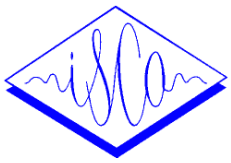

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

THE AUDITORY IMAGE MODEL AS A PREPROCESSOR FOR SPOKEN LANGUAGE

Roy D. Patterson, Timothy R. Anderson** and Michael Allerhand**

* Medical Research Council, Applied Psychology Unit,15 Chaucer Road, Cambridge, England CB2 2EF.
** Armstrong Laboratory, Bioacoustics and Biocommunications Branch,Wright-Patterson AFB, Ohio 45433

ABSTRACT

In the auditory system, the primary fibres that encode the mechanical motion of the basilar partition are phase locked to that motion, and auditory processing in the mid-brain preserves this information, to varying degrees, up to the level of the inferior colliculus. We know that this timing information is used in the localisation of point sources [1] and it is probably also used to separate point sources from more diffuse background noise. The time intervals in these neural patterns are on the order of milliseconds and so traditional speech preprocessors (like MCC and MFCC systems), with frames on the order of 15 milliseconds, remove the time-interval information from the representation. The performance of these systems deteriorates badly when the speaker is in a noisy environment with competing sources. This suggests that we will eventually need to incorporate time-interval processing into speech recognition systems if we are to achieve the kind of noise resistance characterisitic of human speech recognition. In this paper, we describe a) an auditory model designed to stabilise repeating time-interval patterns, b) the 'data-rate problem' associated with auditory models as speech preprocessors, c) a strategy for developing a noise resistant *auditory spectrogram* for speech recognition, and d) recent recognition results with a monaural *auditory spectrogram* .

I. AUDITORY IMAGES AND THE SPACE OF AUDITORY PERCEPTION

When an event occurs in the world around us, we experience an **auditory image** of the event, in the same way that we experience a visual image of the event. The auditory image reveals the pitch and loudness of the source *and* its **sound quality, or timbre.** These latter properties help us identify voices and tell whether a speaker is angry or sad. We have developed an Auditory Image Model (AIM) that simultes the three processes that we believe are essential in the construction of our auditory images and the space of auditory perception (Figure 1). The first two stages simulate the frequency analysis performed in the cochlea and the laterality analysis performed in the midbrain [1,2]. Frequency and laterality are the vertical and horizontal dimensions of the plane in the centre of Figure 1. The activity generated by a compact sound source appears on a vertical line in this representation. The figure illustrates the separation of two sources, one 40 degrees to the right of the listener and containing energy in the mid-frequencies, the other 20 degrees to the left with energy at higher and lower frequencies.

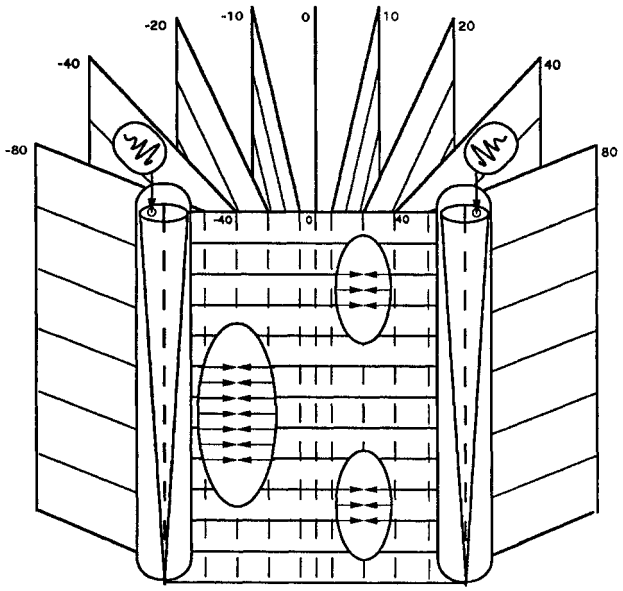

Figure 1 The space of auditory perception in the auditory image model.

This lateralised spectral analysis is often assumed to represent peripheral auditory processing in its entirety and all information concerning sound quality is assumed to be coded by the relative levels of components in the *auditory spectrum* of the sound [2] represented by a vertical line in Figure 1. The distinctive sound qualities we hear in speech and music, however, indicate that auditory image construction also include a sophisticated temporal integration mechanism and an analysis of the time intervals in the resulting neural activity pattern [3]. It is as if the system maintained a Post-Stimulus-Time (PST) histogram [4] for each frequency/laterality combination. The set of histograms activated by a point source form a plane like those in Figure 1 behind the frequency-laterality plane, and the activity that arises in the plane is the auditory image of the sound. When the sound is tonal, the histograms are regular and related as illustrated by the structure in the

lower two thirds of Figure 2 which shows the auditory image of the vowel /ae/ presented concurrently with a noise background. When the sound is noisy, the histograms are irregular and unrelated like those in the upper third of Figure 2. The stabilised time-interval patterns of vowels reveal elaborate structures that are absent in LPC and FFT spectrograms. These structures are referred to as 'auditory figures' and they have led to the suggestion that the performance of speech recognition systems could be improved if their traditional spectrographic preprocessors were replaced by auditory preprocessors that preserve time-interval information. It also seems likely that this representation will help us understand vocal quality and its role in speaker identification.

