APPLICATION OF AN AUDITORY MODEL TO THE COMPUTER SIMULATION OF HEARING IMPAIRMENT: PRELIMINARY RESULTS

C. Giguère, P.C. Woodland and A.J. Robinson

Cambridge University Engineering Department
Trumpington Street, Cambridge CB2 1PZ, England

1. INTRODUCTION

A new computational model of the auditory periphery has been recently reported by the first two authors [1, 2]. The model is particularly attractive for hearing research in that it enables practical simulation of auditory nonlinearities and feedback mechanisms, and the study of the deterioration of these processes in the hearing impaired.

In this article, we present the results of an exploratory study designed to simulate the perceptual consequences associated with cochlear hearing loss. For this purpose, we combine the auditory model to a state-of-the-art recurrent neural network classifier [5] to form an auditory-based automatic speech recognition system. Preliminary speech recognition results for normal and impaired operations of the auditory model are then compared.

The ultimate goal of this work would be to guide the development of signal processing strategies for hearing aids. The idea is to find new ways of processing input speech so that, when passed through a model of impaired cochlea, the observed auditory nerve firing patterns and/or speech recognition scores would be as close as possible to those observed for unprocessed speech passed through a normal cochlea.

2. THE AUDITORY MODEL

The structure of the model closely follows the general architecture of the peripheral ear as shown in Fig. 1.

The different stages of the ascending path can be equivalently represented as lumped-element analog circuits or as wave digital filters (WDFs) [2]. The input $P(t)$ is digitized speech or other incident acoustic wave. This signal is processed by a WDF module representing the sound transformation through the outer ear, middle ear and cochlea. The middle ear stage includes a time-variant capacitive element $C_{st}(t)$ modelling the variable acoustic compliance of the stapes suspension in response to stapedial muscle contractions. The cochlear stage is based on the classical 1-D transmission line model of basilar membrane (BM) motion, extended to account for the mechanical effects of the outer hair cells (OHCs). By injecting energy in phase with BM velocity at low levels, the OHC circuit leads to auditory filters with level-dependent frequency selectivity and sensitivity. At low input levels, the BM is sharply tuned and highly sensitive. At high input levels, the BM is broadly tuned and the characteristic frequency shifts by about half an octave to a lower frequency. Over the full input range, the BM shows 31 dB of dynamic compression near the characteristic frequency. These properties are in broad agreement with physiological observations [3].

The ascending path is completed by the inner hair cell (IHC) transduction model of Meddis [4], implemented as a wave digital filter. There are N WDF inner hair cell modules, one per BM segment, using the parameters of a medium-rate fibre. The input $s_n(t)$ to the IHC indexed n is assumed to be proportional to the velocity $i_n(t)$ of the BM segment n to which it is paired (Fig. 1). The fluid-cilia coupling gain

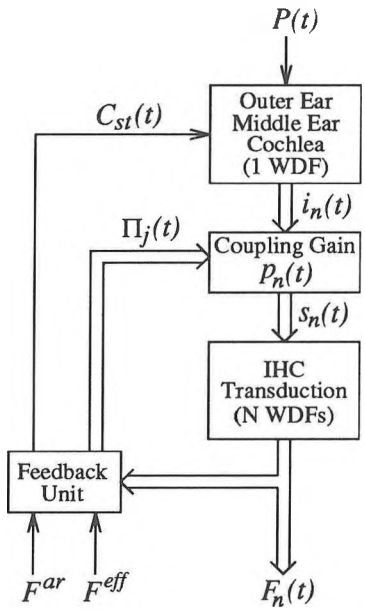

Figure 1. Block diagram of the auditory model.

$p_n(t)$ is made time and space variant as discussed below. The model output $F_n(t)$ is the instantaneous firing rate of the N tonotopically-arrayed IHC afferent fibres.

The model also includes a simple feedback unit simulating the dynamics of the descending paths to the peripheral ear. The acoustic reflex is assumed to be a regulation system whose goal is to maintain the average firing rate to a constant target rate of F^{ar}. The control function is a slow modulation of $C_{st}(t)$, leading to a decrease in middle-ear transmission by up to 15 dB below 1000 Hz. The OHC efferent system is also assumed to be a firing rate regulation system. The control function is taken as a slow modulation of the coupling gain $p_n(t)$. The N fibres at the output of the IHC stage are grouped into J contiguous bands, and regulation is applied independently in each band with a target rate of F^{eff}. The gain control command $\Pi_j(t)$ is then spatially interpolated to yield $p_n(t)$, leading to an inhibition of cochlear output equivalent to a reduction in input level by up to 24 dB.

The auditory model was applied to the analysis of speech data by computing auditory nerve cochleograms [1]. The OHC circuit provides level compression and spectral sharpening in the low energy portions of an utterance. This results in a better resolution of the formant structure for weak vowels and nasals, and an enhancement of fricative noise. The descending paths lead to further dynamic compression.

