The representation of speech in a nonlinear auditory model: time-domain analysis of simulated auditory-nerve firing patterns

Guy J. Brown[1], *Tim Jürgens*[2], *Ray Meddis*[3], *Matthew Robertson*[1], *Nicholas R. Clark*[3]

[1]Department of Computer Science, University of Sheffield, UK
[2]Medizinische Physik, Carl-von-Ossietzky Universität Oldenburg, Germany
[3]Department of Psychology, University of Essex, UK

g.brown@dcs.shef.ac.uk, tim.juergens@uni-oldenburg.de, r.meddis@essex.ac.uk

Abstract

A nonlinear auditory model is appraised in terms of its ability to encode speech formant frequencies in the fine time structure of its output. It is demonstrated that groups of model auditory nerve (AN) fibres with similar interpeak intervals accurately encode the resonances of synthetic three-formant syllables, in close agreement with physiological data. Acoustic features are derived from the interpeak intervals and used as the input to a hidden Markov model-based automatic speech recognition system. In a digits-in-noise recognition task, interval-based features gave a better performance than features based on AN firing rate at every signal-to-noise ratio tested.

Index Terms: auditory model, time interval, automatic speech recognition

1. Introduction

Advances in rapid psychometric testing methods [1] have opened up the possibility of 'personalised' auditory models that are tuned to the hearing profile of particular individuals. In principle, such models can be used to predict a specific listener's speech recognition ability in a wide range of noise backgrounds, by coupling the auditory model with an automatic speech recognition (ASR) system.

However, in order to achieve this goal it is necessary to address the current gap between human and machine performance in speech recognition. One barrier to progress is the mismatch between the characteristics of physiologically-accurate auditory models, and the front-end signal processors used in hidden Markov model (HMM)-based ASR systems. HMM-based recognisers require a low data rate and acoustic features that vary as little as possible with sound level. The components of each feature vector should also be uncorrelated, so that they can be modelled efficiently using Gaussian mixtures with diagonal covariance. These criteria are not met by a physiologically-accurate computer model of the auditory periphery. The output of such a model consists of simulated auditory nerve (AN) activity in a number of parallel frequency channels. The firing rate in adjacent channels tends to be highly correlated, and the data rate is very high. In addition, level-dependent compression in cochlear models causes a signal representation based on AN firing rate to vary considerably with sound level.

One approach to resolving this issue is to look for alternative means of encoding the information from the cochlear model, besides AN firing rate. Physiological [2, 3] and engineering [4] studies suggest that representations based on time-domain analysis of AN firing patterns may provide advantages in terms of noise robustness and level independence.

In the present study, the auditory model of Meddis [5, 6] is appraised in terms of its ability to replicate the physiological data of [2, 3], which show encoding of speech formants in the time intervals of AN firing patterns. Acoustic features based on these time intervals, similar to those proposed by [4], are then employed in a HMM-based ASR system and shown to be advantagous compared to features based on AN firing rate.

2. Computer Model

The current study used the computer model of the auditory periphery proposed by Meddis [5, 6], which consists of a cascade of processing stages. The input signal, sampled at a rate of 44.1 kHz, is first passed to a simulation of the outer/middle ear. The resulting signal (stapes displacement) then provides the input to a bank of dual-resonance nonlinear (DRNL) filters [5], each of which models the displacement of the basilar membrane at one point along the cochlear partition. The output of each DRNL filter is the sum of a linear and nonlinear signal pathway, the latter containing a 'broken stick' function that compresses the stapes displacement when it exceeds a threshold level.

Subsequent stages of the computer model simulate inner hair cell (IHC) stereocilia displacement, the IHC receptor potential, calcium dynamics and neurotransmitter release. The final stages of the model simulate adaptation of the IHC response and the IHC-auditory nerve synapse. The model can be configured either to generate discrete nerve impulses ('spikes') or a probabilistic representation of firing rate in the AN.

