WHAT MAKES SPEECH STICK?

Steven Greenberg

Silicon Speech, Santa Venetia, California, USA
Centre for Applied Hearing Research, Technical University of Denmark
Center for New Music and Audio Technology, University of California, Berkeley, USA
steveng@silicon-speech.com

ABSTRACT

Robustness and reliability is the essence of speech communication. The senses collaborate with memory and other brain mechanisms to decode spoken language in the harsh and often unpredictable environments of the real world.

How the brain makes speech "stick" is the focus of this special session, which examines how the senses and motor system coordinate during speech perception and production.

Keywords: Speech perception and production, hearing, speech-reading, multi-modal synergy

1. HALF OF LIFE IS JUST SHOWING UP

What makes speech "stick" [7]? And, why is it so effective in communicating?

Objectively, the spoken word is nothing more than physical patterns conveyed by sound and sight. At any point in the "speech chain" this information is vulnerable to attack on many fronts – reverberation, background noise, multiple talkers, and so forth. And yet the speaker's intent (the "message") is communicated almost always. How?

The papers in this special session, "Of Mouths, Ears, Eyes and Brains – The Sensory Motor Foundations of Spoken Language," explore some possibilities.

The most fundamental problem confronting any communication system is *reliability*. How can the speaker be certain that his/her message has been received and properly decoded? The problem is non-trivial to solve because of the myriad ways in which the signal can be distorted through its journey to the receiver. Somehow, the message must survive the signal's arduous transit or the effort expended will be wasted.

An illustration of this treacherous path is shown in Figure 1. The signal exiting the mouth is shown at left, while the sound entering the ear is on the right. Clearly, much has changed in transit from sender to receiver. Yet, the linguistic content is the same and often decoded correctly. How is this so?

2. NOISE REDUCTION AND STABILITY

The brain's ability to extract a consistent message from inconstant forms is often taken for granted. We usually pay scant attention to the environment through which the message passes – unless there's some impediment to understanding. Underneath the cerebral "hood" a lot is going on, much of it unconscious.

Oded Ghitza, in his paper "Using Auditory Feedback and Rhythmicity for Diphone Discrimination of Degraded Speech," examines how a specific neural circuit in the auditory brain stem helps to preserve the linguistic integrity of the *acoustic* signal in background noise. The medial, olivo-cochlear bundle (MOC) originates in the medulla, and projects directly to outer hair cells in the cochlea of the inner ear. Two properties make the MOC particularly interesting: (1) it is extremely responsive to noise, and (2) its latency is relatively long, 50 – 100 ms. Effectively, the MOC acts as a slow-acting, noise-reduction circuit that uses temporal integration. The result is a highly enhanced speech representation in noise's presence that provides some linguistic *stability* in an environmentally unpredictable world [3].

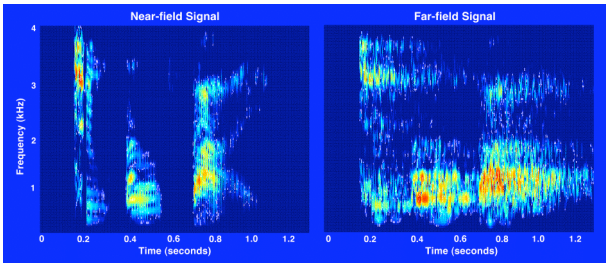
3. THE EYES HAVE IT

The real world can be a very noisy place. However, the need to communicate does not diminish in such circumstances – on the contrary.

The acoustic speech signal is extremely vulnerable to background noise (e.g., [11]). How can such a potentially fragile communication system be so robust in the real world?

Speech is "heard" not only with ears, but also through eyes [10]. Normally, the eyes play an "invisible" role in listening to speech. Only when the going gets tough does their importance become obvious. Under such conditions, speech-reading cues can restore intelligibility to what is otherwise indecipherable babble [5]. The hearing-impaired are especially sensitive to speech-reading cues [5].

