REPRESENTATION OF SPEECH SIGNALS IN THE DISORDERED PERIPHERAL AUDITORY SYSTEM

Donald G. Jamieson, Margaret F. Cheesman and Stefan Krol
Hearing Health Care Research Unit, Communicative Disorders
University of Western Ontario, London, Ontario, Canada

1. INTRODUCTION

The obvious limits on invasive, physiological research with human subjects restrict study of the representation of sounds in the disordered auditory system to two approaches. First, psychophysical techniques can be used to **measure** the response of the human auditory system to speech. Second, computational models can be used to **simulate** speech processing in the human peripheral auditory system. We have applied both approaches, using identical stimuli.

1.1 Physiological Models

Our model of the disordered peripheral auditory system is based closely on that described by Kates (in press) and, for comparison with normal hearing, we have also implemented a model developed by Payton [8;9]. The Kates model is a digital one, operating in the time domain, that concatenates several submodels, each representing a different part of the peripheral auditory system. Thus, the contributions of various parts of the auditory system -- the middle ear, basilar membrane (BM), inner and outer hair cells (IHC and OHC) and their synaptic junctions to auditory-nerve fibers --- are simulated, and can be studied separately. The composite model simulates the signal transformations that occur at each stage of the system, based on biophysical, biomechanical and electrophysiological observations.

The input to the model is an acoustic signal, represented as the pattern of sound pressure at the tympanic membrane. The output of the model is the neural firing rate in the auditory nerve. The processing of the sound by the BM is modeled as a cascade of filter sections. Each section is a third-order bandpass filter with parameters chosen to match the BM tuning. The output at each point on the membrane is the result of processing in all previous filter sections and is further sharpened by a second filter when the BM response is converted into a neural firing pattern by the inner hair cells. In Kates' model, this transduction is represented by a modification of Allen's [1] electrical circuit model. This model is closely connected with calcium ion kinetics during transduction. The inner hair cell (IHC) model transforms the mechanical input to the hair cell into the instantaneous firing rate of a single neural fiber. Four neural fibers are attached to each hair cell: two high-spontaneous rate fibers (50 and 75 spikes/sec) and two low-spontaneous rate fibers (5 and 10 spikes/sec). The output at each point on the BM is the arithmetic mean of the outputs of the four fibers attached to the hair cell. A novel feature of Kates' model is the feedback between the hair cells and the BM; the intensity of firing modifies the parameters of the BM filters, simulating the action of the OHCs. Adjusting this feedback path can mimic the effects of OHC damage; adjusting the resistance of hair-cell circuits can simulate damage to the IHCs.

In the Kates' model the sampling rate is determined, in principle, only by the temporal scale of investigated acoustic phenomena and digital signal processing requirements (Nyquist theorem). Kates used 40 kHz; in our studies we used 42 kHz in order to match the 14 kHz sampling rate signals used psychophysical studies. This is an advantage over Payton's model, which requires that the acoustic input be sampled at least every 0.0065 ms.

In the Payton model, the middle ear portion follows Guinan & Peake [4], who derived an analog circuit having the frequency response characteristics observed in the motion of the middle ear ossicles of anesthetized cats. A two-dimensional BM model and added sharpening mechanism were used to spatially distribute and filter stimulus frequencies.

The equations and solution methods for the cochlear fluid and BM motions are taken from Allen & Sondhi [2], and an FFT is applied. Values of the material and dynamic parameters for the cochlea were chosen to approximate an empirical cochlear place/frequency map based on anatomical measurements from cats [6]:

$$\text{resonance frequency} = 456.*(10**(2.1*(1.-x/L) - 0.8)$$

where x is the distance from the stapes and L is the length of the BM. Using this approach, the active frequencies of the model vary from 119 Hz to 57 kHz -- well beyond the range of useful human hearing. (Similar Greenwood-type formulae with coefficients matching human hearing data are used in Kates' model to place the center frequencies of 112 BM filters). Payton used 20 points on the BM, between $x/L=0.35$ and $x/L=0.825$, covering the frequency interval 270 Hz to 6800 Hz. An improved cochlear map for the present version of this 20-place model was derived from calculations of the synchronization index of the firing rate, with the frequency having the maximal index for a given place being taken as a resonance frequency. The displacement of any point along the BM resembles the output of a sharp bandpass filter with a low frequency tail. This stage of auditory signal processing improves the resolution of tuning curves, making them sharper, to more closely approximate empirical results (the so-called "second filter").

The output of the final stage, that of the excitation of the neural fibers, is modelled in three stages, following Payton [9] and Smith-Brachman [10]: 1) the signal is first passed through a half-wave rectifier; 2) this output is then lowpass filtered; 3) neurotransmitter release into the synaptic cleft is simulated.

The probability of neural firing as a function of time is proportional to the amount of neurotransmitter released by the first of three reservoirs in the model. In our model, the amount of neurotransmitter is transformed to generate a sequence of neural spikes under the assumption that spike generation is a nonstationary Poisson process. This transformation permits period histograms and inter-spike histograms to be generated.

1.2. Models of Disordered Speech

The modelled output of the normal and disordered auditory periphery was computed for synthetic fricative-vowel syllables for which psychophysical estimates of the auditory representation had been obtained previously [3]. The representations of these speech stimuli are of particular interest because the perception of the fricative portion is stongly dependent on the transition and vowel frequencies [7].

2. METHOD

Stimuli were synthetic fricative-vowel (/ʃ/ and /s/; /i/ and /u/) syllables selected based on physical parameters and on subjects' identifications and judgements of the perceptual similarity of a larger set of fricative sounds [3]. These syllables represented conditions under which the vowel context affected the perception of the fricative.