For comparison with the data presented in [3], the auditory model was configured with 178 frequency channels between centre frequencies of 140 Hz and 7520 Hz. Model parameters were tuned to fit the hearing profile of a normal-hearing listener, using psychophysical measurements of absolute thresholds [1], tuning and compression.

2.1. Qualitative features of the model response to speech

The model response to a synthetic three-formant syllable /da/, presented at a level of 69 dB SPL, is shown in Fig. 1. Here, the last stage of the model was configured to produce discrete spikes, and the stimulus was presented 500 times. The spikes elicited by each stimulus presentation were summed, smoothed and normalized by the number of stimulus repetitions to give an instantaneous firing rate (IFR) pattern. The first 45 ms of the IFR are shown, during which F1 rises from 500 Hz towards its steady-state value of 700 Hz, and F2 and F3 fall towards their steady-state values of 1200 Hz and 2400 Hz respectively.

Secker-Walker and Searle [3] studied the representation of

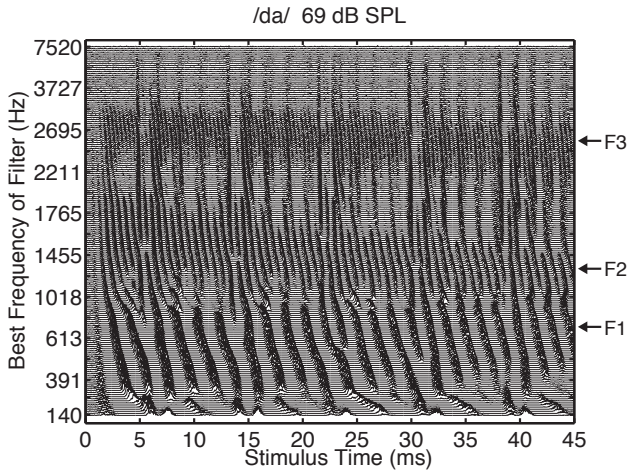

Figure 1: *Response of the auditory model to the first 45 ms of the syllable /da/, presented at a sound level of 69 dB SPL. The figure shows the instantaneous firing rate (IFR) of each frequency channel as a function of time. Formant positions are marked.*

this syllable in AN recordings from the cat, and noted that AN fibres grouped into bands corresponding to each formant, according to the temporal pattern of their activity. We found that some adjustment of the computer model parameters was necessary to reproduce this effect. Specifically, F3 was poorly represented unless the bandwidths of the DRNL filters were reduced from those measured for the human listener. Reducing the DRNL bandwidths by a factor of 0.42 reproduced the findings of [3], as shown in Fig. 1, and gave a reasonable (but suboptimal) fit to the human data. Consistent with the finding of [3], the interpeak intervals in the IFR do not change systematically over frequency – instead, they group into bands representing each formant, and vary over time as the formant frequencies change. This effect can be attributed to frequency-dependent compression, the relatively wide effective bandwidth of auditory filters, and their steep high-frequency cutoff [3]. These properties of the (human and model) auditory filters allow a strong formant to capture the temporal response of nearby AN fibres (so-called 'synchrony capture').

3. Time-domain analysis of auditory nerve firing patterns

A quantitative analysis of timing information in the output of the computer model is now presented. Specifically, an interpeak-interval analysis is used to estimate formant frequencies from the IFR.

3.1. Interpeak-interval analysis

The technique for measuring interpeak intervals described by [3] was applied to the IFR pattern obtained from each channel of the auditory model. A smooth function, in which peaks could be reliably identified, was obtained by autocorrelating 10-ms segments of the IFR of each channel at intervals of 3 ms, to give a sequence of time frames. The square root of each autocorrelation function was smoothed with an 11-point Hamming window, giving a smoothed root autocorrelation (SRA). Peaks were then identified by differentiating the SRA and locating the times at which zero-crossings occurred. Only peaks above the mean value of the SRA were retained for interval analysis.