Figure 1: The effect of reverberation on the acoustic speech signal. The same utterance recorded close to the talker's mouth ("Near-field" on the left) and in the middle of a classroom ("Far-field").



In her paper, "Analysis-by-Synthesis in Auditory-Visual Speech Perception: Multi-Sensory Motor Interfacing," Virginie van Wassenhove examines how the brain combines auditory and visual information to decode the speech signal.

When visual and auditory cues conflict, what happens? This is the premise of the "McGurk effect," in which an acoustic [p] is presented concurrently with a visual [k] (or vice versa) [10]. Under certain conditions, the listener perceives neither [p] nor [k], but [t]. Although [t] is a perceptual illusion (in this context), it is phonetically related to [p] and [k] – its place of articulation is intermediate between the other stop consonants. Somehow, the brain blends the bilabial property of [p] with the velar aspect of [k] and "perceives" the alveolar [t].

In certain regions of the cerebral cortex, the illusory [t] supercedes the auditory [p] and visual [k]. This perceptual "compromise" becomes the dominant response category despite its lack of sensory support. In effect, the brain's template-matching circuitry has taken over, trumping the initial sensory analysis associated with [p] and [k]. This demonstration is extremely important for understanding spoken language.

4. SPEAK, MEMORY

When the senses conflict, or are otherwise confused, a non-sensory system intercedes. This cognitive "white knight" is *memory*, the unsung hero of many linguistic (and other intellectual) endeavors. What is so special about memory? It reflects the individual's real-world experience, which means that it is *intrinsically meaningful*. Memory allows the brain to find the closest match to what it already knows, on the (logical) assumption that the signal is a variant of something previously encountered.

Why is memory so important? Because it allows the brain to classify as "similar," signals that objectively differ. In the harsh environments of the real world, rarely are two instances of anything physically identical. Thus, memory allows for *generalization*, which is the foundation of knowledge in general and language in particular.

In the real world, hardly anything appears as it truly is. This is because the brain filters its sensory input in an effort to make sense of the world. If the brain functioned as a mere transducer of physical "reality" it would be unable to provide a *meaningful* framework with which to interpret the signals received. Because brains evolved to *interpret* the world for *meaningful* action, faithful recording of the physical world is neither relevant nor productive.

The "Template-Matching Circuit" (TMC) described in Ghitza's paper is a form of memory optimized for time-frequency patterns associated with phonetic diphones (consonant-vowel dyads). Unlike most memory systems, this TMC is limited to acoustic patterns associated with short intervals (<100 ms) of speech. It is a memory nevertheless, for it compares what has been previously stored with what is encountered from moment to moment. In other words, the TMC *decodes* the acoustic speech signal based on what's been encountered in the past. In a "noisy" world, the TMC plays a crucial role; it associates the signal with an abstraction linked to higher-level representations such as syllables and words. Such abstractions are crucial for interpreting and categorizing signals that are inherently noisy and ambiguous.

5. STREAMS IN PARALLEL

It is not only memory that helps to decode and interpret the speech signal; the motor system is also important.

In her paper, van Wassenhove demonstrates that parts of the motor cortex are excited during the McGurk task. The motor activation occurs in precisely those regions associated with speech production – an eerie reflection of the "motor theory" of speech perception [9]. Does this result demonstrate that speech is perceived *primarily* through the motor system? Not necessarily.

Just as vision can influence auditory decoding, motor processing may facilitate the speech signal's interpretation. If the sensory streams are noisy or ambiguous, a motor representation may help the brain to deduce the linguistic message. How this

parallel motor representation is computed is unknown. It may be through "mirror" neurons [13] or some mechanism as yet undiscovered.

Of greater importance than motor participation is the presence of *multiple*, parallel streams of processing. It is this multiplicity of representations that probably accounts for speech's robustness and speed of comprehension.

