ON THE ORIGINS OF SPEECH INTELLIGIBILITY IN THE REAL WORLD

Steven Greenberg

University of California, Berkeley
International Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA
steveng@icsi.berkeley.edu

ABSTRACT

Current-generation speech recognition systems seek to identify words via analysis of their underlying phonological constituents. Although this stratagem works well for carefully enunciated speech emanating from a pristine acoustic environment, it has fared less well for recognizing speech spoken under more realistic conditions, such as

(1) moderate to high levels of background noise

(2) moderately reverberant acoustic environments

(3) spontaneous, informal conversation

Under such "real-world" conditions the acoustic properties of speech make it difficult to partition the acoustic stream into readily definable phonological units, thus rendering the process of word recognition highly vulnerable to departures from "canonical" patterns.

Analysis of informal, spontaneous speech indicates that the stability of linguistic representation is more likely to reside on the syllabic and phrasal levels than on the phonological. In consequence, attempts to represent words merely as sequences of phones, and to derive meaning from simple chains of lexical entities, are unlikely to yield high levels of recognition performance under such real-world conditions.

A multi-tiered representation of speech is proposed, one in which only partial information from each of *many* levels of linguistic abstraction is required for sufficient identification of lexical and phrasal elements. Such tiers of linguistic abstraction are unified through a hierarchically organized process of temporal binding and are, in principle, highly tolerant of the sorts of "distortions" imposed on speech in the real world.

1. INTRODUCTION

Human listeners are capable of understanding speech spoken under an exceedingly broad range of acoustic environmental conditions [8, 28]. Such sources of acoustic interference as telephones ringing, computer fan noise, jack hammering and background conversation rarely impede our ability to successfully decode the speech signal.

To date, such environmental versatility has eluded the grasp of automatic speech recognition (ASR) systems. Typically, the performance of ASR systems degrades quite dramatically in the presence of even moderate amounts of background interference. It is not unusual for a system which achieves ca. 85-95% word recognition accuracy for speech spoken under pristine acoustic conditions to recognize only 20-50% of the same material when presented in tandem with either background noise or reverberation [25].

The following report from a recent article [21] on the computerized butler, "Alexander," suggests just how far speech recognition has to go before becoming commonplace in the household:

'Alexander is a voice-recognition system built into a small metal box in the entry way [of the Century Plaza's "Cyber Suite"]. He will, in response to a verbal command, perform a number of menial tasks. Say "Alexander, good night," as you leave, and a robotic voice will bid you adieu as he closes the drapes, turns out the lights and shuts the door behind you. The system also recognizes "good morning," "let's do business," "romance mode" and "party time," and adjusts the ambiance accordingly.

The problem with Alexander is that he's a little too eager to please. *The system picks up ambient noise.* So if you're carrying on a normal conversation across the room you are prone to hear, out of the blue, "Very well, Master," from the British-accented electronic valet. Alexander once interpreted an offhand remark as a command to prepare the room for party mode. There was no stopping him:

"Alexander, NO!!"

"It will be done."

With that, the drapes closed, the lights brightened and disco music on the state-of-the-art Bang and Olufsen stereo system began blasting ...'

1.1 The Central Challenge

Most efforts to improve ASR performance under realistic environmental conditions have focused on ways of enhancing the signal representation through noise suppression [1, 31, 41], echo cancellation [12], and dereveberation [3]. Although these signal processing techniques unquestionably improve recognition performance, the end result still falls far short of human capabilities.

And though continuing efforts to counteract the deleterious effects of background interference will undoubtedly yield some further degree of improvement in recognition performance, it is unlikely that such signal processing techniques, by themselves, will provide the "magic bullet" required to solve the recognition problem.

Such skepticism of current approaches is motivated by several lines of evidence. The history of ASR research over the past two decades is largely one of training systems to effectively handle specific corpora of speech materials. The research effort typically begins by developing acoustic, phonological and grammatical models for a specific body of data, ranging from read sentences (TIMIT), read newspaper text (Wall Street Journal), single digits (Bellcore Digits), street addresses and phone numbers (OGI Numbers), to flight reservations (ATIS) and naval maneuvers (Resource Management). After several years of intensive and costly effort, an ASR system emerges, capable of achieving relatively high levels of performance (85-98% word accuracy), though rarely at the human level for comparable material. Then this system is turned loose on a different corpus and the

Proc. of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels, pp. 23-32, 1997

results are typically discouraging. Years of further effort are expended, developing a new system to handle the current corpus of interest, until such time as the criterial level of performance is achieved.

The wisdom of this "corpus hopping" strategy has recently been called into question by the results of recognition efforts focused on the Switchboard corpus. This corpus comprises spontaneous, informal dialogs recorded over the telephone, between individuals talking about such topics as international politics, vacations, dress codes at work, etc. [14]. After four years of intensive effort the word recognition performance for Switchboard is only ca. 60% correct, on par with ASR performance for simpler speech corpora presented in tandem with background interference.

But the Switchboard corpus is relatively noise and reverberation free. What accounts for the poor ASR performance on this corpus? And might not some of the same factors which conspire to retard performance on the noise-degraded corpora also contribute to the poor results on Switchboard?

1.2 What is Language?

The richness and versatility of human language has so far eluded the most inspired efforts to adequately describe. Edward Sapir perhaps came closest when he characterized language as analogous to an electrical generator sufficiently powerful to run an elevator, but most often used to power a doorbell [34]. This simple analogy captures one of language's most important characteristics - its potentially infinite capacity for expression within a highly restricted domain of abstraction.

Linguistic theory has capitalized on this insight by building elaborate descriptive frameworks for the articulatory-acoustic (phonetic features, phones), phonological (phonemes), grammatical (morphemes, syntactic elements) and semantic (lexical elements, sememes) tiers of organizational abstraction [27]. Each tier is typically treated as an independent level derived from an abstraction of lower organizational levels. Words are thus characterizable as sequences of phonemes analogous to a lexical entry in a dictionary, while phonemes in turn are broken down into their constituent phonetic features and elements. By extension, meaning is derived from a knowledge of the linear sequence of lexical items and the grammatical operations performed on these elements. Many years ago, Charles Hockett warned against characterizing language as merely a sequence of "beads on a string," given its complexity, depth and diversity [23]. Yet his admonition has yet to exert a significant impact on the design of ASR systems.

Current-generation ASR systems typically model a word as a linear sequence of phones [33]. The intent is to recognize as accurately as possible all of the phone elements in the speech signal as a means of increasing word recognition performance. Accurate identification of the entire phone sequence would, in principle, assure perfect performance at the word level as well. In practice, the recognition of lexical elements does not require perfect phone identification, since hidden Markov models (HMMs) are designed, in concert with Viterbi-based techniques, to systematically prune the array of lexical possibilities given imperfect knowledge of the phone constituents.