Following [3], interval analysis was performed by constructing an *inter-peak-interval histogram* (IPIH) for each time frame. Time intervals were pooled across all frequency channels of the model. Specifically, the first three intervals between consecutive peaks in the SRA of each channel were added into a histogram, which had a bin width of 23 μs.

3.2. Results

Secker-Walker and Searle [3] evaluated the IPIH on auditory nerve data recorded by Miller and Sachs [2]. The latter recorded the spiking activity of auditory nerve fibres in anaesthetised cats, for presentations of two synthetic syllables (/da/ and /ba/). Here, we evaluate the computer model using the same speech sounds, and compare the model output to the analysis of [3]. Facsimiles of the stimuli used by [2] were synthesised according to the formant frequencies given in their paper, using the `KlattGrid` parallel speech synthesizer in Praat [7]. Both syllables were 100 ms in length, with formant transitions over the first 50 ms. The steady-state frequencies of F1, F2 and F3 in the last 50 ms of each syllable were 700 Hz, 1200 Hz and 2400 Hz respectively. The IFR elicited by the first 45 ms of the /da/ stimulus is shown in Fig. 1.

Fig. 2 plots the IPIH at each time frame, to give a two-dimensional display. The upper row of the figure shows IPIH analyses computed by [3] from the auditory nerve recordings of [2]. The formant trajectories of /da/ and /ba/ are clearly visible as lines of peaks. For example, F1 is identical in both syllables and appears as the track at the longest interval, initially with a frequency of 500 Hz (2 ms interval) and then rising to its steady-state frequency of 700 Hz (1.4 ms interval). Also note that the IPIHs for /da/ presented at 49 dB SPL and 69 dB SLP are very similar, suggesting that the time interval representation is not strongly dependent upon sound level. Output of the computer model is shown in the lower row of Fig. 2, and closely matches the physiological data.

	Entire syllable			Transition		
Stimulus	F1	F2	F3	F1	F2	F3
/da/ 69 dB SPL	29	35	53	39	48	44
/ba/ 69 dB SPL	19	18	36	26	22	31
/da/ 49 dB SPL	22	21	46	30	28	45

Table 1: *Root-mean-square errors (in Hz) of formant frequencies, estimated from the computer model.*

The differences between the true formant frequencies and their values estimated from the model are shown in Table 1, as root-mean-square (rms) errors. Formant frequency estimates were obtained from the model using the procedure described by [3]. The IPIH was segmented into three bands between 0.31-0.55 ms, 0.55-1.3 ms and 1.3-2.5 ms (as shown by the arrows in the top row of Fig. 2). The maximum within each band was identified, and the reciprocal of the interval at which the maximum occurred was taken as the formant frequency. The errors shown in Table 1 are similar to those found for physiological data by [3].

4. Interpeak interval features for automatic speech recognition

The previous section established that the computer model replicates the interpeak-interval characteristics of auditory nerve recordings [3]. It has previously been noted that the properties