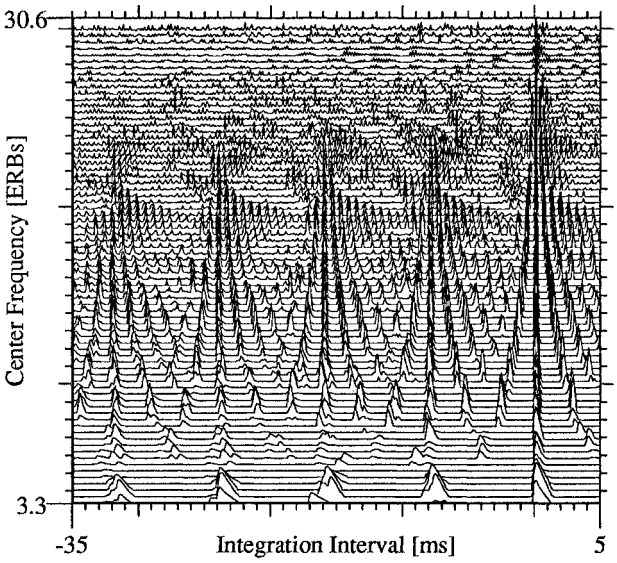

Figure 2. The auditory image of the vowel in 'hat', presented in background noise.

The mechanism that constructs the histogram from phase-locked neural activity is a new form of temporal integration that is intended to stabilise repeating time-interval patterns from quasi-periodic sounds without smearing their fine-structure unduly [3]. Briefly, a bank of delay lines is used to form a buffer store for the neural activity flowing from the cochlea; the activity level decays as it flows down the buffer at the rate of 1.5-2.5 percent/millisecond. Each channel has a strobe unit which monitors the instantaneous activity level and when it encounters a large peak it transfers the entire record in that channel of the buffer to the corresponding channel of a static image buffer, where the record is added, point for point, with whatever is already in that channel of the image buffer. Information in the image buffer decays exponentially with a half-life of about 30 ms. In the case of periodic and quasi-periodic sounds, the strobe unit tends to synchronise to the period of the sound and so generates a regular stream of pulses that initiate temporal integration in synchrony with the repeating neural pattern. This process stabilises repeating patterns of activity in much the same way as a PST histogram reveals a recurring neural firing pattern [4]. The one innovation in AIM is that the pulses that reset time in the construction of this 'auditory PST histogram' are derived from the activity pattern itself, without prior knowledge of the sound.

II. THE DATA-RATE PROBLEM AND AUDITORY SPECTROGRAMS

Although logically feasible, extraction of phonology directly from the auditory image is not yet available. The main problem is the data rate of the auditory model. To ensure that an auditory model is capable of representing all of the discriminations that humans hear, one must digitise the incoming wave with 16-bit accuracy and a sample rate no less than 32 kHz. The filterbank must have no less than 100 channels and so the total data rate is around 3.2 million, 2-byte words per second! In contrast, existing recognition systems typically use some form of LPC or FFT preprocessor which segments the wave into frames about 15 ms in duration and converts the time waveform in that frame into a vector of values that specify the level of activity in a set of different frequency bands. A sequence of such frames is referred to as a spectrogram. Commercial recognisers use 10-20 channels in the analysis and research systems use 20-50 channels. Thus, a fairly high fidelity commercial system, or a moderate fidelity research system might have 10-ms frames and 32 channels for a data rate of 3.2 Kbps -- three orders of magnitude lower than the data rate at the output of an auditory model!

With the advance of computing power, data rates that seemed prohibitive at the start of the last decade should be available by the end of the current decade and so we need not be unduly concerned by the power required to process speech in real time with an auditory model. Indeed, special purpose chips to perform cochlear processing in real time are being developed currently [5]. Nevertheless, the output data rates of auditory models are likely to remain a problem for speech recognizers for the foreseeable future.

