A NEW ALGORITHM FOR VOICING DETECTION AND VOICE PITCH ESTIMATION BASED ON THE NEOCOGNITRON

James R. E. Moxham, Peter A. Jones, Hugh J. McDermott, Graeme M. Clark
Australian Bionic Ear and Hearing Research Institute
384-388 Albert St., East Melbourne, Victoria 3002 Australia.
Tel.: 61 3 283 7500 Fax: 61 3 283 7518

Abstract. Over the last decade cochlear implants have been used increasingly to restore hearing to the profoundly deaf. One of the more widely used implants is the Nucleus multi-electrode implant, developed by the University of Melbourne and Cochlear Pty. Ltd. The speech processor used with this implant is the MSP, programmed with the multipeak strategy. This device incorporates circuits to estimate the fundamental frequency (F0) of speech signals, and to decide whether voicing is present. This paper describes a new F0 estimator and voicing detection algorithm based on the neocognitron; a neural network modelled on the retina and early visual system. Performance was compared with that of three other F0 estimation algorithms: linear predictive coding (LPC), cepstral analysis and the algorithm used in the Multipeak-MSP processor. For the speech samples tested, the neocognitron performed more reliably than the other three systems. On the basis of these results, this work may be able to provide benefits to existing and future cochlear implant users.

I. INTRODUCTION

Cochlear implants, or "bionic ears" have found increasing use over the last decade to restore hearing to the profoundly deaf. A typical bionic ear consists of two parts: an implanted electrode array inserted into the cochlea and connected to a receiver/ stimulator, and an external microphone/speech processor and transmitter.

One popular speech processor used with implants is the Melbourne/Cochlear Pty. Ltd. Multipeak-MSP. Two features of the speech signal that are extracted by the MSP processor are voicing, and if present the pitch of the voicing, or F0 value. Voicing occurs when the vocal cords vibrate, and the F0 value is generally higher for child and female speakers than it is for male speakers. Being able to distinguish voiced sounds, such as /z/ and /b/, from unvoiced sounds, such as /s/ and /p/ is believed to be helpful in enabling cochlear implantees to understand speech.

This present study investigated the potential for using an artificial neural network to improve the reliability of F0 estimator and voicing detector circuits. The neural network model chosen, the neocognitron, was originally proposed by Fukushima et al as a system for recognising handwriting [1]. Parts II and III of this paper describe the neocognitron voicing detector/F0 estimator, and part IV compares the performance with three other algorithms.

0-7803-0557-4/92$03.00 © 1992

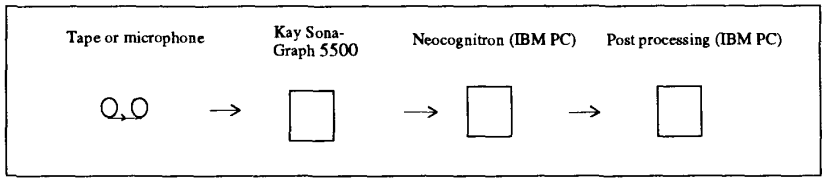

Figure 1. Block diagram showing the neocognitron voicing detector and F0 estimator.

II. STRUCTURE OF THE NEOCOGNITRON NEURAL NETWORK

The voicing detector/F0 estimator consists of three parts; a Fast Fourier Transform (FFT), the neocognitron and a post processing layer. Each of these is shown diagrammatically in figure 1, and will be discussed in turn.

Fast Fourier Transform (FFT)

As shown in figure 2, sounds to be analysed are initially processed by a Kay Sona-Graph model 5500, which is connected, via an interface, to a personal computer. The 5500 calculates a 1024 point FFT from 0 to 1990 Hz in real time. The result is a spectrogram of the sound sample. For analysis by the neocognitron, this spectrogram is divided into 31.25 millisecond slices. Each spectrogram slice is presented to the neocognitron in turn, and if voicing is present, an estimate of the F0 value is made.

The neocognitron

The neocognitron is modelled on the visual system, and recognises patterns that can be drawn using a grid of points. In this case, the input grid measures 199 x 19 points, with frequency represented on the x axis (0-1990 Hz = 199 points) and amplitude on the y axis (approximately 50 dB = 19 points). A typical spectrogram and neocognitron input pattern is shown at the top of figure 3. The neocognitron has been trained to recognise harmonic peaks in the spectrum, and to determine the spacing, which is proportional to F0.