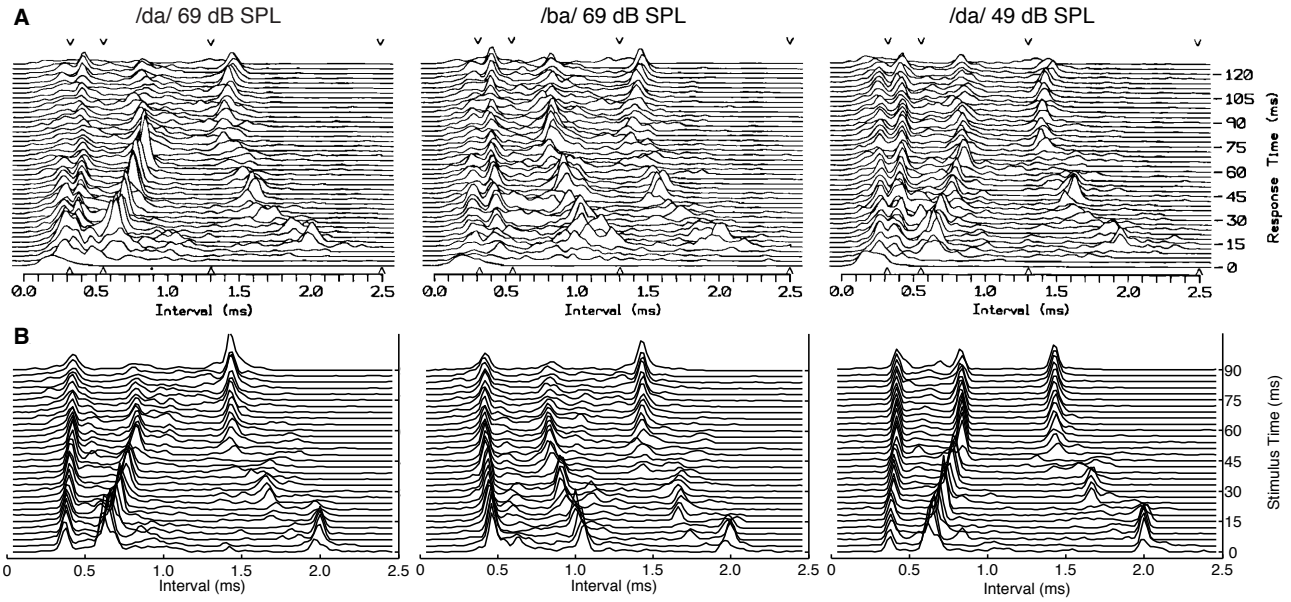

Figure 2: *A. Pooled interpeak interval histograms for the /da/ and /ba/ stimuli, derived from the auditory nerve recordings of Miller and Sachs [2] by Secker-Walker and Searle [3]. Reprinted with permission from H. E. Secker-Walker and C. L. Searle (1990) "Time-domain analysis of auditory-nerve fiber firining rates", J. Acoust. Soc. Am., vol. 88 (3), pp. 1427–1436. Copyright 1990, Acoustic Society of America. B. Output of the computer model for equivalent stimuli, showing that formant tracking is well reproduced.*

of the IPIH (i.e., robust encoding of formant structure across a range of sound levels) make it an effective front-end for ASR [4]. We now evaluate IPIH features and firing rate features derived from the Meddis model, in order to determine which provide the better performance in a speech-in-noise ASR task.

4.1. Acoustic features

Following [4], some modifications were made to the analysis described in Section 3.1 in order to provide acoustic features suitable for a HMM-based ASR system. Time frames were computed at 10 ms intervals (rather than 3 ms) in order to decrease the data rate. The width of the autocorrelation window used by [3] was 10 ms, which was sufficient for the /da/ and /ba/ stimuli because F1 never fell below 500 Hz. However, in general a longer window is required in order to provide sufficient interpeak intervals to give a good representation of utterances with a lower F1. Accordingly, we chose an autocorrelation window length of 25 ms. To reduce computation time, the last module of the Meddis model was configured to give the probability of AN firing, rather than discrete spikes; it was verified that this change did not affect the close match to physiological data shown previously.

To further reduce the dimensionality of the IPIH representation, it was divided into 30 log-spaced bands and the mean was computed within each band. Finally, a discrete cosine transform (DCT) was applied to give features that were approximately decorrelated, and therefore suitable for training HMMs in which observations are modelled by Gaussian mixtures with diagonal covariance. DCT coefficients 2–19 were used, together with their first-order and second-order temporal differences, to give 54 features per time frame.

