PHONETIC CLASSIFICATION OF THE PLOSIVE VOICING CONTRAST USING COMPUTATIONAL MODELLING

A M Darling, M A Huckvale, S Rosen and A Faulkner

University College London, Department of Phonetics and Linguistics, London, NW1 2HE, UK

1. INTRODUCTION

Our knowledge of the processing stages that occur during the transformation of an acoustic speech signal into a "perceived" message is far from complete. One of the issues of concern is the extent to which speech perception is influenced by processes at the auditory level (i.e. involving only the peripheral auditory system (PAS) and not assumed to be unique to speech processing) and the extent to which phonetic processes (i.e. unique to speech) are dominant. In recent years, evidence has been found which indicates that auditory processing may be more influential than was once thought.

One way of investigating the role of auditory factors in a particular perceptual task is to use nonhuman listeners (animal or machine) without any phonetic processing capabilities. Kuhl and Miller [1] for example demonstrated that chinchillas are capable of identifying synthetic plosive consonant-vowel syllables on the basis of voice onset time (VOT) in a similar manner to humans. More strikingly, the phoneme boundary shifted with place of articulation in the same way. Thus this task may perhaps be accomplished at an auditory level, rather than at a phonetic level as was once thought. Damper et al. [2] have obtained similar results for the same task with a synthetic listener in the form of a mathematical model (implemented in software), comprising a model of the PAS coupled to an artificial neural network (ANN) pattern classifier. Their work supports the findings of Kuhl and Miller, and demonstrates the usefulness of a mathematical model as a synthetic listener. While it is doubtful that a mathematical model will ever be constructed that mimicks the human PAS in most respects as closely as the chinchilla's ear does, the use of a mathematical model has obvious advantages over the use of an animal listener. Not only is it quicker and easier to train, but it also allows direct intervention, e.g. alteration of the sharpness of the auditory filters. This ability to manipulate the model parameters permits the robustness of the system's performance to variations of the system parameters to be investigated.

It should be emphasized that the extrapolation of results from nonhuman listeners to human listeners is not necessarily straightforward. The inability of a nonhuman listener to perform a given speech perception task may be due to poor training and/or poor modelling, rather than evidence that a phonetic processing level is required. Similarly, a positive result with a nonhuman listener probably at best indicates the potential of the PAS to perform the required processing, rather than offering definite proof that the PAS is solely responsible. However, despite its shortcomings, at present the mathematical model is our most flexible research tool.

Our goal is to investigate the essential PAS features for carrying out certain speech perception tasks through the use of a synthetic listener in the form of a mathematical model. Our model is not original, but derives from many sources. In view of the work already done by Kuhl and Miller, and Damper et al. cited above, the identification of plosive voicing on the basis of VOT was selected as the first perceptual task to test the model with. This paper reports on the progress we have made to date towards this goal.

2. THE PAS MODEL

Although our model differs in detail from that reported in [2,3], it is essentially a different implementation of the same auditory processes. It thus consists of the same basic components: an outer/middle ear filter, a filterbank, an array of identical haircells, and a pattern recognizer in the form of an ANN. Despite differences in all of these components between the two models, they are expected to be equivalent in all essential aspects. (Note: In [2,3], an extended model including the action of the dorsal acoustic stria is also discusssed. In the following discussion, references to the model in [2,3] are only up to the auditory nerve, and the results obtained therewith.) The main components of our model are as follows:

2.1 Outer/middle ear filter

The combined action of the outer and middle ear produces a power gain in the mid-frequency range, and can be described as a bandpass filter with a flat response between ca. 1kHz and 6kHz. Following Meddis and Hewitt [4], we approximate the outer/middle ear transfer function with a digital high-pass filter of the form $y(n) = 0.887 x(n) - 0.887 x(n-2) - 0.2243 y(n-1) + 0.7757 y(n-2)$, where $y(n)$ is the output and $x(n)$ is the input. For a signal sampled at 20 kHz, this filter has a relatively flat frequency response for frequencies above 1kHz. For signals which have no significant energy above ca. 8 kHz, this high-pass filter should be an adequate approximation to the required bandpass filter. Note: The model in [2,3] attempts to correct for outer/middle ear effects by adjusting the gains of the filters in the filterbank, rather than prefiltering the signals prior to entering the filterbank.

