

FROM HERE TO UTILITY

Melding Phonetic Insight with Speech Technology

Steven Greenberg

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

Abstract Technology and science are often perceived as polar extremes with respect to spoken language. Speech applications rarely incorporate scientific insight and conversely, basic research is often viewed as oblivious to practical concerns of the real world. Melding phonetic insight with speech technology can, however, yield extremely productive results for both applications and basic science if performed within the appropriate theoretical framework. Such an approach is illustrated with respect to the relation between prosodic (stress accent) and phonetic properties of conversational telephone dialogues (American English)) using the Switchboard corpus. Phonetic properties, such as vocalic identity and duration, are shown to reflect prosodic phenomena, and thus could be used to enhance the quality of automatic speech recognition performance, as well as provide detailed insight into the nature of spoken language.

Keywords: Speech technology, automatic speech recognition, prosody, phonetics, spontaneous speech, syllable structure

1. INTRODUCTION

It is twelfth-century Japan, and a nobleman has been killed. A magistrate is charged with establishing the identity of the killer and delineating the sequence of events leading up to the murder. Several witnesses are called to testify – the victim’s wife, the accused (a notorious bandit), a woodsman as well as the victim himself (through a spirit medium). Each witness provides a singular account of the man’s death. They agree on but a single fact – that the nobleman is dead. How he died, and by whose hand, are very much in dispute.

The story of *Rashomon* (Ritchie, 1987) is cited often in philosophical discussions of “truth.” As nothing is known (or knowable) with certainty, all knowledge is relative (and hence ephemeral). The concept of truth is a



Figure 1. A woodcutter, a priest and a peasant ponder the unfathomable nature of “truth” in their attempt to reconstruct the events leading up to a nobleman’s death in twelfth-century Japan. From the film *Rashomon*, directed by Akira Kurosawa (reprinted from Ritchie, 1987).

chimera and therefore unworthy of pursuit.

Yet, there is an alternative interpretation, one that questions not the concept of truth itself, but rather the capacity of its assimilation through a single vantage point. Perhaps the “true” message of *Rashomon* is that deep and ever-lasting knowledge can only be gained through exposure to a variety of perspectives, no single source providing sufficient depth and clarity to comprehend a situation as complex (and as tragic) as the murder of a man.

As in fiction, potentially in science

In *Rashomon* the testimony of each witness acquires new significance in light of alternative accounts (Figure 1). Can an intellectual domain as complex as *spoken language* be fully understood through a single perspective? Or must orthogonal forms of evidence be sought with which to reconstruct the “truth”?

Knowledge gained in the pursuit of “pure” research is often viewed as the pinnacle of scientific endeavor, unsullied by practical concerns of technological application and customer satisfaction. Science unfettered by pragmatic constraints is (from this perspective) the most noble of objectives and should therefore serve as the principal deity in the temple of knowledge.

As in myth, potentially in science....

How does true insight proceed from “objective” study of spoken language? Is it possible to fully comprehend the multivocal nature of a scientific domain from the exclusive vantage point of a laboratory? Or does the spirit of *Rashomon* compel us to seek testimony from a wider variety of sources in the pursuit of objective knowledge?

2. THE STRUCTURE OF SCIENTIFIC EVOLUTION

The course of a discipline’s intellectual evolution is often tortuous and of a curvilinear nature. Where does the domain of speech research lie with respect to its “great chain of being”? Is this community still engaged in determining the number of phonemes *on* a word? Or has the collective unconscious progressed to a higher plane of existence? What will the speech scientists of the *twenty-second* century write concerning the science of the *twenty-first*?

Scientific maturity is often marked by its close relation to technology. The great monuments of any age (whether they be pyramids, cathedrals or casinos) are often based on the most advanced science and technology of the age. And in turn, such monuments usually spur further progress in the domains upon whose foundations they are formed. The synergy between science and technology is simple to discern, for successful products are difficult to build on anything other than a strong and secure scientific foundation. And technology, in turn, provides a rigorous proving ground for the empirical and theoretical precepts of any discipline. Technology may thus serve as a “forcing function,” driving a field beyond the bounds of traditional scientific inquiry, posing challenges to surmount by dint of technical (and often commercial) imperative. In tandem with technology comes a focus on empiricism. It is difficult to divine how well a product is likely to work purely on the basis of theory. For theory needs to be tempered with data representative of the environment in which the technology is deployed. In such fashion a field can mature quite quickly; and thus it may ultimately come to pass with respect to speech technology.

3. THE GALAPAGOS OF SPOKEN LANGUAGE

The voyage of the *Beagle* (Darwin, 1839) provided an effective forcing function for Darwin’s thoughts on the origin of species (Darwin, 1859), particularly his trip to the Galapagos Islands, west of Ecuador. Among the fauna of those islands are many varieties of finch, who by virtue of variation in color, size and shape (particularly of the beak) came to provide crucial clues as to the mechanism of natural selection (Weiner, 1994).

Speech, as a field, is still in search of its Galapagos. Somewhere, off the coast of the intellectual mainstream, lie the finches of language – if only we knew their form and function. Should we wait patiently for their emergence? Or should we embark on our own voyage of discovery, aggressively seeking the critical evidence required to solve the mystery of spoken language?

4. UNOBTRUSIVE MEASURES

Every academic discipline has a favored means of collecting data. Astronomers gaze into the heavens, high-energy physicists smash atoms, ethologists play peeping toms, and linguists either introspect or elicit citation forms from “informants.”

Long ago, marketing researchers discovered some of the pitfalls associated with elicited data. A shopper, upon entering the supermarket, is asked to enumerate the items intended to be purchased in the store. At checkout a video camera enables a comparison of the shopper’s original list with what has actually been bought – intention and deed turn out to bear scant relation to each other; for there is scarcely a product in the shopper’s cart mentioned in the interview only a few minutes before (Ries and Ries, 1998).

Because most spoken-language data are derived from either introspection or elicitation the empirical foundations of linguistics are potentially forged on the scientific equivalent of quicksand. From a distance the foundation appears secure, only to collapse in a nebulous undertow upon closer inspection.

5. THE LINCHPIN OF FUTURE TECHNOLOGY

What is an ambitious field to do? Can a discipline reinvent itself with sufficient zeal and celerity as to accommodate the technological and societal transformations of the twenty-first century?

In this circumstance our *Beagle* (and hence salvation), is likely to emerge in the guise of scientific imperatives driven by the frenetic pace of technology. For speech is destined to serve as a technological linchpin of the twenty-first-century economy by virtue of its ability to facilitate and automate communication between humans and machines (cf. Greenberg, 2001). A unique opportunity potentially arises for a synergistic relationship between the science and technology of spoken language.

A solid empirical and theoretical foundation is generally required to develop reliable technology; speech communication is unlikely to be granted an exemption in this regard. Thus, the science of spoken language is likely to evolve quite rapidly over the coming decades as the demand for speech technology accelerates with the emergence of the “communication age.”