For comparison, acoustic features were also computed directly from the IFR of the auditory model. In this case, the IFR was computed by a configuration of the model with 30 frequency channels, with the last stage of the model configured to give the probability of AN firing. Within each channel, the IFR was summed over 25 ms Hann-windowed segments at intervals of 10 ms. As before, the 30 coefficients (firing rates) in each time frame were DCT-transformed and coefficients 2–19 were used, together with their first- and second-order temporal differences. Finally, a condition was included in which the IPIH and firing rate features were concatenated into a single feature vector (a total of 108 features per frame).

For both the IPIH and AN firing rate features, preliminary experiments showed that optimum performance was obtained when the DCT coefficients were normalised by subtracting their temporal averages. Accordingly, mean-normalised DCT coefficients were used in all of the simulations reported below.

4.2. Corpus and Recogniser

The auditory model was evaluated using a spoken digit test based on the Aurora 2.0 corpus [8], which has previously been used for testing human listeners [9]. An HMM-based digit recogniser was implemented in HTK [10] and used to train word models for each digit, a silence model and a short-pause model. Word models consisted of 16 emitting states, with observations modelled by Gaussian mixture models with 7 components. The silence model and short-pause model had 3 and 1 emitting states respectively. The HMMs were trained on IPIH features computed for each of the 8440 utterances in the Aurora 2.0 clean training corpus. All training utterances were presented to the computer model at a sound level of 60 dB SPL.

The recogniser was tested using the procedure described in [9]. Digit triplets were presented to the auditory model, with the speech scaled to a level of 60 dB SPL and 20-talker babble added at signal-to-noise ratios (SNRs) between -10 dB and 20 dB, in 5 dB steps. 358 triplets were presented at each SNR, and a clean speech condition was also included. The digits "seven" and "zero" were excluded from the test set to ensure that all digits were monosyllabic, and hence that each triplet was of

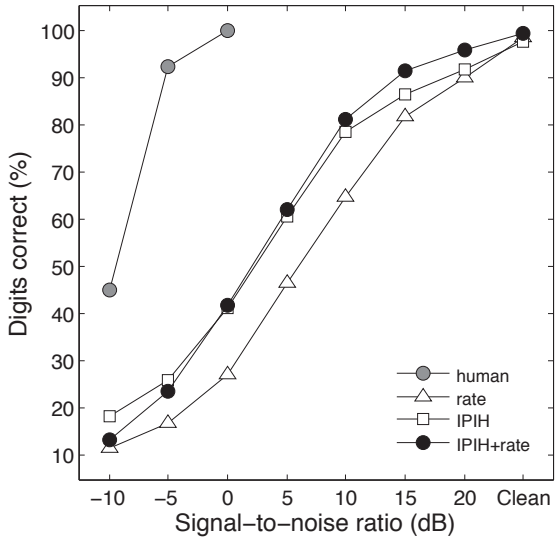

Figure 3: *Digit recognition results for the ASR system using firing rate features, IPIH features, and a combination of both. Data points for human listeners tested at three SNRs (gray circles) are also shown.*

approximately equal difficulty. The recogniser was scored in the same way as human listeners, allowing a comparison between human and machine performance. Specifically, each digit triplet was scored out of three, with a point awarded only when the correct digit was identified in the correct position.

4.3. Results

Results of the ASR experiments are shown in Fig. 3. As expected, recognition performance falls with decreasing SNR in all cases. However, the ASR performance obtained with IPIH features is better at every SNR than that obtained with firing rate features, and the same in the clean condition. Training and testing the recogniser on a combination of IPIH and firing rate features gave the best overall performance at high SNRs, but using IPIH features alone gave the best result below an SNR of 0 dB. Overall, the results suggest that IPIH features provide better noise robustness than features based on firing rate.

Human data for a similar digits-in-noise test were available for three of the SNR conditions, and are plotted in Fig. 3. The data were obtained from a single human listener with normal hearing (see [9]) and are based on a smaller number of test signals (34 digit triplets per SNR condition). Clearly the use of IPIH features helps to close the gap between human and machine performance, but the gap still remains very substantial.

