Proceedings of The Institute of Acoustics

A PERIPHERAL AUDITORY MODEL FOR SPEECH PROCESSING.

Roy B. Gardner and Mahesh K. Uppal

Laboratory of Experimental Psychology, University of Sussex, Brighton BN1 9QG, England.

INTRODUCTION

The Problem: The computational problem facing an auditory model for speech recognition is the same as that faced by the human auditory system: To produce a robust representation of those acoustic features of speech signals necessary for their reliable recognition. These features are likely to consist of formants, periodicity and voicing information etc.. In addition, if the model is to have any use in a practical and general recognition system, the importance and consequences of the constancy of percept acheived by the ear in spite of distortions and additions of sound from other sources must be fully appreciated. Work at our laboratory (Darwin and Gardner [1]) has shown that the raw acoustic input to the ear is first subject to mechanisms that group together components of the sound according to general auditory principles. These include onset and offset times, harmonicity and fundemental frequency. Thus to be of use the model must, at the very least, create a representation not only of the important acoustic features of the speech, but one which also permits the operation of grouping algorithms.

Rationale: The model is based on the assumption that knowledge of the neural processing of speech is useful in the design of the acoustic processing stages of automatic speech recognition systems. It seems a reasonable approach when we consider the excellent performance of the human speech recognition system. The model is not a literal model of the peripheral auditory system; that is, we have not explicity incorporated every known property of peripheral auditory processing in the model. What we have attempted is the simulation of those properties that we feel are the key to solving the problems of representation. These properties are the excitation and suppression of auditory neural activity, and the frequency tuning characteristics and operating characteristics, including thresholds, of these processes. The model consists of an array of channels each of which can be regarded an as auditory-nerve fibre analogue, the output of each channel being a simulation of a post-stimulus time histogram, which is the distribution of nerve fibre firing as a function of time after stimulus onset. Like the auditory system the model retains information about the fine time structure of the signal, as well as information about the distribution of activity across the array of channels (place code). This allows representations of the signal to be created on the basis of place codes and time codes as well as combined place/time codes.

Weintraub [2] has implemented a computational model of auditory sound separation based on analysis of the fine-time information in auditory nerve fibre activity and recent neuro-physiological research has added weight to the view that this information forms the basis of the neural representation of speech sounds. For example Young and Sachs [3] and Delgutte and Kiang [4] have demonstrated that vowel spectra are well represented in the distribution of

activity phase-locked to individual harmonics of the stimulus (the average localised synchronised rate or measure function, ALSR or ALSM). This form of combined rate/time/place code is robust and has a wide dynamic range due in part to non-linear aspects of nerve fibre responses such as suppression phenomena (Sachs and Young [5]). The ALSR representation also encodes consonants, voiceless fricatives and vowels in noise (Delgutte and Kiang [4],[6],[7]). Delgutte [8] has suggested that the ALSR type of representation is the means by which the essential aspects of speech sounds are encoded by the auditory system. We (Palmer et al [9]) have evidence suggesting psychophysical judgements are based on ALSR-type representations.

Suppression: We are primarily concerned with modelling the suppression phenomenon in order to enhance the primary acoustic features of speech signals. The phenomenon of suppression is found in auditory nerve fibres (eg. Javel et al [10]) and hair-cells (Sellick and Russell [11]), and in the auditory nerve takes two forms: (a) Rate suppression. The rate of firing to a single tone, the primary, usually at fibre CF, is found to fall if a second more intense tone is presented simultaneously within a certain range of frequencies beyond the edges of the fibre tuning-curve. This second tone alone would not excite the fibre and is thus said to have a purely suppressive effect. It lies therefore in a frequency region where the excitatory and suppression response areas of the fibre do not overlap. (b) Synchrony suppression. The second tone lies within the excitatory tuning-curve and increases the overall rate of firing of the fibre. Analysis based on the phase-locked temporal response of the fibre shows that the activity synchronised to the primary is reduced in the presence of the second tone. The second tone has thus weakened the representation of the primary in the temporal discharge pattern of the nerve fibre. In general tones falling within the response areas of a fibre will exert a mutually suppressive effect upon each other to a degree dependent upon their intensity and frequency. The effect of this non-linearity is an increase in the relative activity devoted to strong stimulus components (Javel et al [10]), which is not found in linear models of peripheral auditory activity (Sinex and Geisler [12]).