As shown in figure 2, the neocognitron consists of three layers, with each layer made up of four different neuron types. The four types of neurons are respectively S cells, J cells, G cells and C cells.

S cells. S cells are feature extracting cells, and give a response from 0 to 100%, depending on how well the input pattern matches the pattern the cell was trained with. S cells in each layer are arranged in cell planes, with all S cells in a particular cell plane being trained to recognise the same pattern. Cell planes are represented in figure 2 as elongated rectangles.

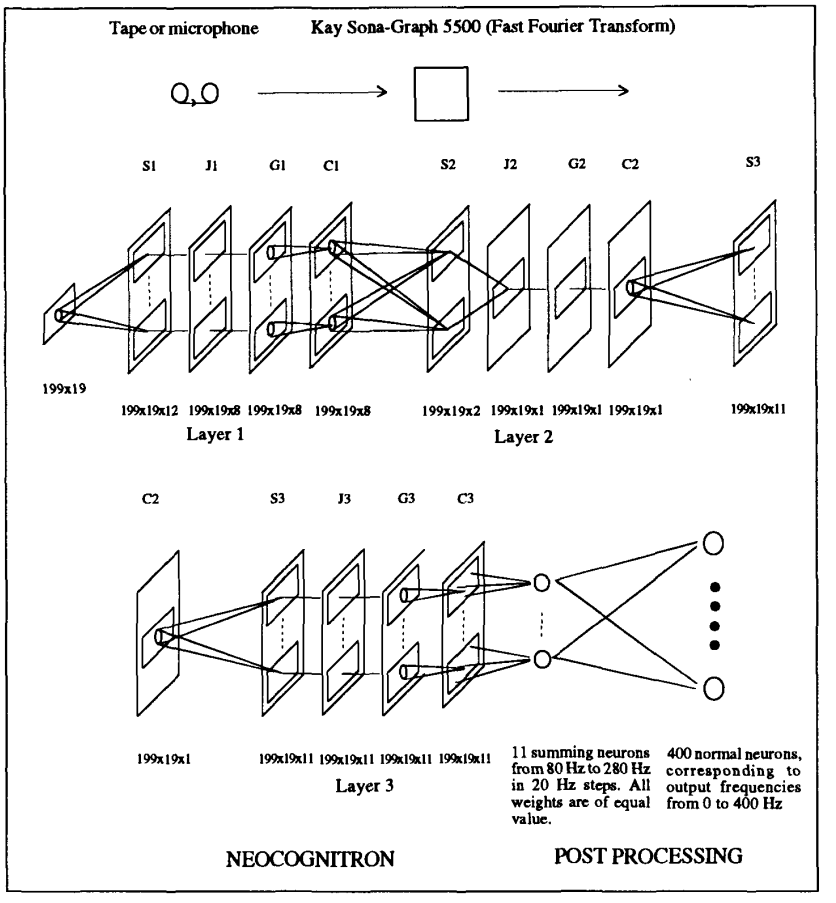

Figure 2. Schematic diagram showing the arrangement of the neocognitron voicing detector/F0 estimator. A 1024 point FFT is performed by the Kay Sona-Graph 5500, the F0 value estimated by the neocognitron and the output interpolated by the post processing layer to give a more accurate F0 estimation. If the greatest normal neuron is above a threshold then this is the estimated F0 value.

J cells. J cells are used to join planes together. Each J cell takes the output from the corresponding S cells in the planes to be joined, and gives a response that is equal to the greatest S cell response. J cells are used to treat two different shapes as the same. For example, referring to figure 3, in layer S2 one plane recognises wide harmonic peaks and the other recognises narrow peaks. J cells are used to treat these two patterns as the same pattern, a harmonic peak.

G cells. G cells determine how selective a layer is for a specific feature. Each G cell takes the output from the corresponding J cell, and applies a sigmoid transformation.

The bottom and top cutoff of the sigmoid curve can be varied, with an increase in the bottom cutoff increasing the selectivity to the training pattern. Increasing the selectivity causes the G cell to respond less to patterns that are not exactly the same as the training pattern.

C cells. C cells have the function of tolerating positional errors of the feature being recognised. Each C cell responds most to G cells near its receptive field centre, and the input weights decrease for synapses that are more distant from the centre. Synapse weights are calculated from the statistical normal distribution, with the centre weight being 1.