All of this leads us to ask a) why the auditory system goes to the trouble of encoding the time intervals in basilar membrane motion, b) what processing is performed on the time-interval patterns, and c) whether these processes could be used to improve machine recognition within a reasonable time frame? There would appear to be at least three good reasons for recording and processing the temporal microstructure of basilar membrane motion; they are binaural source isolation, monaural figure/ground separation, and the extraction of robust sound-quality features. In this section, we briefly describe the binaural process and our strategy for incorporating it into standard recognition systems. Binaural processing is the first step in the chain of time-interval processes and the algorithms for binaural processing are well established (e.g. [6]). Consequently, we will use binaural processing as the primary example in our strategy for incorporating time-interval processes into speech preprocessors.

Models of auditory binaural processing involve correlating the outputs of channels from the left and

right cochleas that have the same center frequency to determine whether the activity on one side is a time delayed version of the activity on the other side. The correlation is performed on the temporal fine-structure because the time delays are on the order of tenths of milliseconds; correlation of time-averaged summaries like LPC coefficients would not reveal such small delays. The process determines the direction of the source whose activity dominates that pair of channels. The set of values can be used to group the channels with a clear, common direction, and to exclude those channel that are either from other directions or which do not show strong directionality, as in the case of many noise sources.

Our strategy for incorporating these processes into speech recognition has three stages: The first stage is to demonstrate that an *auditory spectrogram* from AIM will support good phoneme recognition and speaker identification when interfaced to a well known recognition system. The second stage is to assemble a binaural auditory model using a pair of cochlea simulations and a binaural processor. This binaural AIM would be used to isolate a source presented in a time-varying, direction varying, background noise and produce an *auditory spectrogram* of the isolated speech. In the third stage, the binaural *auditory spectrogram* would be compared with a monaural *auditory spectrogram* of the same speech to determine whether it supports better recognition with the same recognition system.. The individual algorithms required to implement this binaural AIM exist currently. We expect to show that the binaural *auditory spectrogram* is more resistant to the interference produced by extraneous sounds sources. The next section presents the results of two experiments conducted a) to establish that a monaural auditory spectrogram from a standard auditory model will support good speech recognition, and b) to develop an efficient and representative recognition system for evaluating *auditory spectrograms.*

III. SPEECH AND SPEAKER RECOGNITION WITH SELF-ORGANIZING FEATURE MAPS

AIM was used to produce a summary, in spectrographic form, of speech sounds from 151 speakers in the TIMIT data base and Kohonen self-organizing feature maps were used to perform phoneme recognition and speaker identification on the spectrographic summaries. The results of the AIM/Kohonen recogniser are compared with those using LPC features and the same Kohonen recogniser.

The Preprocessors: A 16 millisecond Hamming window was applied to the speech wave to produce a frame, and twenty LPC coefficients were calculated for the frame. Overlapping frames were calculated with a new start point every 5 ms. The LPC coefficients were then converted to mel-cepstral coefficients (MCC) using the bilinear transformation method [7] . AIM was used to calculate the output of a 40-channel cochlea simulation spanning the frequency range 430 to 6641 Hz. The channels were then time averaged with a lowpass filter, whose impulse response had a 16-ms equivalent rectangular duration, and downsampled at 5-ms intervals. Finally the 40 channels were reduced to 20 by averaging adjacent values in a frame. The result is an auditory summary of the speech sound with the same data rate and spectrographic format as that of the traditional MCC preprocessor

The Database: The speech data used to train the Kohonen phoneme recogniser comprised all 10 sentences of 114 talkers (37 female and 77 male from dialect regions 1 and 2) from the TIMIT data base . Testing was performed with all 10 sentences of a different 37 talkers (12 female and 25 male, again from regions 1 and 2). During the development of the SPHINX recognition system [8], the TIMIT phoneme labels were slightly modified. This modified convention was adopted for the present research in order to provide a better means of comparing results with other established systems. This convention yields 39 phones in separate categories

Self-Organizing Feature Maps: A Kohonen network was selected because they have the ability to learn the mapping of an input data space into a pattern space that defines discrimination, or decision, surfaces. This process has been used for phonetic recognition of Finnish and Japanese [9] . The operation of this network resembles the classical vector-quantization method called k-means clustering. Self-organizing feature maps are more general because topologically close nodes are sensitive to inputs that are physically similar. Output nodes will be ordered in a natural manner. Another reason for using Kohonen self-organizing feature maps is that they are efficient computationally and the results are representative of systems with much greater computational loads.

Kohonen's algorithm adjusts weights from common input nodes to output nodes arranged in a two-dimensional grid. Each input node is connected to every output node. Real-valued input vectors are presented sequentially in time to the network without specifying the desired output. After enough input vectors have been presented, each node's weights will specify a cluster center. These cluster centers approximate the probability density function of the input vectors. The weight adjustment is based on a distortion measure. In this work, a mean squared error distortion measure was used, based on the input and stored weights. The codebook size was 256, as is typical of recognition systems employing vector quantization.