2.2 Filterbank

The filterbank consists of a set of fourth order Gammatone bandpass filters, implemented digitally using a multiple pass IIR filter as described in [5,6].

2.3 Haircell and spike generation

The haircell.action is approximated using Meddis's model B [7]. For this test, the haircell model parameters were calibrated so that the haircell response characteristics to a 1 kHz sinusoid were similar to those of the haircell model described in [2,3], and were set as follows: $A = 2.0$, $B = 300.0$, $g = 10000.0$, $y = 5.050505$, $l = 100.0$, $r = 5000.0$, $x = 2000.0$, sampling interval = 0.00005 and scale factor = 1500.0 (excluding the contribution from the sampling interval). The output of the haircell model is the neural firing probability as a function of time, from which neural discharges can be calculated for as many realizations as required with a random number generator, along with the specification of an absolute refractory period (here 1 ms).

2.4 The artificial neural network pattern recognizer

The well documented abilities of ANN's to self-learn and detect general regularities in data make them perhaps an obvious, if not the only, choice as the pattern recognizing element of the model. The ANN software for a feed forward perceptron was developed in house. Other off-the-shelf packages were also tried, and were found to produce essentially the same results, so that any differences between our results and those in [2] are not thought to be simply due to differences in the ANN software. Note: In [2] the ANN software of McClelland and Rumelhart [9] was used.

3. METHOD

For the purpose of this test of our model, where possible and reasonable we have followed the procedure in [2,3] quite closely (any differences are noted). Due to the overall similarities of the models, we would expect similar performances. In addition, we performed listening tests with human listeners for the same stimuli that were presented to the synthetic listener.

3.1 Stimuli

The stimuli consisted of synthetic syllables produced by the Haskins laboratory (see Abramson and Lisker [8]), sampled at 20 kHz. Three series were used, representing English bilabials ("ba"-"pa"), alveolars ("da"-"ta") and velars ("ga"-"ka"). Each series consisted of nine members differing in terms of VOT, which varied in steps of 10 ms from 0 to 80 ms. (The VOT's were confirmed by measurement.) These stimuli should be equivalent to those used in [2], although they may not be identical, having been produced at different times.

3.2 Processing details

The processing steps each stimulus was subjected to are as follows:

Step 1: Level adjustment - by scaling to a simulated rms level of 65 dB SPL, where 0 dB SPL was set to a digital amplitude value of 1.0.

Step 2: Outer/middle ear frequency filtering - with the IIR filter described in Section 2 above.

Step 3: Reduction of signal length - by extracting the 120 ms section starting 25 ms before the burst. (This avoids unnecessary data processing yet includes the temporal region where the acoustic features distinguishing the voicing contrast are known to lie.) Note: In [2] the signals were also shortened to 120 ms starting at 25 ms before the burst, but not until the binning stage. (See Step 7 below.) Data truncation at this earlier stage was done merely to reduce the size of the data sets generated in Step 5 for practical reasons, and is not considered to have a significant effect on system performance.

Step 4: Auditory filtering - in parallel through 128 Gammatone filters, (see Section 2 above) with centre frequencies equally spaced on a Greenwood scale, with the lowest and highest centre frequency (CF) at 50 Hz and 5000 Hz. Although the filter type is different, the number of filters and their CF's were chosen in accordance with [2,3].

Step 5: Generation of neural firing probabilities - using the haircell model (see Section 2 above).

Step 6: Generation of multiple sets of neural firing data from the neural firing probabilities - using a random number generator and a refractory period of 1 ms. From each neural firing probability function, 50 independent data sets were generated.

Step 7: Pattern vector generation - by binning the data into bin widths of 10 ms by 8 channels, producing a 12 (= 120 ms / 10 ms) by 16 (= 128 channels / 8 channels) matrix with 192 bins. The rows of the matrix (of length 12) were then concatenated to form a single vector of length 192, which was used as the input to the ANN.

3.3 Artificial neural network training and testing

The ANN was designed with 192 input nodes, a single output node and one fully connected 16 node hidden layer. Configurations with and without a hidden layer were tried. The inclusion of a hidden layer was not found to have a significant effect on the results, which is consistent with the findings in [2].

