Proceedings of the Institute of Acoustics

SPEECH RECOGNITION IN NOISE: COMPARING AN AUDITORY MODEL WITH HUMAN PERFORMANCE

W.A.Ainsworth and G.F.Meyer[1]

Department of Communication and Neuroscience, Keele University, Keele, Staffordshire ST5 5BG, U.K.

1. INTRODUCTION

It has long been known that background noise interferes with the reception of speech sounds causing perceptual confusions [1]. In order to maintain intelligibility with increasing masking noise the intensity of the speech must be increased [2].

One way of investigating the effects of noise is to examine physiological data. It has been found that the rate-level functions for tone bursts in background noise recorded in the cochlear nerve depend on the level and duration of the noise. Evans [3] reported two important effects of white noise on cochlear nerve responses: the threshold of the receptive fields increases but the spike discharge rate decreases with increasing noise levels. These effects, however, are not found if the noise is presented simultaneously with the test tones for short periods.

The question arises as to how these findings relate to the perception of speech sounds. This has been investigated by presenting speech sounds and noise simultaneously (burst noise) and by presenting speech in continuous noise to a number of listeners. It has also been investigated by incorporating these effects in a model of the auditory system and employing this model as the feature extraction stage of a speech recogniser.

2. PERCEPTION EXPERIMENTS

2.1 Stimuli

The same speech database was employed for both the perception experiments and the machine recognition experiments. It consisted of 300 speech samples, each 100ms of plosive-vowel excised from a longer vowel-plosive-vowel utterance at the plosive burst, spoken by a French speaker. The syllables were /b,d,g/ in combination with 10 French vowels. Three sets of data were produced: the original and two with noise added with signal-to-noise ratios of 6dB and 0dB. The noise was white noise low pass filtered with a first order Butterworth filter with a cut-off point of 500Hz and an equivalent bandwidth of 978Hz. This was chosen because it resembles the long term speech spectrum.

[1]Present Address: Department of Compter Science, Keele University, Keele, Staffs, ST5 5BG, U.K.

2.2 Listeners

Four listeners took part in the experiments. All but one were native French. The other (GFM) was German but his recognition scores nevertheless were slightly above average. The French listeners were paid. None had a history of hearing difficulty.

2.3 Burst noise experiments

Plosive-vowel utterances at the three noise conditions were played at about 70dB(A) in random order under computer control. The listeners were asked to identify the syllable they heard and press an appropriately labelled key. They were given immediate feedback showing the utterance played and whether the answer was correct. An introductory session of 300 utterances was used to familiarise the listeners with the procedure and stimuli.

2.4 Burst noise results

The results showed that recognition of short gated speech and noise bursts is very difficult. The average score was 84.0% for the speech alone, dropping to 47.2% at 6dB S/N and 27.0% at 0dB S/N (Figure 1). If the consonants and vowels are considered separately it can be seen that plosive recognition was much more affected by noise than vowel recognition. At 0dB S/N only 46.4% of the plosives (chance level 33.3%) were recognised correctly compared 50.6% of the vowels (chance level 10%). This is probably a consequence of the lower intensity of the plosives relative to that of the vowels.

2.5 Continuous noise experiments

The same speech samples were used as in the noise burst experiments. The noise was on continuously. It was generated by a hardware thermal noise source then low pass filtered with an analogue first order Butterworth filter with a cut-off of 500Hz. The noise spectrum was checked with a spectrum analyser to ensure that it was the same as the digitally produced noise used in the previous experiment. The speech and noise signals were added together to give a S/N ratio of 6dB or 0dB.

The listeners were asked to keep their headphones on for at least one minute before starting the experiment in order to give the auditory system time to adapt to the background noise. Whereas the noise levels varied randomly during the noise burst experiments, the noise level remained constant during each continuous noise experiment.