This approach works well as long as the speech signal conforms to certain assumptions made by the underlying models. In practice, such models have been tuned primarily on the sort of speech characteristic of formal speaking conditions. But the speech typical of real-world conditions rarely conforms to such an orderly structure. Phone segments are typically transformed or deleted entirely in natural speech, requiring the customization of multiple-pronunciation dictionaries to handle the diverse set of possibilities. Predictably, there are many instances of "unusual" phone sequences that the dictionary is not prepared to accept.

And yet such imperfections in the acoustic stream do not hinder human listeners in their quest to successfully decode the speech signal. What is different?

ASR systems focus on identification of individual elements, be they phones, words or sentences. Humans do not. Indeed, the best ASR systems consistently outperform humans when the latter are restricted to listening to these elements in isolation [10, 16]. How can this be if ASR systems do so much more poorly on recognizing speech within a larger context?

One of the salient properties of human language is its capability of being represented on many organizational tiers concurrently. Traditional linguistic theory implicitly assumes that these tiers are largely autonomous and rarely interact to a significant degree. Sentences are composed of phrases which, in turn, are composed of words, which are formed by syllables, which are broken down into phones, which are themselves decomposable into acoustic-articulatory features. Although this paradigmatic approach is likely to act in concert with syntagmatic factors governing the temporal relations among linguistic elements, the consequences of this complementarity have yet to be systematically explored within the framework of ASR systems.

Such concerns become clearly manifest when human listeners are asked to identify linguistic elements excised from their original sentential context. Even experienced listeners are rarely capable of identifying more than 60% of the phonetic segments (phones) presented in isolation [16, 42] and trained phonetic transcribers typically listen to an entire utterance before attempting to identify the individual phonetic constituents [18]. Listeners typically have trouble identifying even entire words excised from their natural context in spontaneous speech [10]. These observations reinforce the intuition that speech can not adequately be described as a linear sequence of phones or words (or any other linguistic unit, for that matter).

This non-sequential property of language is clearly observed in the reading process. The pattern of foveal scans for experienced readers (of a romanized orthography) is not strictly left-to-right, but rather "dances" around elements of sentence and paragraph length in a highly choreographed fashion [39], focusing on specific "pivot" words around which the surrounding lexical items assume their shape and substance. And though the process of reading certainly differs from that of understanding spoken language, this non-sequential decoding process is likely to be common to both.

Current-generation ASR systems implicitly assume that all words are equally important for the decoding process, and that all phones figure importantly in the extraction of lexical entities. But our everyday listening experience calls these assumptions into question as a model for human speech processing. Most listeners experience difficulty repeating the specific sequence of words recently spoken, particularly if the number of lexical elements exceeds the limits of short-term memory [30].

How then do human listeners extract meaning from the speech stream, and what implications does this process hold for the design of future-generation ASR systems?

The precise mechanisms by which humans process speech are, of course, largely unknown. However, a growing body of evidence suggests that the decoding process is governed by two different types of procedures, neither of which has yet to be substantially incorporated into speech-recognition-system design.

The first entails a coarse segmentation of the speech stream, probably at the syllabic and phrasal levels. Although segmentation at the phone and word level is explicitly built into current-generation ASR systems, there is little evidence to suggest that human listeners segment on these levels, as the observations described above so fully attest. Listeners appear far more sensitive to syllabic [35] and phrasal [4] boundaries than to those imposed by lexical and phonological criteria.

The second procedure involves a dynamic linkage of the representational tiers of language enabling listeners to effectively translate cues and features at one level of analysis into those characteristic of another. Detailed analysis of spontaneous speech illustrates how this is accomplished at the phonetic level. In informal speech many of the spectro-temporal cues (i.e., the formant patterns) for specific phonetic segments are either significantly transformed or altogether missing [18]. However, listeners make sense of speech because such canonical features have either been replaced by other cues (such as temporally appropriate amplitude modulation) or compensated for by a broader phonetic pattern that contains sufficient cues as to pass for a reasonable facsimile of the intended lexico-grammatical element [18]. Speakers appear to have an intuitive understanding of the relationship between cues of different representational tiers and exploit this linkage often.

This implicit knowledge of the relationship between representational tiers is likely to be a key factor in the listener's capability of inferring the linguistic message from partial information. If detailed information pertaining to the phonetic sequence is absent, comparable information is likely to be obtained from analysis of the syllabic and prosodic components of the speech signal. Amplitude modulation, durational information, and pitch contours all function to prune the roster of likely candidates to a manageable number sufficient for unambiguous coding, given some form of prior semantic framework. At present, little of this information is directly encoded into ASR systems, nor is the linkage among the tiers explicitly retrievable.

2. STATISTICAL PROPERTIES OF SPOKEN ENGLISH

The dynamic linkage among the linguistic tiers would not be nearly so useful for decoding the speech stream, were it not for the statistical regularities that characterize each organizational level. These regularities provide an interpretative framework with which to characterize the speech signal and relate these to other representational tiers. Although the statistical patterns observed on any signal level are not in and of themselves definitive, they can serve as a powerful pruning device when combined with statistical knowledge of other organizational levels and the mapping relations which bind them together.

It is possible to characterize such statistical patterns through quantitative analysis of spontaneous speech, such as that contained within the Switchboard corpus. The analyses presented here are intended to serve as a representative sample and do not begin to exhaust the range of organizational levels which exhibit statistically regular trends.

2.1 Word Frequency and Pronunciation Variability

Since the days of Dewey [9] and Zipf [45] it has been known that words differ greatly in terms of their frequency of occurrence in written language. French and colleagues [13] have demonstrated a comparable pattern for spoken discourse.

A frequency analysis of the Switchboard corpus illustrates the magnitude of this effect. The most common words occur much more frequently (by at least several orders of magnitude) than the least common (Figure 1). The frequency plot conforms approximately to a 1/f distribution, which has several interesting implications for linguistic decoding (Section 4). The ten most common words account for approximately 25% of all the lexical instances in the corpus. One hundred words account for fully 66% of the individual tokens (Figure 2). A perusal of these most frequently occurring words (Table 1) indicates that most come from the so-called "closed" or "function" class words such as pronouns, articles, conjunctions and modal/auxiliary verbs. Most of the remainder stem from just a few basic nominal, adjectival or verbal forms. Clearly, mastery of these 100 most common words goes a long way towards facilitating comprehension of spoken discourse, a fact intuitively sensed by many non-native speakers seeking to learn English (or any other foreign language), but which is not *directly* incorporated into most current-generation ASR systems. Clearly, the perceptual criteria for recognizing these common words are very different from those associated with the infrequently occurring lexical elements.