6. AUDITORY-MOTOR COLLABORATION

In everyday communication, the auditory and motor systems collaborate closely – this is the conclusion of Joe Perkell's paper, "Sensory Goals and Control Mechanisms for Phonemic Articulations." The acuity with which the listener perceives speech correlates with that person's articulatory precision. This is a stunning result that parallels what is known about musically gifted individuals. Production and perception work hand in hand (or perhaps, tongue in ear).

Perkell explores the ramification of this insight for the deaf. An intimate, *symmetrical* connection between hearing and production would imply that articulation should be profoundly degraded among the (post-lingual) deaf. And yet, Perkell's study demonstrates that speech production deteriorates little, if any, long after deafness' onset [12]. How can this be so?

7. MODALITY-INDEPENDENT REPRESENTATIONS

Before going deaf, a speaker probably develops many different speech representations – some auditory, others visual or kinesthetic. Among the likely representations are those based on motor patterning. This is not surprising; well-practiced individuals possess a keen sense of what's required to produce a sound (linguistic or otherwise). Some of this knowledge is likely reflected in the brain as "corollary" [14] or mirror [13] discharges.

These representations are correlated with each other, thereby providing a synergistic framework for extracting common patterns across sensory and motor inputs (or outputs). Synergy means that a little bit of information from many different sources results in a huge gain for decoding the signal. For example, in extreme conditions the acoustic signal may provide 20% intelligibility, while the visual cues provide 10% [4]. A product-of-errors model (used in computing the articulation index [1]) would predict ca. 35% intelligibility, when in fact nearly double that is decoded [6].

This synergy is most likely to occur if the base representation is modality-independent. Consider the implications of this conclusion – the underlying representations of phones, syllables, words and phrases are unlikely to be based on formant patterns or other properties specific to speech acoustics. Nor are such linguistic "ür-patterns" likely to be based on articulatory features [8].

8. NEURO-DYNAMICS

What *is* common across the modalities sub-serving speech? Whatever the commonality, it is likely based on some aspect of neural activity. What might this be?

The speech signal is constantly in flux. The changes in acoustic energy across frequency and time are reflections of articulations that also vary from moment to moment. These dynamic variations reflect *information* associated with the speaker's intent. The message is spoken in the manner it is in order to be understood [8]. This means that the underlying representation must be internal to the brain; it cannot be directly observed in either the acoustics or optics associated with the speech signal.

What form might these ür-patterns assume? Sensory and nerve cells respond most intensively to change. Stabilize an image on the retina and it "disappears" after a few seconds. Auditory neurons rapidly adapt to a steady-state tone, and so on. Change is associated with information, stasis with its absence. Individual nerve cells respond to informative signals, which are those that are changing. In this sense, speech is intrinsically dynamic; it is particularly well designed to activate neurons at many levels of the brain. Therefore, it is unsurprising that brain-imaging studies, such as those described by van Wassenhove [15], show excitation across many different regions of the brain, particularly when the talker's face is clearly visible.

If the underlying representations of speech are not directly observable in the acoustics, optics or articulation, how might they be discerned?

9. HIDDEN DIMENSIONS

If the talker's message is ultimately represented in the neuro-dynamics of the brain, a lot can be learned from brain-imaging and modeling studies. And if the parameters controlling the encoding and decoding of speech are largely hidden, experimental neuroscience and computational

modeling may ultimately provide great insight into spoken language understanding and production.

Moreover, speech's durational properties may reflect neural processes in the brain. For example, the syllable's length is close to the range of theta rhythms (3–10 Hz) while gamma oscillations (ca. 25–60 Hz) may be associated with sub-phonemic phenomena. The patterning of speech production and perception may ultimately be manifest in a complex interplay of neural rhythms distributed across time and (cerebral) space [2]. Time will tell.

10. QUE SERA

Let's imagine how speech research may look a decade from now. What would we observe?

First, the discipline is likely to be highly quantitative; there's far too much detail in the speech signal and its underlying processes to rely on rules or simple generalizations alone. Moreover, statistical methods, which currently dominate automatic speech recognition technology, will become increasingly important for characterizing speech and understanding its biological bases.

