normalized to the mean, the distribution is approximately symmetric and Gaussian in shape.

Syllable duration can be conceptualized in terms of "modulation frequency," (e.g., a syllable duration of 200 ms is equivalent to a modulation frequency of 5 Hz, a syllable duration of 125 ms is equivalent to a modulation frequency of 8 Hz, etc.) for comparison with a standard acoustic measure used for studies of speech intelligibility [2]. Such a comparison is illustrated in Figure 1. The modulation spectrum for an octave-wide channel, arithmetically centered at 1.5 kHz, computed from a single speaker's discourse over a two-minute interval is shown and compared with that of the distribution of syllable durations (transformed into equivalent modulation frequencies). The similarity between the two measurements suggests that much of the energy in the modulation spectrum may be derived from syllabic segmentation. This association is of interest in light of recent demonstrations that speech intelligibility is crucially dependent on the preservation of the portion of the modulation spectrum between 2 and 10 Hz [3, 4, 5].

VOWELS	20%	50%	80%	N
Diphthongs				
aw	98.4	133.7	205.3	(147)
ay	87.6	122.8	180.4	(532)
ey	82.0	112.1	159.9	(338)
ow	84.2	123.4	187.2	(528)
oy	112.1	146.1	180.0	(26)
iy	57.7	83.2	122.1	(861)
uw	61.2	94.6	147.3	(259)
Monopthongs				
aa	75.7	110.0	149.4	(404)
ao	68.9	102.0	150.7	(295)
ae	80.5	118.5	185.5	(620)
ih	47.4	69.3	96.7	(1057)
eh	55.2	78.5	107.3	(733)
ah	56.5	83.0	127.8	(682)
uh	40.0	56.0	76.9	(277)
ux	44.5	65.2	88.8	(89)
ax	35.9	51.3	75.5	(956)
ix	35.1	51.4	76.3	(660)
Rotacized				
axr	55.6	80.2	108.0	(54)
er	60.0	92.4	134.9	(423)
Glides				
w	42.5	64.6	95.8	(503)
y	36.5	63.9	100.4	(354)
CONSONANTS				
Liquids				
l	40.8	60.0	84.1	(532)
r	39.5	65.0	95.4	(668)
Fricatives				
sh	92.8	108.7	142.0	(108)
zh	47.1	62.5	85.0	(44)
f	57.1	89.3	118.6	(308)
th	48.9	69.5	90.1	(170)
s	60.0	89.1	124.2	(943)
v	37.7	50.2	69.7	(345)
dh	23.8	40.4	60.5	(534)
z	48.6	69.8	102.0	(462)
hh	40.0	62.2	89.4	(223)
Affricates				
jh	56.4	78.0	111.7	(93)
ch	77.3	107.7	134.0	(106)

	20%	50%	80%	N
Nasals				
m	50.6	68.4	88.6	(569)
n	37.6	56.0	81.9	(1264)
ng	42.8	65.7	104.6	(212)
Stops				
b	11.0	15.7	24.2	(307)
bcl	35.6	55.0	74.1	(358)
b+bcl	46.6	70.7	98.3	
d	12.5	19.0	30.1	(460)
dcl	26.0	41.9	63.0	(459)
d+dcl	38.5	60.9	83.1	
g	20.1	28.2	40.0	(210)
gcl	27.5	41.5	61.1	(184)
g+gcl	47.6	70.7	101.1	
p	22.6	40.0	63.9	(340)
pcl	38.4	54.6	70.4	(347)
p+pcl	61.0	84.6	134.3	
t	20.6	38.9	62.2	(808)
tcl	23.1	39.7	61.0	(783)
t+tcl	43.7	78.6	123.2	
k	27.0	46.3	70.1	(610)
kcl	28.3	45.2	61.0	(588)
k+kcl	55.3	91.5	131.1	
q	18.9	35.4	62.1	(301)
Flaps				
dx	17.9	24.0	31.5	(287)
nx	21.3	26.8	32.9	(131)
SYLLABICS				
Liquid				
el	63.0	89.2	142.3	(113)
Nasals				
em	55.8	70.8	91.6	(19)
en	52.1	79.5	112.6	(92)

Table 2: Durations (in ms) for phone segments transcribed from a portion of the Switchboard corpus, partitioned into the 20th, 50th and 80th percentiles. N = number of instances for each phonetic class. Durational data for [eng] are omitted as a consequence of insufficient number of instances.