2.6 Continuous noise results

These recognition scores are also shown in Figure 1 where the continuous noise data is shown with speech alone data from the burst noise experiments. It can be seen that the scores increased by around 20% for both of the noise conditions. The plosive recognition score deteriorated linearly with increasing noise level while the vowel recognition score was similar to the speech alone score at 6dB but reduced sharply at 0dB.

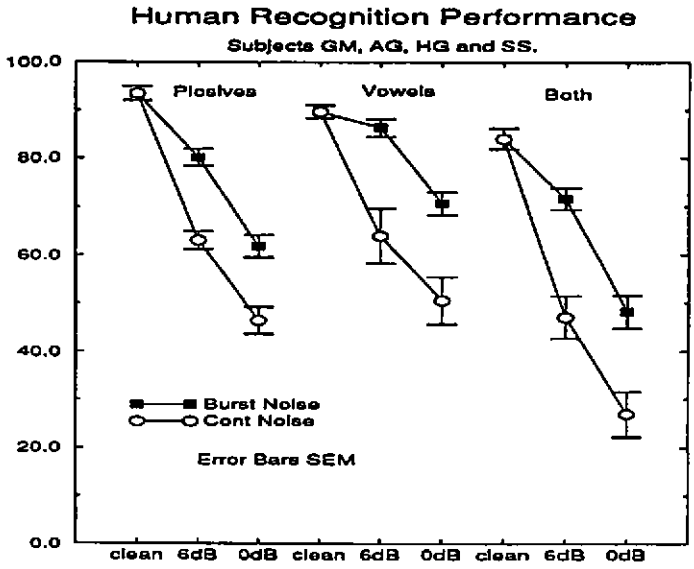

Figure 1: Perception of plosives, vowels and syllables by human listeners in burst noise and in continuous noise

In order to ensure that the improved performance with the continuous noise experiments was not due to learning effects one of the listeners was asked to perform the continuous noise experiments first. The score for the continuous noise experiments was slightly higher than the average of the other listeners, indicating that the results were not due to learning effects.

2.7 Discussion

When the cochlear nerve is stimulated with continuous noise, adaptation of the spike rate is observed. This does not occur with noise bursts as short as 200ms [4]. This adaptation is likely to enhance the coding of spectral information in the cochlear nerve. Cochlear nerve fibres have limited dynamic ranges so synchronous noise plus speech could easily drive them into saturation.

In continuous noise the threshold of the fibre will be raised and its spike rate lowered effectively leaving a window in the dynamic range which can code the speech. This mechanism could be responsible for the improved recognition performance with continuous noise.

3. AUTOMATIC SPEECH RECOGNITION EXPERIMENTS

3.1 Auditory model

A physiologically-based computational model of the cochlear nerve and cochlear nucleus has been developed previously [5,6]. The first stage is an array of filters based on simulated reverse correlation functions as measured in the cochlear nerve [7,8,9]. The output of these filters is then processed by a stage adding a human audiogram and an absolute hearing threshold across the array of channels. The resulting array of signals is then processed by a model of hair cell transduction [10,11] and finally spikes are generated by a probabilistic model of neural firing.

The resulting spike trains are processed by tonotopically organised arrays of models of cochlear nucleus neurones.

Recent work [12] has shown that cochlear nerve fibres have varying thresholds and dynamic ranges. The model takes account of this by generating two populations of cochlear nerve responses: low threshold fibres with thresholds matched to behavioural thresholds, high spontaneous firing rates and a small dynamic range of 30dB and high threshold fibres with thresholds 15dB higher, low spontaneous firing rates and a dynamic range of over 60dB.

3.2 Hidden Markov model

In order to investigate the speech recognition capabilities of the auditory model the output was interfaced to a hidden Markov model recogniser [13]. In all four recognisers were evaluated using the outputs of (a) low threshold auditory nerve units, (b) high threshold cochlear nerve units, (c) cochlear nucleus sustained chopper units, and (d) cochlear nucleus transient chopper units.