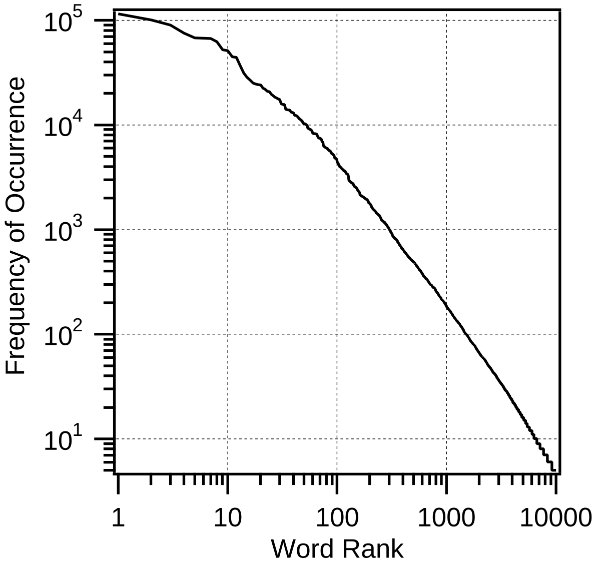

Figure 1. The frequency of occurrence for the 10,000 most frequent words in the Switchboard corpus, organized in rank order of frequency. Total number of distinct words in the corpus is 25,923.

2.2 Syllable Structure and its Relation to the Lexicon

Although such a list of common words does not provide sufficient data to interpret the speech stream by itself, it can be used in conjunction with other knowledge

	Word	N	#Pr.	Most Common Pronunciation	% Tot
1	I	649	**53**	ay	**53**
2	and	521	**87**	ae n	**16**
3	the	475	**76**	dh ax	**27**
4	you	406	**68**	y ix	**20**
5	that	328	**117**	dh ae	**11**
6	a	319	**28**	ax	**64**
7	to	288	**66**	tcl t uw	**14**
8	know	249	**34**	n ow	**56**
9	of	242	**44**	ax v	**21**
10	it	240	**49**	ih	**22**
11	yeah	203	**48**	y ae	**43**
12	in	178	**22**	ih n	**45**
13	they	152	**28**	dh ey	**60**
14	do	131	**30**	dcl d uw	**54**
15	so	130	**14**	s ow	**74**
16	but	123	**45**	bcl b ah tcl t	**12**
17	is	120	**24**	ih z	**50**
18	like	119	**19**	l ay kcl k	**46**
19	have	116	**22**	hh ae v	**54**
20	was	111	**24**	w ah z	**23**
21	we	108	**13**	w iy	**83**
22	it's	101	**14**	ih tcl s	**20**
23	just	101	**34**	jh ix s	**17**
24	on	98	**18**	aa n	**49**
25	or	94	**23**	er	**36**
26	not	92	**24**	m aa q	**24**
27	think	92	**23**	th ih ng kcl k	**32**
28	for	87	**19**	f er	**46**
29	well	84	**49**	w eh l	**23**
30	what	82	**40**	w ah dx	**14**
31	about	77	**46**	ax bcl b aw	**12**
32	all	74	**27**	ao l	**24**
33	that's	74	**19**	dh eh s	**16**
34	oh	74	**17**	ow	**61**
35	really	71	**25**	r ih l iy	**45**
36	one	69	**8**	w ah n	**78**
37	are	68	**19**	er	**42**
38	right	61	**21**	r ay	**28**
39	uh	60	**16**	ah	**41**
40	them	60	**18**	ax m	**23**
41	at	59	**36**	ae dx	**8**
42	there	58	**28**	dh eh r	**22**
43	my	58	**9**	m ay	**66**
44	mean	56	**10**	m iy n	**58**
45	don't	56	**21**	dx ow	**14**
46	no	55	**8**	n ow	**77**
47	with	55	**20**	w ih th	**35**
48	if	55	**18**	ih f	**41**
49	when	54	**18**	w eh n	**31**
50	can	54	**28**	kcl k ae n	**15**
51	then	51	**19**	dh eh n	**38**
52	be	50	**11**	bcl b iy	**76**
53	as	49	**16**	ae z	**18**
54	out	47	**19**	ae dx	**22**
55	kind	47	**17**	kcl k ax nx	**21**
56	because	46	**31**	kcl k ax z	**15**
57	people	45	**21**	pcl p iy pcl l el	**44**
58	go	45	**5**	gcl g ow	**83**
59	got	45	**32**	gcl g aa	**15**
60	this	44	**11**	dh ih s	**47**
61	some	43	**4**	s ah m	**48**
62	i'm	42	**9**	q aa m	**26**
63	would	41	**16**	w ih dcl	**29**
64	things	41	**15**	th ih ng z	**52**
65	now	39	**11**	n aw	**69**
66	lot	39	**9**	l aa dx	**47**
67	had	39	**19**	hh ae dcl	**24**
68	how	39	**11**	hh aw	**53**
69	good	38	**13**	gcl g uh dcl	**27**
70	get	38	**20**	gcl g eh dx	**13**
71	see	37	**6**	s iy	**80**
72	from	36	**10**	f r ah m	**28**
73	he	36	**7**	iy	**39**
74	me	35	**5**	m iy	**87**
75	don't	35	**21**	dx ow	**14**
76	their	33	**19**	dh eh r	**25**
77	more	32	**11**	m ao r	**56**
78	it's	31	**14**	ih tcl s	**20**
79	that's	31	**20**	dh eh s	**16**
80	too	31	**6**	tcl t uw	**60**
81	okay	31	**17**	ow kcl k ey	**45**
82	very	30	**11**	v eh r iy	**36**
83	up	30	**11**	ah pcl p	**34**
84	been	30	**11**	bcl b ih n	**51**
85	guess	29	**8**	gcl g eh s	**42**
86	time	29	**8**	tcl t ay m	**62**
87	going	29	**21**	gcl g ow ih ng	**13**
88	into	28	**20**	ih n tcl t uw	**14**
89	those	27	**12**	dh ow z	**42**
90	here	27	**11**	hh iy er	**25**
91	did	27	**13**	dcl d ih dx	**23**
92	work	25	**8**	w er kcl k	**66**
93	other	25	**14**	ah dh er	**26**
94	an	25	**12**	ax n	**28**
95	I've	25	**7**	ay v	**46**
96	thing	24	**9**	th ih ng	**52**
97	even	24	**7**	iy v ix n	**40**
98	our	23	**9**	aa r	**33**
99	any	23	**11**	ix n iy	**23**
100	I'm	23	**9**	q aa m	**26**

Table 1. Pronunciation variability for the 100 most common words in the Switchboard Transcription Corpus. "N" is the number of instances each word appears in the 72-minute 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 the Arpabet orthography. Further details concerning both the pronunciation data and the transcription orthography may be found in [18].

sources to considerably reduce the uncertainty. One way in which this is potentially achieved is to characterize these most common words in terms of other representational units, such as the syllable.