Psychophysical masking patterns were obtained using 10-ms pure tones with 5-ms raised-cosine rise and fall ramps. Seventeen probe frequencies were used, ranging from 1 kHz to 5 kHz in 0.25 kHz steps. Probes were positioned at several temporal locations in the masker; probe delays are expressed relative to the end of the fricative/onset of the vowel.

Monaural probe thresholds in the presence of the three synthetic speech maskers were obtained using a method of adjustment. Further details of the procedure are available in Cheesman [3].

2.3 Application of the Models

2.3.1 Payton model

2.3.1.1 Interpolation of Sampled Data As the speech signals of interest were sampled at 14kHz, when applying the Payton model the sampling period was adjusted to be 0.071 ms, rather than the required $\leq$0.0065 ms. To satisfy this requirement, 12 points were interpolated between each pair of sampled data points using linear interpolation. This action reduced the sampling period to 0.00595 ms, with 4200 samples within each 25 ms segment of the signal.

2.3.1.2 Stimulus Scaling Because the intensity of firing depends on the level of the input signal, it is important to ensure that each portion of the signal is appropriately scaled. This was accomplished by calculating the root mean square amplitude across the signal and using this value to normalize each sample, by dividing by the root mean square value.

2.3.1.3 Processing. To approximate the psychophysical procedure as closely as possible, the first 250 ms of the signal were processed through the model, and the firing rate calculated at each of six specific time points. Specifically, the **mean** firing rate was calculated for each of the following 10 ms intervals, measured from signal onset: 95 to 105 ms, 120 to 130 ms, 145 to 155 ms, 170 to 180 ms, 195 to 205 ms, and 220 to 230 ms. These calculations were made at each of 20 frequencies (i.e., for the output associated with each of 20 basilar membrane locations), to generate neural analogs to the psychophysical masking patterns.

2.3.2 Kates model

Application Kates' model to our stimuli was more straightforward. It required the interpolation of 3 points in order to match 42 kHz sampling rate.

3. RESULTS AND DISCUSSION

3.1. Masking Patterns

Several consistent patterns were evident in each of the masking patterns. The masking patterns showed good representation of the frication noise and the vocalic portion of the syllables. At the junction of the fricative and vowel, portions of the lower frequency vowel and fricative information were present in the patterns.

3.2. Model Output

The model described above was applied to process the speech signals described above, in order to compare psychophysical measures of the auditory representation of speech with firing rates patterns. In general we have seen good correspondence between the data obtained through the psychophysical and modelling approaches, with the auditory model appearing to be somewhat more sensitive than the normal ear to the acoustic properties of the speech signal. As one example, the transition from fricative noise to vowel region is very sharp in the model.

In the normal ear, firing rate is synchronized with the first and second formant (F1 and F2) for voiced speech sounds. When the OHCs are damaged, the firing rate decreases and most timing information is lost, with synchronization being observed only for F1. In addition, a strong onset response in the low frequency fibres disappears when OHCs are damaged, suggesting that the OHC/BM feadback is responsible for the onset response.

When the stereocillia are damaged in a significant proportion of the IHCs in a particular region of the cochlea, timing information appears to remain intact in the neural output, but the intensity of firing is substantially reduced, possibly below the difference threshold.

4. REFERENCES

[1] ALLEN, J.B.(198), "A hair-cell model of neural response", in Peripheral Auditory Mechanisms, edited by E. de Boer and M.A.Viergever, the Hague: Martinus Nijhoff Publishers, 1983.

[2] ALLEN, J.B., & SONDHI, M.M. (1979), "Cochlear macromechanics: Time domain solutions", *Journal of the Acoustical Society of America, 66*, 123-132.

[3] CHEESMAN, M.F. (1989), "The auditory representation of context-conditioned fricatives", Doctoral Dissertation, University of Minnesota.

[4] GUINAN, J.J. & PEAKE, W.T. (1967), "Middle-ear characteristics of anaesthetized cats", *Journal of the Acoustical Society of America, 41*, 1237-1261.

[5] KATES, J.M. (1991), "A time domain digital cochlear model",IEEE Transactions on Acoustics Speech and Signal Processing, in press.

[6] LIBERMAN, M.C. (1982), "The cochlear frequency map for the cat: Labeling auditory-nerve fibers of known characteristics frequency", *Journal of the Acoustical Society of America, 72*, 1441-1449.

[7] MANN, V.A. & REPP, B.H. (1980), "Influence of the vocalic context on perception of the [ʃ]-[s] continuum", *Perception & Psychophysics, 28*, 213-228.

[8] PAYTON, K.L. (1986), "Vowel processing by a model of the auditory periphery", *Ph. D. Thesis, John Hopkins University*, Baltimore.

[9] PAYTON, K.L. (1988), "Vowel processing by a model of the auditory periphery: A comparison to eight-nerve responses", *Journal of the Acoustical Society of America, 83*, 145-162.

[10] SMITH, R.L., BRACHMAN, M.L. (1982), "Adaptation in auditory-nerve fibers: A revised model", *Biological Cybernetics, 44*, 107-120.

6. ACKNOWLEDGEMENTS

We are grateful to Ketan Ramji and Lucy Kieffer for technical assistance and to J. Kates and K. Payton for providing code for the auditory model. The work was supported, in part, by grants from NSERC to MFC and from NSERC, URIF and Unitron Industries Ltd. to DGJ.

Correspondence should be addressed to Dr. D.G. Jamieson, Hearing Health Care Research Unit, Department of Communicative Disorders, University of Western Ontario, London, ON, CANADA, N6G 1H1.