

**RESEARCH ON SPOKEN LANGUAGE PROCESSING**  
Progress Report No. 28 (2007)  
*Indiana University*

**New Directions in Speech Research<sup>1</sup>**

**Adam Buchwald, Tessa C. Bent, Christopher M. Conway,  
Susannah V. Levi and Jeremy L. Loebach**

*Speech Research Laboratory  
Department of Psychological and Brain Sciences  
Indiana University  
Bloomington, Indiana 47405*

---

<sup>1</sup> We would like to acknowledge NIH NIDCD grant number DC00012 for supporting the production of this paper.

## New Directions in Speech Research

**Abstract.** In October 2006, many of the top scholars of speech perception research gathered in Bloomington, Indiana for a conference focused on new directions in speech research. This short paper provides a summary of the talks that were presented at this conference, which discussed the use of methodological innovations, novel theoretical frameworks, and the use of a variety of research populations in speech research.

### Introduction

This paper provides a summary of talks presented at “PisoniFest,” a conference held in Bloomington, Indiana on October 20-22, 2006 which explored new avenues and topics for research on speech perception. The talks at this conference focused on methodological innovations and concerns, the application of new – and used – theoretical frameworks to the study of speech perception, and the use of a variety of under-examined research populations in speech research. This paper provides synopses of each presented paper, organized by focus.

### Methodological Innovations and Concerns in Speech Research

**Cynthia Clopper**, Ohio State University, and **Janet Pierrehumbert**, Northwestern University, examined the effects of dialect variation on spoken word recognition and lexical access. Previous research has shown a benefit of local over non-local dialect, as well as standard over non-standard dialect in word recognition. Clopper and Pierrehumbert extended this past work by examining potential phonological sources for the interaction between dialect variation and word recognition. Two dialects – Northern Cities (non-standard) and Midland (standard) – which exhibit acoustic-phonetic overlap of different vowel categories were selected for the studies. In the first experiment, listeners performed an open-set word recognition task with monosyllabic stimuli drawn from these two dialects. Systematic lexical confusions based on dialect differences were found, where words with greater acoustic-phonetic overlap between the dialects resulted in an increased number of lexical confusions. In the second experiment, listeners performed a speeded classification task (“bad” vs. “bed”) for stimuli from these two dialects. Listeners were faster at categorizing the words from the Midland dialect than the Northern Cities dialect, confirming findings of previous studies which showed better performance on a standard dialect. The two studies reported by Clopper and Pierrehumbert provide an account of certain vowel confusions that listeners make when listening to non-standard dialects by considering acoustic-phonetic similarity of different vowel categories.

**Mitchell S. Sommers**, Washington University, presented a new measure of listening comprehension, LISN (lectures, interviews, spoken narratives), which can be used to assess listening comprehension in diverse populations including clinical and non-clinical listeners from a variety of age groups. LISN is part of a larger project investigating changes in cognitive abilities across the lifespan including measures of working memory, speech processing, and auditory processing. LISN includes three types of passages: lectures from the BBC, interviews from CSpan, and spoken narratives. Three types of questions were used to assess a listener’s comprehension: information, integration, and inference. Thus far Sommers has used the test with normal hearing young and older adults, and hearing-impaired older adults. He presented three studies which used the LISN to assess listening comprehension in these populations. The first study revealed that younger listeners were better than older listeners in the listening comprehension tests but there was a great deal of variability depending on passage and question

type. In a second study, hearing impairment contributed marginally to a decline in listening comprehension. Lastly, an audiovisual (AV) task was used to assess the benefit of lipreading and AV integration in these three populations. In the AV task all participants performed similarly, suggesting that the addition of visual information eliminated the age differences shown in audio-only conditions. In the future, Sommers hopes to use the test with people with hearing loss, Alzheimer's disease, and aphasics.

**Kevin Munhall**, Queen's University, discussed "hot problems" in audiovisual speech perception. In particular, he focused his discussion on studies examining gaze fixation and duration during audiovisual (AV) speech perception. He also investigated audiovisual perception of animated speech to address the intelligibility "gain" an observer gets from receiving visual speech information in addition to auditory information. With respect to gaze duration and fixation, Munhall reported that these vary depending on the task as well as on viewer-specific biases. Munhall also reported that dynamic realism in animation is a necessary condition to generate AV gain in animated AV speech perception.