Step 1: Artificial neural network training - by separate training for each of the three syllable series (bilabial, alveolar and velar) on the endpoints (0 ms VOT and 80 ms VOT) using 1000 trials, where each trial consists of the presentation of 50 different pairs of endpoint data sets. Although fewer trials than reported in [2] (1000 as compared to 3000), we found that with our software the error typically dropped to 1.0 % of its starting value after 1000 iterations, after which the results did not change significantly with further iterations. It is difficult to compare the error values directly with those given in [2], as the error measures are not the same.

Step 2: Artificial neural network testing - each series was tested using 50 representations (the same number used in [2]) of each series member. The output of the ANN is in the form of an "activation function", which in this case takes on values between 0 and 1, where 0 and 1 represent the endpoints the net was trained on. Mean activation functions and labelling functions (LFs) were derived. To derive the LF's, the activation value dividing the two categories in each series was set to 0.5.

3.4 Human listening tests

Separate tests were carried out for each syllable series (bilabial, alveolar, and velar). Five native English speakers (drawn from members of the department) were presented with 10 copies of each member of the given test series in random order, and asked to label the stimuli according to a two alternative forced choice paradigm.

4. RESULTS AND DISCUSSION

Figures 1a and 1c show the mean activation functions and LF's obtained with our synthetic listener. The mean activation functions of Damper et al.'s synthetic listener [2, Fig. 9 upper section] have been redrawn in Fig. 1b for comparison. Damper et al.'s results were also obtained with an ANN with a fully connected hidden layer of 16 nodes. Neither Damper et al. nor we found any significant differences between results obtained with and without a hidden layer. Figure 1d shows the mean LFs (averaged over listeners) of our human subjects. The 50% category boundaries, obtained by simple linear interpolation, are summarized in Table 1, where UCL and DAMPER refer to our results and those of Damper et al. respectively. Note: The values quoted in the table for Damper et al.'s results differ from those quoted in [2], as the latter were derived from an inappropriate probit analysis of the mean activation function.

The human listener mean LF's for bilabial, alveolar and velar continua varying in VOT reported in the literature (see e.g. [1]) typically display similar sigmoidal shapes with approximately equispaced, parallel straight line sections, reflecting a shift of phoneme boundary with place of articulation. The 50% category boundary for the bilabial, alveolar and velar series is typically in the region of 25 ms, 35 ms and 45 ms respectively. An examination of Table 1 shows that all the phoneme boundaries obtained agree with these values for the alveolar and velar places. Only for the bilabial continuum do significant differences occur. Whereas our human listeners evidence behaviour close to that reported in the literature, our computational model shows a bilabial boundary about 20 ms too long. On the other hand, Damper et al.'s computational results show a bilabial boundary about 10 ms too short. However, our results do not exhibit the correct relative positions of the three phoneme boundaries with place of articulation, whereas Damper et al.'s do. It is interesting to note that just as Table 1 shows most variability for the bilabial boundary, there was also more inter-individual variability for this continuum than for the other two.

The following separate modifications to the training procedure were also tried without significantly affecting the results: 1) increase of the number of trials used in the training by a factor of 2, 2) increase of the number of realizations used in each trial from 50 to 300 and 3) replacement of the ANN software with an alternative package. We are currently investigating various model details to help explain the observed differences between our model and that of [2,3].

Table 1. Phoneme boundaries for three plosive voicing continua obtained from human and synthetic listeners.

	Bilabial	Alveolar	Velar
UCL activation	47.0 ms	31.6 ms	46.5 ms
UCL labeling	46.3 ms	32.1 ms	45.6 ms
UCL labeling (human)	27.9 ms	32.6 ms	44.7 ms
DAMPER activation	15.8 ms	30.7 ms	43.3 ms

5. CONCLUSIONS AND WORK IN PROGRESS

While it is not the purpose of this investigation to perform a comprehensive comparative study of the various different PAS models that exist in the literature, it is worth noting that these models can not be considered functionally equivalent unless it can be demonstrated that they can reproduce equivalent results on a given perceptual task. Our current results hint that there may be very subtle aspects of auditory processing that give rise to the variation in phoneme boundary with place of articulation found for the Haskins stimuli. Clearly it is important to know how sensitive the system performance is to variations of the model parameters, not only for the purpose of developing useful PAS models, but also for understanding their influence on perceptual phenomena. We are currently still investigating the cause(s) of the performance differences between our model and that of Damper et al. Once this has been resolved, we intend to use the model to investigate the role of the PAS in various speech perception tasks, including the perception of affricates and fricatives, and to relate diminished perceptual performance with degradations of PAS model parameters from their 'optimum conditions'.