Feature extraction. In the entire network the process of feature extraction by the S cells, joining by J cells, selecting features above a certain threshold by G cells, and tolerating deformities by C cells is repeated for each layer. This process of tolerating positional errors a little at a time in each layer is effective for recognising deformed patterns. Like the human visual system, local features are extracted in lower stages, and are integrated into more global features by higher stages [7].

Post-processing layer

As shown in figure 2, the output of the neocognitron appears in 11 planes, with each plane containing 199x19 neurons. The outputs of every neuron in each of the 11 planes are added together by 11 summing neurons. Each summing neuron has 199x19 synapses, each synapse being of equal weight. The output of these 11 neurons represents the estimated F0 from 80 Hz to 280 Hz in 20 Hz steps.

In order to improve the precision, the outputs of these 11 summing neurons are interpolated by 400 "normal" neurons, so called because the weights of the synapses were calculated from the statistical normal distribution. Each normal neuron represents an output frequency from 1 to 400 Hz, and has a synapse from each of the summing neurons. The F0 of a voiced utterance is determined by selecting the normal neuron with the largest response. A speech segment is deemed to be unvoiced if none of the 400 normal neurons respond above a certain threshold. Using several sound samples, the actual threshold value was set to give a reliable output under both quiet and noisy conditions.

III. RESPONSE OF THE NETWORK

Figure 3 shows the response of the neocognitron to a 31.25 ms time slice of a female speaker saying /afa/. The spectrogram at the top of the figure was generated by the Kay Sona-Graph 5500. The spectrogram for slice 4 is typical of an input pattern that may be presented to the neocognitron. Training patterns for each layer are shown on the left hand side of the figure. Responses of the C layer neurons are shown, with darkness representing the output from 0 to 5.

Layer 1 recognises line segments of different orientation. As can be seen, the top cell plane was trained to recognise horizontal line segments, and the middle planes vertical