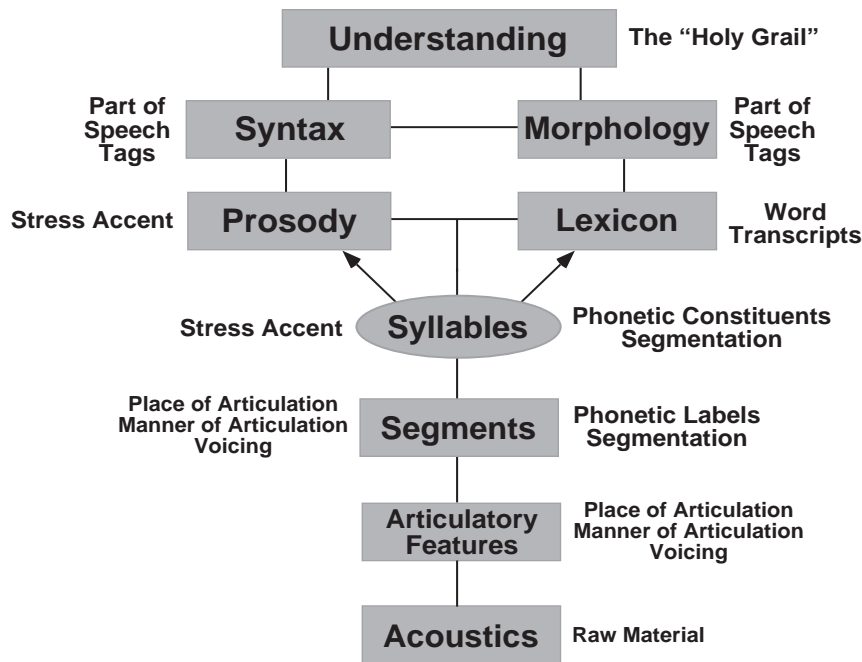


Figure 2. Corpus-centric perspective on spoken language. Manually annotated material forms the basis for statistical characterization of speech, as well as for training systems to perform automatic labeling for speech recognition. Currently, most manual annotation focuses on the lexical level and seeks to derive labels and segmentation for the lower tiers (particularly segments) via automatic methods using some form of Viterbi decoding. The quality of such automatically generated labels and segmentation boundaries is poor when applied to spontaneous corpora such as Switchboard (cf. Greenberg and Chang, 2000). There is precious little manually annotated material associated with non-lexical tiers for any language.

Sophisticated technology depends on “getting the details right” to a degree that far exceeds what passes for knowledge and insight within the domain of “pure” science (which is why applied technology research is so much more costly than basic research). With respect to speech the contrast between “pure” and “applied” research is stark indeed. Linguists and phoneticians often view spoken language through a “glass menagerie” of abstract forms, which often bear but the faintest resemblance to language spoken in the “real” world. Current speech technology (whether it be in the form of automatic speech recognition or text-to-speech synthesis) relies heavily on training materials representative of the task domain for this very reason (cf. Figure 2). Such a training-intensive approach offers many advantages over a more abstract, rule-governed framework, particularly with respect to performance. But an emphasis on machine-learning algorithms and training regimes often comes at the expense of genuine insight into the nature of spoken language and not infrequently violates the precepts of the hypothetico-deductive method (cf. Greenberg, 1998; Popper, 1959).

Speech technology can proudly point to its *apparent* success with speech recognition and concatenative synthesis in defense of its machine-learning-centric approach. And indeed, imperfect science is capable of providing an effective foundation for technology – as long as the demands of the market place are not exceedingly stringent or profound. However, as commercial expectations rise, immature science is unlikely to suffice as the empirical and theoretical foundation of future-generation technology (Greenberg, 2001).

6. THE SCIENCES OF THE SUPERFICIAL

The academic perspective on language differs markedly from that of the technologist. The linguist is primarily concerned with abstraction and structure of what is normally hidden from view, while the technologist focuses on the more superficial aspects of language (such as the acoustic signal) most amenable to computation (Figure 2); each perspective has its pros and cons.

The linguist can use extensive knowledge to make great leaps of intuition that can, on occasion, derive significant insight into spoken language (e.g., Jakobson et al., 1961). But typically such insight is of limited utility to the technologist, saddled with the gory details of daily chatter. Under such circumstances it is unsurprising that speech technology relies mainly on methods designed to automatically divine structure through statistical analysis of surface forms. Does there somewhere lie a path, between the surface and the deep, that provides a plane of mediation between linguistics and technology?

7. INTO THE WILDS (OF SPONTANEOUS SPEECH)

Scholars of medieval Europe sought, in vain, to determine the number of angels residing on the head of a pin (Lovejoy, 1939), their efforts stymied through want of empirical data.

In the realm of spoken language we are more fortunate, for the world literally reeks of material with which to quantify virtually any (superficial) aspect of human discourse; it is merely a matter of recording an appropriate mix of speakers talking in ways representative of the “real” world and then taking the time to annotate the material for statistical characterization (cf. Figure 3).

Two corpora of spoken language are particularly germane to the present discussion. “Switchboard” (Godfrey et al., 1992) has served as a development corpus for evaluation of automatic speech recognition systems for nearly a decade. The corpus contains hundreds of brief (5-10 minute) telephone

dialogues representative of casual conversation, and is thus of great use in characterizing properties of spontaneous (American English) speech. A subset (ca. five hours) of this material has been phonetically annotated by linguistically knowledgeable transcribers at the International Computer Science Institute (Greenberg, 1999) and is electronically accessible over the web (<http://www.icsi.berkeley.edu/real/stp>).

A one-hour subset of Switchboard has also been manually labeled with respect to stress-accent by two individuals not involved in the phonetic annotation. The remaining four hours has been automatically labeled using an algorithm trained on hand-labeled material (cf. Greenberg et al., 2001).

These same two individuals also labeled two and a half hours of stress-accent material from a separate (phonetically annotated) corpus, “OGI Stories” (Cole et al., 1994), containing hundreds of telephone *monologues* (of ca. 60-seconds each). These two annotated corpora provide (but) one means with which to characterize spoken language (and thereby serve to bridge the gap between linguistics and technology).

8. THE ACOUSTIC BASIS OF STRESS ACCENT

Prosodic accent is an integral component of speech, particularly for languages, such as English, that so heavily depend on it for lexical, syntactic and semantic disambiguation (thereby providing important information concerning the focus of a speaker’s attention). Languages mark accent in a variety of ways, utilizing such acoustic properties as duration, amplitude and fundamental frequency (f_0). Some languages, such as Japanese, tend to mark accent primarily in terms of fundamental frequency variation (“*pitch* accent” systems), while others, such as English and German, accentuate using a *constellation* of features (i.e., *stress*) including vocalic duration and identity, as well as fundamental frequency and other acoustic properties associated with the patterning of syllables within an utterance (Beckman, 1986; Clark and Yallup, 1990).

Traditionally, f_0 (and its perceptual correlate, pitch) has been thought to serve as a primary cue for stress accent in English (Fry, 1955; Fudge, 1984; Gimson, 1980; Lehiste, 1970):

“Pitch is widely regarded, at least in English, as the most salient determinant of prominence.... when a syllable or word is perceived as ‘stressed,’.... it is pitch height or a change in pitch, more than length or loudness that is likely to be mainly responsible....”

(Clark and Yallup, 1990; p. 280)

However, it is unclear whether such statements truly apply to language spoken in the “real” world, free from constraints imposed by scripted or non-

meaningful material recorded in the laboratory).

In an effort to resolve this thorny issue the acoustic basis of stress accent was examined as part of a project to incorporate such information into automatic speech recognition systems focused on spontaneous material from the OGI Stories corpus (Silipo and Greenberg, 1999; Silipo and Greenberg, 2000). These studies suggest that duration and amplitude appear to play a far more important role than f_0 in accounting for the stress-accent patterns observed in this corpus. Several different automatic methods (based on neural networks, fuzzy logic, and signal-detection theory melded with a threshold model) were developed for simulating the stress-accent patterns labeled in the manual transcription of the prosodic patterns. Each computational method weighted duration and amplitude far more heavily than f_0 in order to provide a faithful simulation of the stress-accent annotation (Silipo and Greenberg, 2000), consistent with recent studies examining this issue from the perspective of (American English) telephone voicemail (Koumpis and Renals, 2001) and Dutch spontaneous phone material (van Kuijk and Boves, 1999). Together, such studies suggest that pitch variation plays a much smaller role in the stress-accent pattern of spontaneous speech than has been generally believed (cf. Figure 11 and Table I, as well as Section 12, for additional material germane to this issue); thus caution is warranted in extending the conclusions of laboratory studies on stress-accent to the real world, particularly if technology is viewed as the ultimate arbiter of “truth.”