**Jennifer Pardo**, Barnard College, presented some recent research on the nature of phonetic convergence during conversational interaction. Pardo discussed the notion of accommodation as phonetic convergence (e.g., shifting vowel targets) towards an interlocutor, and suggested that this seemingly unconscious process may actually be a choice, as participants are likely to diverge away from an insulting experimenter. Pardo's research investigated phonetic convergence by comparing the pre- and post-interaction utterances of a speaker with that of their interlocutor, and found that a naïve set of participants were more likely to judge the post-interaction utterances to be similar to the interlocutor, with some interesting gender and task effects.

**John Sidtis**, New York University School of Medicine, discussed the limitations of fMRI research with respect to understanding more about speech processing. The central point in Sidtis' claim was that complex behaviors are not reliably decomposed by contrasting tasks, and imaging research often relies on this method of "cognitive subtraction." In particular, Sidtis warned against the use of "resting states" as controls for complex cognitive functions such as speech production or perception. Further, he presented his own research indicating that more blood flow may not always be a reliable indicator that a particular brain region area performs a specific function.

**J.D. Trout**, Loyola University – Chicago, spoke about the use of animal models in understanding human cognition. He focused on the dangers of the "possibility proof" methodology, in which scholars argue that something is not unique to humans because other animals can show the same behavior (e.g., Gentner, Fenn, Margoliash, & Nusbaum, 2006). Trout's critique centered on the claim that it is not clear that animal studies tap into the same skills that humans use when they are performing linguistic tasks. Trout cited common discrepancies in findings as evidence that experiments with animal populations and with humans may be tapping into different skills.

**Luis Hernandez**, Indiana University, gave an illuminating presentation on the reliability of collecting reaction time data on modern computer systems. Systems that rely on multi-tasking can be unreliable because of differences between the onset of execution and when the physical presentation occurs. An external microcontroller that is not system specific and does not rely on computer resource management was proposed as the best, most accurate and cost effective solution for experiments requiring fine temporal resolution.

## Theoretical Frameworks for Studying Speech Perception: New and Used

**Olaf Sporns**, Indiana University, presented research on the connection between information theory and embodiment and demonstrated how such ideas provide a new understanding of how artificial and biological systems interact with the environment. Although there are different varieties of embodiment, Sporns suggested they all have in common at least three core concepts: the rejection of the idea that cognition is the processing of symbols; an emphasis on the dynamic coupling of organism to the environment; and a focus on development and self-organization. Pursuing work in robotics, Sporns' research investigates how an embodied agent interacting with the environment affects perceptual development. Using mathematically-defined information metrics such as entropy, mutual information, integration, and complexity, Sporns shows that embodied interactions affect the statistical structure of the organism's environment. That is, through its actions, an organism can "shape" its own environmental structure. As an example, Sporns demonstrated a simple robotic active vision system, in which a camera samples visual information, actively adjusting the camera to focus on particular salient parts of the scene (e.g., the color red). The coupling between the robot's action and perception systems was manipulated, with the results showing that decoupling produces less information structure. In sum, understanding how embodied systems benefit from environmental interaction and the coupling of perception/action systems can provide important new insights into the nature of speech perception, which has traditionally been dominated by a classic, information-processing view of perception.

**Geoff Bingham**, Indiana University, gave an interesting presentation on the underlying tenets of Gibson's theory of direct perception, and demonstrated how it can account for many aspects of perception. In this framework, events in the world are conceived of as spatio-temporal objects that are constrained by the environment. Under Gibson's theory, both humans and animals detect events and objects by recognizing patterns of information specified in the dynamics in the environment. Using point light displays, Bingham illustrated that we can recognize a variety of events, both animate and inanimate, based solely on the dynamics of their movement. Moreover, recognition accuracy is disrupted when the dynamics are altered such that the information specified by them becomes inconsistent with our ecological point of view. Bingham argued that the perception of biological motion, therefore, is not special, in that we can recognize the motion of a variety of objects (both animate and inanimate) even though we may not be able to produce the actions ourselves. In addition, referencing others motion to our own motion is inadequate in that we cannot ourselves witness the motions that we produce under normal circumstances. Moreover, theories that specify a motor code in the recognition of events are incomplete because they would apply only to humans (not other animals or inanimate objects), overestimate the role of the motor code in generating movement (such a code is merely correlated with the motions), and severely underestimate the role of perception. Bingham concluded that we perceive information in our environment, not our motor systems, as is argued by proponents of the Motor Theory of speech perception.