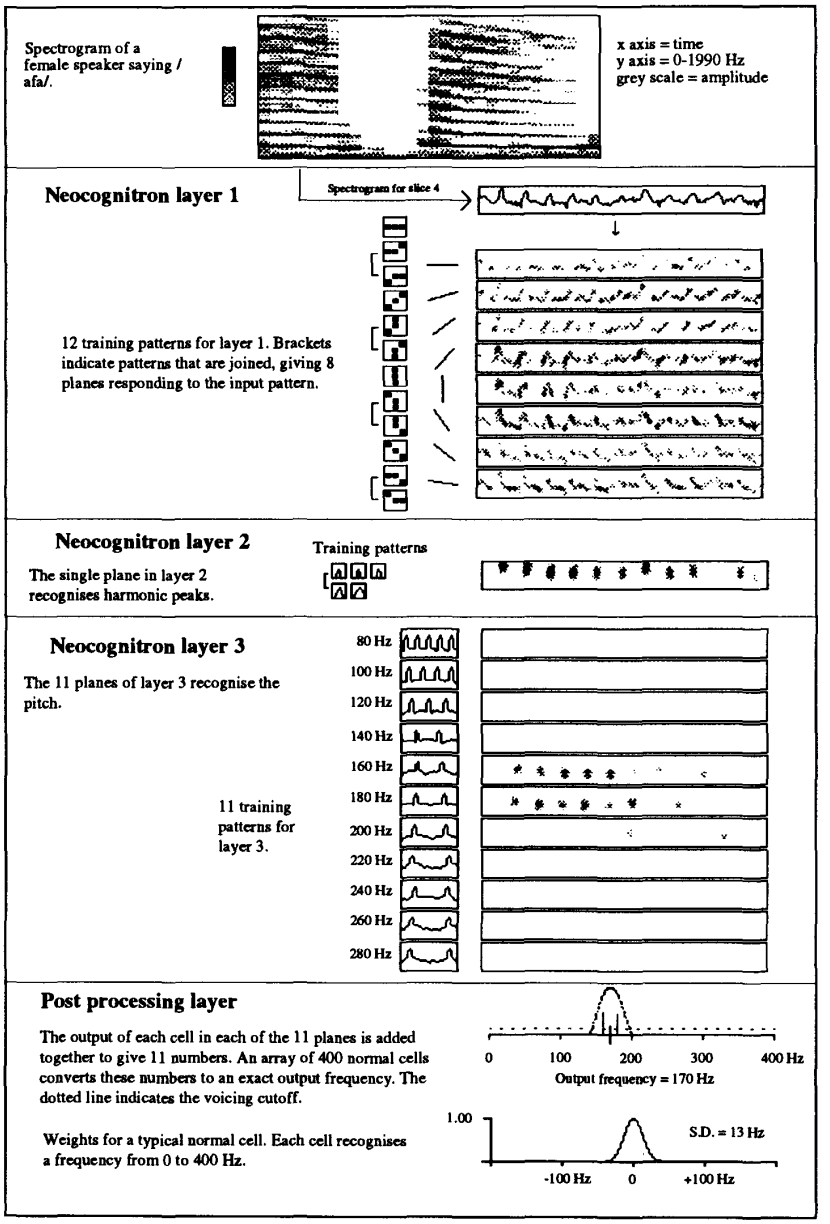

Figure 3. Complete analysis for slice 4 of a female saying /afa/. The neocognitron gives an output F0 value of 170 Hz

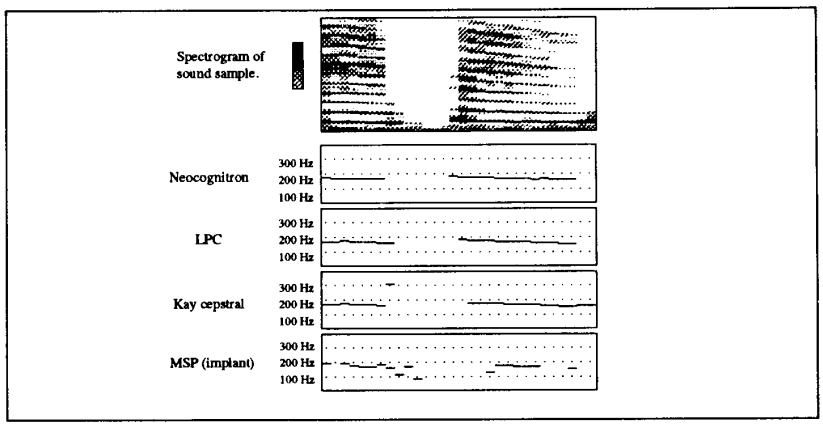

Figure 4. Spectrogram and responses of the neocognitron, LPC, cepstral analysis and the MSP for a female speaker saying /afa/.

lines.

Layer 2 recognises the harmonic peaks. All the peaks are recognised correctly, apart from the peak second from the right. This peak is too wide to be matched with any of the training patterns. A training pattern to recognise this peak was not generated because the network would then also respond to noise.

Layer 3 recognises F0 from 80 to 280 Hz, in 20 Hz steps. As can be seen the 160 Hz and 180 Hz planes are responding to a similar extent, indicating the F0 of this sample is close to 170 Hz. Training patterns for layer 3 were generated artificially by a computer program that added together sine waves. The waveform was treated as a speech sample, being analysed by the 5500 and the first two layers of the neocognitron in turn. All the training patterns for the neocognitron were thus generated artificially.

The top part of the post processing section shows a frequency scale from 0 to 400 Hz. The two bars at 160 and 180 Hz represent the output of the summing neurons. The curve over the bars shows that normal neurons from around 130 to 230 Hz are responding. The neuron that is responding the most is at 170 Hz, as indicated by the narrow line on the frequency axis. As a comparison, for this particular slice the output of the other three systems were respectively LPC = 168 Hz, cepstral = 170 Hz and MSP = 178 Hz. The dotted line indicates the voicing threshold. If no normal neuron responds above this line then the sound slice is deemed to be unvoiced.

IV. PERFORMANCE

The performance of the neocognitron was compared with that of three other systems currently used for F0 estimation; cepstral analysis, linear predictive coding (LPC) and the MSP algorithm used with the Nucleus implant.

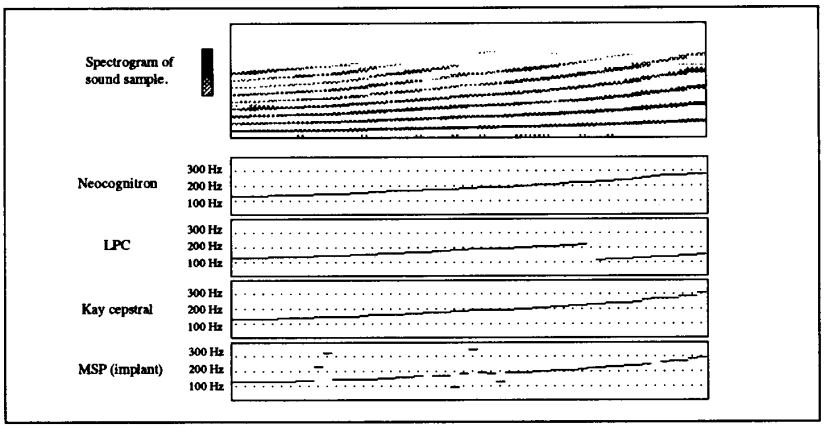

Figure 5. Spectrogram and responses for a male speaker saying /a/ with rising intonation.

