

The relative importance of temporal and spectral cues for recognition of speech and music

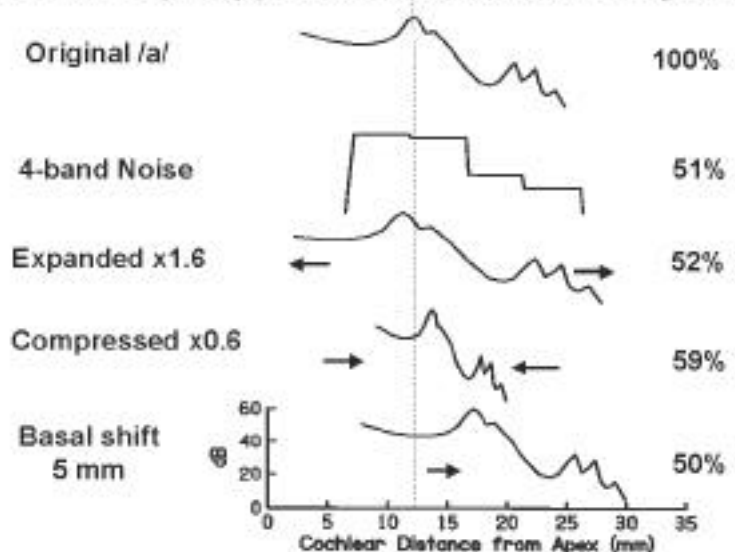
Robert V. Shannon, House Ear Institute, 2100 W. Third St., Los Angeles, CA 90057

A traditional area of auditory research has attempted to quantitatively relate physiological responses of the auditory nerve to perceptual phenomena. This has proven to be a difficult task even for such simple auditory perception as loudness and intensity discrimination. We propose that task specific demands dictate the use of information from the auditory nerve. Some tasks, like binaural localization and music appreciation, require much of the fine structure of information on timing and tonotopic place contained in the auditory nerve. Other tasks, like speech recognition, are higher-order pattern recognition tasks, which require only global aspects of the temporal and tonotopic pattern of information available on the auditory nerve. Models relating perceptual events to underlying physiological properties should take into account the differing needs of the central nervous system for the required task.

Speech Recognition in Quiet. It has long been known that speech recognition is highly robust to distortion in the acoustic signal. Recent studies related to hearing impairment and cochlear implants have quantified this robustness, and the results demonstrate the relative roles of the ear and brain in auditory pattern recognition. Experiments have shown that speech recognition is preserved even when the auditory signal is reduced to as few as four spectral channels (Remez *et al.*, 1981; ter Keurs *et al.*, 1992, 1993; Shannon *et al.*, 1995; Boothroyd *et al.*, 1996; Dorman *et al.*, 1997), when the spectral extent is severely restricted (Warren *et al.*, 1995; Lippman, 1996), when the temporal envelope is low-pass filtered below 20 Hz (Fu and Shannon, 2000; Drullman, 1995), when the cross-spectral temporal synchrony is altered by more than 100 ms (Fu *et al.*, 2001; Greenberg and Arai, 1998, Arai and Greenberg, 1998), and even when the temporal waveform is reversed in time in 50-100 ms segments (Saber *et al.*, 1999). These studies demonstrate that speech recognition is robust to the loss of spectral detail, to the loss of fine time structure, and to temporal distortions.

In contrast, speech intelligibility is devastated by distortions in the mapping of frequency to tonotopic place, even when spectral and temporal resolution are excellent (Shannon *et al.*, 1998; Fu and Shannon, 1999). Figure 1 shows a summary of the effect of frequency-place distortions on vowel recognition. The figure shows a schematic spectrum of

Effect of frequency/place distortion on vowel recognition



the original vowel /a/, and altered spectra that reduce vowel recognition to about 50% correct. Although a basal shift of 5 mm preserves the shape and detail of the spectrum, recognition is reduced by the same amount as a severe quantization of the spectrum (4 channels) when the spectral information is in the correct tonotopic location. Expansion or compression of the vowel spectral pattern in frequency/place by 60% also reduces recognition to about 50%. This pattern of results suggests that speech recognition is a higher-order pattern recognition task that requires little temporal and spectral detail, but does require the spectral information to be in the “right” cochlear location.

Speech Recognition in Difficult Listening Conditions. When the listening conditions are difficult, or when finer distinctions are required than word recognition, more spectral and temporal information are needed. For example, more spectral channels are required for word recognition in noisy listening conditions (Fu and Shannon, 1999) than in quiet. Identification of talker gender and intonation contour also require a higher level of spectral resolution than simple word identification.

Music Recognition. When music is processed with a “noise-band vocoder” (Shannon *et al.*, 1995) the melody is not recognizable until the processor contains at least 16 spectral bands, and complex melodies require 32 or more bands. Music with a familiar rhythm and vocal lyrics can be identified with fewer bands. Audio examples of different types of music processed with a noise-band vocoder will be presented.

