Topic 4:

Opening Comments for Speech Perception and Production Session

Jim Flanagan, Rutgers University CAIP Center

We have reached an era where most communication – whether between humans and complex information systems, or with other humans – is ***machine-mediated***. An overarching objective is ***usability***. Humans appreciate the familiarity and comfort enjoyed in face-to-face exchange. Multimodal interaction, based upon simultaneous use of sight, sound and touch, emulates naturalness and contributes to usability. But of these modes, voice bears a primary burden. Hence we aspire to provide machines with intelligence that embraces the attributes of both Perception and Production of speech.

A fundamental issue is how to achieve a compact, parsimonious representation of spoken information. And, in achieving this representation we are drawn to the ***constraints*** that characterize Perception and Production of speech. These constraints are directly reflected in Language – the protocol that communicating individuals have chosen for spoken information exchange.

Speech is not an arbitrary acoustic signal. It is generated by a unique, largely deterministic ***source***, and in the human, is comprehended by an equally unique, and, so far, incompletely understood ***receiver***. Both source and receiver exhibit constraints that can be studied, and to increasing extent, quantified. A continuing challenge is to exploit these constraints to achieve parsimonious, information-preserving representations that are appropriate for human/machine communication.

Historically, speech representation has been influenced by specific applications. These, traditionally, have been ***coding*** (for efficient transmission), ***synthesis*** (for voice answer back, dialogue generation and reconstitution from compressed data), and ***recognition*** (for voice actuation of automata). Each regime has developed specialized technologies for representation to suit the different purposes. But basically they are all the same problem -- which if understood in exquisite enough detail might coalesce the three regimes into one. This exquisite detail must capture the elements of Perception and Production – our topics here.

Some progress has been made in establishing the necessary detail. On the side of perceptual constraints, we long understand the central importance of the short-time amplitude spectrum in preserving intelligence. We exploit simultaneous masking in frequency, and forward and backward masking in time to achieve high-quality perceptual coding. And, we recognize the contributions of prosody in synthesis and reconstitution from compressed data. But, deep knowledge of information processing in the auditory system, and certainly in central comprehension, has at best been elusive.

Similarly, on the side of production constraints, improved knowledge of articulatory dynamics, together with virtually unlimited computation and accurate characterization of wave behavior in 3D enclosures, permit rapid iteration of Navier-Stokes equations to “mimic” arbitrary input speech and synthesize replicas from vocal cord/vocal tract simulations based upon first principles of fluid flow. Even so, we remain a distant way from, say, speech recognition by vector quantization of articulatory descriptors inferred from the original acoustic signal, or from the ultimate low bit-rate vocoder.

Our speakers in this Session, Professor Jont Allen from the University of Illinois, and Dr. Li Deng from Microsoft, bring a wealth of expertise from their personal research to address the issues of Perception and Production. We welcome their presentations of these topics – topics which likely will figure prominently in next-generation speech recognizers.