RECOGNITION IN A NEW KEY - TOWARDS A SCIENCE OF SPOKEN LANGUAGE

Steven Greenberg

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

ABSTRACT

Automatic speech recognition in the twenty-first century will strive to emulate many properties of human speech understanding that currently lie beyond the capability of present-day systems. Such future-generation recognition will require massive amounts of empirical data in order to derive the organizational principles underlying the generation and decoding of spoken language. Such data can be efficiently collected through systematic computational experimentation designed to identify the important building blocks of speech and delineate the nature of the structural interactions among linguistic tiers associated with the extraction of semantic information.

1. INTRODUCTION

Human listeners are capable of understanding spoken language under an exceedingly broad range of acoustic conditions [16] and interactional contexts [15,18]. Such contextual versatility has thus far eluded the grasp of automatic speech recognition (ASR) systems which routinely degrade in the presence of background noise or when the conversational style is informal. Why should this be so?

Over the past two decades ASR research has focused on statistically training systems to 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. This new, improved system is then turned loose upon a different corpus de jour, with predictably discouraging results, occasioning yet another round of research and development.

The wisdom of this "corpus hopping" strategy is called into question by the ASR community's recent experience with the Switchboard corpus, which comprises spontaneous, informal dialogs recorded over the telephone between individuals discussing such topics as international politics, workplace dress codes and the like [9]. This corpus is among the first to incorporate a realistic speaking style with the sorts of filled pauses, hesitations, corrections, speech errors, phonetic/lexical deletions and transformations typical of spoken discourse. After four years of intensive effort, word recognition performance for Switchboard remains mired at the 50-70% correct level [3,7], far below human performance. Why should this be so?

2. ALL SEGMENTS ARE CREATED EQUAL

The roots of our community's discontent can be traced to two distinct, yet related problems. One lies in our conception of spoken language. Linguistic theory has traditionally built elaborate descriptive frameworks for each "level" of the language hierarchy, ranging from the articulatory-acoustic (phonetic features, phones) to the phonological (phonemes) and grammatical (morphemes, syntactic elements), as well as the semantic (lexical elements, sememes) tiers of organizational abstraction [15]. Each tier is typically treated as an independent level derived from an abstraction of lower organizational levels, enabling words to be characterizable as sequences of phonemes, analogous to a lexical entry in a dictionary, while phonemes, in turn, are decomposable into constituent phonetic elements and features. Within this framework, all segments, whether they occur at the phonetic, lexical or morphological level, possess equal status within the statistical (Hidden-Markov-Model-based) machinery designed to go from sound to meaning (where "meaning" is the product of the lexical identities derived from HMMs). Although this "beads on a string" approach has worked relatively well for recognition of highly stylized [8] or lexically circumscribed corpora, it has been less successful with spoken language characteristic of the real world. In order to understand why this is so let us first consider some of the statistical properties of spontaneous speech.

3. HOWEVER, SOME SEGMENTS ARE MORE EQUAL THAN OTHERS

3.1 In the beginning there was the word...

The most common words occur much more frequently (by at least several orders of magnitude) than the least common [13], the profile conforming (in approximate fashion) to a 1/f distribution. 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 [13]. The most frequently occurring words generally come from the so-called "closed" or "function" class words such as pronouns, articles, conjunctions and modal/auxiliary verbs and the majority of the remainder stem from just a few basic nominal, adjectival or verbal forms [13]. Mastery of these hundred most common words almost assuredly facilitates comprehension of spoken discourse.

3.2 Let there be .... syllables

Although a list of common words does not provide a sufficient basis with which to interpret the speech stream by itself, it can be used in conjunction with *other knowledge sources* to considerably reduce the uncertainty. One means by which to accomplish this objective 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 contain but a single syllable. Of the 100 most frequent lexical items only ten are polysyllabic (and all of these exceptions contain just two syllables). This lexical preference for syllabic brevity is consistent with Zipf's law [23] and has potentially important implications for decoding the speech signal [11].

Although only 22% of the Switchboard lexicon is composed of monosyllabic forms, fully 81% of the corpus tokens are just one syllable in length [13]. This statistical skew towards short syllabic forms provides a potential interpretative constraint on the decoding of the speech stream. Knowing the number of syllables in a word provides some degree of

Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, Seattle, pp. 1041-1045, 1998

grammatical information as 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 an 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 [4].