The 30 most common words in the Switchboard corpus are monosyllabic, and of the 100 most frequent lexical items only ten are not (and all of these are disyllabic). This lexical preference for syllabic brevity is consistent with Zipf's law (originally formulated in terms of word length based on the orthographic sequence of characters) and has potentially important implications for decoding the speech signal.

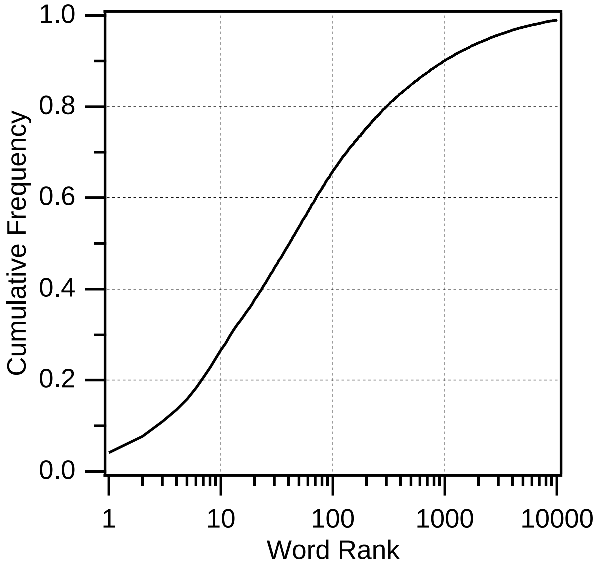

Figure 2. Cumulative frequency of occurrence as a function of word frequency rank for the 10,000 most frequent lexical items in the Switchboard corpus.

In spontaneous English discourse there is a decided preference for words of a single syllable. Although only 22% of the Switchboard lexicon is composed of monosyllabic forms, fully 81% of the corpus tokens are just one syllable in length (Table 2). This statistical skew towards short syllabic forms provides yet another interpretative constraint on the decoding of the speech stream.

#Syllables	*Usage (%)*	*Lexicon (%)*
1	81.04	22.39
2	14.30	39.76
3	3.50	24.26
4	0.96	9.91
5	0.18	3.21
6	0.02	0.40

Table 2. The proportion of words consisting of n-syllables for the entire Switchboard corpus (i.e., tokens) and lexicon (i.e., type). Comparable data from a telephone dialog corpus study performed in the 1920's [13] shows a virtually identical frequency pattern as a function of syllabic length for lexical items.

Knowing the number of syllables in a word also provides some degree of grammatical information. This is a consequence of the tendency for polysyllabic words (particularly those containing three or more syllables) to be either a noun (66% of the time) or adjective (15%). In contrast, verbs are rarely longer than two syllables in length. Speakers of English appear to be well aware of such statistical regularities and use syllable count as an effective strategy for pruning grammatical class candidates [5]. Impairment of the ability to accurately count syllabic units is one of the consequences of information degradation in the auditory frequency channels above 3 kHz [15] and it is therefore not surprising that individuals with a sensorineural hearing loss (which generally affects these high-frequency channels most severely) exhibit particular difficulty in understanding speech in noisy and reverberant environments where such knowledge would prove especially useful for pruning lexical candidates.

The syllable has often been dismissed in linguistic circles as a viable candidate for decoding of spoken English [7] (and Dutch [43]) due to its heterogeneous phonological structure. Statistical analysis of the Switchboard corpus demonstrates, however, that the syllabic composition of spontaneous English is generally much closer to the relatively transparent structure characteristic of so-called syllable-timed languages than would otherwise be imagined. In contrast to languages such as Japanese, where the syllable generally assumes only one of a few potential phonological forms (such as consonant+vowel [CV], vowel [V] or consonanant+vowel+consonant [CVC]), English syllables can assume a wide range of patterns due to the occurrence of consonant clusters (e.g., "strengths" = CCCVCCC). According to this logic, the heterogeneity of syllabic forms makes it difficult for accurate syllabification to proceed in real time.

However, the syllable structure of English is far more homogeneous than one would initially imagine. Over 83% of the corpus syllables are of CV, CVC, VC or V form (Table 3). This is the case whether the statistics are based on the canonical phonological representation ("Corpus") or on the phonetic realization ("Phn. Tr."). The primary distinction between the two concerns the tendency for CVC phonological forms to reduce to CV structure in spoken discourse (as indexed by the reciprocal relation between the frequency of occurrence for CVC and CV syllables). The remaining, more complex syllabic forms constitute only a small proportion of the corpus tokens. This relatively predictable syllabic structure (nearly half of the syllables in spoken discourse are of the CV form) is likely to facilitate the process of syllabic segmentation and to aid in the identification of the constituent phones (since one has *a priori* knowledge concerning both the number and gross class of the phonetic segments). Deviations from the canonical syllable patterns also provide important information since these tend to be nouns or adjectives of relatively infrequent occurrence.

In spontaneous, informal speech the phonetic realization often differs markedly from the canonical phonological form. Entire phone elements are frequently dropped ("deletions") or transformed into other phonetic segments ("substitutions"). At first glance the patterns of deletions and substitutions appear rather complex and somewhat arbitrary when analyzed on the phonological level. Current-generation ASR systems attempt to handle such phonetic variation through multiple-pronunciation dictionaries that include the most common forms.

However, this strategy is unable to capture the entire range of variability, which is often quite broad. It is not uncommon for frequently occurring words to be phonetically realized in dozens of different ways, with the most popular variant often accounting for only 10-15% of the forms (Table 1). However, the patterns of phonetic variation are relatively straightforward to describe within a syllabic framework. The syllable can be divided into three components, the onset, nucleus and coda. For example, the word "cat" can be represented phonetically as three segments [k] [ae] [t], each of which is identified with one of these components. The onset, [k], is typically the most well-preserved portion of the syllable, while the coda [t] is most likely to delete in fluent discourse and the nucleus is most prone to substitution (e.g., [ae] > [I] or [ɛ]). In fast, running speech the syllable can reduce to just the onset (as in "t'day" for "today"). Syllables beginning with vocalic segments (i.e., where the onset and nucleus are one and the same) often convert into a CV(C) form if the preceding syllable contains a consonant coda (e.g., "four" [f ao r] + "eight" [ey t] > [f ao^r] [r ey t]). This "resyllabification" is quite common when contiguous syllables are phonologically of the (C)VC + V(C) form and the syllables belong to the same phrasal unit [18]. At present, most ASR systems do not explicitly model such trans-syllabic phenomena, nor do they explicitly encode lexical information into atomic elements of the syllable, despite the relatively systematic behavior they engender in spontaneous discourse.

