Proceedings of The Institute of Acoustics

# EXPLOITING PHYSIOLOGICAL CODES IN ASR

Ray Meddis

Department of Human Sciences, Loughborough.

## INTRODUCTION

Since human listeners are acknowledged to be experts at recognising speech, connected or unconnected, the strategy of simulating human listening function would appear to be potentially very fruitful. The problem is that we have only loose notions of how this human ability actually works. We are, however, not totally ignorant and it is a serious research question as to how much progress we can make by implementing existing knowledge in ASR technology. When a sound engineer, intent upon making a stereophonic recording, places his microphones in a dummy head with pinnae in position corresponding to the eardrums [1, p252], he is engaged in just such an implementation. Similarly, the use of an array of band-pass filters in a recognition device is a passable analogue of the mechanical filtering effect of the basilar membrane in the mammalian ear. When these filters have their centre frequencies equally spaced on a Bark scale [2], the analogy is made explicit. Future developments could include simulation of the characteristics of the travelling wave along the basilar membrane [3] and the generation of spike trains such as are found in fibres of the auditory nerve [4].

The possible benefits of these strategies are various. The use of realistic stereophonic recordings should help with the separation of auditory figure from ground, i.e. the isolation of the focal stimulus from a noisy, distracting background. The use of carefully spaced band-pass filters should be an efficient arrangement for extracting frequency information, if we assume that the auditory system is optimally configured for this purpose. Modelling the travelling wave introduces phase shifts in components of the signal frequencies which lie close to the center frequency of the bandpass filter. This can be used to emphasise dominant frequencies making them, easier to extract from the total stimulus pattern.

Modelling spike trains represents an uncertain and relatively unexplored possibility. Certainly, such trains show a dramatic increase in density at the onset of a stimulus and a period of abnormally low activity immediately after stimulus effect. These properties may prove useful in marking changes in stimulus quality. If information concerning the identity of a word or phrase is carried in the pattern of change of stimulus parameters rather than during steady state portions then this could be a very valuable property indeed.

Extracting useful information from stereophonic inputs depends partly on relative amplitude at the two ears and partly on timing differences. Because spike activity is finely located in time, it has considerable potential for the extraction of information concerning the timing component of the spatial localisation of the sound source. As a consequence, spike trains may be useful in the separation of an auditory image from its background.

Because auditory nerve spikes are only nonlinearly and probablistically related to the instantaneous amplitude of the signal, it may appear to be an unpromising basis for extracting frequency or pitch information. However, we do know that human listeners can achieve just that and some theories [1, p140; 5], do exist as speculations as to how this can be achieved. These theories involve computing the time intervals between spikes. One advantage of this technique is that the system is sensitive to amplitude modulation as well as simple frequencies thus reflecting the human sensitivity to the pitch of harmonic complexes (common in speech stimuli). It also provides a basis for understanding how harmonics can be grouped (on the basis of common time intervals) and subsequently isolated from competing harmonically structured sounds. This would be a second method for achieving some of the benefits of auditory selective attention, a problem which has proved most challenging to devices working in the frequency domain.

Time intervals between spikes (which occur most commonly at times associated with the peaks of the waveform) appear to share many features with zero-crossing measures. We know [6], however, that a great deal of frequency information is preserved in the zero-crossings of a bandpass signal, if the filter width is not too great. Because the mechanical properties of the basilar membrane ensure that individual fibres are responding to bandpass filtered signals, we have good grounds for optimism that spike trains will carry enough frequency information for our purposes.

There are difficulties to be overcome, however. For example, auditory nerve fibres have a very restricted dynamic range - often less than 30dB - and are easily saturated. It is also the case that the tight relationship between signal phase and spike timing has only been observed at low frequencies (less than 5 kHz). Spike activity also shows refractory effects whereby, loosely speaking, spikes closer than 1 msec apart are never observed. How the system generates human hearing achievements given these limitations, remains a matter of controversy and these problems will need to be directly addressed by anyone choosing to use these techniques in a recognition device.

## DESIGN CRITERIA

Considerable progress has been made in recent years in devising methods for generating spike trains from acoustic stimuli and the rest of this paper will be devoted to looking at some of the methods used as well as the criteria employed in choosing between them. No theory has yet been published which definitely meets even the current short list of criteria and some circumspection is called for. However, a number of the existing methods are already good enough for most of the purposes of ASR and we can start exploring immediately.