The calibration process was similar to that used by Kohonen; once trained, learning was turned off (the weights were fixed) and the training data was presented to the feature map a second time. The node that responded to each training token was associated with that token label; the token label with the largest number of responses was deemed the label for that node. Learning Vector Quantization (LVQ) was used on the calibrated codebook to adjust the codewords for improved performance [10]. In this work LVQ3 was used.

Phoneme Recognition Performance: We compared the recognition performance of the Kohonen net using spectrograms from AIM and MCC and showed that AIM performed significantly better

than MCC in terms of phoneme-recognition accuracy. In the table below, the results are broken down into average recognition, substitution, deletion, insertion rates, and broad class recognition rates.

Summary Results

	MCC	AIM
Correct	49.18	50.87
Deletions	11.67	7.50
Insertions	2.20	1.61
Substitutions	36.95	40.02

Broad Class Results

	MCC	AIM
Fricatives	62.42	72.57
Glides	42.13	34.13
Nasals	52.08	43.65
Silence	91.19	93.65
Stops	0.00	38.05
Vowels	82.52	87.01
Total	71.31	76.11

The *auditory spectrogram* supported significantly better phoneme-recognition performance than the MCC spectrogram in the broad class categories of Fricatives, Silence, Stops and Total broad class performance.

Speaker Recognition: Sambur [11] has shown that, in general, vowels are the best broad phoneme class to extract features for speaker recognition. Using this as a guideline, speaker recognition experiments were performed using the vowel sounds on their own. Speaker dependent codebooks were created using Kohonen learning as described above. Each codebook contained 64 element; the codebooks were trained with 40 epochs of the data. The vowel data used for training the speaker dependent codebooks were from 7 sentences (the si and sx sentences) from the test set above. This provided a 37 speaker data base for speaker identification. The test set consisted of vowel data from 1 sentence (the sa1 sentence) from each speaker.

Speaker recognition was based on minimum average distortion defined over all speaker codebooks and over N frames. For each speaker, in the database, the mean-square-error (the distortion) was calculated as the difference between the input and the codeword closest to the input, averaged over all input vectors. The speaker recognized was the one with the minimum distortion. The recognition results for the 37 speakers showed that the performance of AIM (91%) compares well with that of MCC (94%).

IV. CONCLUSIONS

AIM shows promise as a preprocessor for phoneme and speaker recognition. Phoneme and speaker recognition provide a good pair of tests for an auditory representation beers the same spectrographic output can be used to test the representation on both source quality and phonetic content.

Whereas distortion metrics and signal processing methods have been extensively developed for LPC and cepstral representations, these currently do not exist for auditory model representations. Improvements in auditory modelling should continue to be exploited for speech and speaker recognition. Future research will examine temporal aspects of the auditory periphery models, such as the strobed triggered-temporal integration of AIM, and its use in speech and speaker recognition.

Acknowledgements

The authors would like to thank Janet Slifka for developing the Kohonen code and analysis tools. This work was supported by the Air Force Office of Scientific Research (AFOSR) through its Window-on-Science Program, through AFOSR Task 2313V3, and through the AAM HAP project of the UK Defence Research Agency, Farnborough.

References

[1] Blauert, J. (1983). Spatial Hearing. MIT Press, Cambridge, Massachusetts.

[2] Patterson, R.D. (1994a) The sound of a sinusoid: Spectral models. J. Acoust. Soc. Am., (revision submitted January 1994)

[3] Patterson, R.D. (1994b) The sound of a sinusoid: Time-interval models. *J. Acoust. Soc. Am.*, (revision submitted January 1994)

[4] Pickles, J.O.(1988). An introduction to the physiology of hearing. Academic Press, London.

[5] Lyon, R.F and Mead, C. (1988) "An analogue electronic cochlea." IEEE Trans of ASSP, 36, 1119-1134.

[6] Bodden, M. (1993) "Modelling human sound-source localization and the cocktail-party-effect." *Acta Acoustica*, **1**, 43-55.

[7] Lee, K.F and Hon, H.W. (1989) "Speaker independent phoneme recognition using Hidden Markov Models." IEEE Trans on ASSP, 37, 1621-1648.

[8] Lee, K.F (1989) Automatic speaker recognition: the development of the SPHINX system. Kluwer Academic, Boston.

[9] Kohonen, T. (1988) "The neural phonetic typewriter." IEEE computer magazine, **21**, 11-22.

[10] Kohonen, T. (1989) Self-organisation and associative memory, 3rd edition. Springer-Verlag, Berlin.

[11] Sambur, M.R. (1975) "Selection of acoustic features for speaker identification." IEEE Trans of ASSP, 23, 176-182.