Proceedings of the Institute of Acoustics

THE PROCESSING OF PLOSIVES BY MODELS OF CELLS IN THE COCHLEAR NUCLEUS

G.F. Meyer(1), A.C. Morris(2), W.A. Ainsworth(1) and J.-L. Schwartz(2)

(1) Dept of Communication and Neuroscience, University of Keele, Keele, Staffs., ST5 5BJ, England.

(2) Institut de la Communication Parlee, ENSERG-INPG, 46 Ave Felix Viallet, 38031 Grenoble, France.

1. INTRODUCTION

An attractive model of human speech processing is a multistage system. The first stage extracts acoustical features from the signal representation in the auditory nerve, the second stage maps this acoustical representation into symbolic units which may then be mapped into word or semantic units.

We are particularly interested in the features the cochlear nucleus, the first stage in the auditory system, can extract from speech sounds. The responses of different cochlear nucleus neurons to pure tones, modulated and transient sounds show that the nucleus is particularly well suited for extracting temporal features, fig 1.

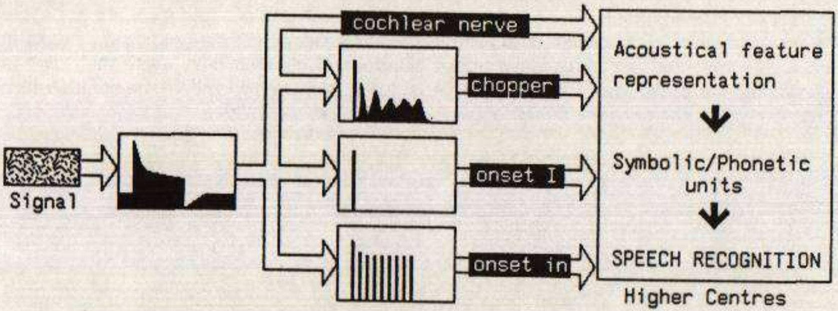

Fig. 1: Diagrammatic view of signal processing in cochlear nerve and nucleus with idealised temporal response patterns to pure tone stimuli.

This paper addresses two main problems:

What information is extracted at the level of the cochlear nucleus?

We will discuss the acoustical features which three types of physiologically plausible units extract from the signal representation in the cochlear nerve. The model has been discussed in detail elsewhere [1,2], so that only the features critically important to the encoding of speech in the model units will be mentioned here.

A first unit type responds with regular spike trains (chopper response) whenever the cochlear nerves feeding it are activated.

The other units discussed here are two types of onset units: one marking tone onsets, another encoding the fundamental frequency in speech.

How can these acoustical features be mapped into a phonetic representation?

We discuss some exploratory work in combining streams of phasic and tonic features extracted by the cochlear nucleus for the recognition of voiced plosives. A window of spectral information is read whenever a tone onset is detected and mapped into a phonetic representation using a statistical recognition model. The results are compared with those derived from more traditional preprocessing techniques.

All sounds are processed by tonotopically organised arrays of cells in the cochlear nucleus which receive input from cochlear nerve fibres with a limited characteristic frequency range.
In the context of plosives this means that, for instance, tone onset units encode 'plosive burst in frequency space' information which can be used as a cue to categorise plosive sounds [2]. While this cue is particularly important for the discrimination of plosives it is well known that listeners use a range of cues such as the burst spectrum, formant transitions and VOT for the categorisation of plosives [3,12,13].
All recognition experiments are based on a spectral window following the plosive burst alone.

2. CHOPPER UNITS - DETECTING SPEECH PRESENCE

A large proportion of units in the cochlear nucleus (48%) respond with regular discharges to any stimulus in their receptive field [4]. The discharge frequency is largely unrelated to the stimulus frequency, so that the neurons (Stellate cells) are unlikely to encode temporal information. With dynamic ranges of typically less than 25dB the stimulus amplitude, too, is only poorly encoded [4]. The receptive fields of chopper neurons are similar to those of cochlear nerve fibres, so that the frequency representation is relatively good [5]. The units encode *signal presence* in a narrow frequency range.