Syllable Type	*Lexicon(%)*	*Corpus(%)*	*Phn. Tr(%)*
CV	36.2	34.0	47.2
CVC	28.8	31.6	22.1
VC	5.3	11.7	4.8
V	4.8	6.3	11.2
Subtotal	*75.1*	*83.6*	*85.3*
"Complex"			
CVCC	7.3	6.3	2.9
VCC	0.5	4.3	0.5
CCV	7.4	2.6	5.1
CCVC	5.0	2.2	2.5
CCVCC	2.2	0.6	0.4
CVCCC	1.0	0.4	0.2
CCCVC	0.5	<0.1	0.1
CCCV	0.4	<0.1	0.3
CCVCCC	0.3	< 0.1	< 0.1
CCCVCC	0.2	< 0.1	< 0.1
VCCC	< 0.1	< 0.1	< 0.1
CCCVCCC	< 0.1	< 0.1	< 0.1

Table 3. The relative frequency of occurrence for various forms of syllable structure in both the lexicon and in the actual usage for the entire Switchboard corpus. These data are derived from canonical pronunciations of dictionary sources, and are compared with the syllable structure for actual pronunciation derived from phonetic transcription (Phn. Tr.).

2.3 Syllable Frequency

Another common objection to the syllable as an organizational unit in English is the large number of potential entries in the language's repertoire (over 8000 distinct syllables according to one authoritative account [32]). Within the Switchboard corpus alone there are nearly 5000 syllabic forms (in terms of their phonological representation). However, the distribution of these syllables is far from uniform, either in the lexicon as a whole, or in terms of their frequency of occurrence (Figure 3). Approximately 63% of the corpus is accounted for by just 100 syllables. An additional 170 syllables is required to cover 80% of the syllable occurrences in the Switchboard corpus (Figure 3), suggesting that it may indeed be feasible for both humans and machines to encode English in terms of syllabic units.

The "Corpus" statistics pertain to actual instances of occurrence, while the "Lexicon" statistics refer to syllable frequency within the dictionary, independent of their frequency of occurrence.

A comparison of the cumulative frequencies for syllabic and lexical elements in the Switchboard corpus (Figure 4) indicates the essential similarity of the distributions for the 1000 most common words.

3. SEGMENTATION OF THE SYLLABLE

Anyone who has attempted to partition the speech stream into phone-level elements is aware of how difficult a task this is to perform. It requires an intimate knowledge of acoustic and articulatory phonetics as well as a large expenditure of time. Syllables are considerably easier to segment. In contrast to phone elements there is typically a convergence of waveform, spectrographic and audio cues upon which to rely, resulting in relatively seamless segmentation. This coherence of physical cues provides a basis for the syllable to function as a fundamental grouping element for speech, reflecting the acoustic realization of the articulatory gesture.

Given the potential utility of syllabic information, how practical is it to segment the speech signal into such units automatically? Shire [44] has shown that it is possible to accurately determine the syllabic onsets about 85-90% of the time, using relatively simple signal processing techniques that look for patterns of energy over contiguous spectral channels. And Kingsbury [20] has shown that a representation of the speech signal based on relatively long time windows of 250 ms is especially sensitive to syllabic units.

3.1 Syllable-length Analysis Required For Representational Stability

The conventional window for spectral analysis is ca. 20-30 ms, an interval corresponding roughly to the time constant of the auditory periphery [17]. Although it is currently fashionable to model the auditory response to speech and other complex signals with such a short time constant [6, 37] there are many reasons to believe that human listening performance also requires an analysis at considerably longer time constants to account for the perceptual stability characteristic of auditory experience.

The average length of a syllable in English discourse (as represented by the Switchboard corpus) is ca. 190 ms. Half of the syllables have durations greater than 167 ms. Twenty percent are longer than 260 ms, while anther 20 percent are shorter than 107 ms [19], as illustrated in Figure 5. This heterogeneity reflects a combination of factors, including differential speaking rate, linguistic

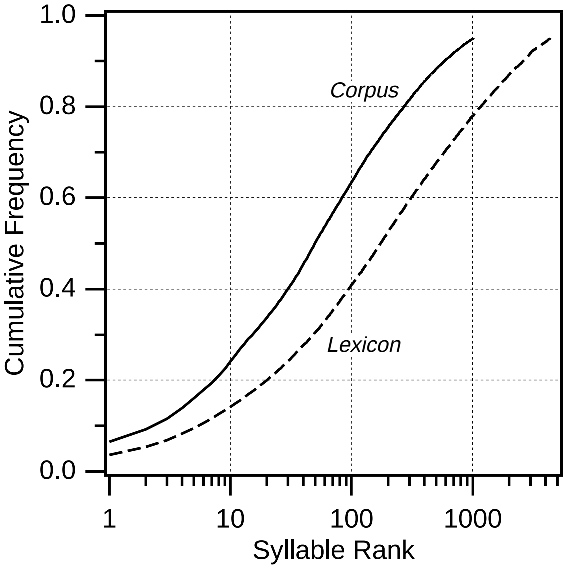

Figure 3. The cumulative frequency of syllables in the entire Switchboard corpus as a function of syllable frequency rank. The "Corpus" statistics pertain to actual instances of occurrence, while the "Lexicon" statistics pertain to syllable frequency within the dictionary, independent of their frequency of occurrence.

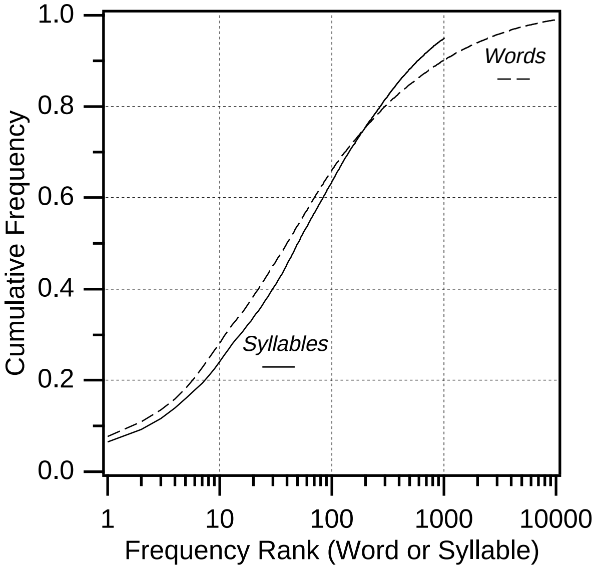

Figure 4. The cumulative frequency of syllables in the entire Switchboard corpus as a function of syllable frequency rank (data from Figure 3) compared with the cumulative frequency of occurrence for words in the same corpus (data from Figure 2).

stress (which tends to lengthen syllables) and the differential number of phonetic constituents (as described in Section 2.2 and in Table 3). The variability in syllable duration is reflected in the long-term modulation spectrum which shows a peak at ca. 4 Hz (Figure 5), but which also contains significant energy between 2 and 8 Hz. This is the modulation spectral range over which information germane to phonetic and lexical identity is contained [24] and whose modification significantly degrades speech intelligibility (Dutch [11], English [16] and Japanese [2]). It is also the frequency range over which the speech articulators move [38].