6. REFERENCES

[1] P K KUHL & J D MILLER,'Speech perception by the chinchilla: Identification functions for synthetic VOT stimuli', J.Acoust.Soc.Am.63, pp 905-917 (1978)

[2] R DAMPER, M PONT & K ELENIUS,'Representation of initial stop consonants in a computional model of the dorsal cochlear nucleus', STL-QPSR 4/1990 (1990)

[3] M PONT & R DAMPER,'Software for a computational model of an afferent neural activity from the cochlea to dorsal acoustic stria', VSSP technical report 89/TR2 (1989)

[4] R MEDDIS & M J HEWITT,'Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification', J.Acoust.Soc.Am.89, pp 2866-2882 (1991)

[5] J HOLDSWORTH, I NIMMO-SMITH, R PATTERSON & P RICE,'Implementing a GammaTone filter bank', in SVOS Final Report - Part A: The auditory filter bank, MRC Applied Psychology Unit, Cambridge, England (1988)

[6] A M DARLING,`Implementing a Gammatone filter', in Speech Hearing and Language: work in progress, vol. 5, University College London, Department of Phonetics and Linguistics, pp 43-61 (1991)
[7] R MEDDIS,`Simulation of mechanical to neural transduction in the auditory receptor', J.Acoust.Soc.Am.79, pp 702-711 (1986)
[8] A ABRAMSON & L LISKER,`Discrimination along the voicing continuum: cross language tests', in Proceedings of 6th International Congress of Phonetic Sciences, Prague 1967, pp 569-573 (1970)
[9] J L McCLELLAND & D E RUMELHART,`Explorations in parallel distributed processing: A handbook of models, programs and exercises', MIT Press/Bradford Books, Cambridge,MA (1987)

7. ACKNOWLEDGEMENTS

This work is part of a study on computational modelling of speech perception supported by the Joint Council Initiative in Cognitive Science and Human Computer Interaction. Thanks to Richard Baker for his assistance in preparing this manuscript.

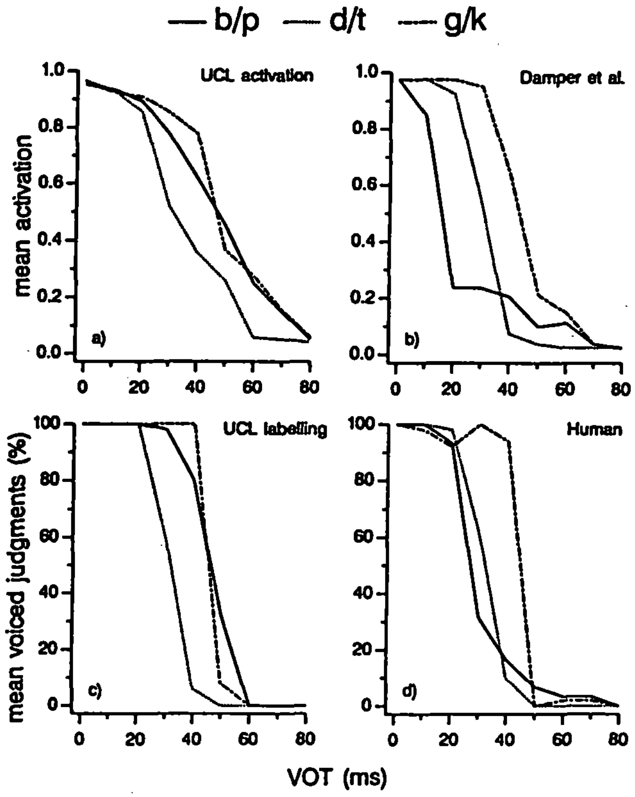
— b/p — d/t --- g/k
UCL activation
Damper et al.
UCL labelling
Human
a)
b)
c)
d)
mean activation
mean voiced judgments (%)
1.0
0.8
0.6
0.4
0.2
0.0
100
80
60
40
20
0
0
20
40
60
80
VOT (ms)

Figure 1