When speech is put through an array of chopper units, responses are evoked in channels which receive input where the intensity exceeds 30dB. In contrast to cochlear nerve fibres chopper units are not spontaneously active so that a very clear picture of activity in the cochlear nerve emerges, fig 2. Peaks in the spectrogram are represented in the response tracks of arrays of chopper units. Note that there is relatively little activation in the low frequency range. This is due to the cochlear nerve thresholds which are set in accordance with human hearing thresholds.

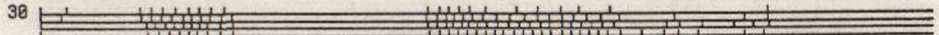

Fig. 2: Chopper unit response to /adi/ and /aki/ spoken by a male English speaker.

3. ONSET UNITS - DETECTING EVENTS IN SPEECH

Models of two types of onset units are described here. Both units produce onset responses by very different mechanisms and are based on different cell types. The first unit is tuned to respond preferably to tone onsets. The second onset type is tuned to encode the fundamental frequency.

3.1 ONSET $_I$ - TONE ONSET DETECTION

The units which we use as tone onset detectors are based on the physiology and anatomy of octopus cells in the PVCN (for a functional model of onset detection, see [6]). The model is based on a mechanism suppressing action potentials which is inherent in the action potential generator. After each action potential the intracellular potential has to fall to close to the cell's resting potential to re-enable the spike generator [7]. This means that for continuous excitation of the units spikes are reliably suppressed after the onset spike. The parameters of the model are adjusted to model a unit which does not phase lock into click frequencies exceeding 100Hz. This value seems very low when compared with physiological data from bats [8], or rats [9], but neither of the two animals would benefit from an ability to resolve such low frequencies.

Speech in arrays of onset I units is shown in fig 3. Each tone onset causes a spike in most of the channels. Here the following events are encoded:

-A- The initial vowel, /a/

-B- The plosive release, /g/ in trace 1 and /k/ in trace 2.

-C- The second vowel onset in trace 2, the time difference between the two onset positions is the voice onset time (VOT).

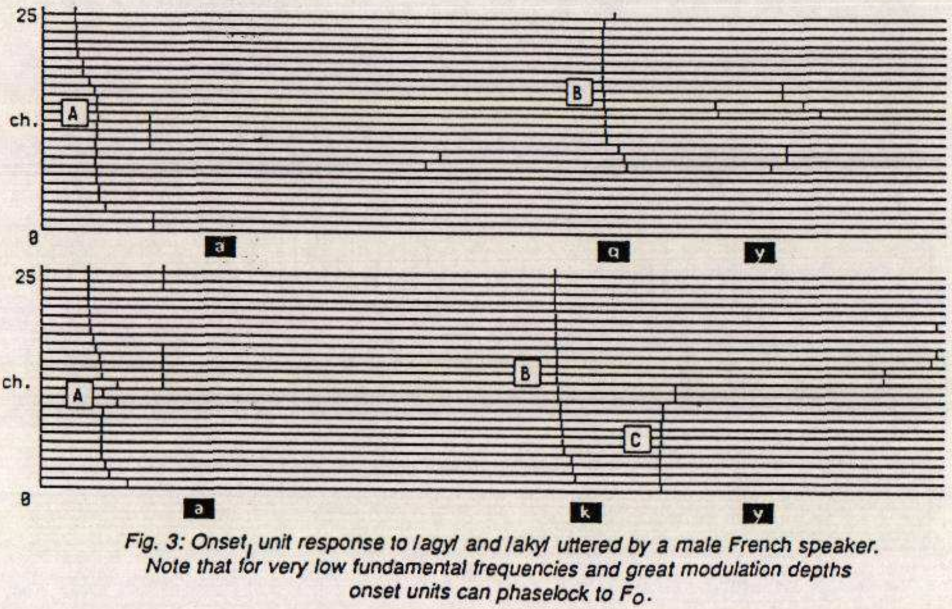

Fig. 3: Onset$_I$ unit response to /agy/ and /aky/ uttered by a male French speaker. Note that for very low fundamental frequencies and great modulation depths onset units can phaselock to F_O.

