From Here to Utility – Melding Phonetic Insight with Speech Technology

Steven Greenberg
1nternational Computer Science Institute
1947 Center Street, Berkeley, CA 94704 USA

Abstract

An historic 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 solved through the concerted efforts of scientists and engineers focused on the creation of "flawless" speech technology.

1. Introduction

It is the twelfth-century in Japan, and a nobleman has died a violent death. A magistrate is charged with establishing the identity of the killer and delineating the sequence of events leading up to the murder. During the formal hearing several witnesses are called to testify – the victim's wife, the accused (a notorious bandit), a woodsman and 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, indeed, dead. How he died and by whose hand are very much in dispute.

The story of *Rashomon* [28] 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 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 detail to comprehend a situation as complex (and as tragic) as the murder of a man.

As in film, perhaps in science. In *Rashomon* the testimony of each witness acquires new significance in light of the other accounts. Can an intellectual domain as complex as *spoken language* be fully understood through the testimony of a single perspective? Or must orthogonal varieties of evidence be sought with which to reconstruct the "truth"?

Knowledge gained in the pursuit of "pure" research is often viewed as the ultimate form of scientific endeavor, one 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 principle deity in the temple of knowledge.

As in myth, perhaps 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 purely the vantage point of a laboratory? Or does the spirit of *Rashomon* compel us to seek testimony from other sources in the pursuit of objective knowledge?

2. On the Path to Enlightenment

The path to enlightenment is often curvilinear. AT&T built a radio telescope in the mid-1960's to assist in the company's efforts to develop satellite communications. As part of the development effort two physicists from Bell Labs (Arno Penzias and Robert Wilson) calibrated the instrument using liquid helium and found to their distress that the peak in the telescope's radio spectrum was not 0˚ K (i.e., absolute zero or –459˚ F) as it should have been, but rather 3.5˚ K. Over the course of a year all potential sources of contamination were considered and ruled out, but still the 3.5˚ K peak remained. Penzias, in desperation, called a colleague at MIT to ask for advice who, in turn, suggested talking with Bob Dicke, a Princeton astronomer. Dicke had been building a radio telescope for basic research, but hadn't progressed very far. His interest lay not in satellite communications but in cosmology. Dicke's calculations predicted that the "big bang" associated with the universe's beginning, some 12 billion years ago, should have left a "signature" in the cosmic background radiation slightly above 0˚ K. When Penzias called and asked whether he had any idea where this pesky 3.5˚ K radiation was coming from, Dicke replied "I believe you have found the origin of the universe" (Penzias and Wilson were awarded the 1978 Nobel Prize in Physics for their "discovery") [3].

Shortly after the first World War, a young engineer working for the Hungarian telephone company, was asked to design a receiver superior to those used in phones of the past. After pondering the problem for awhile the engineer concluded that the receiver's design should be informed by knowledge of the human listener's reception capabilities, whereupon he devoted the remainder of his life to the study of the inner ear [2] (Georg von Békésy was awarded the 1961 Nobel Prize in Physiology and Medicine for his efforts).

Around the same time, four thousand miles to the East, a young physicist, newly graduated from the University of Chicago, joins the research group at Western Electric (soon to become Bell Labs when acquired by AT&T). He is asked to ascertain the most narrow passband than can be used for effective transmission of speech over the telephone (his answer: 300-3400 Hz). Over the coming decades Harvey Fletcher would develop novel methods for computing speech intelligibility under a wide range of environmental conditions, inventing the Articulation Index along the way [8]. Like von Békésy, Fletcher concludes that improved telephony-based technology requires deep knowledge of auditory function and therefore launches an intensive study of hearing that ultimately results in the development of the "critical band" concept [8], as well as the earliest stereophonic recordings. In contrast to his Hungarian counterpart, Fletcher does not abandon the engineering trade, but continues his research within the confines of the world's (then) leading technology company. For his efforts Fletcher does *not* receive the Nobel Prize. Instead, his doctoral supervisor (Robert Millikan) receives the call to Stockholm in recognition of the work that Fletcher performed for his doctoral thesis on measuring the electron charge using a drop of oil [9].

Submitted to the 7th European Conference on Speech Communication and Technology (Eurospeech-2001)

3. The Structure of Scientific Evolution

The course of a discipline's intellectual evolution is often tortuous and rarely linear. Where does speech research lie with respect to its "great chain of being"? Is our field still engaged in determining the number of phonemes *on* a word? Or has the collective discussion progressed to a higher plane? What will the speech scientists of the *twenty-second* century write about the science of the twenty-first?

Scientific maturity is often marked by its close relation to technology. The great monuments of any age (be they pyramids, cathedrals or theme parks) are often based upon the most advanced science and technology of the time. And in turn, such monuments usually spur further progress in the domains upon whose foundations they are built.

