AN AUDITORY NEURAL SYSTEM FOR SPEECH PROCESSING AND RECOGNITION

V. Rodellar[1], P. Gómez[1], M. Hermida[1], and R. W. Newcomb[2]

[1]Departamento de Arquitectura y Tecnología de Sistemas Informáticos
Universidad Politécnica de Madrid
Campus de Montegancedo, s/n
Boadilla del Monte
28660 Madrid SPAIN

[2]Microsystems Laboratory
Electrical Engineering Department
University of Maryland
College Park, MD 20742
USA

Abstract

Through the present paper a Speech Feature Encoder for Speech Training and Recognition is presented. The system is composed of a Time-Domain Digital Filter to emulate an Auditory Model, and of a Neural Network, which associates the output of the filter with a given code word from a Phonetic Encoding Scheme. The design of the Digital Filter relies on Pade's Approximation to implement the involved irrational functions, and details are given on its deduction. The Neural Network is based on a Time-Domain Back-Propagation Network, and details are also given on the structure being used, and on its dimensionality. The Phonetic Encoding Scheme proposed is also presented, and its associated graph for the most used sounds in Spanish is given. Results from practical simulations with the Auditory Model are exposed and discussed, and an experiment to check the associative features of the Neural Network is examined, to conclude that a structure of 18:11:8 nodes may code the densest subset of the 16 closest sounds in Spanish yielding a neglectable error rate.

Introduction

The main purpose of the research described in the present paper is aimed to design an automatic Speech Feature Encoder (SFE), amenable of operating in Real Time. Feature Encoding is a technique which consists in assigning a special given label from a Phonetic Code to a fragment of Speech. The Phonetic Code is a set of code words, which describes the presence of certain Phonetic Features in the target fragment of Speech. The main applications of Feature Encoding are to be found in Speech Learning and Correction, and in the design of Aids for the Hearing Impaired, among others. Many different schemes may be used to build Feature Encoders, a specific one being shown in Fig. 1.

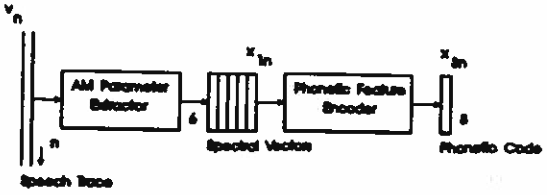


Fig. 1. Typical Scheme of a Speech Feature Encoder

The Speech Feature Encoder shown in Fig. 1 is composed of two main blocks. The first block, labelled as AM (Auditory Model) Parameter Extractor converts the Speech Trace v_n, to a set of 6-dimensional Spectral Vectors x_{1n}, which give the main spectral characteristics of the trace being processed. The operation of the Auditory Model Parameter Extractor is based on a Time-Domain Filter which emulates the behaviour of the Auditory Model to a given extent [1]. Besides, the Auditory Model Parameter Extractor reduces the redundancy present in the spectral characteristics of speech by bank-filtering the Speech Trace and detecting the envelope of its main spectral components present in the filtered outputs. These envelopes once re-sampled constitute the 6-dimensional output vector x_{1n}. This model will be described in more detail later. On the other hand, the block labelled as Phonetic Feature Encoder is based on the Associative Section of a Time-Delay Neural Network [2] trained for this purpose. This section associates every three time-consecutive sets of the spectral vectors with a Phonetic Code word x_{3n}, which is designed to give a more or less accurate description of the phonetic features present within the last speech fragment encoded, according with a Phonetic Encoding Scheme, which will also be described later. Other possibilities could also be used for the Spectral Extractor, as common LPC techniques, for example [3]. The main reasons to recommend a simplified Auditory Model are the following:

> The Auditory Model being used has been especially designed to assume a sensibly low computational cost. It may be supported by simple arithmetics and on a small area Integrated Circuit [4].
>
> It produces reasonably small Spectral Descriptors. As it has been shown [5], no more than 6-8 element vectors (x_{1n}) are enough give a complete spectral information of speech.
>
> Its frequency selectivity may be easily changed or adapted to different transmission patterns, by altering the adequate parameter values.
>
> It may be used to represent certain hearing disorders when treating deafness, or help in training cochlear implantees in improving their ability to identify spectrally time-varying sounds and to reproduce the associated articulatory phonetics.

In any case, the Auditory Model may be seen as a 6-section filter with an associated frequency selectivity, which reproduces the time-frequency behaviour of the Inner Ear at a reasonably low computational cost. The main characteristics of the whole system will be described in the next sections. The applications of the Speech Feature Encoder may be found in Computer Assisted Speech Training for the Hearing Impaired or in Language Learning.

