12  Spatiotemporal Encoding of Vowels in Noise Studied with the Responses of Individual Auditory-Nerve Fibers

MICHAEL G. HEINZ

1  Introduction

The neural basis for robust speech perception exhibited by human listeners (e.g., across sound levels or background noises) remains unknown. The encoding of spectral shape based on auditory-nerve (AN) discharge rate degrades significantly at high sound levels, particularly in high spontaneous-rate (SR) fibers (Sachs and Young 1979). However, continued support for rate coding has come from the observations that robust spectral coding occurs in some low-SR fibers for vowels in quiet and that rate-difference profiles provide enough information to account for behavioral discrimination of vowels (Conley and Keilson 1995; May, Huang, Le Prell, and Hienz 1996). Despite this support, it is clear that temporal codes are more robust than rate (Young and Sachs 1979), especially in noise (Delgutte and Kiang 1984; Sachs, Voigt, and Young 1983). Sachs et al. (1983) showed that rate coding in low-SR fibers was significantly degraded at a moderate signal-to-noise ratio for which human perception is robust. In contrast, temporal coding based on the average-localized-synchronized-rate (ALSR) remained robust.

Although temporal coding based on ALSR is often shown to be robust, evidence for neural mechanisms to decode these cues is limited. Spatiotemporal mechanisms have been proposed for decoding these types of cues (e.g., Carney, Heinz, Evilsizer, Gilkey, and Colburn 2002; Deng and Geisler 1987; Shamma 1985). However, the detailed evaluation of spatiotemporal mechanisms has been limited primarily to modeling studies due to difficulties associated with the large population responses that are required to study spatiotemporal coding (e.g., see Palmer 1990). For example, Deng and Geisler (1987) used a transmission-line based AN model to suggest that spectral coding based on the peak cross-correlation between adjacent best-frequency (BF) channels was robust in the presence of background noise. In the present study, spectral coding of vowels in noise based on rate, ALSR, and a simple cross-BF coincidence detection scheme is evaluated from the responses of single AN fibers. By using data from a single AN fiber, many of the difficulties associated with large-population studies are eliminated.
2 Methods

AN recordings were made from pentobarbital anesthetized cats using standard methods (see Heinz and Young 2004). Spikes were measured with 10-µs resolution. Each fiber was characterized using an automated tuning-curve algorithm to determine BF, Q10, and SR.

All vowels were created using a cascade synthesizer and were scaled versions of the vowel /eh/, which has its first two formants at F1=0.5 kHz and F2=1.7 kHz, with the intermediate trough at T1=1.2 kHz. To maintain F0 within the voice-pitch range for each BF, a baseline steady-state vowel was re-synthesized for each AN fiber. The baseline vowel had F0=75 Hz and was created with F2 at BF. The other formant frequencies and all bandwidths were scaled based on the frequency shift from the nominal F2 value for /eh/. The baseline vowel and a baseline broadband noise token were both 400 ms in duration and were sampled at 33000 Hz. The vowel-in-noise conditions with F1 and T1 near BF were produced via changes in sampling rate for the vowel and noise. Signal-to-noise ratio in dB was defined as the difference in overall vowel level and noise level within the frequency range from 0 Hz to the trough between the 3rd and 4th formant of the baseline vowel.

Spectral coding was evaluated based on individual-neuron responses in a manner similar to the spectrum manipulation procedure (SMP), which was developed to study rate-based spectral coding (e.g., May et al. 1996). In the SMP, spectral coding is evaluated by comparing responses to vowels with formants and troughs placed at BF via changes in sampling rate. The slope of the discharge rate as a function of vowel feature level is used to quantify spectral coding, with robust coding indicated by a constant slope across vowel level (or SNR). Although the SMP is useful for evaluating rate coding, changes in the temporal waveform with changes in sampling rate do not allow spatiotemporal coding to be evaluated.

The spectro-temporal manipulation procedure (STMP) was developed to study spatiotemporal coding by estimating the responses of several neurons with nearby BFs to the same stimulus waveform using the responses of a single neuron to different stimuli with a spectral feature shifted nearby BF (Heinz 2005). Based on a neuron with BF0 and a vowel with F1=BF0, the response of a neuron with BF<BF0 can be predicted by playing the vowel at a higher sampling frequency, thus increasing the frequency of all vowel features. The response of the below-F1 neuron to the baseline vowel waveform (F1 at BF0) is estimated by scaling up the measured spike times in response to the shifted vowel by the same factor used to increase the sampling rate, thus reducing the effective vowel feature frequencies to baseline values. To account for the fixed neural delay, an offset of 1 ms was subtracted from all spike times prior to time scaling and then added back afterwards. Temporal scaling changes the overall discharge rate of the estimated fiber response, which is accounted for by scaling the resulting period histograms by the temporal scaling factor. A computational AN model has been used to test the STMP approach (Heinz 2005; Zhang, Heinz, Bruce, and Carney 2001).