THE MODEL

Stage 1 - Initial frequency analysis: This consists of an array of 64 equally spaced, symmetrical, Butterworth 4th order linear band-pass filters in the range 50 - 5000 Hz. These provide an analysis of the input waveform into frequency components. The bandwidths of these filters are equal and are constrained by the need for an analysis detailed enough to separate the harmonics of voiced speech without losing signal onsets. We are currently testing a number of bandwidths for a range of representative speech sounds.

Stage 2 - The model. The stimulus for the development of the model came from the suggestion of Javel et al [10] that the excitatory potentials of hair cells are produced by the positive going polarity of a stimulating waveform and suppression by the negative going polarity. We were not concerned with the exact mechanism of hair cell transduction but with the possibility of implementing the sense of the suggestion as the basis for our suppressive and excitatory processes, at the same time moving closer to a more realistic modelling of the suppression phenomenon. Our approach differs radically from models of suppression using the RPNL approach (eg. Cooke [12]) which consist

of a compressive non-linearity placed between two band-pass filters, the second of these having a narrower bandwidth than the first.

It follows from our scheme that at any time an input component will exercise an excitatory or suppressive effect on a channel but never both. If a signal contains a number of frequency components then at any time, t, some of these components will excite the channel and some will exert an opposite suppressive influence. If the total excitatory influence is greater than the combined suppressive effects, and exceeds a threshold value, then the channel output becomes non-zero at that time with a value given by the total resultant excitation. The output of a channel can only be positive-going. In this, initial implementation of the model, it is assumed that excitation and suppression are instantaneous and do not extend to adjacent time samples.

Each channel has two band-pass filter functions associated with it; an excitatory function and a wider (by a factor of 1.5) suppressive function. These are of skewed Gaussian form with steeper high frequency slopes. Each channel is defined by a characteristic frequency (CF) which is the frequency of component to which the channel produces maximum excitation or suppression at a particular intensity and a bandwidth. The general form of the function is as follows (see Figure 1):

$$y = x^{-c(cf_j - F_i)^2/2\pi b^2} \qquad \begin{matrix} c=1 & \text{if } F_i \leq CF_j \\ c=4 & \text{if } F_i > CF_j \end{matrix}$$

where x is the amplitude of a component in dB
y is the output in dB
CF_j is the characteristic frequency of the channel in Hz
F_i is the frequency of the component
b is the bandwidth of the channel in Hz
c is the skew factor

In the current implementation the formal bandwidth of the excitatory channels is constant up to 1.5 kHz and from 1.5 KHz to 5 KHz is defined by (see Figure 1):

$$b_e = (CF_j \times B_e)/1500$$

$$b_s = 1.5[(CF_j \times B_e)/1500]$$

where b_e is the excitatory bandwidth in Hz
b_s is the suppression bandwidth in Hz
B_e is the excitatory bandwidth below 1.5 kHz

and

$$B_s = 1.5\,B_e$$

where B_s is the suppressive bandwidth below 1.5 kHz

The bandwidths thus increase with increasing CF. The initial bandwidths and the slope of the CF-frequency function are variable.

The computation of excitation and suppression is also determined by operating characteristics which define the output, in dB, of a channel as a function of the input intensity, also in dB. In the current implementation these are

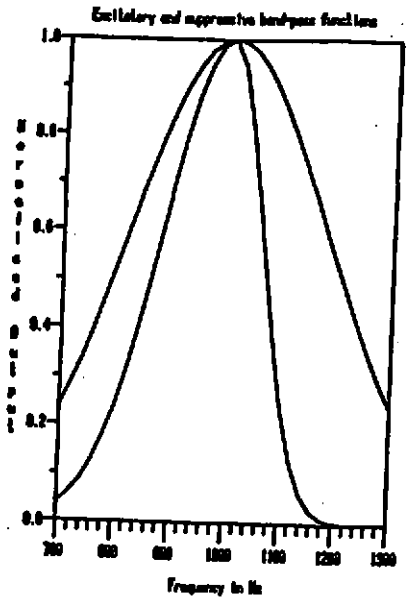

Figure 1. The form of the band-pass functions for excitation and suppression are shown for a channel with a CF of 1.0 kHz.