The signals used for the training and testing were the 300 plosive-vowel utterances of 100ms duration described earlier. They were scaled to give an average amplitude of 60dB above the threshold at 1kHz for the speech alone and noise was added at an appropriate level to give 6dB or 0dB signal-to-noise ratios. The HMMs were trained on 180 utterances (10 repetitions of each plosive in combination with the vowels /a,e,i,o,ɛ,ɔ/) and tested on the remainder. All results given are for the test set only.

3.3 Low threshold cochlear nerve fibres

3.3.1 Burst noise experiments

As the cochlear nerve shows no adaptation to burst noise, using an array of cochlear nerve fibres is a good approximation to human processing. Most of the low threshold units are operating near the top of their dynamic ranges with the speech alone signals so additive noise drives the units into saturation. It was therefore predicted that the recognition scores would decrease.

Figure 2 shows that this was the case. It can be seen that the results are similar to those obtained with human listeners. The model outperforms human listeners by around 5% for the vowel part while human listeners perform better by a similar amount for the plosive component.

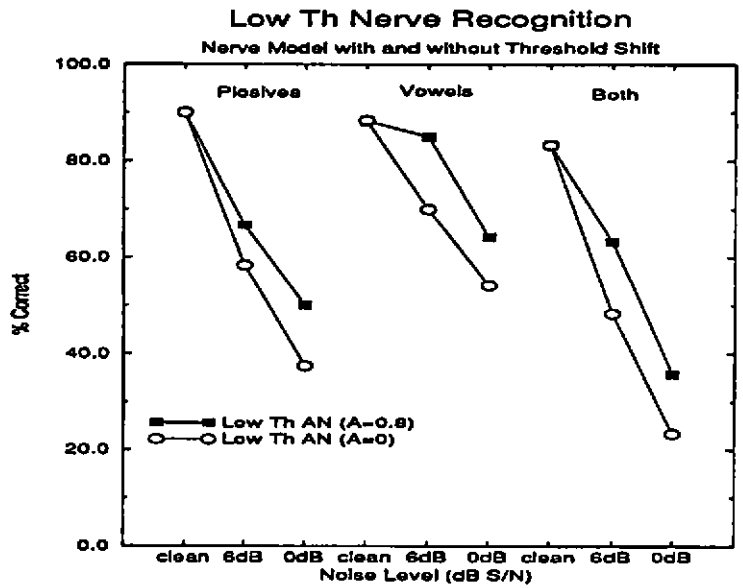

Figure 2: Recognition scores with low threshold cochlear nerve units in burst noise and in continuous noise

3.3.2 Continuous noise experiments

Gibson et al [4] performed physiological experiments similar to those of Evans [3] but recorded from both the cochlear nerve and cochlear nucleus. Their results can be expressed as a shift in saturation intensity ΔI_{sat} given by:

$$\Delta I_{sat} = 0 \quad I < \theta$$

$$\Delta I_{sat} = A(I - \theta) \quad I \geq \theta$$

where I is the background intensity, θ is the intensity at which the shift first occurs and A is the ratio of the shift relative to θ. Gibson et al found a value of A for noise backgrounds of about 0.8 for both the cochlear nerve and the ventral cochlear nucleus. They concluded that the mechanism responsible for the shift is located in or before the cochlear nerve.

In order to simulate these effects the dynamic range of the cochlear nerve units in the model were shifted according to the above formula with a value of A of 0.8 and θ of 20dB. No attempt was made to model the time course of adaptation. It was assumed that the model was fully adapted to the particular noise level. The same signals were used as before.