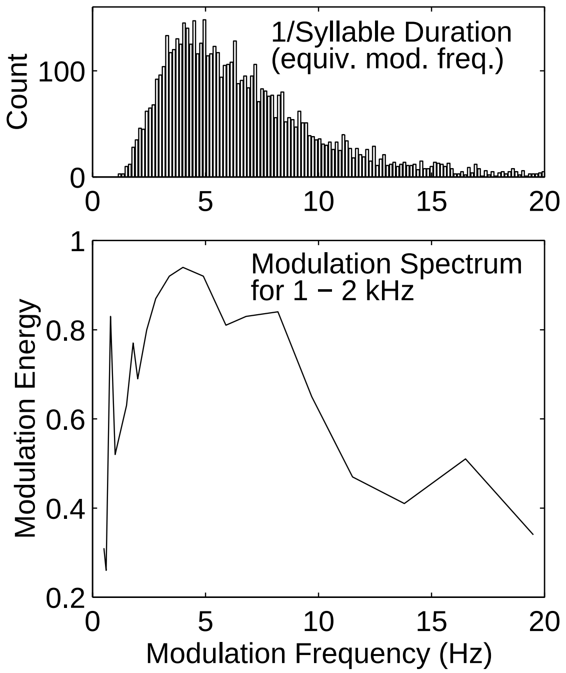

Figure 5. Frequency histogram for 2925 syllables from a portion of the Switchboard corpus (upper illustration). The modulation spectrum for two minutes of spoken discourse from a single speaker is shown in the lower illustration. Reprinted from [19].

It is possible to capture this syllabic information in terms of a representation known as the modulation spectrogram [20, 25]. This representation decomposes the speech signal into 1/4 octave channels and extracts the magnitude of energy in the low-modulation-frequency range, using a filter whose shape is similar to the long-term temporal modulation characteristics of running speech (Figure 6). The representation exhibits a high degree of stability in both noisy and reverberant environments ([20], for a color representation, see http://www..icsi.berkeley.edu/~bedk/ICASSP97_fig2_color.gif). Recognition performance using this representational format results in a significant reduction in the word error rate for moderately reverberated speech, in comparison to the more conventional representations based on static spectral features (Table 4).

Representation	*Clean(%)*	*Reverberant(%)*
PLP	8.5	42.6
ModSpec	10.2	29.3

Table 4. Recognition results from the OGI Numbers95 corpus. The PLP recognizer uses 8th-order PLP and delta PLP features, without energy, and an MLP with 512 hidden units. The modulation spectrogram based recognizer uses cube-root compressed modulation spectrographic features, with the real and imaginary filters separated, no per-channel normalization and no thresholding, and an MLP with 328 hidden units. The reverberation condition, T60, is 0.5 s, with a direct-to-reverberant energy ratio of 0 dB.

4. SPEECH UNDERSTANDING THROUGH CORRELATIVE DEDUCTION

Despite the impressive reduction in error rate effected by such signal processing techniques as the modulation spectrogram, it is unlikely that such methods, alone, can yield the sort of fault-tolerant performance typical of human listeners. This limitation is largely a consequence of current-generation ASR systems' focus on deriving word elements from the acoustic stream, based on the dubious assumption that such lexical items are easily modeled as a linear sequence of phones.

A more principled approach to automatic speech recognition is required, one that draws inspiration from strategies used by human listeners to successfully decode the speech signal under real-world conditions, and which de-emphasizes the extraction of rigidly defined elements (such as words and phones) which often bear but a tangential relationship to the *information* transmitted during the course of human discourse. Speakers possess keen intuitions concerning the information valence of their speech and sculpt their pronunciation and timing in anticipation of their interlocutor's decoding requirements. Thus, vocal effort increases under severely noisy conditions (the so-called "Lombard" effect [22,26]), and both the precision and temporal properties of the utterance will vary according to the emphasis the speaker believes is required for the listener to accurately decode the intended message. Linguistic stress, syllabic reduction, lexical selection all reflect such informationally guided processes in the normal conduct of vocal communication. At present, ASR systems do not explicitly model these effects, and are thus consigned to modeling each novel environmental and communication condition largely from scratch.

One of the hallmarks of human discourse is its flexible format and expression. There are many different ways to pronounce common words (Table 1) and embed these within a grammatical framework that is itself rather fluidly defined. Despite such variability in linguistic expression, listeners readily interpret the speech stream. How can this be so?

Despite the seemingly infinite capacity for novelty in linguistic discourse, most of the words expressed derive from an extremely restricted lexical pool (Figure 2). This tendency towards predictability provides a relatively stable statistical framework with which to interpret the high-information-content words.

Operating in tandem with this low entropic background are a variety of prosodic cues based on such physical quantities as the fundamental frequency contour, amplitude fluctuations and segmental duration, as well as a range of grammatical constraints (such as word order and morphological markings), all of which serve to reduce the degree of interpretive uncertainty. This mixture of high and low entropic elements within the linguistic stream may be likened to the use of impurities to control the electron flow within a semiconductor-based circuit.

Time functions as a crucial parameter in this constraint-based decoding process, providing a infrastructure within which to embed and interpret the abstracted features derived from the speech signal. Statistically based knowledge of the durational properties of syllabic and phonetic segments, as well as of lexical and phrasal constituents, all serve to facilitate the parsing of the speech stream and to bind features derived from heterogeneous levels of abstraction into coherent informational units. This interpretative adaptability is likely to underlie the ability of human listeners to successfully decode speech derived from minimal spectral cues [36].

The temporal binding process appears to operate on a time constant of 40-250 ms at the most basic level of speech analysis. Listeners are capable of tolerating random temporal shifts in spectrally partitioned (e.g., quarter-octave channels) speech, as long as these shifts occur over less than ca. 120 ms [16]. And the visual speech signal can be desynchronized from the acoustic stream by as much as 250 ms without significant degradation of the intelligibility derived from combining these two distinct sources of linguistic information [29].