The synergy between science and technology is simple to discern, for it is difficult to build a successful product on anything other than a secure scientific foundation. And technology, in turn, provides a rigorous proving ground for the empirical and theoretical precepts of a discipline. In this sense, technology may 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 will work purely on the basis of theory (as audiences in Lincoln Center's original Philharmonic Hall discovered, much to their aesthetic chagrin). Thus, theory needs to be tempered with data representative of the environment in which the technology is deployed. Under such circumstances a field can mature quite quickly; and so it may ultimately come to pass with respect to speech technology.

4. The Galapagos of Spoken Language

The voyage of the *Beagle* provided an effective forcing function for Darwin's thoughts on the origin of species [7], particularly his sojourn in the Galapagos Islands, west of Ecuador. Among the fauna of those islands were many varieties of finch, who by dint of variation in color, size and shape (particularly of the beak) were to provide crucial clues as to the mechanism of natural selection [7].

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 evidence required to solve the mystery of spoken language?

5. Unobtrusive Measures

Every discipline has a favorite means of collecting data. Astronomers gaze into the heavens with their telescopes, high-energy physicists smash atoms, ethologists play peeping toms, and linguists either introspect or elicit data from "informants."

Marketing researchers discovered, long ago, the pitfalls of elicited data. A shopper, upon entering the market, is asked to enumerate produce and products to be purchased shortly. At checkout the marketer compares the shopper's original list with what has actually been bought, only to discover that intention and deed bear scant relation to each other; for there is nary a product in the shopper's cart that was mentioned during the interview a few minutes prior [27]. Because most spoken-language data are derived from either introspection or elicitation the empirical foundations of our field are built largely on the scientific equivalent of quicksand. From a distance the foundation appears secure, only to collapse in a nebulous undertow upon closer inspection.

6. Speech as a Linchpin of Future Technology

What is an ambitious field to do? Can a discipline reinvent itself with sufficient 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 economy by virtue of its ability to facilitate and automate communication between humans and machines (cf. [17]). Under such circumstances a unique opportunity 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, and speech communication is unlikely to be granted an exemption in this regard. Thus, the science of spoken language will probably evolve quite rapidly over the coming decades as the demand for speech technology accelerates with the emergence of the "communication age."

Sophisticated technology depends on "getting the details right" to a degree that far exceeds what passes for knowledge and insight within the domain of science (this 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 only the faintest resemblance to speech spoken in the "real" world. This is one of the reasons why current speech technology (whether it be in the form of automatic speech recognition (ASR) or text-to-speech synthesis) relies so heavily on training materials representative of the task domain. 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 deep insight into the nature of spoken language and not infrequently violates precepts of the hypothetico-deductive method (cf. [14][26]).

Speech technology can proudly point to its *apparent* success with automatic speech recognition and concantenative synthesis in defense of its machine-learning approach. And imperfect science is indeed capable of providing an effective foundation for technology – as long as the demands of the market place are not exceedingly stringent. However, as commercial expectations of the technology increase, immature science is unlikely to suffice as the empirical and theoretical foundation of future-generation applications [17].

7. 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; 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 [20]. 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, capable of providing a plane of mediation between linguistics and technology?

8. 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 [25], 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.

Two corpora of spoken language are particularly germane to the present discussion. "Switchboard" [13] 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 at the International Computer Science Institute [15] and is electronically accessible over the web (http://www.icsi.berkeley.edu/real/stp).

A one-hour subset of Switchboard has also been labeled with respect to stress-accent by two individuals not involved in the phonetic annotation. These individuals also labeled two and a half hours of stress-accent material from a separate (phonetically annotated) corpus, "OGI Stories" [6], 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).

9. The Acoustic Basis of Stress Accent

Stress 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, and thereby provides important information concerning the focus of a speaker's attention. Stress-related information is derived from a complex constellation of acoustic cues associated with the duration, amplitude, and fundamental frequency (f_o) of syllabic sequences within an utterance [1][5]. Traditionally, f_o (and its perceptual correlate, pitch) has been thought to serve as the primary cue for stress in English [10][11][12][24]:

> "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 (see, for example, Fry 1958, Grimson 1980, pp. 222-226, Lehiste 1976, Fudge, 1984, ch. 1) ...." ([5] p. 280)

However, it is unclear whether such statements truly apply to spontaneous speech (as opposed to scripted and non-meaningful material). For this reason 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 [30][31]. During the course of the study it was found that duration and amplitude play a far more important role than f_o in accounting for the stress patterns observed in the OGI Stories 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 of the human transcribers. Each method weighted duration and amplitude far more heavily than f_o in order to provide an accurate simulation of the stress-accent annotation [31]. These findings are similar to those of a recent study examining a related issue from the perspective of Dutch [22]. Together, these studies suggest that pitch variation plays a much smaller role in the stress-accent pattern of spontaneous speech than many believe, and thus caution is warranted in extending the conclusions of laboratory studies to the real world, particularly if technology is the ultimate arbiter of the "truth."