**Nelson Cowan**, University of Missouri, presented work investigating short-term memory (STM) and forgetting, where STM is informally defined as the small amount of information one can hold in mind for a short period of time. Cowan described a seminal paper by Pisoni (1973) that led Cowan to investigate several important questions about the nature of STM, especially for acoustic and phonetic input: What happens to STM codes over time? What is the role of attention in STM? Is STM memory lost through decay or interference? To investigate the first question, Cowan described research examining memory for vowels, which suggests that forgetting results in an expansion of the uncertainty of the sound. That is, the representation of a particular vowel "slides" toward the average vowel sound located in the middle of vowel space. Thus, Cowan argues that forgetting involves a shift of the memory code toward the average or prototypical representation of that class of sounds. To explore the second

question, Cowan presented work showing that attention is necessary in order to get a stable representation of a phonetic code. Finally, to address the third question, Cowan presented evidence arguing that forgetting may involve a combination of proactive interference and a “sudden” loss of the memory code after a particular amount of time (as opposed to a gradual decay). This work investigating STM and forgetting is important because it helps to clarify the role of memory and cognition in the perception and representation of speech sounds.

**Robert Port**, Indiana University, discussed his new proposal of “phonology with rich memory.” Port argued against the traditional notion of language as a symbol system in which we store mental representations corresponding to sound structure units such as phonemes, phones or segmental features. Instead, Port contended that our mental representations of language consist of the exemplars that we have encountered and encoded. Evidence for this assertion comes from a variety of studies demonstrating that we store and are able to use episodic information. Examples of this occur when participants perform better in a recognition memory task when a word is produced by the same speaker during familiarization and testing (e.g., Palmeri, Goldinger, & Pisoni, 1993). Port strongly argued that the existence of an episodic store of phonological events is incompatible with the traditional linguistic view that we store abstractions over those exemplars. He claimed that part of the reason we are drawn to the notions of these abstract units that compose words is our alphabetic training (an argument famously proposed by Ladefoged, 1980), and cited work on non-literate individuals suggesting that their “phonological awareness” (as defined by segmental awareness) is impoverished compared to literate individuals. He ended with the assertion that traditional phonology is necessary to describe socially agreed-upon linguistic conventions (“social phonology”) but that understanding the language processing system can only be done by examining episodic memory.

**Robert E. Remez**, Columbia University and Barnard College, discussed a neglected problem in speech perception research: How do human listeners determine which auditory inputs should be processed as speech? Most theories of speech perception start with a listener’s analysis of the speech signal rather than starting with an analysis of the complete auditory input. Remez argued that deciding which inputs to process as speech is not a trivial problem/task. For example, listeners must determine which parts of the auditory signal are relevant speech samples to be analyzed, which are complex non-speech signals, and which are irrelevant speech samples (e.g., background talkers). Certain technologies (e.g., ViaVoice) and models of speech perception (such as TRACE) match all auditory signals to stored speech templates thereby translating non-speech sounds to the closest speech equivalents. However, early perceptual processes must help listeners distinguish auditory signals from known languages, unknown languages, and non-speech sources. This early stage of processing is necessary for listeners to know what to process as speech. Furthermore, Remez pointed out that the perceptual system is highly flexible; even inputs that do not closely match stored speech templates can be perceived as speech if listeners are told that the signal is speech. For example, sinewave speech is often initially perceived as non-speech but listeners can also extract linguistic and extra-linguistic content from signals once they are told that the signal is speech. Remez concluded that theories of speech perception must consider how listeners determine which auditory inputs to process as speech. Listeners must simultaneously exclude complex speech-like auditory signals that are not speech and include signals that differ from naturally produced speech but can be processed as speech.