The average recognition score increased markedly when the dynamic range was shifted with the noise (Figure 2). The model predicts human continuous noise background per-

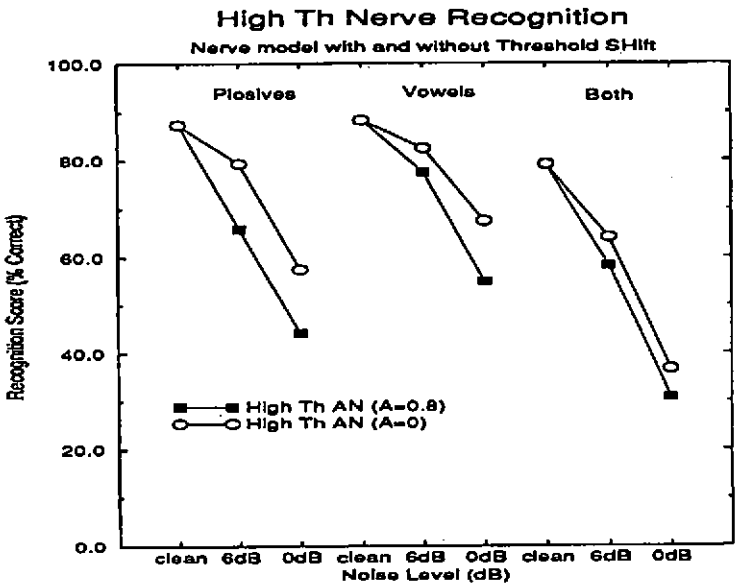

Figure 3: Recognition scores with high threshold cochlear nerve units in burst noise and in continuous noise

formance well for vowels but the model's score is consistently lower for plosives. As with the human experiments the vowel recognition score is largely unaffected by the 6dB S/N condition.

3.4 High threshold cochlear nerve fibres

High threshold fibres have dynamic ranges of about 60dB and their thresholds are about 15dB above those of low threshold fibres. Consequently these cover the whole amplitude range of the sounds used in the experiments. High threshold fibre models simulate well the behaviour of humans for continuous noise. The scores for vowels are similar to those of adapted low threshold fibres but for plosives they were found to be higher, approaching those of human listeners (Figure 3).

When the dynamic range of high threshold fibres is shifted up to simulate the continuous noise condition the performance decreases rather than increases. Again it is the plosive recognition performance which mainly deteriorates. With the raised thresholds it is likely that some of the plosives are too weak to excite these fibres.

3.5 Cochlear nucleus units

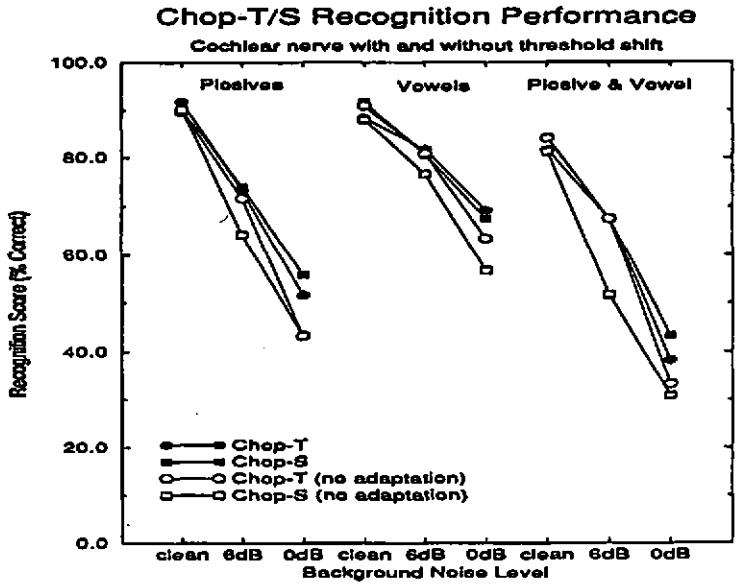

Figure 4: Recognition scores with sustained and transient chopper units in burst noise and in continuous noise

It is believed that both transient and sustained chopper units receive a convergent input from both high and low threshold cochlear nerve units, enabling spectral cues to be coded over a wide intensity range. Transient choppers are better in this respect than sustained choppers [14]. If dynamic range is the limitation for speech recognition in noise, a good performance should be expected from chopper units.

For the continuous noise simulations the low threshold cochlear nerve units by the above formula with A=0.8 and θ=20dB as before but the high threshold units employed A=0.8 and θ=30dB.