Recent work by Delgutte and colleagues (Delgutte and McKinney, 2000; Delgutte and Oxenham, 2001) has clearly demonstrated the differences in peripheral sensory information required for speech and music perception. They created “auditory chimeras” by combining speech and music signals. They filtered speech and music into multiple spectral bands. The envelope from each band of speech was used to modulate the spectral fine structure from the same band of music. As the number of bands increased the perception changed from music to speech. Auditory chimeras directly demonstrate the different requirements for speech and music perception. Speech recognition only requires slowly changing temporal envelope information from a few spectral regions, while music perception requires spectral fine structure from a broad spectral region.

Binaural Processing. Binaural processing is a specialized function of the auditory system that places high demands on the temporal information from the two ears. Both speech and music can be perceived well monaurally, but the binaural information aids recognition in noise considerably.

Summary. Not all sensory information coded at the level of the auditory nerve is required for complex auditory pattern recognition. The aspects of the sensory pattern of information necessary for successful recognition are determined by the perceptual task. Speech recognition requires only low-frequency temporal information and relatively coarse spectral information, as long as the information is presented to the correct tonotopic place. In contrast, music recognition and binaural localization require much greater spectral and temporal resolution.

References.

- Arai, T. and Greenberg, S. (1998). Speech intelligibility in the presence of cross-channel spectral asynchrony, Proc. IEEE/ICASSP, Seattle, pp. 933-936.
- Boothroyd, A., Mulhearn, B., Gong, J., and Ostroff, J. (1996). Effects of spectral smearing on phoneme and word recognition, J. Acoust. Soc. Am., 100, 1807-1818.
- Delgutte, B. and McKinney, M. (2000). Coding of speech and music in the auditory midbrain: Low-frequency temporal modulations, Abstracts of the 23rd Annual Midwinter Research Meeting, G. Popelka (Ed.), Association for Research in Otolaryngology, Mt. Royal, NJ, p. 68.
- Delgutte, B. and Oxenham, A. (2001). Auditory chimeras, Abstracts of the 24th Annual Midwinter Research Meeting, P. Santi (Ed.), Association for Research in Otolaryngology, Mt. Royal, NJ, p. 175.
- Dorman, M.F., Loizou, P.C. and Rainey, D. (1997b). "Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs", J. Acoust. Soc. Amer., 102(4), 2403-2411.
- Drullman, R. (1995). "Temporal envelope and fine structure cues for speech intelligibility", J. Acoust. Soc. Amer., 97, 585-592.
- Fu, Q.-J., Shannon, R.V., and Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing, Journal of the Acoustical Society of America, 104(6), 3586-3596.
- Fu, Q.-J. and Shannon, R.V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing, J. Acoust. Soc. Amer., 105(3), 1889-1900.
- Fu, Q.-J. and Shannon, R.V. (2000). Effect of stimulation rate on phoneme recognition in cochlear implants, Journal of the Acoustical Society of America, 107(1), 589-597.
- Fu, Q.-J. and Galvin, J. (2001). Recognition of spectrally asynchronous speech by normal-hearing listeners and Nucleus-22 cochlear implant users, J. Acoust. Soc. Amer., 109(3), 1166-1172.
- Greenberg, S. and Arai, T. (1998). Speech intelligibility is highly tolerant of cross-channel spectral asynchrony, Proc. Int. Cong. Acoust., Seattle, pp. 2677-2678.
- Lippmann, R.P. (1996). Accurate consonant perception without mid-frequency energy, IEEE Trans. Speech Audio. Proc., 4, 66-69.
- Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). "Speech perception without traditional speech cues", Science, 212, 947-950.
- Shannon, R.V., Zeng, F.-G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. Science, 270, 303-304.
- Shannon, R.V., Zeng, F.-G., and Wygonski, J. (1998). Speech recognition with altered spectral distribution of envelope cues, J. Acoust. Soc. Amer., 104(4), 2467-2476.
- Saberi, K. and Perrott, D.R. (1999). Cognitive restoration of reversed speech, Nature, 398, 760.
- ter Keurs, M., Festen, J.M., and Plomp, R. (1992). "Effect of spectral envelope smearing on speech reception. I", J. Acoust. Soc. Amer., 91, 2872-2880.
- ter Keurs, M., Festen, J.M., and Plomp, R. (1993). "Effect of spectral envelope smearing on speech reception. II", J. Acoust. Soc. Amer., 93, 1547-1552.
- Warren, R.M., Reiner, K.R., Bashford, J.A. Jr., and Brubaker, B.S. (1995). Spectral redundancy: Intelligibility of sentences heard through narrow spectral slits, Perception and Psychophysics, 57(2), 175-182.