### **Examining Under-Examined Research Populations**

**Robert Shannon**, House Ear Institute, described some of his recent work that patients with Auditory Brainstem Implants. Although cochlear implants have been extraordinarily successful, not everyone with severe hearing impairment is a candidate for cochlear implantation. In some cases, severe

head trauma can sever the auditory nerve, making cochlear implantation impossible. In other cases a genetic disorder can lead to the growth of NF2 tumors on the vestibular branch of the auditory nerve, requiring surgical removal of both tumor and nerve. For such individuals, Auditory Brainstem Implants (ABIs) may be an option. Rather than inserting electrodes into the cochlea, electrode arrays are inserted into the peripheral layers of the cochlear nucleus, the first stop in the ascending auditory pathway. ABI recipients, however, show a mixed pattern of results depending on etiology. Individuals with head trauma show levels of speech recognition comparable to many cochlear implant users. NF2 patients, however, show very poor speech recognition abilities (0-20% correct). In both cases, subjects have access to all of the necessary perceptual information (what Shannon calls “bits”), but only the patients without NF2 tumors can correctly assemble such information to provide high levels of speech recognition. Shannon argued that the progressive growth of NF2 tumors destroy neurons that are vital to the organization of the cochlear nucleus. These low spontaneous rate/high threshold neurons respond to sound over a large dynamic range, and provide input to the small cell cap area of the cochlear nucleus, an area that is particularly sensitive to temporal modulations. Both NF2 and head trauma patients show normal frequency mapping along the tonotopic axis in the cochlear nucleus under ABI stimulation, suggesting that high spontaneous rate/low threshold neurons may be more involved in pitch perception and sound localization. Shannon concluded that the gradual destruction of the auditory nerve due to NF2 tumor growth disrupts the organization of the ascending auditory pathway, limiting patients’ capacity to develop high levels of speech perception abilities.

**Diana Van Lancker Sidtis**, New York University, discussed the perception of voice characteristics. She argued that the representation and processing of voice information is more similar to face processing than speech processing. For example, voice and face information both contain keys to personal identity and show hemispheric specialization. Processing prosodic information – crucial for the perception of voice identification – involves the perception of both timing and pitch. Van Lancker Sidtis presented data from two patients who showed differential loss of timing and pitch information. The first patient produced appropriate timing in both singing and speech, but was unable to produce pitch differences in either task. In contrast, the second patient produced accurate timing, and also produced accurate pitch when singing, though not in speech. Van Lancker Sidtis argued that data from these patients confirmed hemispheric specialization for processing these two aspects of prosody; timing information is processed by the left hemisphere, whereas and pitch information is processed by the right hemisphere. She also summarized data from fMRI, ERP, and lesion studies showing a right hemisphere advantage in processing familiar voices. Using these data, Van Lancker Sidtis concluded that voice information and speech information are processed in different hemispheres.

**Rosalie Uchanski** presented work on the identification and discrimination of emotions in American English speaking cochlear implant users compared to normal hearing adults and children. Cochlear implant (CI) users had more difficulty than the normal hearing adults or children. In particular, the CI-users frequently confused fearful and happy productions.

**Mario Svirsky**, New York University, discussed his work on frequency mismatch and spectral degradation in cochlear implant simulations with normal hearing adults. Adaptation to spectral shift was facilitated when the shift was gradually introduced over a series of training sessions. This work suggests that CI-users may benefit from self-selected tuning of electrode mapping.

## References

- Gentner, T. Q., Fenn, K. M., Margoliash, D., & Nusbaum, H. C. (2006). Recursive syntactic pattern learning by songbirds. *Nature*, *440*, 1204-1207.
- Ladefoged, P. (1980). What are linguistic sounds made of? *Language*, *56*, 485-502.
- Palmeri, T. J., Goldinger, S. R., & Pisoni, D. B. (1993). Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*, 309-328.
- Pisoni, D.B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, *13*, 253-260.