AM Parameter Extractor

The Auditory Model Parameter Extractor has been designed to preserve the main characteristics of frequency selectivity within the Inner Ear. One important aspect of its features is the place effect, which assigns a given central frequency to each section, the first sections being sensitive to high frequencies, in contrast to the last sections, which are sensitive to low frequencies. Another important feature is that time relations are preserved within the different signals, as low frequencies appear later due to the propagation delay present in the system, which affect positively to phase relations. Finally, frequency selectivity has been emulated to certain extent [1], to maintain a given trade-off between high selectivity and low computational complexity. This means that the spectral peaks in the transfer functions for high frequencies are sligtly broader than in a sharp Inner Ear, but this effect affects only to the first section of the filter. On the other hand, the model reproduces reasonably well not only the peaks, but the flat sections in transfer functions, the relative broadening of the peaks as we move toward lower frequencies, and the phase relations. These features are granted because the set of Digital Filters reproducing the frequency selectivity of the Inner Ear has been deducted through the re-elaboration of a Transmission Line Model [6], preserving the main critical aspects to render it realizable using Digital Signal Processing Techniques. In fact, the model may be described by a set of coupled differential equations relating the differential pressure P on both sides of the Basilar Membrane, and the volumetric velocity inside one of the scales U in the domain of Laplace as follows:

$$\frac{\delta P}{\delta x} = - Z_s(s,x)\, U \qquad (1)$$

$$\frac{\delta U}{\delta x} = - Y_p(s,x)\, P \qquad (2)$$

where $Z_s(s,x)$ and $Y_p(s,x)$ are the impedance and admittance associated to the mechanics of the system, in terms of the frequency of Laplace s, and of the distance measured from the stapes towards the interior of the cochlea x. These equations must appear associated to the terminating conditions at both ends of the system. The termination at the stapes takes the form of an exciting generator of volumetric velocity and a loading admittance. The terminating condition at the oposite side (helicotrema) takes roughly the shape of a short circuit. Both conditions are transformed accordingly into the definitive set of Digital Filters. To accomplish a recursive solution amenable of a digital implementation, a corresponding transformation associated to the following set of equations is carried out:

$$P = Z_c\,(F + G) \qquad (3)$$

$$U = F - G \qquad (4)$$

$$Z_c = \sqrt[2]{\frac{Z_s}{Y_p}} \qquad (5)$$

where Z_c is the Characteristic Impedance of the Transmission Line, and F and G are the forward and backward components of the volumetric velocity U. The initial set of equations given by (1) and (2) may be thus transformed as follows:

$$\frac{\delta F}{\delta x} = - F\,(\gamma + \rho) - G\,\rho \qquad (6)$$

$$\frac{\delta G}{\delta x} = - F\,\rho + G\,(\gamma - \rho) \qquad (7)$$

where γ and ρ are two functions of the propagating characteristics of the media, known as the Propagation and Reflection Functions, given by expressions (8) and (9). The influence of the mechanical characteristics on the transmission process which will condition the frequency selectivity are to be found in these functions as follows:

$$\rho = \frac{1}{2Z_c}\,\frac{\delta Z_c}{\delta x} \qquad (8)$$

$$\gamma = \sqrt[2]{Z_s\,Y_p} \qquad (9)$$

Equations (6) and (7) may be numerically solved assuming that the Transmission Line is divided into a finite number of sections, inside one of which, the transmission functions are supposed to remain reasonable constant, compatibilizing the results between neighbour sections to grant the continuity in the dynamic variables. The resulting incremental structure of the Auditory Model may be seen in Fig. 2.

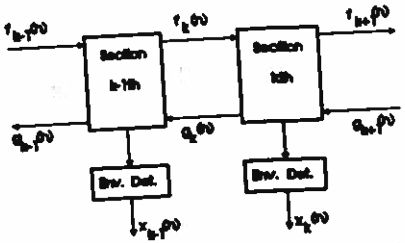


Fig. 2. Incremental structure of the model.

The incremental solution of equations (6) and (7) for one of the sections, when the continuity conditions are established may be written as follows:

$$F'_{k-1} = F_{k-1}\,\Gamma_k \qquad (10)$$

$$G_{k-1} = G'_{k-1}\,\Gamma_k \qquad (11)$$

$$F_k = (1 - \rho_k)\,F'_{k-1} - \rho_k\,G_k \qquad (12)$$

$$G'_{k-1} = \rho_k\,F'_{k-1} + (1 + \rho_k)\,G_k \qquad (13)$$

where Γ_k and ρ_k are functions of γ and ρ given by (8) and (9). The above set of equations may be reflected in the structure given in the flowgraph shown in Fig. 3.