Stress-accent may be of importance for future-generation speech recognition systems. Not only does it provide a means of determining key words in an utterance, but also appears to be highly correlated with certain types of word error in current-generation ASR systems [18].

10. Stress-Accent, Duration and Vowel Height

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 prosodic prominence should be relatively uniform across vowels. However, a rather different pattern emerges from analysis of the Switchboard corpus. 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 vowels, such as [eh], [ey], [ah] and [ow] exhibit a stress-accent pattern intermediate between their low and high vocalic counterparts [19].

The relation between vocalic identity and stress-accent appears to go far deeper than a mere statistical association between parameters. Vowel duration is highly correlated with stress-accent. Stressed nuclei are often 50% to 100% longer in duration than their unstressed counterparts. In consequence, duration and vowel height are highly correlated. Duration may thus serve as a secondary (and under certain circumstances, even as a primary) cue to vowel height. In some sense stress-accent and vowel height are not easily distinguishable. Vocalic distinctiveness is, in principle, based on the pattern associated with formants one, two and three [23]. Yet duration (bound with stress-accent) appears to play an important role as well, reflected, perhaps, in the pattern of vocalic reduction observed in spontaneous speech.

Such knowledge may be of utility for automatic speech recognition, particularly under conditions of background interference where the low-frequency portion of the spectrum is degraded.

11. All Articulatory Features are Created Equal (but some features are created more equal than others)

Articulatory-acoustic features (such as voicing, lip-rounding, place and manner of articulation) are considered by many to function as essential (and independent, cf. [8][29]) building blocks of the phonetic constituents of language. In principle each segment is decomposable into a unique cluster of articulatory-acoustic features (AFs), potentially yielding a more parsimonious description of "underlying" phonological patterns than conventional segmental accounts permit [5]. Moreover, AFs may provide a relatively stable lexical representation under conditions of acoustic interference, and therefore are potentially relevant to the development of robust automatic speech recognition systems (cf. [21]).

A four-hour subset of the Switchboard corpus was analyzed with respect to AF patterns observed relative to the *canonical* (i.e., dictionary) form associated with each lexical item. In an earlier study it had been shown that phonetic phenomena in spontaneous speech are highly structured at the syllabic level [15], and for this reason AF patterns were analyzed relative to their segmental position within the syllable (i.e., onset, nucleus and coda).

There is relatively little deviation from the canonical in onset position. The few deviants observed generally involve *manner* (e.g., stops vs. fricatives) rather than place (e.g., bilabial, alveolar and velar) or voicing features. In those rare instances where place and voicing do deviate from the canonical, such forms are usually accompanied by deviation in manner as well [16]. A similar pattern is observed in syllable coda position, the primary difference pertaining to the frequency of deviation observed (much higher for codas than for onsets). Overall, place features appear remarkably stable in both onset and coda position, seldom deviating from the canonical (and when they do, it is usually in tandem with manner deviation). Although voicing deviates more frequently from the canonical than place, its pattern of expression is linked to manner in a manner comparable to place.

The pattern of AF deviation from canonical is quite different for syllabic nuclei. In such instances manner features are quite stable (unsurprising, given the vocalic nature of most nuclei), while place varies from the canonical with frequency. Moreover, features associated with manner, voicing and rounding rarely deviate from canonical form, *except* in the company of place deviation [16].

Voicing and rounding behave as secondary features in all syllabic positions, yoked to either manner or place in terms of their phonetic realization [16]. For this reason, their phonological status is likely to be of a different order than that accorded to place and manner.

Articulatory-acoustic feature classification may figure importantly in the development of future-generation ASR systems. The current methods for AF classification are less than ideal, in part because the conventional wisdom pertaining to partitioning of the phonetic-segment space in terms of articulatory-based features (e.g., [23]) is not quite right. Re-partitioning the phonetic space can improve AF classification performance under certain conditions [4][32].

12. Coda (and Reification)

Spoken language, as seen through the “eyes” of phonetics and technology, appears 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 into a mature field of scientific endeavor.

13. References

[1] Beckman, M., *Stress and Non-Stress Accent.* Dordrecht: Fortis, 1986.

[2] Békésy, G. von, *Experiments in Hearing.* New York: McGraw-Hill, 1960.