9. STRESS ACCENT AND AUTOMATIC SPEECH RECOGNITION PERFORMANCE

Stress accent is likely to prove of critical importance for future-generation speech recognition systems. Not only does it provide a potential means of identifying key words in an utterance, but such material may also help to enhance recognition performance overall. In a diagnostic study of the linguistic and acoustic factors associated with recognition performance in ASR systems using the Switchboard corpus (telephone dialogues – cf. Godfrey et al., 1992) it was determined that the stress-accent pattern is highly correlated with a specific form of recognition error, namely word deletion (Greenberg and Chang, 2000). If a word contains a primary accent it is far less likely to sustain a deletion error in recognition than if it contains only unaccented syllables (Figure 3). This pattern, observed across all eight recognition systems examined, suggests that stress-accent information could be used to improve recognition performance (particularly for large-vocabulary task domains, which generally contain a significant proportion of unaccented words) by utilizing such knowledge to interpret the acoustic signal with respect to phonetic classification and lexical segmentation.

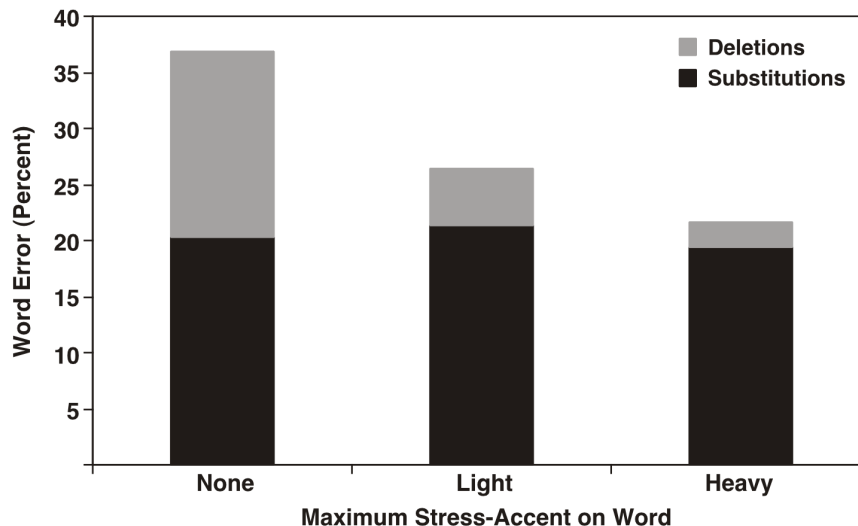


Figure 3. The relation between stress-accent level and word error in the Switchboard corpus for eight separate speech recognition systems (the data have been pooled, given the common pattern exhibited across sites). Word-deletion errors are highly correlated with stress accent level. In contrast, word-substitution errors appear unaffected by stress-accent level. Over 80% of the words are monosyllabic. Three quarters of the remainder consist of just two syllables. In polysyllabic words the maximum stress-accent level pertains to the syllable with the highest degree of accent, irrespective of the stress pattern associated with the other syllables in the word. From Greenberg and Chang (2000).

Currently, stress accent is not commonly incorporated into ASR system design. Moreover, there is no general consensus as to the specific form and nature of the prosodic parameter, especially its acoustic correlates. Perhaps there is another property of the speech signal that garners a higher degree of agreement as to its linguistic manifestation and which bears a close affinity to stress accent?

10. SYLLABLE STRUCTURE AND AUTOMATIC SPEECH RECOGNITION PERFORMANCE

Words may be classified in terms of their constituent syllable structure. Most words in English are monosyllabic and their structure is likely to be one of several forms – consonant + vowel + consonant (CVC), consonant + vowel (CV), vowel + consonant (VC) and vowel (V). Together, these syllable types account for ca. 85% of the structural forms found in spontaneous (American) English (cf. Figure 5 and Greenberg, 1999). Consonant clusters occasionally occur at either the syllable onset (e.g., CCVC) or coda (e.g., CVCC), but such forms account for only ca. 15% of the syllable types in spontaneous English (Greenberg, 1999). And a relatively small proportion of words (ca. 19% in the Switchboard corpus) contain more than a single syllable (of this number,

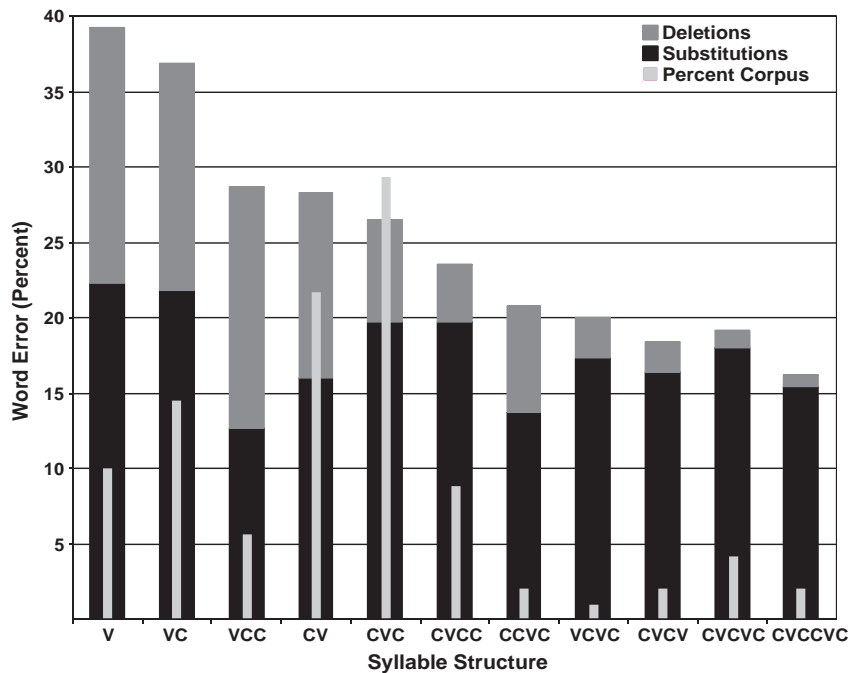


Figure 4. Relationship between word-error rate and syllable structure for Switchboard speech recognition systems. Monosyllabic words beginning with a vowel are far more likely to be mis-recognized in terms of word deletions than words beginning with a consonant or containing two or more syllables. From Greenberg and Chang, 2000).

approximately three quarters are disyllabic in form).

Of interest, in the current context, is the relation between syllable structure and word-deletion errors for the Switchboard speech recognition systems. Monosyllabic words beginning with a vowel (i.e., V, VC and VCC forms) are far more likely to be mis-recognized in terms of word deletions than other syllable forms. The governing parameter does not appear to be vocalic-initial lexical forms *per se*, as VCVC words (such as “about”) are rarely associated with word-deletion errors (Figure 4). Rather, the word-deletion rate appears linked to the stress-accent pattern associated with each syllabic form. Disyllabic words usually carry a heavily accented syllable, typically the second when the initial syllable begins with a vowel. Words with consonantal onsets also tend to carry some measure of accent. Thus, syllable structure and accent pattern are in some sense inextricably linked – two sides of the same linguistic coin. Perhaps the philosophy of Rashomon is also relevant to understanding spoken language; the phenomena under study are multifaceted and far too complex to yield their secrets viewed from just a single perspective. And there may be other perspectives (such as vocalic identity) that are equally germane.