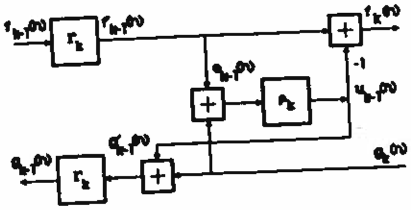


Fig. 3. Structure of one of the incremental sections.

For the sake of brevity, we will not extend too deeply into th mathematical details necessary to deduct the adequate expressions t implement each section, referring the interested reader to the preliminar work [4, 5, 6]. The most important details to be presented are relate with the structure rendered after rewritting the different propagatio functions from the domain of Laplace to that of the z-transform, usin for such the Bilinear Transformation [7]. The expression for ρ_k given (14), shows that a simple digital filter of second order, as the one in Fi 4 may easily support the reflection effects on the line:

$$\rho_k = c_{k0}\,\frac{1 + 2\,z^{-1} + z^{-2}}{a_{k0} + a_{k1}\,z^{-1} + a_{k2}\,z^{-2}} \qquad (14)$$

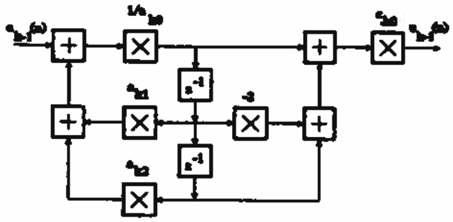


Fig. 4. Structure of the digital filter implementing ρ_k.

The case is not the same for the propagating function Γ_k, whose expression, given in (15) and (16) show an intrinsic irrationality:

$$\Gamma_k = e^{-\gamma_k \Delta x_k} \qquad (15)$$

$$\gamma^2_k = \frac{4 D l_k f^2_m (1 - z^{-1})}{a_{k0} + a_{k1} z^{-1} + a_{k2} z^{-2}} \qquad (16)$$

The synthesis of these functions as digital filters can not be done in a simple way. Series expansion techniques require a great amount of terms to reproduce the main spectral characteristics. A different approach was used in this case, by means of Pade's Approximation [8], which allowed us to represent the spectral behaviour of the module of γ_k from (16) using a third order polynomial in z as follows:

$$\gamma_k = \frac{a_{l0k} + a_{l1k} z^{-1} + a_{l2k} z^{-2} + a_{l3k} z^{-3}}{a_{m0k} + a_{m1k} z^{-1} + a_{m2k} z^{-2} + a_{m3k} z^{-3}} \qquad (17)$$

having used the following expression to approximate (15):

$$\Gamma_k \approx \frac{2 - \gamma_k \Delta x_k}{2 + \gamma_k \Delta x_k} \qquad (18)$$

The pair of expressions (17) and (18) allow us to extract a realizable structure for γ_k as given in Fig. 5.

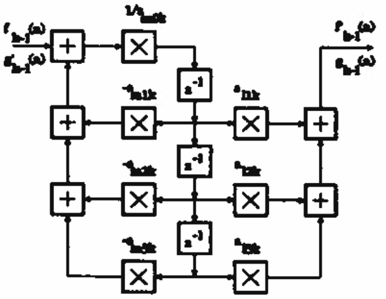


Fig. 5. Structure of the digital filter implementing γ_k.

where the sets of filter parameters $\{a_{l0-3k}\}$ and $\{a_{m0-3k}\}$ are given as functions of the general transmission and reflection properties of the structure, as is also the case with $\{a_{k0-3}\}$. By means of these structures, a 6-section digital filter may be easily built, its behaviour having been checked, as discussed later. The outputs of the filter may be obtained directly from the forward or the backward propagating signals, f_k or g_k, although, as shown in Fig. 8, these signals must be processed in order to obtain their envelopes, as the vectors to feed the Neural Network. The algorithm for envelope extraction is rather simple, and is based on a sliding average low-pass filter, because of its simplicity. Finally, the envelope of the signals at the different sections of the filter, is re-sampled at a rate of 1 sample every 2.5-5 msec., which comprises the highest frequencies present in the modulations of the formants of speech. In this way, a big reduction in information is carried out, the throughput changing from one sample each 100 µsec. to this rate, which substantially simplifies the task of the Neural Network. Each of the outputs of the 6-section filter configures an element of the 6-element vector x_{1n} used as input to the Phonetic Feature Encoder as explained before.

Neural Network

As it was mentioned, the Phonetic Feature Encoder is configured as the Associative Section of a Time-Delay Neural Network [2]. The structure of the TDNN in use is that of a Back-Propagation, and may be seen in Fig. 6.