Second, speech research is likely to be far more inter-disciplinary than it is now. As we learn more about spoken language, the need for a multi-faceted approach becomes increasingly clear. Spoken language is not just about hearing, seeing, articulation, memory and thinking. It involves these sub-disciplines and many, many more. In order to understand spoken language in its full context we, as a field, need to expand our horizons to encompass disciplines as diverse as physics, statistics, cybernetics, engineering, psychology, biology, economics and anthropology (among others) – a true consilience [16] of language.

Finally, most speech research will be conducted in the service of specific applications. What might these be?

11. WHAT'S THE USE

The days of "pure" research (i.e., performed for its own sake, irrespective of potential payoffs) are gone and unlikely to return. Our society's needs are so great and resources too limited to foster "blue sky" research in spoken language (or most other fields). Rather than bemoan this fate, it is far better to embrace the future by developing speech applications that require expanding our knowledge of spoken language and its associated processes. What might these be?

Three applications come readily to mind: (1) machine understanding of speech, (2) creation of artificial voices, and (3) restoration of speech comprehension among the hearing-impaired. Each is technically challenging and requires a lot more knowledge (and insight) about spoken language to be fully effective. However, history has a habit of mocking predictions of the future. The most important applications of speech-based knowledge are likely to be those not envisioned in this paper or elsewhere.

12. REFERENCES

[1] Allen, J.B. 2005. *Articulation and Intelligibility*. San Rafael, CA: Morgan and Claypool.

[2] Buszaki, G. 2006. *Rhythms of the Brain*. New York: Oxford University Press.

[3] Ghitza, O. 2007. Using auditory feedback and rhythmicity for diphone discrimination of degraded speech. *Proc. 16th ICPhS,* Saarbrücken, this volume.

[4] Grant, K., Greenberg, S., Poeppel, D., van Wassenhove, V. 2003. Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. *Seminars in Hearing* 25, 241-255.

[5] Grant, K.W., Walden, B.E., Seitz, P.F. 1998. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *J. Acoust. Soc. Am.* 103, 2677-2690.

[6] Greenberg, S., Arai, T., Grant, K. 2006. The role of temporal dynamics in understanding spoken language. In: Divenyi, P., Greenberg, S., Meyer, G. (eds). *Dynamics of Speech Production and Perception*. Amsterdam: IOS Press, 171-190.

[7] Heath, C., Heath, D. 2007. *Made to Stick: Why Some Ideas Survice and Others Die.* New York: Random House.

[8] Jakobson, R., Fant, G., Halle, M. 1952/1963. *Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates*. Cambridge, MA: MIT Press.

[9] Liberman, A.M., Cooper, F.S., Shankweiler, D.P., Studdert-Kennedy, M. 1967. Perception of the speech code. *Psych. Rev.* 74, 431-361.

[10] McGurk, H., MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264, 746-748.

[11] Miller, G.A., Nicely, P. 1955. An analysis of perceptual confusions among some English consonants. *J. Acoust. Soc. Am.* 27, 338-352.

[12] Perkell, J. 2007. Sensory goals and control mechanisms for phonemic articulations. *Proc. 16th ICPhS,* Saarbrücken, this volume.

[13] Rizzolatti G., Craighero L. 2004. The mirror-neuron system. *Ann. Rev. Neurosci.* 27, 169-192.

[14] Sperry, R.W. 1950. Neural basis of the spontaneous optokinetic response produced by visual inversion. *J. Comp. Physiol. Psychol*. 43, 482-489.

[15] van Wassenhove, V. 2007. Analysis-by-synthesis in auditory-visual speech perception: Multi-sensory motor interfacing. *Proc. 16th ICPhS,* Saarbrücken, this volume.

[16] Wilson, E.O. 1998. *Consilience: The Unity of Knowledge*. New York: Knopf.