EXPLOITING PHYSIOLOGICAL CODES IN ASR

Nerve fibre response has a number of peculiar properties which need to be simulated by any successful spike generator embodying a model of the process:
(1) The simplest is the rate-intensity function given in Fig. 1.
(2) In the absence of any stimulus, the generator should produce intermittent activity with inter-spike intervals broadly consistent with a Poisson process with the exception of very short intervals which are limited by refractory effects.
(3) Because the maximum rate of fixing is often considerably less than the frequency of the stimulating tone, spikes cannot occur on every cycle but neither are the spikes equally spaced in time but follow only probablistic tendencies. However, if the response to a low frequency tone (say 1 kHz) is averaged over many cycles, a period histogram (Fig. 2) shows that the firing density reflects the amplitude of the positive half of the signal while showing no relationship to the negative half (i.e. the period histogram is a halfwave rectified version of the stimulus). At very low amplitudes, the halfwave rectification is not present and the period histogram is modulated across the complete cycle. This relationship between stimulus amplitude and period histogram modulation can be observed below the fibre's firing threshold, i.e. before the firing rate shows any sign of change.
(4) At high signal frequencies (>5 kHz) the period histogram shows no relationship to the fine structure of the stimulus.
(5) Stimuli with a rapid onset against a relatively quiet background produce a most characteristic onset firing peak which rapidly fades as responding settles down to an adapted level (Fig. 3). Detailed analysis [7] has shown that the early phase of this adaptation process is composed of at least two simultaneous exponential decay functions having time constants in the region of 3 msec and 40 msec respectively. This analysis has also shown that the saturation effect so characteristic of the rate intensity function for adapted responding (Fig. 1) is much less marked for onset responding (Fig. 4).
(6) Smith et al. [8] have also shown that the increase in firing which occurs when a stimulating tone is increased in intensity is quite independent of the state of adaptation of the fibre. This particular property is known as the principle of additivity. The converse, it appears, does not apply to decrements in stimulation intensity.
(7) Following stimulus offset, firing ceases briefly then returns gradually to spontaneous firing levels with a time constant in the region of 50 msec. Westerman [9] has, however, shown that the capacity of the system to respond to a new stimulus recovers at different rates for onset and post onset responding rates.

## SPIKE GENERATORS

With so may criteria to satisfy, it is not surprising that no system has yet been published and shown to achieve all the goals. However, those that do exist are certainly good enough for exploring their potential in a speech recognition context. Most of the models currently being developed are digital simulations of processes which might conceivably be taking

place in the inner hair cell in the organ of Corti. Because spike activity in auditory fibres is believed to be initiated by transmitter substance released from the hair cell into the junction with the fibre, it is this process which receives closest attention.

The models are variations on a general theme which is illustrated in Fig. 5. Transmitter substance arrives from a large global store or 'factory' and is delivered to the space adjacent to the cell membrane. Before arriving there it may be passed through one or more 'reservoirs'. The stimulus directly modifies the probability that transmitter in the immediate store will be released across the membrane into the cleft. This can either be achieved by changing the global permeability of the membrane, i.e. changing the rate of release or changing the number of sites on the membrane which release transmitter at a fixed rate. The amount of transmitter released into the cleft then influences the probability of a spike being generated in the attached auditory fibre. At this point the transmitter is either lost or is taken back into the cell and reutilised after reprocessing. References [10-15] give typical variations of this theme and give access to most of the relevant literature.

The digital simulation is in the form of differential equations which are evaluated, say, 20,000 times per second of simulated events. These equations are effectively updating the amount of transmitter in the various locations of the model (the boxes in Fig. 5) as it flows through or around the system. There is no reason why the models could not be handled as well or better using analogue devices. When models are perfected this may ultimately be the best method of using spike generators in speech recognition devices.

From a purely computational point of view, the fewer boxes the better because each box requires the evaluation of a differential equation for each epoch of the simulation. The Smith and Brachman model [10] is able to demonstrate the additivity principle but requires 500 immediate stores. Schwid and Geisler [12] have a compromise by using only six stores. Meddis [13, 15] gets very close using only one immediate store and a reuptake mechanism. Cooke's ingenious 'state-partition' model [14] simulates large numbers of immediate sites very economically using a small number of variables.

So far no comparative study has been attempted to evaluate the respective merits of these models. This would be an interesting exercise but probably academic from the point of view of the ASR community since they all generate spikes which are tolerably similar to those coming from a recording electrode. The first priority is to get them into service on an exploratory basis as 'plug in' modules which can be replaced later when a better model of auditory neural transduction is forthcoming.