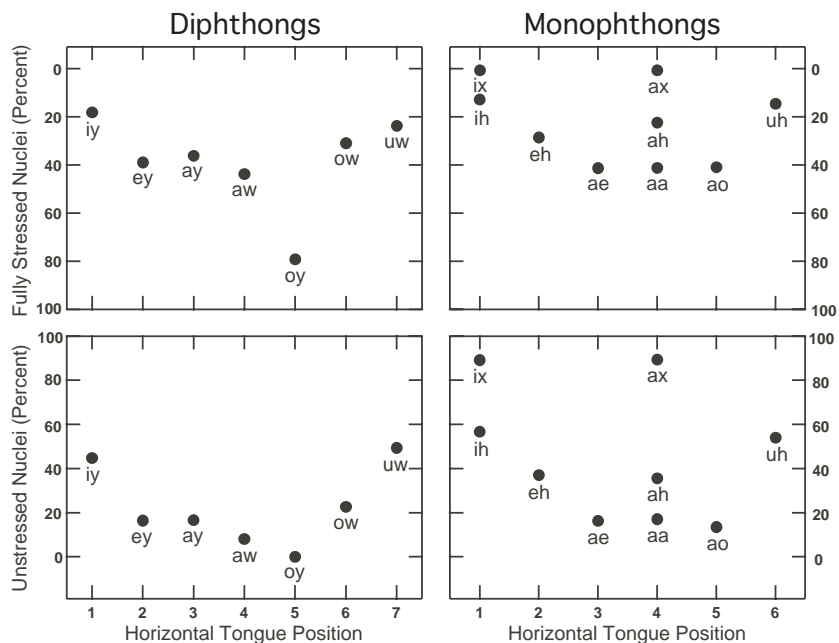


Figure 5. The proportion (in percent) of tokens for each vocalic class labeled as either completely accented (level-1 accent, top panels) or entirely unaccented (level-0 accent, bottom panels), partitioned into two broad classes, diphthongs and monophthongs (for clarity of illustration). Note reversal of scale for the ordinates associated with the top and bottom panels. This scale reversal is required to maintain the spatial relationship between vowel height and proportion of heavily accented (or unaccented) syllables. Adapted from Hitchcock and Greenberg (2001).

11. STRESS-ACCENT AND VOCALIC IDENTITY

In principle, stress accent is independent of vowel quality (with each vocalic segment capable of assuming any degree of stress), and therefore the distribution of accent should be relatively uniform across the vocalic inventory. From this perspective, stress accent is largely a lexical phenomenon, where each word has its distinctive accent pattern (as defined in a pronouncing dictionary) that is only marginally influenced by embedding within the context of spoken discourse. And as there is an arbitrary relation between sound (in this instance, vowels) and symbol (i.e., words) there should be little evidence of a systematic relationship between stress-accent and lexical form.

However, a rather different pattern emerges from analysis of the Switchboard corpus (cf. Figure 5). High vowels (e.g., [ih], [uh]) are far more likely to be unstressed than low vowels (e.g., [ae], [aa], [ao]); this relation between vowel height and stress accent extends to diphthongs as well. Thus, [iy] and [uw] are much less frequently accented than [aw] and [ay]. Moreover, the relation between vowel height and stress accent is graded. Mid-height

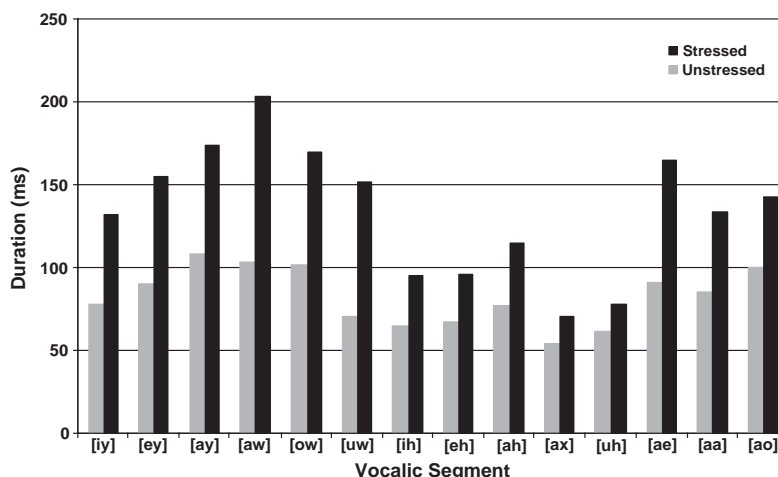


Figure 6. The relationship between segment duration and vocalic identity. Stressed nuclei are consistently longer in duration than their unstressed counterparts. The difference in duration is particularly marked for diphthongs and low monophthongs, and is smallest for the high monophthongs (which are rarely heavily accented). Only segments consistently labeled as fully stressed or entirely unstressed are included in the analysis. Fully stressed [ix] segments were too few to include in the analysis. From Greenberg et al., 2001.

vowels, such as [eh], [ey], [ah] and [ow] exhibit a stress-accent pattern intermediate between their low and high vocalic counterparts (Hitchcock and Greenberg, 2001; Greenberg et al., 2001).

The relation between vocalic identity and stress accent appears to go far deeper than a mere statistical association between parameters. Vocalic duration, for example, is highly correlated with stress accent. Stressed nuclei are often 50% to 100% longer in duration than their unstressed counterparts (cf. Figure 6). In consequence, duration and stress accent are highly correlated in spontaneous discourse (cf. Figure 6). Moreover, there is a close association between duration and vowel height (Figure 7; Hitchcock and Greenberg, 2001; Peterson and Lehiste, 1960) that is likely to be linked to stress accent as well. Duration may hence serve as a secondary (and under certain circumstances, even as a primary) cue to vowel height.

Vocalic amplitude is also correlated with vowel height (Figure 7), though not *at first glance* to the degree exhibited by duration. Vowel height is directly correlated with the frequency of the first formant; “high” vowels are associated with a low-frequency F_1 (225 - 350 Hz) while “low” vowels have a high F_1 (700 - 800 Hz). The audibility function for human hearing changes markedly over this range, so that a component at 800 Hz is likely to be as much as 20 dB louder than one at 250 Hz. Thus, the seemingly small disparity in amplitude between high and low vowels may actually be considerably larger when perceptually relevant factors are taken into account.