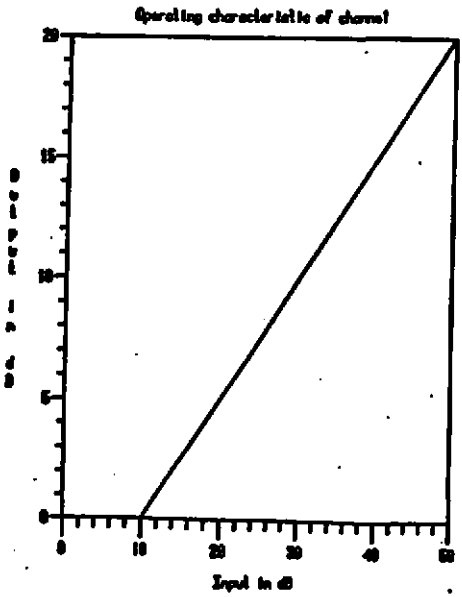

Figure 2. This shows the general form of the operating characteristics of excitation and suppression. Slope and threshold are variables.

linear functions with variable slope and threshold (see Figure 2):

$$y = a(x - [thresh])$$

where $[thresh]$ is the threshold in dB
a is the slope

The calculation of the output of a particular channel is done on a sample by sample basis in 50 msec. blocks as follows:

(a) The amplitude, $[amp]_{ik}$ at sample k of each component f is converted to dB.

$$\text{if } [amp]_{ik} > 0 \quad x_{ik} = 20 \log_{10} [amp]_{ik}$$
$$\text{if } [amp]_{ik} < 0 \quad s_{ik} = 20 \log_{10} |[amp]_{ik}|$$

where x_{ik} is an excitatory sample
s_{ik} is a suppressive sample

(b) The total suppression at sample k for channel j, S_{jk}, is calculated. That is, the suppressive contributions of the individual negative- going components at sample k are calculated and summed. The contribution of an individual harmonic to this total is a function both of the distance in frequency of the component from the characteristic frequency of the channel, and of the intensity of the component in dB.

$$S_{jk} = \sum_{i=1}^{i=nsup} 10^{[sup]_{ik}/20}$$

$$[sup]_{ik} = a_s \left\{ \left(s_{ik} e^{-c(CF_j - F_i)^2 / 2\pi (b_s)_j^2} \right) - [thresh]_s \right\}$$

where $nsup$ is the number of suppressive components
$[thresh]_s$ is the suppression threshold
a_s is the slope of the suppression operating characteristic

(c) The resultant excitation, E_{jk}, produced by a single positive-going component at sample k is calculated, as a function of the distance in frequency of the component from the CF of the channel. The total excitation is then calculated.

$$[excit]_{ik} = a_e \left\{ x_{ik} e^{-c(CF_j - F_i)^2 / 2\pi (b_e)_j^2} - [thresh]_e \right\}$$

$$E_{jk} = \sum_{i=1}^{i=nexct} 10^{[excit]_{ik}/20}$$

where $[thresh]_e$ is the excitation threshold
a_e is the slope of the excitation operating characteristic
$nexct$ is the number of excitatory components

(d) The total resultant excitation, R_{jk}, for the channel at sample k is is calculated from the excitatory operating characteristic.

$$R_{jk} = E_{jk} / S_{jk}$$

(e) If the total resultant excitation at sample k is greater than zero the value is assigned to the channel output at that sample.