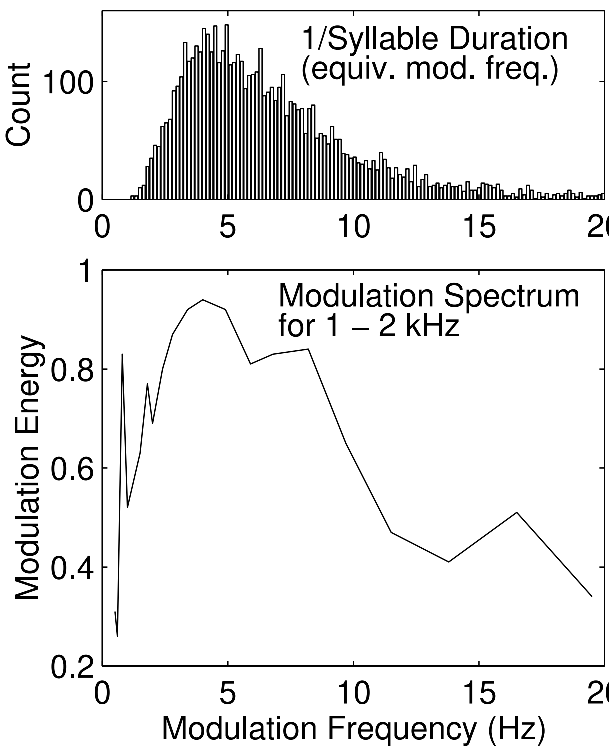

Figure 1: Frequency histogram for 2925 syllables derived from the Switchboard corpus (upper illustration). The modulation spectrum for two minutes of spoken discourse from a single speaker is shown in the lower illustration.

6. CONCLUSIONS

Detailed phonetic transcription of a spontaneous-speech corpus indicates that the spectral properties of many phonetic elements deviate significantly from their canonical form. Despite these spectral "abnormalities" such speech is almost always understandable, suggesting that other properties of the signal, such as segmental duration, may provide significant cues for intelligibility. Measurements of phonetic segment and syllable durations reveal a degree of temporal regularity that may serve as an important basis for understanding speech under a wide range of speaking conditions.

7. ACKNOWLEDGMENTS

The Switchboard Transcription Project (STP) was funded as part of the Speech Recognition Workshop held at the Center for Language and Speech Processing, Johns Hopkins University during the summer of 1996. STP was the product of many individuals' time and effort. We extend our appreciation and gratitude to Candace Cardinale, Melinda Chen, Rachel Coulston, Charles Gotcher, Mike McDaid, Diane Moffit, Colleen Richey and Gail Solomon who painstakingly transcribed the spoken-language materials. We would also like to extend our thanks to the following individuals for their efforts - John Ohala served as the "phonetician in residence" for the project and helped develop the transcription system and supervise the transcribers. Bob Weide provided useful advice concerning transcription systems. Eric Fosler and Jeff Bilmes provided valuable programming expertise during various stages of the project. Brian Kingsbury computed the modulation spectrum shown in Figure 1. Terri Durham, of the Oregon Graduate Institute, transcribed a portion of the corpus for us to compare and contrast with that of the Berkeley transcribers. Valuable discussion and advice was provided by many members of the ICSI Realization Group. Particular thanks are owed to Nelson Morgan and John Ohala for their counsel during the course of the project.

REFERENCES

1. Fletcher, H. (1953), *Speech and Hearing in Communication*, Princeton: van Nostrand.

2. Houtgast, T and Steeneken, H. (1985) 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.

3. Drullman R; Festen J. M. and Plomp, R. (1994) Effect of temporal envelope smearing on speech reception. *J. Acoust. Soc. Am.*, 95, 1053-1064.

4. Arai, T., Hermansky, H. Pavel, M. and Avendado, C. (1996) Intelligibility of speech with high-pass filtered time trajectories of spectral envelopes. Proc. ICSLP.

5. Greenberg, S. (1996) Understanding speech understanding: Towards a unified theory of speech perception. Proceedings of the ESCA Workshop on The Auditory Basis of Speech Perception, Keele University, pp. 1-8.