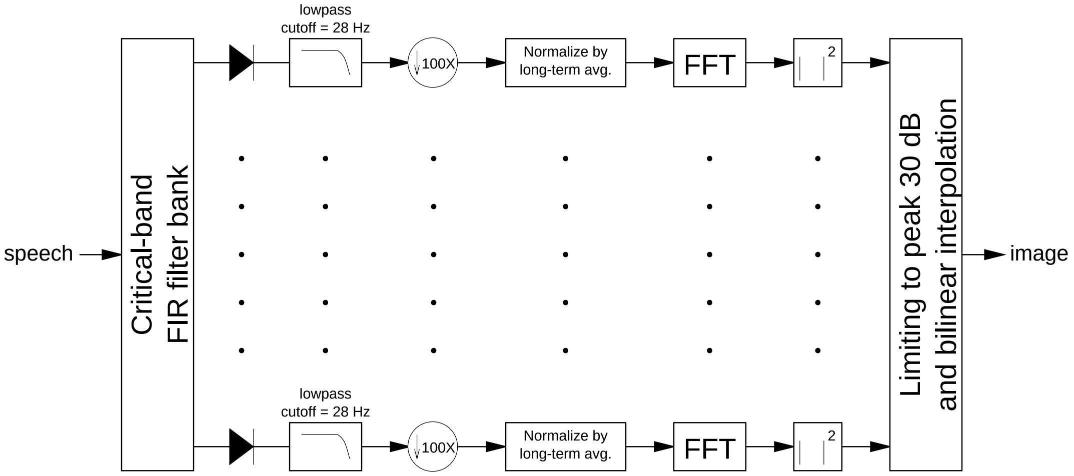
speech
Critical-band FIR filter bank
lowpass cutoff = 28 Hz
100X
Normalize by long-term avg.
FFT
2
lowpass cutoff = 28 Hz
100X
Normalize by long-term avg.
FFT
2
Limiting to peak 30 dB and bilinear interpolation
image

Figure 6. A schematic diagram of the signal processing underlying the computation of the modulation spectrogram. The spectrum is divided into 18 1/4-octave channels, the output of each is half-wave rectified, down-sampled by a factor of 100 and an FFT computed and this magnitude squared to obtain the power in the 4-Hz-centered tap (which comprises significant energy [10 dB down] between 0 and 8 Hz). This magnitude is mapped to a color table and thresholded to remove energy more than 30 dB down from the peak energy level during the utterance. Reprinted from [20].

This relative tolerance of temporal asynchrony provides an important foundation for the seamless, transparent nature of speech decoding under all but the most highly reverberant conditions. Figure 7 schematically illustrates the nature and relation of the variety of temporal analyses germane to speech processing that are likely to characterize the human listener's search for meaning when engaged in spontaneous discourse.

The actual process of speech decoding is likely to involve dozens, if not hundreds of parametric analyses, all proceeding in parallel. The extraction of information and its interpretative framework can be likened to a process of "hyper-triangulation" in an n-dimensional space through time, where n is likely to exceed 50. None of these dimensions is encoded with sufficient precision to provide a comprehensive, robust representation of the linguistic information contained within the speech signal. The process of speech understanding involves, rather, a complex process of deduction, whereby patterns of convergence across some proportion of these dimensional analyses provides the interpretative specificity and precision absent from any single representational tier.

Current-generation ASR systems focus on delineation of just a few of the linguistically relevant tiers (typically, sequences of phones and words, and their co-occurrence behavior). Such a limited number of representational tiers has proven sufficient to successfully decode speech under highly artificial and constrained communication environments, but is unlikely to yield the sort of quantal improvements required to recognize speech under the sorts of environmental and speaking conditions that characterize the majority of linguistic interactions among humans. In order to achieve truly robust, flexible speech recognition, a new paradigm is required, one that focuses on *extraction of linguistic information* derived from coarse specification of many representational dimensions rather than on the mere recognition of lexical elements derived from sequences of phones. However, an information-based approach will necessarily transcend current recognition strategies, and in so doing, bring us a step closer towards true speech understanding.

ACKNOWLEDGMENTS

Many individuals have contributed to the data and concepts presented in this paper. Dan Ellis, Joy Hollenback and Su-Lin Wu computed the data pertaining to syllable frequencies and segmentation. Su-Lin Wu also provided the data pertaining to word frequencies. Brian Kingsbury developed the implementation of the modulation spectrogram described in the text and provided the data concerning the recognition of speech in a reverberant environment. Carlos Avendano (Oregon Graduate Institute), in collaboration with Jim West and Gary Elko of Bell Laboratories, provided the impulse response used in the recognition experiments. Grateful appreciation is also extended to members of the ICSI Realization Group who have discussed and provided feedback on many of the ideas presented in this paper.

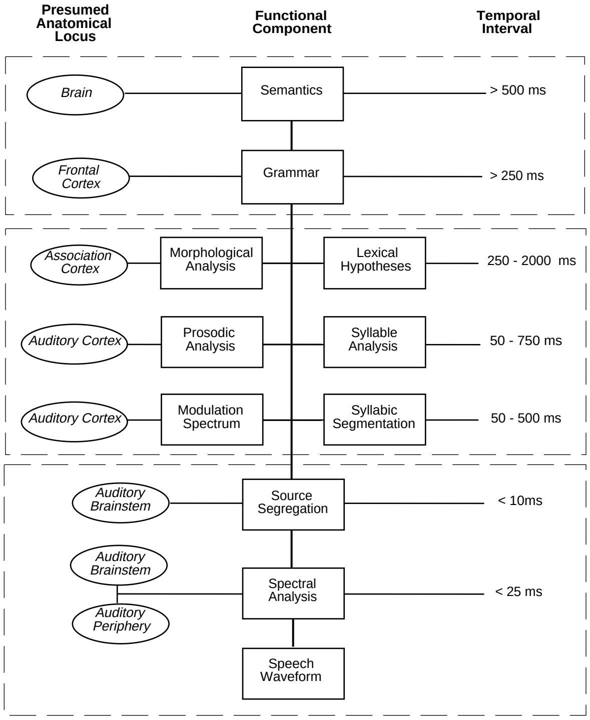

Figure 7. A schematic illustration of the functional speech analysis performed by human listeners, and its relation to the neurological bases of the temporal binding across the tiers of linguistic organization.

REFERENCES

[1] Acero, A. (1990) *Acoustical and Environmental Robustness in Automatic Speech Recognition*. Ph.D. Dissertation, ECE Department, Carnegie-Mellon University.

[2] Arai, T., Hermansky, H. Pavel, M. and Avendano, C. (1996) Intelligibility of speech with high-pass filtered time trajectories of spectral envelopes, in *ICSLP-96, Fourth International Conference on Spoken Language Processing*, Philadelphia, pp. 2490-2493.

[3] Avendano, C. and Hermansky, H. (1996) Study on the dereverberation of speech based on temporal envelope filtering, in *ICSLP-96, Fourth International Conference on Spoken Language Processing*, Philadelphia, pp. 889-892.