Comparison of the four algorithms

Several different utterances were used to assess performance. Default parameters were used for both the LPC and cepstral analysis. Once set, parameters were not altered for each of the different sounds being analysed.

The sound samples used were a female saying /afa/, a male /a/ with rising intonation and a child saying /aba/ with multi-speaker babble added in 5 dB increments.

The neocognitron, cepstral analysis and LPC analysed signals directly from cassette tape. For the MSP, a dummy head was used with a microphone placed behind the ear. Speech samples were played through speakers to give a measured level at the dummy head of 65 dBA.

Response to a female speaker saying /afa/. The response of the four systems to the female /afa/ is shown in figure 4. The first three systems correctly identified F0, and gave no output during the unvoiced segment. The MSP correctly identified F0 for most of the first /a/, but failed to respond for around half of the second /a/. This example was typical of samples of voiced speech with no noise in that the estimates of F0 of each of the four systems were generally within 5 Hz of each other.

Response to a male saying /a/ with rising intonation. The analysis of a male /a/ with rising intonation is shown in figure 5. This rises from 120 Hz to 280 Hz, and tests the frequency response of each system. Above 220 Hz the LPC program gave an output that was half the actual F0 value. The algorithm for LPC only allows for a narrow range of frequencies to be analysed. The LPC program was used with the default settings, but Kay [6] show how the program can be optimised for child voices. This is, however, at the expense of being able to analyse low pitched male voices. The MSP works over the entire frequency range tested, but made several octave errors (x2 and /2). Cepstral analysis works over the frequency range of 78 to 350 Hz [6], and the neocognitron works from 70 Hz to 290 Hz. Below 220 Hz, LPC, cepstral and the neocognitron all agree on

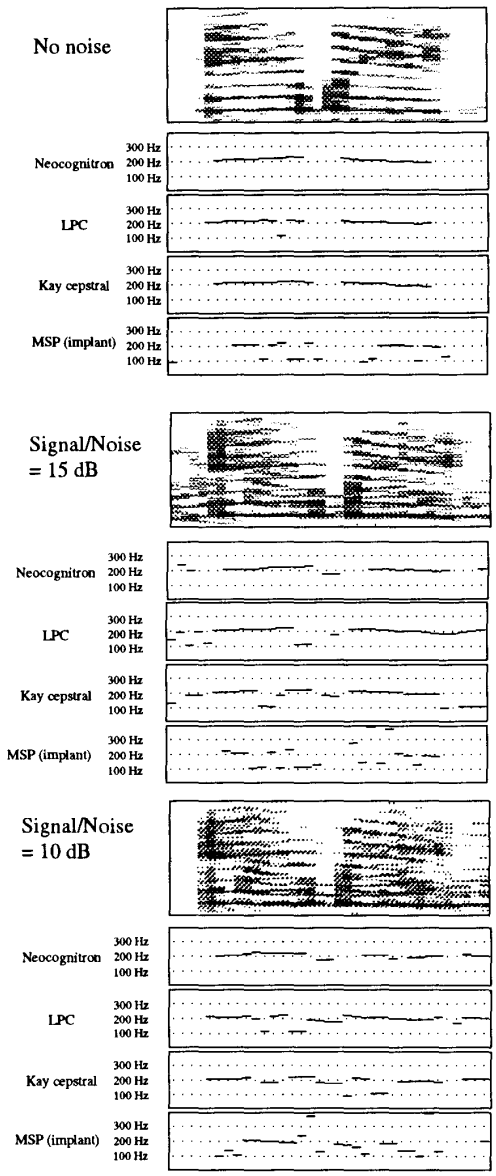

Figure 6. Performance of the neocognitron, LPC, cepstral analysis and MSP implant for a child saying /aba/ with four speaker babble added in 5 dB increments.

the F0 value to within 5 Hz.