The sustained and transient chopper units behaved similarly in all respects (Figure 4). Units connected to adapted cochlear nerve fibres (continuous noise condition) produced higher scores than the unadapted (burst noise condition). Overall the chopper models outperformed the cochlear nerve fibres and reached levels similar to human performance.

4. CONCLUSIONS

Experiments have been performed to determine the effects of noise coincident with speech sounds (burst noise) and continuous noise on the perception of plosive-vowel syllables by human listeners. It was found that while performance decreased with increasing noise levels in

both conditions this deterioration was less with continuous background noise. This suggests some form of adaptation to the noise.

In order to investigate possible mechanisms underlying this phenomena an auditory model was developed which incorporated response threshold shifts in cochlear nerve fibres which have been observed in physiological experiments. This auditory model, which contained simulations of units found in both the cochlear nerve and cochlear nucleus, was interfaced to an HMM recogniser and used to process the same sounds that were used in the perceptual experiments. It was found that the recognition scores obtained with the auditory model recogniser showed the same trends as were found with the human perception experiments.

5. ACKNOWLEDGEMENTS

This work was supported by Contract SCI.0044.C(H) under the EC Science programme.

6. REFERENCES

[1] G A MILLER & P E NICELY, 'An Analysis of Some Perceptual Confusions Among Some English Consonants',*J.Acoust.Soc.Am.*,27 p338-352 (1955)

[2] H FLETCHER, 'Speech and Hearing in Communication', *Van Nostrand, New York* (1953).

[3] E F Evans, 'Auditory Frequency Selectivity and the Cochlear Nerve', in Zwicker and Terhard: Facts and Models of Hearing, p 118-132 (1975)

[4] D J GIBSON, E D YOUNG & J A COSTALUPES, 'Similarity of Dynamic Range Adjustment in Auditory Nerve and Cochlear Nuclei', *J Neurophys* 63 No.4 p940-958 (1985)

[5] G F MEYER & W A AINSWORTH, 'Modelling Response Patterns in the Cochlear Nucleus Using Simple Units', *Advances in Speech, Hearing and Language Processing*, 3 JAI Press (in press)

[6] W A AINSWORTH & G F MEYER, 'Speech Analysis by Means of a Physiologically-Based Model the Cochlear Nerve and Cochlear Nucleus', *Visual Representations of Speech Analysis, ED M Cooke and S Beet*, Wiley p1-6 (in press)

[7] E de BOER, 'Reverse Correlation II: Initialisation of Nerve Impulses in the Inner Ear", *Proc Kon Ned Acad Wet*, 72 p129-151 (1969)

[8] E F EVANS 'Cochlear Nerve and Cochlear Nucleus', *Handbook of Sensory Physiology* 5(2) (1975)

[9] L J CARNEY & T C T YIN, 'Temporal Coding of Resonances By Low-Frequency Auditory Nerve Fibres: Single-Fiber Responses and a Population Model', *J Neurophys*, 60 p1653-1677 (1988)

[10] R MEDDIS, 'Simulation of Mechanical to Neural Transduction in the Auditory Receptor', *J Acoust Soc Am* 79 (3) p702-711 (1986)

[11] R MEDDIS, 'Simulation of the Auditory-Neural Transduction: Further Studies', *J Acoust Soc Am* 83 1056-1063 (1988).

[12] G K Yates, I M WINTER & D ROBERTSON, 'Basilar Membrane Nonlinearity Determines Auditory Nerve Rate-Intensity Functions and Cochlear Dynamic Range', *Hearing Res*, 45 p203-220 (1990)

[13] S YOUNG 'HTK: Hidden Markov Toolkit V1.2' *Speech Group*, CUED (1990)

[14] C.C. BLACKBURN & M.B. SACHS, 'The Representation of the Steady-State Vowel Sound ɛ in the Discharge Patterns of Cat Anteroventral Cochlear Nucleus Neurons', *J. Neurophys* 63 p1191-1212 (1990)