Plosive and vowel onsets are not distinguishable if nothing is known about the spectrum following the onset. Cues such as formant transitions (frequency shifts) are not encoded systematically at the level of the cochlear nucleus, although it is known that some cells in the auditory cortex react exclusively to frequency modulated signals [10].

3.2 ONSET$_{In}$ - PITCH ENCODING

Onset$_{in}$ units occur throughout the cochlear nucleus but the physiological response pattern has not yet been correlated to any particular cell type, even though evidence exists that at least one of the units producing onset responses are Giant cells in the DCN (Zhao, pers comm). The onset response is generated by the interaction of excitation and delayed tone evoked inhibition.

The units act as coincidence detectors, that is a number of presynaptic spikes have to occur within a small time window to elicit an action potential. The effect of this is a suppression of spontaneous activity and an enhancement of the pitch encoding [11].

When onset$_{in}$ units are stimulated with repetitive clicks, phase locking can be observed up to 400Hz, but for higher frequencies only the tone onset is encoded. In terms of average response rate this unit type can best be described as a band-pass filter with a characteristic frequency of 200Hz.

When speech is processed by an array of units a good representation of the fundamental frequency emerges, fig. 4. The bottom trace is the summed output of all channels. The channels have been 'dephased' [11], that is delays due to basilar membrane delays which occur throughout the low level auditory system have been subtracted out.

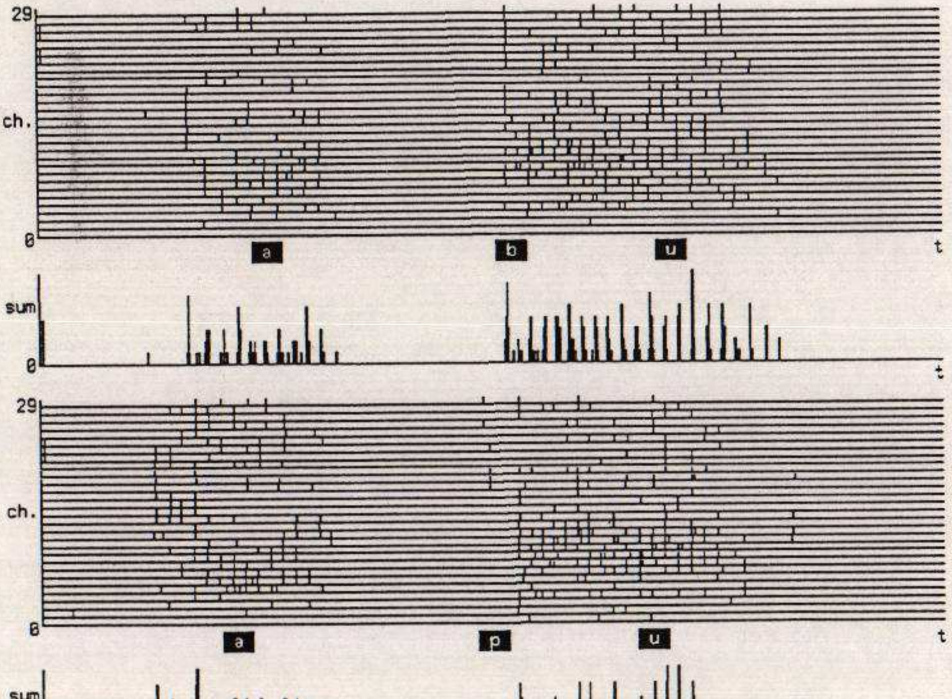

Fig. 4: Onset$_{in}$ unit response to /abu/ and /apu/ uttered by a male English speaker.

So far the examples were chosen to cover all plosives and as many final position vowels as possible. Fig 5. shows a full set of traces for a two utterances /ada/ and /ata/.

