社団法人 電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.
SP2003-43 (2003-06)

Speech Dynamics by Ear, Eye, Mouth and Machine

Takayuki ARAI† and Steven GREENBERG‡

† Sophia University, Tokyo, Japan
‡ The Speech Institute, Oakland, California, USA

Speech communication is often modeled as a "Speech Chain" (Denes and Pinson, 1993). A speaker's message that he/she wants to transmit to a listener is first converted into a sequence of words via a linguistic process; speech organs move by motor control from the brain and speech sounds are produced. On the other hand, the sounds reaching the listener are first processed in the auditory periphery, and eventually we understand the intended message in the brain.

One of the important properties of this Speech Chain is that "change" carries a lot of important speech information. This sequence of change is controlled by the brain through articulatory movements, and as a consequence, the vocal tract's acoustic output is rich in such change. The auditory system detects these changes, and the brain decodes the information. Visual information from associated facial movement also helps to decode the speech signal.

Thus, speech is dynamic, and studies of speech dynamics by ear, eye, mouth and the brain constitutes the essence of speech study. The ramifications of this approach are important not only for science, but also for such technical fields as human-computer interface design where such knowledge can help develop more human-like machines.

During the summer of 2002, the NATO Advanced Study Institute on Dynamics of Speech Production and Perception was held at Il Ciocco in Italy. At this ASI there are many tutorial lectures and reports of state-of-the-art studies describing the processes associated with speech production and perception. The preface of the ASI Proceedings (Divenyi and Vicsi, 2002) states "The study of dynamic processes in speech has lead to a reexamination of fundamental questions in phonetics, linguistics, neuroscience, and speech technology..."

One may ask the following questions concerning speech dynamics:

> "What rate of change in the speech signal is most important for human speech recognition?" and
> "Is this rate similar to the optimum rate characteristic of automatic speech recognition (ASR)?"

Figure 1 provides a provisional answer to these questions. This figure shows the contribution to speech recognition performance as function of rate of change (or the modulation frequency) for (a) human speech recognition based on a perceptual experiment performed by Arai et al. (1999), and (b) automatic speech recognition (Kanedera et al., 1999). From this figure it is apparent that very slow changes of less than 1 Hz or very fast changes greater than 16 Hz contribute far less to speech recognition by both humans and machines than modulation frequencies in the core range of 1-16 Hz.

This bandpass characteristic of the modulation spectrum has been observed in many studies, such as in Viemeister's earlier work on sensitivity to amplitude-modulated signals (Viemeister, 1988) and the bandpass characteristics of primary auditory cortex (Schreiner and Urbas, 1988). Houtgast and Steeneken (1985) observed similar bandpass characteristics in the modulation spectrum of intelligible speech. RASTA processing in ASR also emphasizes this range of modulation frequencies for robust automatic speech recognition (Hermansky and Morgan, 1994).

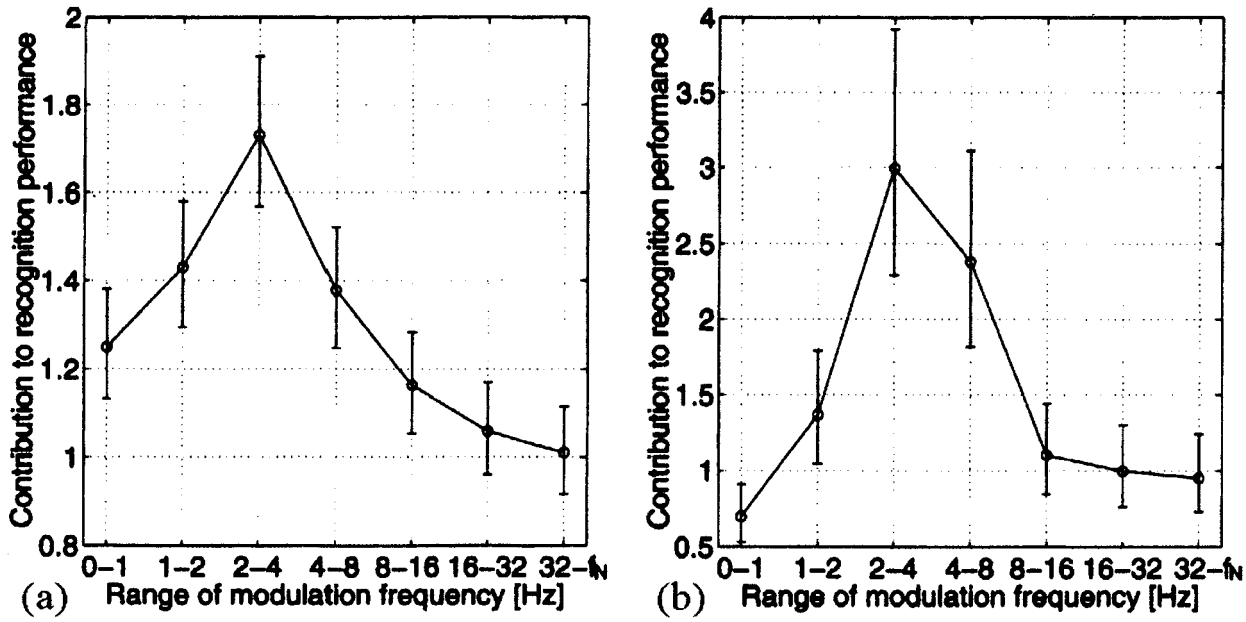

Fig. 1: Contribution to speech recognition performance by different components of the modulation spectrum; (a) human speech recognition based on the perceptual experiment in Arai et al. (1999), and (b) automatic speech recognition (Kanedera et al., 1999).

It is the purpose of the current workshop to shed some additional light on the role of speech dynamics through a broad range of scientific and technical perspectives. We would like to thank the participants of this workshop for helping to answer some of the very ancient questions posed by the dynamics of speech.

References

[1] Arai, T., Pavel, M., Hermansky, H. and Avendano, C. (1999) Syllable intelligibility for temporally filtered LPC cepstral trajectories. J. Acoust. Soc. Am. 105: 2783-2791.

[2] Denes, P. and Pinson, E. (1993) The Speech Chain: The Physics and Biology of Spoken Language (2nd ed.). San Francisco: W.H. Freeman.

[3] Divenyi, P. and Vicsi, K. (eds.) (2002) Proceedings of the NATO Advanced Study Institute on Dynamics of Speech Perception and Production, Il Ciocco, Italy.

[4] Hermansky and Morgan (1994) RASTA processing of speech. IEEE Transactions on Speech and Audio 2: 578-589.

[5] Houtgast T. and Steeneken H. (1985) "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria." J. Acoust. Soc. Am. 77: 1069-1077.

[6] Kanedera, N., Arai, T., Hermansky, H. and Pavel, M. (1999) On the relative importance of various components of the modulation spectrum for automatic speech recognition. Speech Comm. 28: 43-55.

[7] Schreiner CE, Urbas JV (1988) Representation of amplitude modulation in the auditory cortex of the cat. I. The anterior auditory field (AAF). Hearing Res 21: 227-241.

[8] Viemeister N.F. (1988) Psychophysical aspects of auditory intensity coding. In: Edelman G, Gall W and Cowan W (eds) Auditory Function. New York: Wiley, pp. 213-241.