A PERIPHERAL AUDITORY MODEL FOR SPEECH PROCESSING

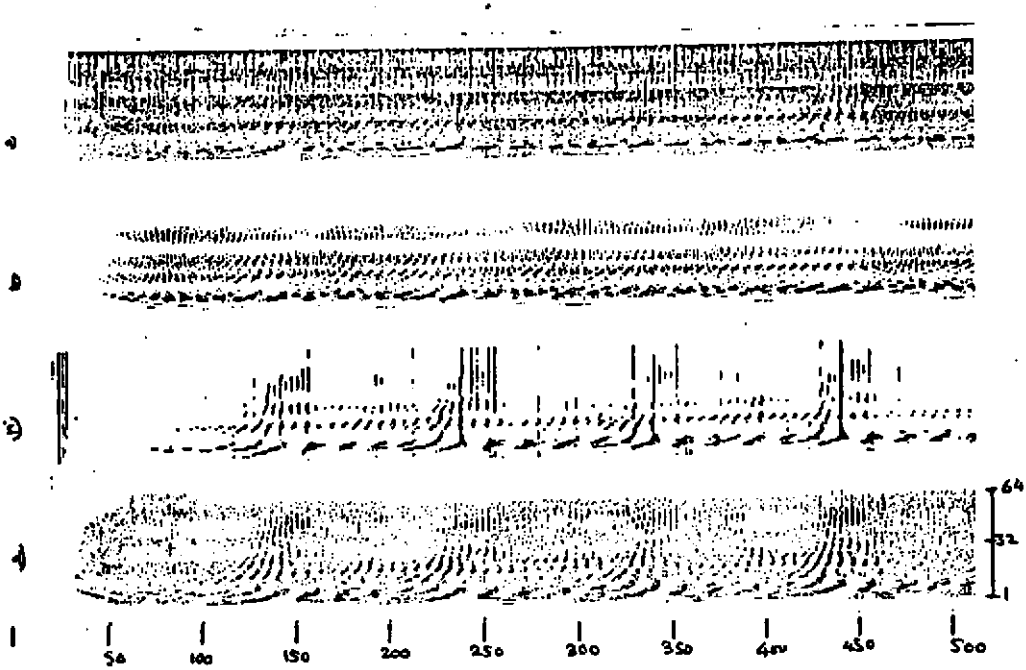

Figure3. Plots of model output for the vowel /e/ spoken by Roy Gardner. The darker the gray-scale the greater the activity in the channel at that point. The vertical axis represents the frequency range 50 - 5000 Hz on a linear scale, the horizontal axis 0 - 50 msecs.Bandwidth for initial linear filters is 120 Hz Bandwidth B of excitatory filters is 100 Hz. (a) Raw model output without any suppression, (b) again no suppression but with the the excitatory threshold set at 20 dB. Both plots show clearer the positions of the main spectral features of the vowel. Analysis of the temporal pattern of activity in channels shows dominance of these patterns by components nearest the channel CF. (c) and (d) show output with suppression. In both cases the threshold of suppression is set to zero as is that of the excitation. The slopes of the suppressive operating characteristic are 1.0 and 0.5 respectively. (c) shows an extreme case of suppression leaving only the bare bones of the spectrum of the vowel. F4 has disappeared channels in the this region responding to the F3 component if at all. The general picture is that temporal patterns of activity become dominated by the dominant spectral components (at the formants) of the signal. This spread of the influence of these dominant components is a function of the bandwidth of the channels. Channel outputs also code the f0 of the signal.

A PERIPHERAL AUDITORY MODEL FOR SPEECH PROCESSING

$$[out]_{jk} = R_{jk} \quad \text{if } R_{jk} > 0$$
$$[out]_{jk} = 0 \quad \text{if } R_{jk} \leq 0$$

where $[out]_{jk}$ is the activity in channel j at sample k

(f) The channel output is low-pass filtered (moving average) to simulate the loss of phase-locking to high frequency components found in auditory nerve fibres.

PROCESSING OF THE MODEL OUTPUT:

At the time of writing we are assessing various ways of processing the raw model output to create representations appropriate to our objectives. These include:

(a) Place code processing - the position and bandwidth of maximum activity in the channel array is defined at some point in time. This gives estimates of formant frequencies over time.