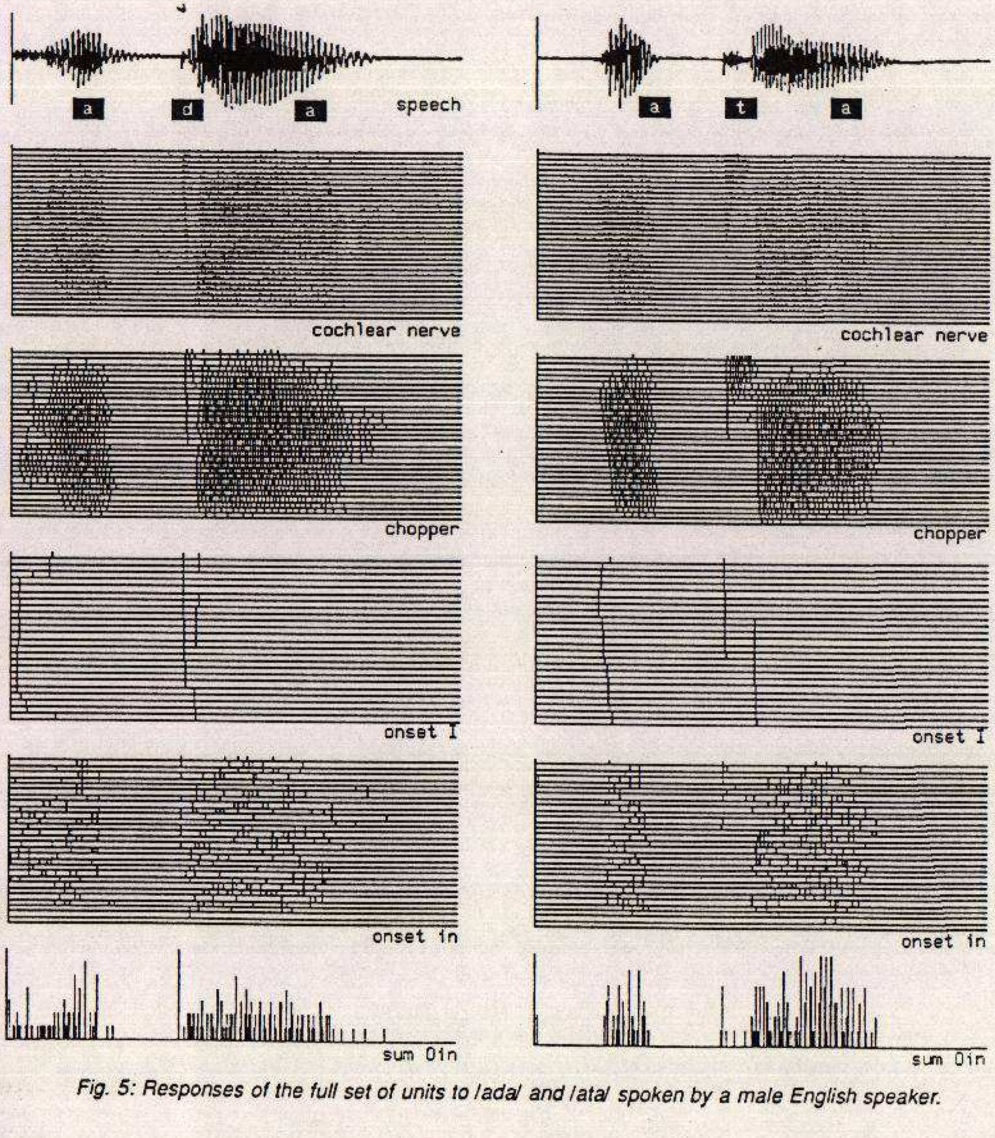

Fig. 5: Responses of the full set of units to /ada/ and /ata/ spoken by a male English speaker.

4. PLOSIVE RECOGNITION BASED ON OUTPUT FROM THE COCHLEAR NUCLEUS

It is known from psychoacoustic experimentation [12,13], that speaker and vowel- context independent cues sufficient for distinguishing all syllable initial stop consonants are contained in the gross form of the burst release spectrum and the rate and duration of the subsequent formant transitions over 20-40ms following burst release.

A possible hypothesis is that onset_I cells are used to focus the attention of higher centres on those parts of the signal which contain concentrations of information relevant to plosive identification.

In order to test this hypothesis we have used the output from a model of onset detection units [2] to trigger the extraction of a window of spectral domain information from each of a set of vowel-plosive-vowel signals. 3/4 of this set of windows was then used to train a statistical pattern recognition model, while the remaining 1/4 was used to test recognition performance.