Performance in noise. The sample used to compare performance in noise was a child saying /aba/. Four speaker babble was added at a signal/noise ratio of both 15 dB and 10 dB to simulate conditions that cochlear implant patients may experience.

Three recordings were made; speech with no noise and speech / noise (SNR) = 15 and 10 dB. Levels were set with a sound level meter and recordings were made from a master tape so the noise in each sample was the same.

Figure 6 shows the response of each of the systems for different signal to noise ratios. As can be seen, the first three systems performed comparably for sound with no noise. The MSP made several errors, the most common being to analyse the fundamental frequency as being lower than the correct value.

With a SNR of 15 dB the neocognitron, LPC and cepstral all detected someone in the babble saying "joy" between the first and second /a/. The neocognitron also detected some of the babble at the beginning of the sample. LPC detected the babble at the beginning and at the end, and also has an octave jump before the child finishes the first /a/. The cepstral also detected the babble at the beginning and the end, and has an octave jump in the middle of the first /a/. More than half of the MSP outputs were one octave too low in the first /a/ and for the second /a/ some samples are one octave too high, and some are 40 Hz too low. Unlike the other three systems, the MSP was not confused by the noise after the second /a/.

With a SNR of 10 dB the neocognitron was late in detecting the second /a/. It also had a 31.25 ms "dropout" during the second /a/ and along with the other three systems detected some babble at the end of the sample. The LPC had several octave jumps during the first /a/ and responded almost continuously from this point on. The cepstral had several "dropouts" during the first /a/ and gave an F0 value that is about 40 Hz too low in the middle of this first /a/. An octave error was also made during the second / a/. The MSP detected the first /a/, but analysed F0 as falling when it actually rose slightly. The F0 value of the second /a/ was not estimated accurately at all. However, unlike the other three systems the MSP was not confused by the noise at the beginning and end of the sound sample. As a voicing detector, but not as an F0 estimator, in this sample the MSP was performing equal to, if not better than LPC or cepstral analysis.

In summary the performance of linear predictive coding and the cepstral analysis were comparable for noisy conditions such as 15 dB and 10 dB SNR. Under these conditions the neocognitron performed equal to, if not better than the LPC or the cepstral analysis. For these samples the MSP performed worse than the other systems as an F0 estimator. However, as a voicing detector alone the performance was better than the cepstral analysis or LPC, and equal to, if not better than the neocognitron.

V. CONCLUSIONS

Originally proposed as a system for recognising handwriting [1], the neocognitron neural network has been shown to perform competitively with established voicing detector/F0 estimator algorithms. As the network is capable of performing more accurately than the algorithm used in the popular MSP speech processor, improvements in the intelligibility of speech may be possible if the neocognitron is implemented as a voicing detector and F0 estimator in a cochlear implant speech processor.

REFERENCES

[1] K. Fukushima, N. Wake, "Handwritten character recognition by the neocognitron," IEEE Trans. Neural Networks, vol. 2, no. 3, pp. 355-365, May 1990.

[2] I. S. Howard, M. A. Huckvale, "Speech Fundamental period estimation using a trainable pattern classifier", 7th FASE Symposium, Speech-88, Edinburgh, 1988.

[3] M. W. Skinner et al , "Performance of Postlingually Deaf Adults with the Wearable Speech Processor (WSPIII) and MINI Speech Processor (MSP) of the Nucleus Multi-Electrode Cochlear Implant", Ear and Hearing Vol. 12, No. 1., 1991.

[4] E. Barnard, R. Cole, F. Alleva, "Pitch detection with a neural-net classifier," IEEE Trans. Sig. Processing, vol. 39, no. 2, pp. 298-306, Feb 1991.

[5] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, "Phoneme recognition using time-delay neural networks," IEEE Trans. Accoustics, Speech & Signal Processing vol. 37, no. 3, pp. 328-339, March 1989.

[6] Kay Elemetrics manual for the DSP Sonagraph model 5500, volumes 1 and 2.

[7] R. A. Moses, Adler's Physiology of the Eye, THE C. V. MOSBY COMPANY, St. Louis, Missouri, 1981.

[8] G. M. Clark, Y. C. Tong, J. F. Patrick, Cochlear Prostheses, CHURCHILL LIVINGSTONE, 1990.

[9] G. M. Clark et al, The University of Melbourne-Nucleus Multi-Electrode Cochlear Implant, Basel (Switzerland), KARGER, 1987.