3. THE RECURRENT NEURAL NETWORK

The structure of the recurrent neural network is shown in Fig. 2 and described in detail in [5]. The input and output

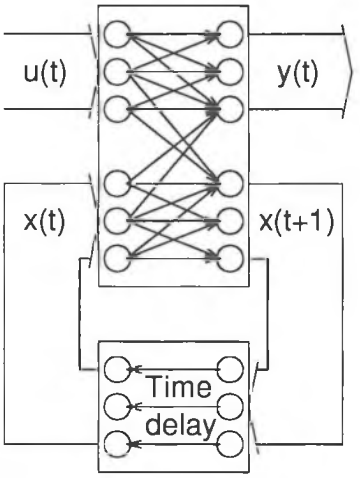

Figure 2. Block diagram of the recurrent neural network.

vectors are divided into external and internal portions. The external input vector $u(t)$ is a sequence of frames (16 ms duration) of parameterized speech of 48 dimensions from the auditory model. The components of $u(t)$ are as follows:

$$u_i(t) = \begin{cases} C_{st}(t) & \text{for } i = 0 \\ \Pi_i(t) & \text{for } 1 \leq i \leq \mathrm{J} \\ F_{i-\mathrm{J}}(t) & \text{for } \mathrm{J}+1 \leq i \leq \mathrm{J}+\mathrm{N} \end{cases}$$

where J=3 and N=44. Thus, both the tonotopic distribution of firing activity in the ascending path and the control commands from the descending paths are fed to the recognition stage. The external output vector $y(t)$ has 61 dimensions, one per symbol of the phone set of the TIMIT speech database [7]. The internal output forms a state vector $x(t)$ of 160 dimensions and is fed back to the input in the next time frame. These recurrent connections allow contextual information to be accumulated over time in the state vector. The recognition process can then use this information to make a more accurate classification.

The training data consists of 8 utterances (the si and sx sentences) from each of the 420 different speakers of the training portion of the TIMIT database. Training proceeds by unfolding the network in time over several frames of speech, comparing the external outputs to the target hand-labelled phone symbols, and adjusting the weights of the network so as to maximize a log-likelyhood cost function.

The test data consists of 8 utterances from each of the 210 different speakers of the test portion of the TIMIT database. The external outputs $y(t)$ are interpreted as phone probabilities for the specified speech frame. The most likely sequence of phone symbols is then computed from the frame by frame phone probabilities using dynamic programming. Final phone recognition results are obtained by comparing these machine-labelled symbols to the target hand-labelled symbols.

4. RECOGNITION RESULTS

Recognition results based on two modes of operation of the auditory model, normal and impaired, are reported in Table 1. The first number is the percentage of hand-labelled phone symbols correctly detected. The last number is the recognition accuracy, defined as 100% minus the percentage of insertion, substitution and deletion errors, and is the most important performance measure in this table.

When the auditory model is operated in its normal mode, the recognition accuracy is 61.8%. Over 2/3 of all errors

mode	correct	insert.	subst.	delet.	accur.
normal	66.1%	4.2%	26.5%	7.4%	61.8%
impaired	58.7%	4.9%	32.4%	8.9%	53.8%

Table 1. Recognition results

are substitution errors. Inspection of the confusion matrix revealed that the most common substitution errors are between phones from the same broad class (e.g. /z/ vs /s/, /m/ vs /n/) and often involves nearby vowels on the vowel triangle (e.g. /ih/ vs /ix/, /ax/ vs /ix/). There are relatively few errors across broad classes.

The impaired mode of operation of the auditory model consisted of disconnecting the OHC circuit, in effect simulating a total loss of OHCs. There is a loss of sensitivity and frequency resolution at low levels. The descending paths have also been cut off in this mode, although the control commands have been calculated and fed to the recognition stage for a fairer comparison with the normal mode. The neural network has been re-trained and re-tested with this new data. The recognition accuracy has dropped to 53.8%, a decrease of 8.0% in absolute terms with respect to the normal mode. All major types of errors have increased, but particularly substitution errors. Inspection of the confusion matrix revealed that recognition performance has decreased for all classes of phones. Nasals and non-sibilant fricatives are the most affected. Vowels, affricates and sibilant fricatives are the least affected. Amongst vowels, there is a tendency for the confusions to occur between vowels having their first formant in similar frequency regions, and this is also observed in hearing-impaired listeners [6].

5. FUTURE WORK

There remains many aspects to consider before we can achieve the ultimate goal of guiding the development of signal-processing hearing aids. In the short-term, recognition results will have to be repeated for more selective cochlear lesions than used here, both in quiet and in noise. In the longer term, further validation of the auditory model is needed by comparing the model output to physiological data for normal and impaired ears. The recognition stage will also have to be reviewed to ensure that the observed decrement in scores in the impaired mode closely follows the pattern seen in hearing-impaired listeners in psychoacoustical experiments.

REFERENCES

[1] C. Giguère & P.C. Woodland (1993). Chap. 25 in: *Visual Representations of Speech Signals*, edited by M. Cooke, S. Beet and M. Crawford (Wiley, London), pp. 257–264.

[2] C. Giguère & P.C. Woodland (1993). Proc. of ICASSP-93 (Minneapolis), Vol.II, 708–711.

[3] B. M. Johnstone, R. Patuzzi & G. K. Yates (1986). Hear. Res. **22**: 147–153.

[4] R. Meddis *et al.* (1990). J. Acoust. Soc. Am. **87**: 1813–1816.

[5] T. Robinson & F. Fallside (1991). Computer Speech and Language **5**: 259–274.

[6] S.G. Revoile & J.M. Pickett (1982). Chap. 2 in: *Deafness and Communication*, edited by D.G. Sims, G.G. Walter and R.L. Whitehead (Williams, Baltimore), pp. 25–39.

[7] *The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus*, National Inst. of Standards and Technology, NIST Speech Disc CD1-1.1, Gaithersburgh, MD.