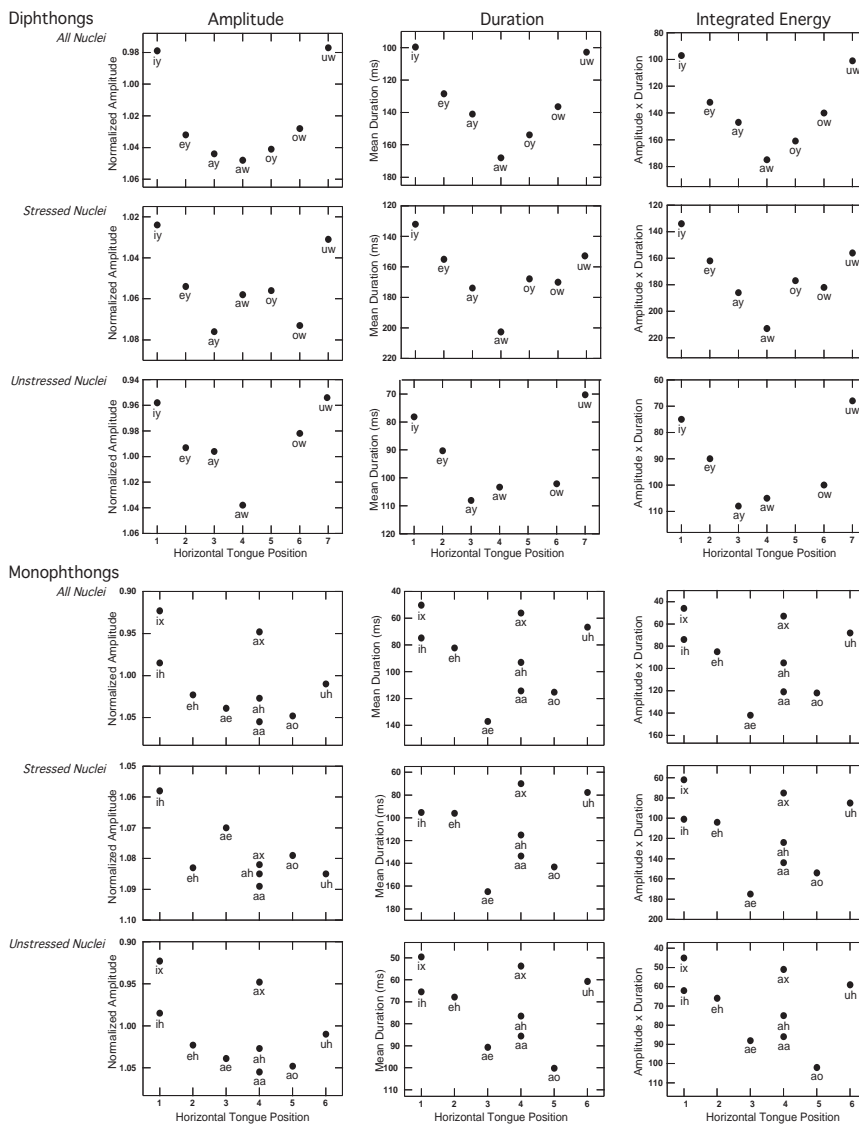


Figure 7. Spatial patterning of the duration, amplitude and integrated energy of vocalic nuclei as a function of stress level (heavy or none), as well as for occurrences averaged across all levels of accent. The data are partitioned into two classes, diphthongs and monophthongs, in order to highlight the patterns. The data points represent averages for each vocalic class. The standard deviations were relatively uniform and are therefore omitted (but are provided in a more extended account – Hitchcock, 2001). The vocalic labels are derived from the Arpabet orthography (cf. Greenberg, 1997 for a description of the phonetic inventory). Horizontal tongue position is schematic in nature and is not intended to denote articulatory measurement (but is *roughly* correlated with the frequency of the second formant). From Hitchcock and Greenberg (2001).

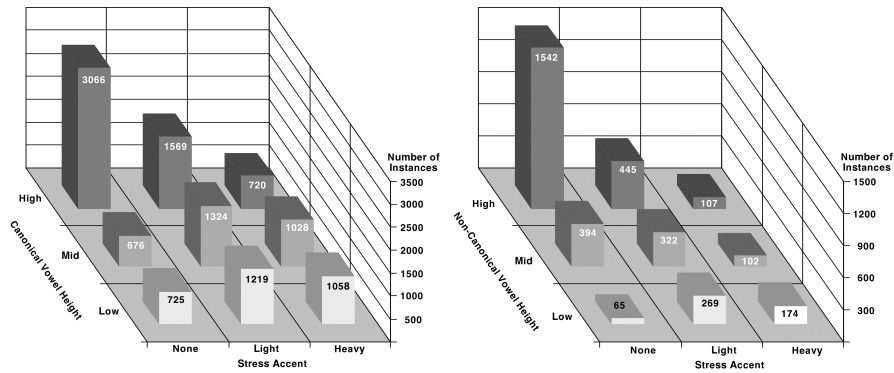


Figure 8. The impact of stress accent on the number of vocalic segments associated with high, mid and low articulatory height (cf. Figure 10 for the relation between segmental *identity* and vowel height), partitioned into canonical (left panel) and non-canonical forms (right panel). Note the difference in scale between the two panels. There is a pronounced skew towards the high vowels for both the canonical and non-canonical forms associated with unaccented syllables. From Greenberg et al. (2002).

In some very real sense stress-accent and vowel height may not be entirely distinguishable. Vocalic distinctiveness is, in principle, based on a pattern associated with formants one, two and three (Ladefoged, 1993); yet duration (bound with stress-accent) appears to play an important role as well (cf. Figures 6 and 7), reflected, perhaps, in the pattern of vocalic reduction observed in spontaneous speech (cf. Lindblom, 1990).

The consequence of such patterns is a systematic relation between vowel height and stress-accent pattern. Tongue height associated with vocalic forms in unaccented syllables is *far* more likely to be high than mid or low, for both canonical and non-canonical realizations of syllables and words (Figure 8). The distribution of vowels with respect to tongue height is of a far more even nature for syllables with some degree of stress accent (either light or heavy) relative to those without.

As a consequence of this relation between stress accent and vowel height the overall distribution of unaccented vocalic forms differs dramatically from those associated with heavily accented syllables (Figure 9). The overwhelming majority of vocalic forms in unaccented syllables are in the high-front and high-central regions of the vowel space. The number of low and mid vowels associated with such syllables is rather small. Many (but not all) of the words incorporating such unaccented syllables are “function” words (such as conjunctions, articles, pronouns and demonstratives) which occur with great frequency in conversational speech. Thus, a listener may be “primed” to interpret unaccented syllables as function words under many circumstances (barring evidence to the contrary).