The utility of the syllable as a hypothesized unit of spoken language becomes even more apparent when considering pronunciation variation. In spontaneous, informal speech the phonetic realization often differs markedly from the canonical phonological form. Entire phone elements are frequently dropped or transformed into other phonetic segments. 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 [13]. However, the patterns of phonetic variation are relatively straightforward to describe within a syllabic framework, when the syllable (e.g., [k] [ae] [t], "cat") is divided into three components, the onset [k], nucleus [ae] and coda [t]. The onset is typically the most well-preserved portion of the syllable, while the coda 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 [13,14]. 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.

4. WHAT IS THE ESSENCE OF SPOKEN LANGUAGE?

ASR systems currently assume that the path to "meaning" is paved with words and that the route to the word goes through the phone(me) (and some believe that the most direct course to the phone lies with articulatory-based features). Yet this linear hierarchy makes certain assumptions about both the organization and decoding of spoken language that are neither empirically substantiated, nor lend themselves to the sorts of sophisticated models required to derive meaning from the speech stream. Future-generation models will necessarily incorporate a dynamic linkage among the linguistic (e.g., the syllabic and lexical) tiers, based on both statistical and abstract categorical criteria. The statistical regularities observed at each tier provide an interpretative framework with which to characterize the speech signal and relate these to other representational levels. Although the statistical patterns observed on any single tier 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. Meaning can be likened to an "emergent" property, derived from the analysis of many different representational tiers, both observed and hidden. Because speakers rarely talk in "words" or "phones" per se, except in so far as a medium for communicating their intent, it is perilous to derive the meaning of a spoken utterance merely from a reconstruction of the lexical or phone sequence, as is currently done in ASR systems.

If the path to meaning does not lie through words, from whence does it come? And what is truly being communicated through speaking, if not words or phones? The answers are not immediately apparent, but may be of more than passing interest and utility for building ASR systems focused on "understanding" rather than on lexical transcription.

5. THE LAND OF MAKE BELIEVE

Yet another source of discontent among the ASR community concerns the current methods used in developing quantitative models for automatic recognition of spoken language. Currently, the statistical models and units of representation are based on traditional linguistic theory (as described above), but neither the conventional units nor the HMM-based framework is fundamentally rooted in spoken-language behavior. The limited success demonstrated by HMM systems to date more likely reflects engineering ingenuity than some inherent capability of the statistical models for characterizing spoken language.

A potentially more fruitful path to spoken language recognition may be found through the precepts of the hypothetico-deductive method [19], successfully tested and refined by students of various scientific fields over the past several hundred years. A typical state sequence for such an endeavor is illustrated in Figure 1. This regime of observation, hypothesis formulation, falsification and hypothesis refinement lies at the heart of the scientific method. Central to this endeavor is the practice of controlled experiments in which as many factors pertaining to the phenomenon of interest are "held constant," while a single parameter is varied systematically. This methodology can be applied to spoken language in such a fashion as to pose the following questions:

(1) What are the basic "building blocks" of speech? Articulatory feature? Phone? Syllable? Word? Phrase? Other?
(2) How are these linguistic "elements" bound together into the organic-like "compounds" associated with speech?

ASR systems currently display a predilection for the phone and the word as the basic units with which to model the speech stream. It would be logical to assume that these units have emerged as victors after a hotly contested trial in which all contenders have been scrutinized and evaluated. However, it appears that the special status accorded the phone and the word are derived from linguistic "theory," rather than from experiment. Those in favor of such units, based on their "self-evident" nature (typically derived from orthographic sources) may wish to consult a dictionary of the English language issued prior to the time of Samuel Johnson [10] or attempt to partition medieval text into lexical forms on the basis of spatial segmentation [5].