5. Discussion

This paper has evaluated the ability of the auditory model of Meddis [5, 6] to reproduce physiological data relating to the encoding of speech formants in the time intervals of auditory nerve firing patterns. With some adjustment to the filter bandwidths of the model, we were able to obtain a close match to the data of [3]. Specifically, the Meddis model accurately encodes formant frequencies in broad bands of auditory nerve activity, distinguished by the pattern of their interpeak intervals. Formant frequency estimates from the model were of a similar accuracy to those obtained from physiological preparations. We have also demonstrated that interpeak intervals are a promising way of encoding the output of the auditory model, allowing it to be coupled with a HMM-based ASR system. In a digits-in-babble test, the best performance was obtained when interpeak interval features were used.

In Section 2.1, it was noted that it was necessary to reduce the bandwidths of the DRNL filters in order to match the findings of [3], leading to a suboptimal fit to auditory filter widths obtained from forward masking tests with a human listener. Using broader filters, F3 was poorly represented because the temporal response of simulated AN fibres in the region of F3 tended to be captured by F1, the most prominent formant. Although the study of [2] is widely cited in support of the role of timing information in human speech recognition, their data was recorded from cats, not humans. However, unpublished data from Meddis suggests that an auditory model configured with human bandwidths also provides an excellent match to bandwidths determined from AN fibres of the cat. The current result (i.e., that narrower filters are required to simulate the data of [3]) cannot therefore be explained in terms of inter-species differences, and remains an issue for further research.

The simulations conducted in this paper used an auditory model configured to match the profile of a normal hearing listener. In future research, we will also investigate the IPIH representation derived from hearing impaired models. Such studies may lend further insight into the influence of specific hearing impairments on speech recognition in a noisy background.

6. Acknowledgments

The authors gratefully acknowledge the support of EPSRC (Brown, Meddis), RNID (Robertson) and SFB TRR 31 “The active auditory system” (Jürgens). The model software can be downloaded from `http://tinyurl.com/5rmexxo`

7. References

[1] Lecluyse, W., and Meddis, R., “A simple single-interval adaptive procedure for estimating thresholds in normal and impaired listeners”, J. Acoust. Soc. Am., 126:2570–2579, 2009.

[2] Miller, M. I. and Sachs, M. B., “Representation of stop-consonants in the discharge patterns of auditory-nerve fibres”, J. Acoust. Soc. Am., 74:502–517, 1983.

[3] Secker-Walker, H.E. and Searle, C. L., “Time-domain analysis of auditory-nerve-fiber firing rates”, J. Acoust. Soc. Am., 88:1427–1436, 1990.

[4] Sheikhzadeh, H. and Deng, K., “Speech analysis and recognition using interval statistics generated from a composite auditory model”, IEEE Trans. Speech. Audio. Proc., 6:90–94, 1998.

[5] Lopez-Poveda, E. A. and Meddis, R., “A human nonlinear cochlear filterbank”, J. Acoust. Soc. Am. 110:3107–3118, 2001.

[6] Meddis, R. “Auditory-nerve first-spike latency and auditory absolute threshold: A computer model”, J. Acoust. Soc. Am. 119:406–417, 2006.

[7] Boersma, P. and Weenink, D., “Praat: doing phonetics by computer (version 5.2.19)”. Online: http://www.praat.org/.

[8] Pearce, D. and Hirsch, H.-G., “The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions”, in Proc. ICSLP, IV:29–32, 2000.

[9] Robertson, M., Brown, G. J., Lecluyse, W., Panda, M. and Tan, C. M. “A speech-in-noise test based on spoken digits: comparison of normal and impaired listeners using a computer model”, Proc. INTERSPEECH, 2470-2473, 2010.

[10] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., “The Hidden Markov Model Toolkit (HTK)”. Online: http://htk.eng.cam.ac.uk/.