There is a relatively even distribution of vocalic forms associated with

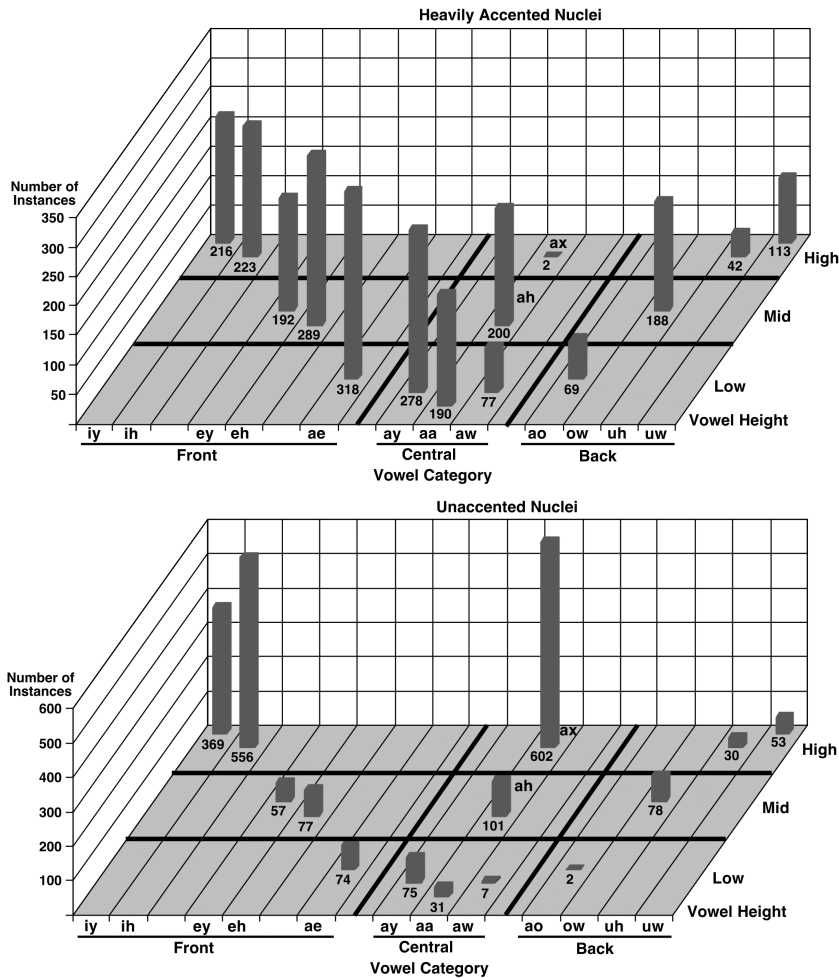


Figure 9. The impact of stress accent (“Heavily Accented” versus “Unaccented”) on the number of instances of each vocalic segment type in the corpus. The vowels are partitioned into their articulatory configuration in terms of horizontal tongue position (“Front,” “Central” and “Back”) as well as tongue height (“High,” “Mid” and “Low”). Note the concentration of vocalic instances among the “Front” vowels associated with “Heavy” accent and the association of high-front and high-central vowels with unaccented syllables. The data shown pertain solely to canonical forms realized as such in the corpus. The skew in the distributions would be even greater if non-canonical forms were included (cf. Figure 9 for additional data pertaining to this issue). From Greenberg et al., 2002).

fully accented syllables (particularly among the front and low/mid-central vowels). Certain vowels, such as [ao], [oy], [aa] [ae] and [aw], rarely occur in unaccented syllables and are typically associated with “content” words (such as nouns and their adjectival complements), particularly those that are relatively uncommon (and hence highly “informative” from a mathematical perspective).

The phonetic realization of vocalic forms is shaped to a certain degree by the (negative) entropy (or “information”) associated with the syllables and words in which they are contained. The stress accent pattern can thus be thought of as the surface manifestation of local variation in information associated with the sequence of words and syllables within an utterance.

The intimate relationship between stress accent and vocalic identity in spontaneous discourse suggests that the two may also not be readily distinguishable at some (relatively high) level of linguistic abstraction. Accent may be as integral a component of vocalic identity as tongue height and horizontal tongue position (if not more so). Diphthongs are rarely found in unaccented syllables, regardless of the underlying canonical form, nor are low or back vowels frequently encountered in such contexts. In this sense the absence of accent is accompanied by a constriction of the articulatory space to mostly high-front and high-central vowels. Such a constriction is probably associated with the reduction in duration associated with unaccented syllables and is likely to reflect the “undershoot” phenomenon described by Lindblom (1963) and others (e.g., Öhman, 1966).

The phonetic forms associated with consonantal segments in both onset and coda constituents of the syllable exhibit a comparable (though quite different) dependence on stress accent (Greenberg et al., 2002). The durational properties of onset (but not coda) consonants are highly sensitive to stress accent – the onsets of heavily accented syllables tend to be 50-60% longer than their unaccented counterparts. And coda constituents are far more likely to be “deleted” (or at least phonetically unrealized) in unaccented syllables than in syllables with some degree of stress accent (relative to their “canonical” pronunciation), particularly for alveolar and liquid segments. Such patterns of pronunciation variation provide yet additional evidence that prosodic factors are extremely important in understanding the phonetic properties of spoken language.

12. THE UTILITY OF PHONETIC INSIGHT

Knowledge of the relation between pronunciation and stress accent may be of utility for automatic speech recognition, particularly under conditions of acoustic interference where the low-frequency portion of the spectrum is degraded. For such knowledge to be of utility in technology applications automatic methods are required to computationally embed the kernel of insight within the confines of a functioning system.

Such an automatic stress accent labeling (AutoSAL) system has been developed for the Switchboard corpus. Multilayer perceptron (MLP) neural networks were trained on 45 minutes of manually labeled material and then applied to an additional four hours of data from the same corpus. The training material contains five distinct levels of stress accent (from entirely unaccented at one end of the spectrum to heavily accented at the other). The degree of

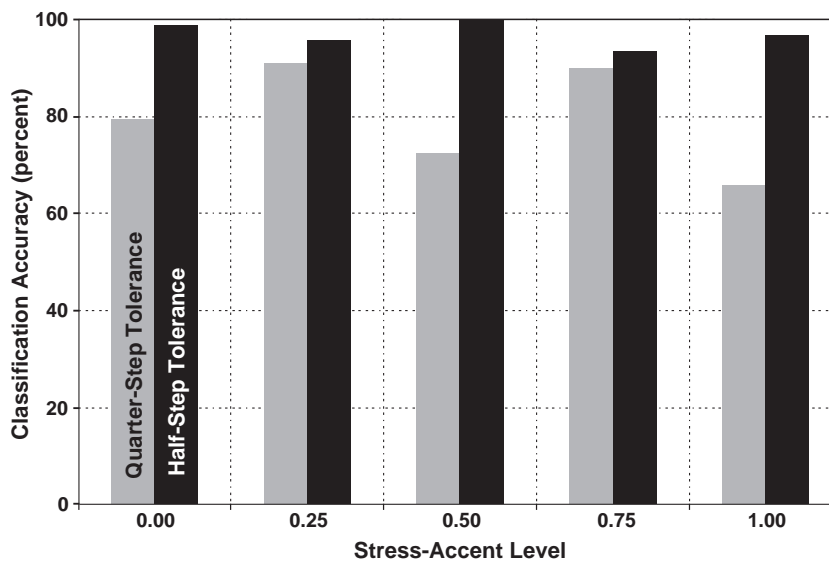


Figure 10. Classification accuracy of the automatic (MLP-based) stress-accent labeling (AutoSAL) system for the Switchboard corpus using two degrees of accent-level tolerance – quarter-step and half-step. The reference accent level is derived from the (average of the) material manually labeled by two transcribers. A syllable is scored as correctly labeled if the ASAL system output is within the designated tolerance limit. Such a metric is required to compensate for the inherent “fuzziness” of stress accent in spontaneous material, particularly for syllables with some measure of accent. For accented syllables there appears to be a gradation in stress; in contrast, unaccented syllables behave as a relatively homogeneous class. From Greenberg et al. (2001).

machine-human concordance depends on the granularity of the accent labeling. For a very strict metric of concordance (an exact match between human and machine labels) there is precise agreement for 67.5% of the syllables. When the concordance metric is relaxed to a single level of accent disparity the concordance rises to 78%. And when the concordance criterion is further relaxed to 2 accent levels of disparity the agreement between human and machine is nearly 98%. Because the human transcribers were using a *three*-level system to mark accent (i.e., fully accented – 1, completely unaccented – 0, an accent in between the extremes – 0.5), the most realistic concordance metric to assess the reliability of AutoSAL provides for two levels of accent disparity. In this sense, the machine labels are as reliable (and as consistent) as those generated by highly trained human transcribers (Figure 10).