REFERENCES

[1] B.C.J. Moore, 'An introduction to the psychology of hearing', Academic, (1982).
[2] A. Sekey and B.A. Hauson, 'Improved 1-Bark bandwidth auditory filter', J.A.S.A., Vol. 75, (1984).
[3] S.A. Shamma, 'Speech processing in the auditory system', J.A.S.A., Vol. 78, 1612-1632.
[4] J.A. Pickles, 'An introduction to the physiology of Hearing', Academic.
[5] J.C.R. Licklider, 'Three auditory theories', in 'Psychology: a study of a science', ed. S. Koch, (1959).
[6] B.F. Logan, Jr., 'Information in the zero-crossings of bandpass signals', Bell.Syst.Tech.J., Vol. 56, 487-510, (1977).
[7] L.A. Westerman and R.L. Smith, 'Rapid and short term adaptation in auditory nerve responses', Hearing Research, Vol. 15, 249-260, (1984).
[8] R.L. Smith, M.L. Brachman and R.D. Fusinia, 'Sensitivity of auditory-nerve fibres to changes in intensity', J.A.S.A., Vol. 78, 1310-1316, (1985).
[9] L.A. Westerman, 'Adaptation and recovery of auditory nerve responses', Ph.D. thesis Syracuse University (1985).
[10] R.L. Smith and M.L. Brachman, 'Adaptation in auditory nerve fibres: A revised model', Biol.Cybernet., Vol. 44, 107-120, (1982).
[11] S. Ross, A model of the hair cell - primary fibre complex', J.A.S.A., Vol. 71, 926-941, (1982).
[12] H.A. Schwid and C.D. Geisler, 'Multiple reservoir model of neuro-transmitter release', J.A.S.A., Vol. 72, 1435-1440, (1982).
[13] R. Meddis, 'Simulation of mechanical to neural transduction in the auditory receptor', J.A.S.A., Vol. 79, 702-711, (1986).
[14] M.P. Cooke, 'A computer model of peripheral auditory processing', National Physical Laboratory Report DITC 58/85 May 1985.
[15] R. Meddis, 'Simulation of auditory-neural transduction: further studies', submitted for publication.

FIG. 1 Rate-intensity function (schematic).

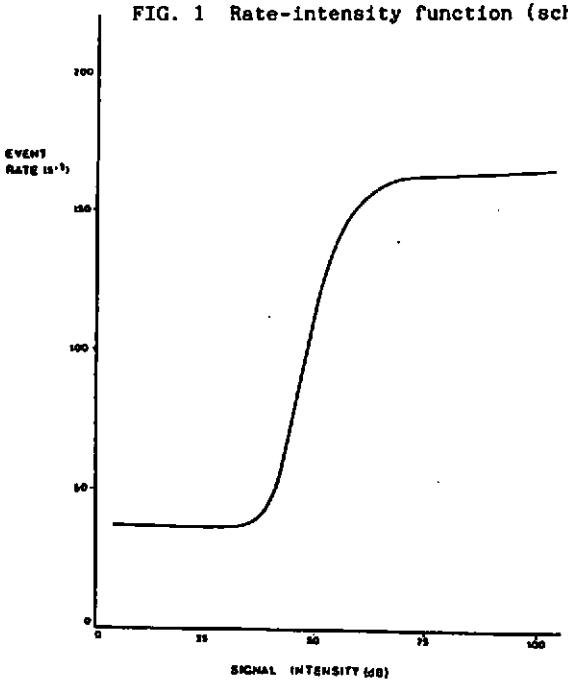



FIG. 2 Period histogram (schematic) for 1 kHz tone.

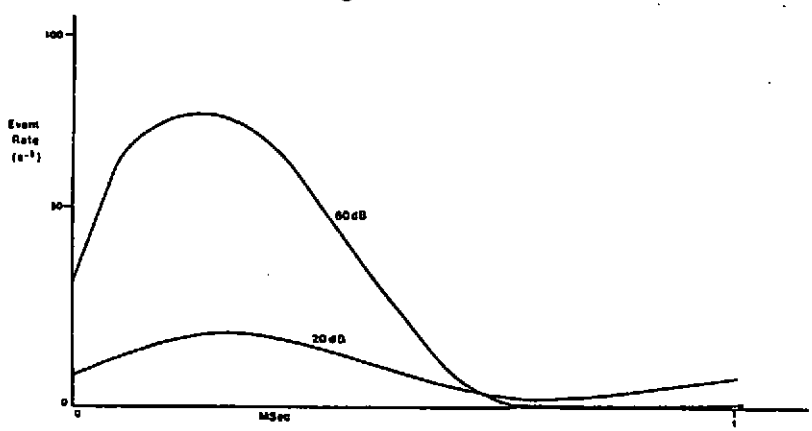


FIG. 3 Adaptation of response to a 300 msec tone burst.