[4] Bever, T. G. and Hurtig, R. R. (1975) Detection of a nonlinguistic stimulus is poorest at the end of a clause. *J. Psycholing. Res.*, 4, 1-7.

[5] Cassidy, K.W. and Kelly, M.H. (1991) Phonological information for grammatical category assignments. *Journal of Memory and Language*, 30, 348-369.

[6] Cooke, M. (1993) *Modeling Auditory Processing and Organisation*. Cambridge: Cambridge University Press.

[7] Cutler, A. (1996) The comparative study of spoken-language processing, in *ICSLP-96, Fourth International Conference on Spoken Language Processing*, Philadelphia, p. 1.

[8] Dermody, P. (1992) Human capabilities for speech processing in noise, in *Proceedings of the ESCA Workshop on Speech Processing in Adverse Conditions*, M. Grenie and J.C. Junqua (eds.), Cannes-Mandelieu, France, pp. 11-20.

[9] Dewey, G. (1923) *Relative Frequency of English Speech Sounds*. Cambridge: Harvard University Press.

[10] Doddington, G. (1996) Personal communication.

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

[12] Flanagan, J., Johnston, J., Zahn, R. and Elko, G. (1985) Computer-steered microphone arrays for sound transduction in large rooms. *J. Acoust. Soc. Am.*, 78, 1508-1518.

[13] French, N. R., Carter, C. W. and Koenig, W. (1930) The words and sounds of telephone conversations. *Bell System Technical Journal*, 9, 290-324.

[14] Godfrey, J. J., Holliman, E. C. and McDaniel, J. (1992) SWITCHBOARD: Telephone speech corpus for research and development, *ICASSP-92, IEEE International Conference on Acoustics, Speech and Signal Processing*, 1, pp. 517-520.

[15] Grant, K. W. and Walden, B. E. (1996) Spectral distribution of prosodic information. *J. Speech Hearing Res*., 39, 228-238.

[16] Greenberg, S. (1996) Understanding speech understanding - towards a unified theory of speech perception, in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, W.A. Ainsworth and S. Greenberg (eds.), Keele, England, pp. 1-8.

[17] Greenberg, S. (1996) Auditory processing of speech, in *Principles of Experimental Phonetics*, N. Lass (ed.), St. Louis: Mosby, pp. 362-407.

[18] Greenberg, S. (1996) *The Switchboard Transcription Project. Johns Hopkins Workshop on Innovative Techniques for Large Vocabulary Continuous Speech Recognition Technical Report*, Baltimore.

[19] Greenberg, S., Hollenback, J. and Ellis, D. (1996) Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus, in *ICSLP-96, Proceedings of the Fourth International Conference on Spoken Language*, Philadelphia, S32-35.

[20] Greenberg, S. and Kingsbury, B. (1997) The modulation spectrogram: In pursuit of an invariant representation of speech, in I*CASSP-97, IEEE International Conference on Acoustics, Speech and Signal Processing*, Munich.

[21] Hafner, K. (1997) Stop, Alexander, Stop! *Newsweek* February 17th issue, p. 72.

[22] Hansen, J. H. L. and Bria, O. N. (1990) Lombard effect compensation for robust automatic speech recognition in noise, in *ICSLP-90, First International Conference on Spoken Language Processing*, Tokyo, 1, pp. 1125-1128.

[23] Hockett, C. (1960) The origin of speech, *Scientific American*, September issue.

[24] 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.

[25] Kingsbury, B., Morgan, N. and Greenberg, S. (1997) Improving ASR performance for reverberant speech, in *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, this volume.

[26] Lane, H. and Tranel, B. (1971) The Lombard sign and the role of hearing in speech. *J. Speech Hear. Res.*, 14, 677-709.

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

[28] Lippman, R. (1996) Speech perception by humans and machines, in *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, W.A. Ainsworth and S. Greenberg (eds.), Keele, England, pp. 309-316.

[29] Massaro, D.W. and Cohen, M.M. (1993) Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication*, 13, 127-134.

[30] Miller, G. (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psych. Rev.*, 63, 81-97.

[31] Moreno, P.J. (1996) *Speech Recognition in Noisy Environments*. Ph.D. Dissertation, ECE Department, Carnegie-Mellon University.

[32] O'Shaughnessy, D. (1987) *Speech Communication*. Reading, MA: Addison-Wesley.

[33] Rabiner, L.R. and Juang, B.-H. (1993) *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice Hall.

[34] Sapir, E. (1921) *Language: An Introduction to the Study of Speech*. New York: Harcourt.

[35] Segui, J. Dupoux, E. and Mehler, J. (1990) The role of the syllable in speech segmentation, phoneme identification, and lexical access, in *Cognitive Models of Speech Processing: Psycholinguistic and Computational Perspectives,* G. Altmann, (Ed.), Cambridge: MIT Press, pp. 263-280.

[36] Shannon, R. V., Zeng, F.-G., Kamath, V. and Wygonski, J. (1995) Speech recognition with primarily temporal cues. *Science*, 270, 303-304.

[37] Slaney, M. and Lyon, R. F. (1993) On the importance of time - a temporal representation of sound, in *Visual Representations of Speech*, M. Cooke, S. Beet, and M. Crawford (eds.), Chichester: Wiley, pp. 95-116.

[38] Smith, C., Browman, C., McGowan, R. and Kay, B. (1993) Extracting dynamic parameters from speech movement data. *J. Acoust. Soc. Am.*, 93, 1580-1588.

[39] Smith, F. (1985) *Reading*, 2nd ed., New York: Cambridge University Press,

[40] Smoorenburg, G. F. (1992) Speech reception in quiet and in noisy conditions by individuals with noise-induced hearing loss in relation to their pure tone audiogram. *J. Acoust. Soc. Am.*, 91, 421-437.

[41] Stern, R., Acero, A., Liu, F. H and Ohshima, Y. (1996) Signal processing for robust speech recognition, in *Automatic Speech and Speaker Recognition*, C.H. Lee, F.K. Soong and K.K. Paliwal (eds.), Boston: Kluwer.

[42] Strange, W. (1989) Dynamic specification of coarticulated vowels spoken in sentence context. *J. Acoust. Soc. Am.*, 85, 2135-2153.

[43] Vroomen, J. and Gelder, B. de (1994) Speech segmentation in Dutch: No role for the syllable, in *ICSLP-94, Third International Conference on Spoken Language Processing*, Yokohama, pp. 1135-1138.

[44] Wu, S-L., Shire, M., Greenberg, S. and Morgan, N. (1997) Integrating syllable boundary information into speech recognition, in ICASSP-97, IEEE International Conference on Acoustics, Speech and Signal Processing, Munich.

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