Results were produced for spectral domain information obtained from three different sources:

1. directly from the cochlear nerve model, which is equivalent to output from CN 'primary-like' cells;
2. the cochlear nerve model, followed by CN 'chopper' cell model;
3. from 'psy' coefficients produced using classical speech processing procedures [14] (preemphasis, FT, Bark-scaling, log-compression);

In one further experiment results were obtained from the 'psy' model using spectra from the preceding vowel offset as well as from the plosive burst position [13,14]. Information concerning syllable final stop consonants is strongly concentrated at the voicing termination (off position). There is no evidence for units in the cochlear nucleus encoding tone offsets, so that tone offset positions have not been used in experiments where speech has been preprocessed in cochlear nucleus models.

4.1 THE SPEECH CORPUS

The training corpus consists of 72 logatomes, 24 examples of each plosive:
/a/, /b,d,g/, /a,i,u,y/ from six male French speakers
The test corpus consists of 24 logatomes, 8 examples of each plosive:
/i,o/, /b,d,g/, /a,i,u,y/ from one of these speakers.
The four final vowels [a,i,u,y] in the French set were selected to have maximally different formant characteristics so that recognition would be based only on cues which are independent of vowel context.

4.2 THE RECOGNITION MODEL

Each of a fixed number N_s of spectra following plosion onset is concatenated into a single pattern vector, with dimension N.

The recognition model [14], is trained by estimating for each plosive (i) a mean vector $M(i)$ and covariance matrix $V(i)$ over each example in the training set for this plosive. These provide estimates for parameters which completely specify the probability density function for the multivariate Gaussian distribution $G(i)$ most likely to have given rise to these observed patterns.

Recognition is then performed on each example in the test set by finding the probability that the pattern vector for this example has arisen from each of these distributions, and choosing the identification which has the maximum probability. Probability that X is from $G(i)$ given that X is from $G(b)$, $G(d)$ or $G(g)$

$$prob = \frac{G(i,X)}{G(b,X)+G(d,X)+G(g,X)}$$

where probability density at X in distribution $G(i)$ is

$$G(i,X) = \frac{\exp\left[-\frac{1}{2}(X-M(i))'.W(i).(X-M(i))\right]}{\sqrt{(2\pi)^N.det(V)}}$$

where W is the inverse of V.

Note that since we have only 24 training examples for each plosive, and V is not invertible when N is greater than the number of training examples (which was the case here whenever more that one spectrum was taken together as a pattern vector), it was necessary for $N_s>2$ to alter the model in some way so that V would become invertible. This was done by setting all covariances to zero. If the number of training examples had been sufficiently large that this was not necessary, then covariance information would have been retained and the model should have been expected to perform better.

5. RECOGNITION RESULTS

Confusion matrices for experiments 1 to 4 are given below. These show the average probabilities for phoneme (row) being classified as phoneme (col), together with identification-counts and misidentifications.

cochlear nerve model 8 windows (5ms), no overlap, total 40ms							
	b	d	g	b	d	g	misident.
b	0.50	0.38	0.12	4	3	1	ibi iby oby obi
d	0	0.70	0.30	0	6	2	odi ody
g	0.13	0.38	0.50	1	3	4	igu iga oga ogi

cochlear nerve model → chopper units, 8 windows (5ms), no overlap, total 40ms							
	b	d	g	b	d	g	misident.
b	0.50	0.44	0.07	4	4	0	ibi iby oby obi
d	0	0.88	0.12	0	8	0	odi ody
g	0.09	0.25	0.65	1	2	5	igu iga igy

'psy' model, 7 windows (12.8ms), 6.4ms overlap, total 44.8ms							
	b	d	g	b	d	g	misident.
b	0.62	0.25	0.13	5	2	1	ibi oby obi
d	0.01	0.87	0.12	0	7	1	odi
g	0	0.38	0.62	0	3	5	iga oga ogi

'psy' model, 2 windows (25.6ms) at off position 7 windows at onset pos. total 38.4+102.4ms							
	b	d	g	b	d	g	misident.
b	0.70	0.21	0.09	7	1	0	iba
d	0.20	0.73	0.06	0	8	0	
g	0.14	0.38	0.68	1	0	7	oga

The first two tables show the recognition performance when data has been preprocessed using the physiologically plausible model, while the third and fourth table are produced with more traditional preprocessing models. The performance of the two models is roughly equivalent for similar window numbers and sizes. Note that when the tone onset as well as the tone offset position is used the number of correctly 'guessed' plosives rises to 92% while the number 'expected' correct is still only 70%.