It is of interest to ascertain the specific acoustic, phonetic and linguistic features required to simulate stress-accent assignment performed by the human transcribers in order to understand the nature of the cues potentially used by human listeners when decoding spoken language. Forty-five distinct

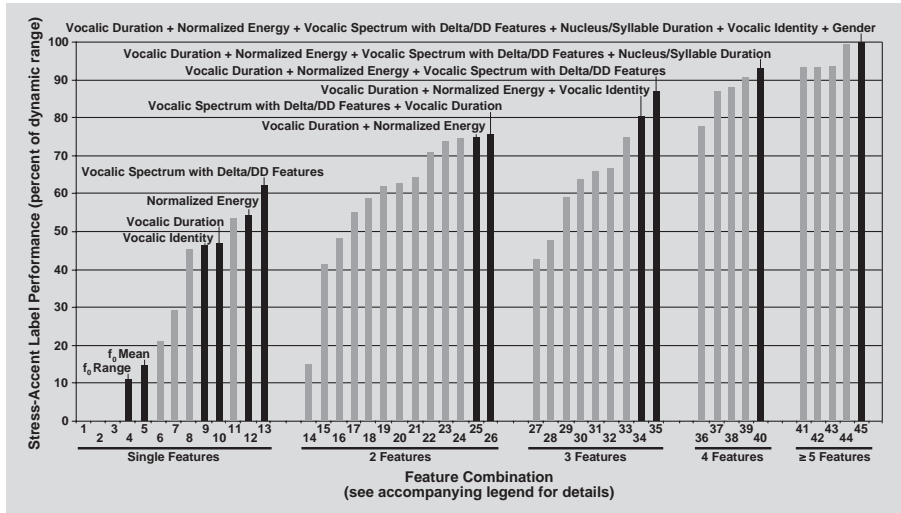


Figure 11. Features used in developing the automatic stress-accent labeling (AutoSAL) system. The final version is based on the features associated with set #45, hereafter defined as the baseline (100 percent performance), achieving performance equivalent to that of a human transcriber. The most poorly performing feature sets are those whose labeling accuracy is close to chance (40%; hereafter 0% of the dynamic range), equivalent to the prior probability of the most common stress-accent label (level-0). The magnitude associated with each feature set is the label accuracy transformed into dynamic-range-normalized units. The best performing feature combination (#45) achieves an accuracy of 67.5% with respect to *five* distinct levels of stress accent, comparable to the *overall* concordance between the two human transcribers. These results are based on an analysis using a tolerance step of 0 (i.e., an *exact* match between human and machine accent labels was required for a “hit” to be scored) and a five-accent-level system. The concordance between machine and human labels is 78% for the five-level system, and is 97.5% for a three-level version of the same system. The feature set is detailed in Table I. Revised version of a figure from Greenberg et al. (2001), in which additional details about the AutoSAL system are described.

feature combinations were used as input to the AutoSAL system in an effort to determine the features mostly closely associated with human-like, stress-accent labeling performance (Figure 11 and Table I). These feature sets were derived from a variety of acoustic, phonetic and linguistic parameters thought to be of relevance to the perception of stress accent (e.g., Fry, 1955; Lehiste, 1970; Lehiste, 1996; Silipo and Greenberg, 1999) – duration and amplitude of the syllabic nucleus, the fundamental frequency contour across syllables, as well as parameters believed to be germane to the task through statistical analysis of the Switchboard corpus (Hitchcock and Greenberg, 2001), such as the height and forward position of the tongue during vocalic articulation, vocalic identity and the dynamic properties of the nucleus (i.e., whether the segment is a diphthong or monophthong).

The traditional perspective on stress accent ascribes a prominent role to pitch (i.e., fundamental frequency) variation across syllables in a phrase (e.g., Fry, 1955; Fudge, 1984; Gimson, 1980); yet the AutoSAL system does not

Feature Legend

1. Vocalic place (front-central-back) [Voc-Place]
2. Nucleus/Syllable Duration Ratio [N_S-Dur-Ratio]
3. Speaker gender [Gender]
4. Minimum-maximum (dynamic range) of vocalic f_0 [f_0 -Range]
5. Mean vocalic f_0 [f_0 -Mean]
6. Static/Dynamic Property of Nucleus (Diphthong/Monophthong) [Voc-Dyn]
7. Vocalic height (high-mid-low) [Voc-Height]
8. Average vocalic-segment spectrum [Voc-Spec]
9. Vocalic identity [Voc-ID]
10. Vocalic-segment duration [Voc-Dur]
11. Voc-Spec + delta features [Voc-Spec_D]
12. Normalized energy (of the nucleus relative to the entire utterance) [Z-Energy]
13. Voc-Spec + delta and double-delta features [Voc-Spec_D_DD]
14. f_0 -Mean + f_0 -Range
15. Voc-Height + Voc-Place
16. Voc-ID + f_0 -Range
17. Voc-Dur + f_0 -Range
18. Z-Energy + f_0 -Range
19. Voc-Dur + Voc-ID
20. Voc-Dur + N_S-Dur-Ratio
21. Voc-Spec_D_DD + f_0 -Range
22. Voc-ID + Z-Energy
23. Voc-ID + Voc-Spec_D_DD
24. Voc-Spec_D_DD + Z-Energy
25. Voc-Dur + Z-Energy
26. Voc-Dur + Voc-Spec_D_DD
27. Voc-Height + Voc-Place + Voc-Dyn
28. Voc-Height + Voc-Place + Voc-ID
29. Voc-Height + Voc-Place + Voc-Dur
30. Voc-Height + Voc-Place + Z-Energy
31. Voc-Height + Voc-Place + Voc-Spec_D_DD
32. Voc-Dur + N_S-Dur-Ratio + f_0 -Range
33. Voc-Dur + Z-Energy + f_0 -Range
34. Voc-Dur + Voc-ID + Z-Energy
35. Voc-Dur + Z-Energy + Voc-Spec_D_DD
36. Voc-Dur + Z-Energy + Voc-Height + Voc-Place
37. Voc-Dur + Z-Energy + Voc-Spec_D_DD + f_0 -Range
38. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Gender
39. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Voc-ID
40. Voc-Dur + Z-Energy + Voc-Spec_D_DD + N_S-Dur-Ratio
41. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Voc-ID + Gender
42. Voc-Dur + Z-Energy + Voc-ID + N_S-Dur-Ratio + f_0 -Range
43. Voc-Dur + Z-Energy + Voc-ID + N_S-Dur-Ratio + Gender
44. Voc-Dur + Z-Energy + Voc-Sp_D_DD + Voc-ID + N/S-Dur + Gen + f_0 -Mean + f_0 -Range
45. Voc-Dur + Z-Energy + Voc-Spec_D_DD + Voc-ID + N_S-Dur-Ratio + Gender

Table I. Features used in developing the automatic stress-accent labeling (AutoSAL) system. Delta features refer to the *first* temporal derivative of the spectrum, while double-delta features are associated with the *second* temporal derivative of the same representation. Vocalic energy is normalized in terms of standard-deviation (Z) units relative to the mean. Features listed pertain to those associated with labeling performance shown in Figure 11.