(b) Dominant component analysis - The dominant component in the temporal pattern of activity of a particular channel is defined by auto-correlation. Plots of dominant component against channel CF give estimates of the dominant spectral components of the signal (see Degutte [9]).

(c) ALSM analysis - a measure of a component's amplitude is calculated by taking the average measure of the activity synchronised to the component in a range of channels with CFs close to the component frequency.

(d) Time code processing - the level of the CF component of a particular channel is determined from auto-correlation functions.

(e) Extraction of FO - the output of almost all channnels is modulated by the FO component of voiced speech as evidenced from the vertical bands in the raw model output. This allows a running measure of FO to be calculated, and a discrimination of the voiced-unvoiced didtinction.

(f) The spectral-temporal pattern of activity in the array of channels is subjected to lateral inhibitory processing (cf Shamma [15]) to enhance dominant spectral features of the signal.

(g) Auto-correlation and coincidence functions are calculated to provide the material for Weintraub's [2] grouping algorithms.

IMPLEMENTATION:

All programs are written in FORTRAN 77 and run on a DEC VAX-11/780 machine.

REFERENCES

[1] Darwin,C.J. and Gardner,R.B. , 'Perceptual separation of speech from concurrent sounds', in M.E.H.Schouten (Ed.), The Psychophysics of Speech Perception, NATO ASI Series, M.Nijhoff, Netherlands, (in press).

[2] Weintraub,M., 'A theory and computational model of auditory sound separation', PhD. thesis, Stanford University, (1985).
[3] Young,E.D. and Sachs,M.B., 'Representation of steady-state vowels in the temporal aspects of the discharge paterns of populations of auditory nerve fibers', J.Acoust.Soc.Amer., 66, 1381-1403, (1979).
[4] Delgutte,B. and Kiang,N.Y.S., 'Speech coding in the auditory nerve: I. Vowel-like sounds', J.Acoust.Soc.Amer., 75, 866-878, (1984).
[5] Sachs,M.B. and Young,E.D., 'Effects of nonlinearities on speech encoding in the auditory nerve', J.Acoust.Soc.Amer., 68, 858-875, (1980).
[6] Delgutte,B. and Kiang,N.Y.S., 'Speech coding in the auditory nerve: III. Voiceless fricative consonants', J.Acoust.Soc.Amer., 75, 887-896, (1984).
[7] Delgutte,B. and Kiang,N.Y.S., 'Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics', J.Acoust.Soc.Amer., 75, 897-907, (1984).
[8] Delgutte,B. and Kiang,N.Y.S., 'Speech coding in the auditory nerve: V. Vowels in background noise', J.Acoust.Soc.Amer., 75, 908-918, (1984).
[9] Delgutte,B., 'Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds', J.Acoust.Soc.Amer., 75, 879-886, (1984).
[10] Palmer,A.R., Winter,I.M., Gardner,R.B. and Darwin,C.J., 'Changes in the phonemic quality and neural representation of a vowel by alteration of the relative phase of harmonics near F1', in M.E.H. Schouten (Ed.), The Psychophysics of Speech Perception, NATO ASI Series, M.Nijhoff, Netherlands, (in press).
[11] Javel,E., McGee,J., Walsh,E.J., Farley,G.R. and Gorga,M.P., 'Suppression of auditory-nerve responses. II. Suppression threshold and growth, iso-suppression contours', J.Acoust.Soc.Amer., 66, 801-813, (1985).
[12] Sellick,P.M. and Russell,I.J., 'Two-tone suppression in cochlear hair cells', Hearing Research, 1, 227-236, (1979).
[13] Sinex,D.G. and Geisler,C.D., 'Comparison of the responses of auditory nerve fibers to consonant-vowel syllables with predictions from linear models', J.Acoust.Soc.Amer., 76, 116-121, (1984).
[14] Cooke,M.P., 'A computer model of peripheral auditory processing', NPL Report DITC 58/85, (1985).
[15] Shamma,S.A., 'Speech processing in the auditory system II: Lateral inhibition and central processing of speech evoked activity in the auditory nerve', J.Acoust.Soc.Amer., 78, 1622-1632, (1985).