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. Speech understanding involves rather, a complex process of deduction whereby patterns of convergence across some proportion of

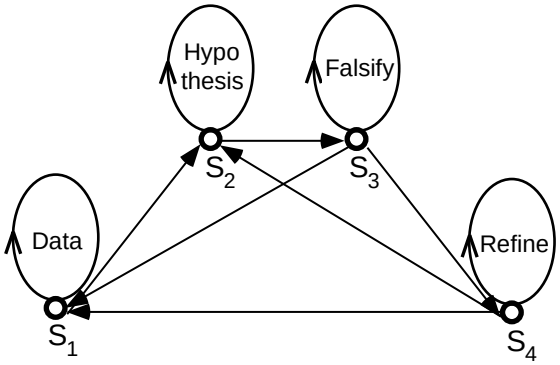

Figure 1. The hypothetico-deductive method, as applied to automatic speech recognition. Each stage of the scientific process incorporates a recursive potential, as well as interactions with the other stages.

---

these multi-dimensional analyses provides the interpretative specificity and precision absent from any single representational tier. The process is analogous, in certain respects, to the framework described by Zadeh for "information granulation" [22] with respect to "computing with words" [21]. There are many aspects of spoken language that are not readily amenable to precise quantitative characterization. This degree of uncertainty is usually expressed in terms of probabilities, but it is not entirely clear that such a stochastic framework captures the essence of the ambiguity inherent in spoken language. ASR systems currently 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 [6, 11]. How can this be if ASR systems do so much more poorly on recognizing speech within a larger context?

One potential explanation lies in the dynamic linkage among the representational tiers of language, that enables 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 [14]. 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 [12, 14]. Speakers appear to have an intuitive understanding of the relationship among 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 [17], 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.

6. THE GOD THAT FAILED

The current *practice* of automatic speech recognition is not readily compatible with the hypothetico-deductive method. Clearly defined end points for each stage of the analysis are rarely delineated, nor is there a clear sense of the precision required at any level of linguistic analysis for the successful decoding of the speech signal at a different tier. Thus, there is no clearly specified criteria for successful performance except in terms of the "bottom line," namely word recognition accuracy (though a counter current is beginning to emerge, cf. [1]). For example, it is clear that some degree of detail is required at the phonetic level for accurate word recognition to occur, but how much is sufficient? Are there instances where an overabundance of phonetic detail actually impedes lexical recognition? Are there other sources of linguistic representation, such as grammatical part of speech or syllable type, that could be used, in conjunction with more conventional categories, to more accurately infer the words spoken or the meaning intended?

Combining information from separate streams of the speech signal is becoming increasingly popular (e.g., [2, 20]). And yet there is relatively little effort expended so far toward the manner in which the information should be combined, or at what level (e.g., acoustic frame, phone, syllable). A more efficient and economical strategy is in order, one that focuses on discovery of underlying principles of spoken language structure, not just on statistical pattern recognition.

7. PLAYING DICE WITH THE (LINGUISTIC) UNIVERSE

A more principled approach for speech recognition may lie in "playing dice" with the linguistic universe. Imagine playing "overlord" of a linguistic terrain. Assume that you have "perfect" knowledge of the articulatory features associated with the speech stream during a finite-duration dialog. Would such knowledge be sufficient to reconstruct the sequence of phonetic or lexical units in the utterance? To the extent not, what other sorts of information (e.g., prosody, duration [17] grammar) would help? Or assume that you have "imperfect" knowledge of the articulatory features (and have "perfect" knowledge of these imperfections), how would this situation affect recognition performance?

Such information could be obtained by running systematic "cheating" experiments in which much of the linguistic detail is already "known," both in kind and degree. How much detail is actually required from various linguistic levels for accurate reconstruction of the word sequence or of the underlying message? Is there a trade-off between the number of different representational tiers utilized and the precision of the representation at each level? Without such knowledge it will be difficult to ascertain precisely how much benefit there is to be gained from developing better models for specific components of the language, nor will it be easy to understand the nature of the interaction among different linguistic tiers without such empirical support. ASR systems of the future will need to be both adaptive and flexible with respect to environmental and contextual factors. Understanding the basic principles underlying the specification of linguistic information may provide the most efficient means with which to design robust, reliable ASR systems in the century ahead.

8. THE SOUL OF A NEW (ASR) MACHINE

Controlled experiments can provide the basis for designing new machinery for automatic recognition (and some day, understanding) of spoken language. Current HMM-based systems may not be particularly efficient (or adept) at exploiting the diversity of information derivable from more

than a few organizational tiers of linguistic representation. Their limits can be ascertained through experiment.