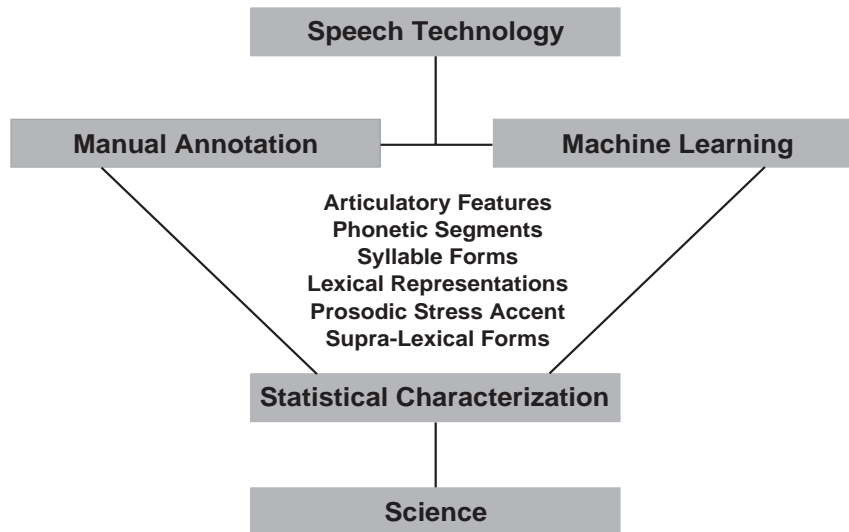


Figure 12. The “eternal pentangle” illustrates the essential tension between science and technology. Although the two poles are often considered exclusive domains, they are in fact complements of each other, providing synergistic relations that further the goals of both. Great technology generally depends on first-rate science and conversely, cutting-edge science often requires superb engineering. Moreover, insights garnered from activity in one pole often help to elucidate problems in the other.

require such f_0 -based features to achieve performance on par with an experienced human transcriber. Of the 45 feature-combination sets tested (Table I), parameters associated with vocalic identity and the attendant spectrum (in terms of the spectral contour over the duration of the segment) are consistently among the most effective cues, along with the duration and normalized energy associated with the syllabic nucleus. Thus, statistical analysis of a spoken-language corpus has proven to be a far better guide for developing classification algorithms of stress accent than perceptual studies using (rather) artificial stimuli. In this fashion speech technology can provide the sort of insight into the nature of spoken language that complements and extends knowledge gained from more traditional sources of scientific experimentation (cf. Figure 12).

13. THE ONCE AND FUTURE KINGDOM OF SPOKEN LANGUAGE RESEARCH

Many aspects of spoken language can be likened to the unicorns of yore – mythical in nature, with their sanctity especially esteemed. These mythical (and languid) creatures are often “sighted,” yet ever fail to materialize, the ephemeral pot of gold at the edge of the linguistic rainbow. Thus, spoken language, as seen through the “eyes” of phonetics and technology, may

appear as a chimera, its form and substance in perpetual mutation, and its reification dependent on circumstance rather than on principle.

Scientific insight often stems from necessity, and in such circumstance technological imperatives are likely to serve as an effective catalyst in transforming phonetics (and the rest of linguistics) into a mature field of scientific endeavor. An essential tension exists between science and technology with respect to spoken language. Over the coming decades this tension is likely to dissolve into a collaborative relationship melding linguistic knowledge with machine-learning and statistical methods as a means of developing mature science and technology pertaining to human-machine communication. In the process many mysteries surrounding the form and substance of spoken language are likely to be resolved through the concerted efforts of scientists and engineers focused on the creation of “flawless” speech technology.

14. ACKNOWLEDGEMENTS

The author wishes to thank Hannah Carvey, Shuangyu Chang, Jeff Good, Leah Hitchcock and Rosaria Silipo for important contributions to the research described. The research was funded by the U.S. Department of Defense and the National Science Foundation.

15. REFERENCES

- Beckman, M., (1986) *Stress and Non-Stress Accent*. Dordrecht: Fortis.
- Clark, J. and Yallup, C. (1990) *Introduction to Phonology and Phonetics*. Oxford: Blackwell.
- Cole, R., Fanty, M., Noel, M. and Lander, T. (1994) “Telephone speech corpus development at CSLU,” *Proceeding of the Third International Conference on Spoken Language Processing*.
- Darwin, C. (1839) *Voyage of the Beagle*. New York: Collier [reprinted, 1909]
- Darwin, C. (1859) *On the Origin of Species*. Cambridge, MA: Harvard University Press (facsimile of the 1st edition, 1964).
- Fry, D. (1955) “Experiments in the perception of stress,” *Language and Speech*, 1: 126-152.
- Fudge, E. *English Word-Stress*. London: Allen and Unwin, 1984.
- Gimson, A. (1980) *An Introduction to the Pronunciation of English (3rd ed.)*. London: Edward Arnold.

- Godfrey, J.J., Holliman, E.C., and McDaniel, J. (1992) "SWITCHBOARD: Telephone speech corpus for research and development," *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, pp. 517-520.
- Greenberg, S. (1997) "The Switchboard Transcription Project," in *Research Report #24, 1996 Large Vocabulary Continuous Speech Recognition Summer Research Workshop Technical Report Series*. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, MD.
- Greenberg, S. (1998) "Recognition in a new key – Towards a science of spoken language," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1041-1045.
- Greenberg, S. (1999) "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, 29: 159-176.
- Greenberg, S. (2001) "Whither speech technology? – A twenty-first century perspective," *Proceedings of the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, pp. 3-6.
- Greenberg, S., Carvey, H. and Hitchcock, L. (2002) "The relation between stress accent and pronunciation variation in spontaneous American English discourse," *Proceedings of the International Conference on Speech Prosody-2002*.
- Greenberg, S and Chang, S. (2000) "Linguistic dissection of switchboard-corpus automatic speech recognition systems," *Proceedings of the ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, pp. 195-202.
- Hitchcock, L., *Acoustic Properties of Vocalic Nuclei Associated with Prosodic Stress Accent in Spontaneous American English Discourse*, Undergraduate Honors Thesis, Department of Linguistics, University of California, Berkeley, 2001 (available from <http://www.icsi.berkeley.edu/steveng/prosody>).
- Hitchcock, L. and Greenberg, S. (2001) "Vowel height is intimately associated with stress-accent in spontaneous American English discourse," *7th European Conference on Speech Communication and Technology (Eurospeech-2001)*, pp.79-82.
- Jakobson, R., Fant, G. and Halle, M. (1961) *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press.

- Koumpis, and Renals, S. (2001) "The role of prosody in a voicemail summarization system," *Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, pp. 93-98.
- Kuijk, D. van and Boves, L. (1999) "Acoustic characteristics of lexical prominence in continuous telephone speech," *Speech Communication*, 27: 95-111.
- Ladefoged, P. (1993) *A Course in Phonetics* (3rd ed.). New York: Harcourt.
- Lehiste, I., (1970) *Suprasegmentals*. Cambridge, MA: MIT Press.
- Lehiste, I. (1996) "Suprasegmental features of speech," in *Principles of Experimental Phonetics*, N. Lass (ed.). St. Louis: Mosby, pp. 226-244.
- Lindblom, B. (1963) "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, 35: 1773-1781.
- Lindblom, B. (1990) "Explaining phonetic variation: A sketch of the H and H theory," in *Speech Production and Speech Modelling*, W.J. Hardcastle and A. Marchal (eds.), Dordrecht: Kluwer, pp. 403-439.
- Lovejoy, A.O. (1939) *The Great Chain of Being*. Cambridge, MA: Harvard University Press.
- Öhman, S.E.G. (1965) "Coarticulation in VCV-utterances: Spectrographic measurements," *Journal of the Acoustical Society of America*, 39: 151-168.
- Popper, K. (1959) *The Logic of Scientific Discovery*. London: Hutchinson. [originally published in German, 1934]
- Ries, A. and Ries, L. (1998) *The 22 Immutable Laws of Branding*. New York: Harper.
- Ritchie, D. (ed.) (1987) *Rashomon*. New Brunswick, NJ: Rutgers University Press.
- Silipo, R. and Greenberg, S., (1999) "Automatic transcription of prosodic prominence for spontaneous English discourse," *Proceedings of the XIVth International Congress of Phonetic Sciences*, pp. 2351-2354.
- Silipo, R. and Greenberg, S. (2000) "Prosodic stress revisited: Reassessing the role of fundamental frequency," *Proceedings of the NIST Speech Transcription Workshop*.
- Weiner, J. (1994) *The Beak of the Finch*. New York: Knopf.