6. CONCLUSION

The results reported here are encouraging insofar as they show that simple models of neurons in the cochlear nucleus can extract a range of acoustical features from speech sounds. The recognition results show an improvement when cochlear nucleus units as well as, rather than just cochlear nerve fibres to encode speech signals. Recognition performances for both cochlear nucleus and traditional preprocessing techniques are equivalent when similar window locations and sizes are used.

The "psy" model above obtained 100% recognition for voiced plosives when trained and tested on examples from a single speaker [14]. It is clear that some form of speaker normalisation is required for multi-speaker plosive recognition, although we believe that a larger training set would significantly enhance all of the above results.

7. ACKNOWLEDGEMENTS

The work presented in this paper was supported by European Community Science contract SC1.0044.c(H). Thanks to Neil Blackwood for providing the cochlear nerve model. We are also grateful to Ted Evans, Carole Hackney, John Holdsworth and Frederique Berthommier for valuable discussions.

REFERENCES

[1] G.F. Meyer and W.A. Ainsworth, "Modelling Response Patterns in the Cochlear Nucleus Using Simple Units", *Advances in Speech, Hearing and Language Processing*, Vol. 3, JAI Press (in press).

[2] N. Blackwood, G. Meyer and W.A. Ainsworth, 'A Model of the Processing of Voiced Plosives in the Auditory Nerve and Cochlear Nucleus', *Proc. I.O.A.* V12(10) p423-430 (1990).

[3] W.A. Ainsworth, 'Perception of Stop Consonants in Synthetic CV Syllables', *Language and Speech*, V11(3) p139-155 (1968).

[4] W.S. Rhode and D.H. Smith, 'Encoding Timing and Intensity in the Ventral Cochlear Nucleus of the cat', *J. Neurophys* V56 p261-286 (1986).

[5] E.F. Evans, 'Cochlear Nerve and Cochlear Nucleus', *Handbook of Sensory Physiology, V5(2), Springer Verlag, Berlin* (1975).

[6] Z.L. Wu, J.-L. Schwartz and P. Escudier, 'Physiologically Plausible Modules for the detection of Articulatory-Acoustic Events.', *Advances in Speech, Hearing and Language Processing*, Vol. 3, JAI Press (in press).

[7] G.M. Shepherd, 'Neurobiology', *Oxford Univ. Press, 2nd ed.* p111 (1988).

[8] M. Vater, 'Single Unit Responses in the Cochlear Nucleus of Horseshoe Bats to sinusoidal frequency and amplitude modulated signals.', *J. Comp. Physiol. A*, V149 p369-388 (1982).

[9] A.R. Moller, 'Unit Responses in the Rat Cochlear Nucleus to Repetitive, Transient Sounds', *Acta physiol, scand.* V75 p542-551 (1969).

[10] E.F. Evans and I.C. Whitfield, 'Classification of unit Responses in the cortex of the unanaesthetised and unrestrained cat', *J.Physiol.(Lond)* V171 p476-493 (1964).

[11] F. Berthommier, 'Reseaux de neurones et traitement des signaux dans le systeme auditif' *Rapport Technique, Institut IMAG TIM3*, RT62 (1990).

[12] P. Lieberman and S.E. Blumstein, 'Speech physiology, speech perception, and acoustic phonetics', *Cambridge Studies in Speech Science and Communication*, p224-225 (1988).

[13] K.N. Stevens and S.E. Blumstein, 'Invariant cues for place of articulation in stop consonants", *J.Acoust.Soc.Am.*, V64, p1358-1368 (1979)

[14] A.C. Morris and J-L. Schwartz, 'Internal technical report for EC project SC1.0044.C(H) on the Cochlear nucleus', p10-26 (1990)