To the extent that their capability is found lacking, it should be possible to design new, statistically oriented machinery that incorporates a "convergence" principle of information extraction. This principle exploits the statistically systematic relationship among different linguistic tiers to enable accurate inferences to be made concerning the information associated with a sequence of coarsely specified elements on *many different* representational tiers. Even though no single level is specified with precision, the convergence of coarsely granulated representations yields a unique (or near-unique) solution.

"Missing" data (at the phonetic, lexical or grammatical levels) often does not impede successful information extraction, since the message is distributed across many tiers concurrently and in such a fashion as to be relatively impervious to environmental degradations or variation in speaking style and pronunciation. This is language's way of handling the potentially "hostile forces of nature" and the acoustic variability associated with communication by a variety of sources under a wide range of often unpredictable conditions.

ACKNOWLEDGMENTS

Support by the U.S. Department of Defense, the National Science Foundation and the Center for Language and Speech Processing (Johns Hopkins) is gratefully acknowledged, as are fruitful discussions with Nelson Morgan, Hynek Hermansky, Hervé Bourlard and other members of the Realization Group at ICSI on topics germane to this paper. However, the opinions expressed reflect solely those of the author.

REFERENCES

[1] Bourlard, H., Hermansky, H. and Morgan, N. "Towards increasing speech recognition error rates." *Speech Communication,* 18: 205-231, 1996.

[2] Bourlard, H. and Dupont, S. "Subband-based speech recognition," in ICASSP-97, *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1997, pp. 1251-1254.

[3] Byrne, W., Finke, M., Khudanpur, S., McDonnough, J., Nock, H., Saraclar, M., Wooters, C. and Zavaliagkos, G. "Pronunciation modelling for conversational speech recognition - A status report from WS97," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 26-33.

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

[5] Crosby, A. *The Measure of Reality, Quantification and Western Society 1250-1600*. Cambridge: Cambridge University Press, 1997.

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

[7] Ganapathiraju, A., Goel, V., Picone, J., Corrada, A., Doddington, G., Kirchhoff, K., Ordowski, M. and Wheatley, B. "Syllable - A promising recognition unit for LVCSR," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 1996, pp. 207-214.

[8] Gauvain, J.L., Lamel, L.F., Adda, G. and Adda-Decker, M. "The LIMSI continuous speech dictation system: evaluation on the ARPA Wall Street Journal task," ICASSP-94,*IEEE International Conference on Acoustics, Speech and Signal Processing*, 1994, pp. 557-560.

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

[10] Green, J. *Chasing the Sun, Dictionary Makers and the Dictionaries They Made*. New York: Henry Holt, 1996.

[11] Greenberg, S. "Understanding speech understanding - towards a unified theory of speech perception." *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, 1996, pp. 1-8.

[12] Greenberg, S., Hollenback, J. and Ellis, D. "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus." *International Conference on Spoken Language Processing*, Philadelphia, 1996, pp. S32-35.

[13] Greenberg, S. "On the origins of speech intelligibility in the real world," ESCA Workshop on *Robust Speech Recognition for Unknown Communication Channels*, Pont-a-Mousson, France, 1997, pp. 23-32.

[14] Greenberg, S. "The Switchboard Transcription." Project 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, 1997.

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

[16] Lippman, R. "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, 1996, pp. 309-316.

[17] Pols, L., Wang, X. and ten Bosch, L. "Modelling of phone duration (using the TIMIT database) and its potential benefit for ASR." *Speech Communication*, 19, 161-176, 1996.

[18] Pols, L. "Flexible human speech recognition." *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pp. 273-283.

[19] Popper, K. (1959) *The Logic of Scientific Discovery*. New York: Basic Books. [German edition, 1934]

[20] Tibrewala, S. and Hermansky, H. "Sub-band based recognition of noisy speech," *ICASSP-97, IEEE International Conference on Acoustics, Speech, and Signal Processing* 2, 1997, pp. 1255-1258.

[21] Zadeh, L.A. "Fuzzy logic = computing with words." *IEEE Transactions on Fuzzy Systems*, 4: 103-111, 1996.

[22] Zadeh, L.A. "Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic." *Fuzzy Sets and Systems*, 90: 111-127, 1997.

[23] Zipf, G. K. "The meaning-frequency relationship of words." *J. Gen. Psych.*, 33: 251-256, 1945.