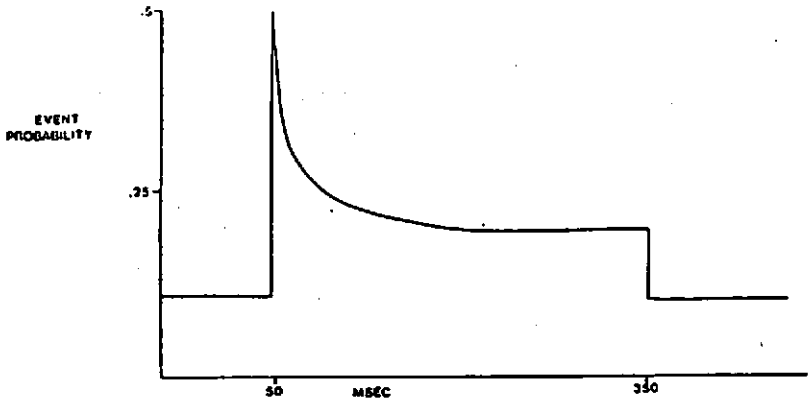
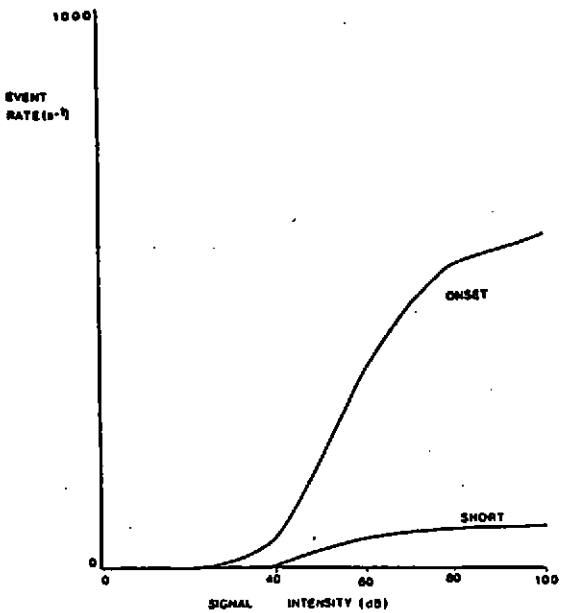



FIG. 4 Rate-intensity function for onset and adapted responding.



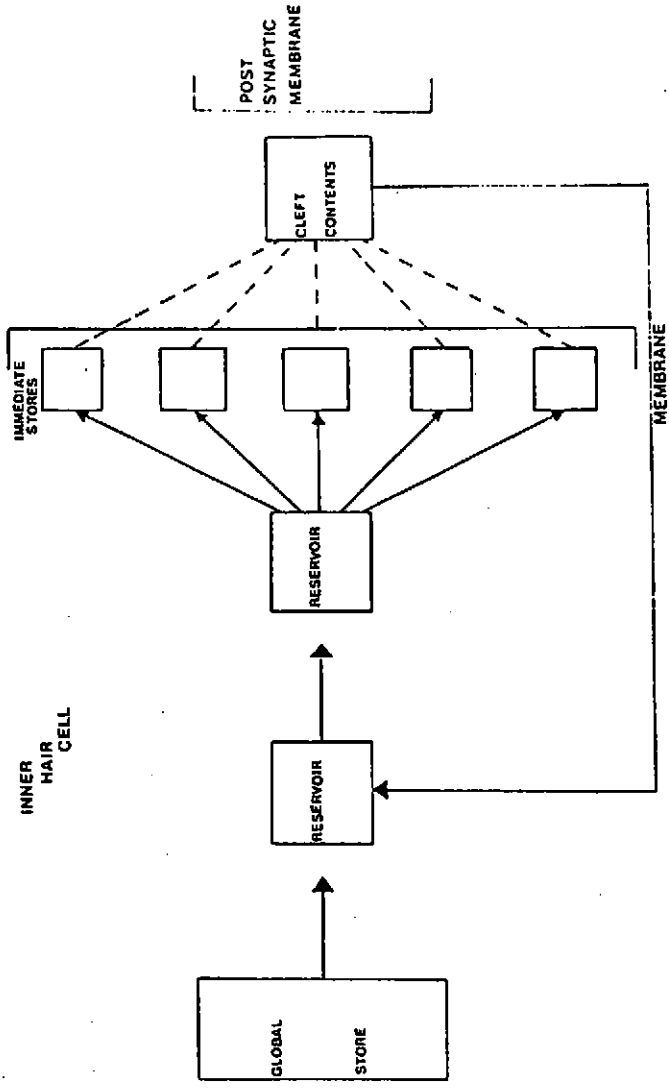



FIG. 5 Generalised representation of spike generating models.