[3] Bernstein, J. *Three Degrees Above Zero: Bell Labs in the Information Age*. New York: Scribners, 1984.

[4] Chang, S., Greenberg, S. and Wester, M. “An elitist approach to articulatory-acoustic feature classification,” *Proc. Eurospeech*, 2001.

[5] Clark, J. and Yallup, C., *Introduction to Phonology and Phonetics*. Oxford: Blackwell, 1990.

[6] Cole, R., Fanty, M., Noel, M. and Lander, T., “Telephone speech corpus development at CSLU,” *Proc. Int. Conf. Spoken Lang. Proc.*, 1994.

[7] Darwin, C. *Voyage of the Beagle.* New York: Collier, 1909 [first edition, 1839].

[8] Fletcher, H. *Speech and Hearing in Communication.* New York: Van Nostrand, 1953.

[9] Fletcher, H. “My work with Millikan on the oil-drop experiment,” *Physics Today* (June), pp. 43-47, 1982.

[10] Fry, D., “Experiments in the perception of stress,” *Lang. Speech*, 1: 126-152.

[11] Fudge, E., *English Word-Stress.* London: Allen and Unwin, 1984.

[12] Gimson, A., *An Introduction to the Pronunciation of English (3rd ed.).* London: Edward Arnold, 1980.

[13] Godfrey, J.J., Holliman, E.C., and McDaniel, J., “SWITCHBOARD: Telephone speech corpus for research and development,” *Proc. IEEE Int. Conf. Acoust. Speech Sig. Proc*., pp. 517-520, 1992.

[14] Greenberg, S. “Recognition in a new key – Towards a science of spoken language,” *Proc. IEEE ICASSP,* pp. 1041-1045, 1998.

[15] Greenberg, S. “Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation,” *Speech Communication*, 29: 159-176, 1999.

[16] Greenberg, S. “Understanding spoken language using statistical and computational methods,” presentation at *Patterns of Speech Sounds in Unscripted Communication – Production, Perception, Phonology,* Akademie Sankelmark, October 11, 2000.

[17] Greenberg, S. “Whither speech technology? – A twenty-first century perspective,” *Proc. Eurospeech*, 2001.

[18] Greenberg, S and Chang, S. “Linguistic dissection of switchboard-corpus automatic speech recognition systems,” *Proc. ISCA Workshop on Automatic Speech Recognition: Challenges for the New Millennium*, pp. 195-202, 2000.

[19] Hitchcock, L. and Greenberg, S. “Vowel height is intimately associated with stress-accent in spontaneous American English discourse,” submitted to *Eurospeech-2001.*

[20] Jakobson, R., Fant, G. and Halle, M. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press, 1961.

[21] Kirchhoff, K. *Robust Speech Recognition Using Articulatory Information,* Ph.D. Thesis, University of Bielefeld, 1999.

[22] Kuijk, D. van and Boves, L., “Acoustic characteristics of lexical prominence in continuous telephone speech,” *Speech Communication*, 27: 95-111, 1999.

[23] Ladefoged, P. *A Course in Phonetics* (3rd ed.). New York: Harcourt, 1993.

[24] Lehiste, I., *Suprasegmentals*. Cambridge, MA: MIT Press, 1970.

[25] Lovejoy, A.O. *The Great Chain of Being.* Cambridge, MA: Harvard University Press, 1939.

[26] Popper, K. *The Logic of Scientific Discovery.* London: Hutchinson, 1959. [originally published in German, 1934]

[27] Ries, A. and Ries, L. *The 22 Immutable Laws of Branding*. New York: Harper, 1998.

[28] Ritchie, D. (ed.) *Rashomon*. New Brunswick, NJ: Rutgers University Press, 1987. [contains the screenplay by Akira Kurosawa and Shinobu Hashimoto, as well as the stories by Ryunosuke Akutagawa, “Rashomon” and “In the Forest,” upon which the screenplay is based]

[29] Saul, L.K., Rahim, M.G. and Allen, J.B. “A statistical model for robust integration of narrowband cues in speech,” *Computer Speech and Language*, in press.

[30] Silipo, R. and Greenberg, S., “Automatic transcription of prosodic prominence for spontaneous English discourse,” *Proc. XIVth Int. Cong. Phon. Sci.*, pp. 2351-2354, 1999.

[31] Silipo, R. and Greenberg, S., *Automatic Detection of Prosodic Stress in American English Discourse*, Technical Report TR-00-001 (29 pages), International Computer Science Institute, Berkeley, 2000.

[32] Wester, M., Greenberg, S. and Chang, S. “A Dutch treatment of an elitist approach to articulatory-acoustic feature classification,” *Proc. Eurospeech*, 2001.