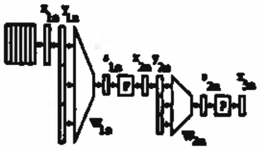


Fig. 6. Associative section of the TDBPNN being used.

The first input layer to the network is being built using 3 consecutive samples of x_{1n}, which means that the input vector is given as:

$$y_{1n} = \begin{pmatrix} x_{1n} \\ x_{1n-1} \\ x_{1n-2} \end{pmatrix} \qquad (19)$$

This time-composite vector is projected then on the 8x18 matrix of weights given by W_{1n}, which results in a first projection 8-dimensional vector s_{1n}. As this vector gives an unnormalized measure of the projection, a nonlinear function F{.} is used to restore the measure to the interval [-1,+1], giving as a result a new 8-dimensional vector x_{2n}. Now, the time composition may be applied again, in a similar way as in (19):

$$y_{2n} = \begin{pmatrix} x_{2n} \\ x_{2n-1} \\ x_{2n-2} \end{pmatrix} \qquad (20)$$

this new vector being of dimension 24. Projecting it on the 8x24 matrix of hidden weights W_{2n} we will obtain a new 8-dimensional vector s_{2n}, which when corrected to the interval [-1,+1] results in the projection measurement x_{3n}, which may be considered as a tentative Phonetic Code. The first projection with the weight matrix W_{1n} produced a first grouping of time-evolving Spectral Vectors in different clusters. The second projection with W_{2n} produced a re-grouping of these basic clusters into the definitive code words. More details on the operation of Back-Propagation Networks are considered irrelevant at this point, and can be found in the specialized literature [9]. The resulting tentative Phonetic Code, is a vector composed by 8 real numbers in the interval [-1,+1], which try to resemble a NRZ Code, in which Boolean 0's and 1's are represented by -1 and +1, respectively. As the real numbers in the elements of x_{3n} will approach these limits without completely matching them, the resulting code-word is considered "tentative". The proximity of an element to +1 or -1 indicates that the associated phonetic feature (for example voicing) is present or absent in the speech trace. In what follows, we will deal with the standard procedure used to train the Neural Network. Figure 7 illustrates the procedure, which starts with the recording (r_{jn}) and fragmenting (r_{jmn}) a set of reference sounds.

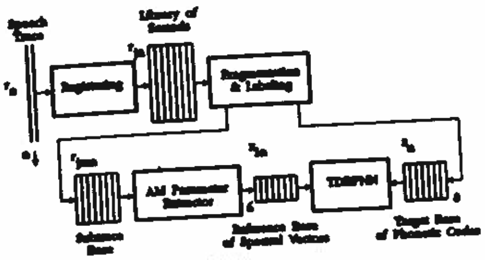


Fig. 7. Training procedure.

These fragments are filtered by the Auditory Model, and according with their nature, a label z_n corresponding to their Phonetic Code is associated with every x_{1n} produced by the Auditory Model. These data are stored in a Data Base, and used to be presented to the TDBPNN, according with a training strategy which will be discussed in a later section. Everytime a vector x_{1n} is presented to the TDBPNN, its associated code z_n is presented too, and the weights of the matrices W_{1n} and W_{2n} are changed according with standard Back-Propagation Rules [9], this operation being referred to as the Training Step. A set of 16 syllables of the kind consonant+vowel (C-V) or vowel+consonant+vowel (V-C-V) have been used in the training [10]. We refer to this set as the Training Set. The eficiency of the Training Process is measured by the number of times the Training Set has been completely presented to the Network for its convergence. After this, the weights W_{1n} and W_{2n} may be used by the Associative Section to establish the tentative Phonetic Codes.

Phonetic Encoding

The next topic we have to discuss is the establishment of the Phonetic Codes according with the Phonetic Features we want to consider. The Encoding Scheme is based on an 8-bit Hamming Distance Code related to the Phonetic Features of Spanish as shown in Table 1, and may be projected on an 8-dimensional hypercubic graph, in which a given node may be associated with a given sound.

Bit#	Feature	Comment
0	Oral/Nasal	b0=1 oral; b0=0 nasal
1	Voiced/Unvoiced	b1=1 voiced; b1=0 unvoiced
2	Degree of Closure	b2=0, b3=0 plosive
3		b2=0, b3=1 fricative
		b2=1, b3=0 vowel
		b2=1, b3=1 semiconsonant, glide
4	Articulatory Place	b4=0, b5=0 palatal
5		b4=0, b5=1 dental
		b4=1, b5=0 velar
		b4=1, b5=1 labial
6	In Vowels:	b6=1 rounded; b6=0 oral
	Round/Oval	
	In Consonants:	b6=1 frontal; b6=0 lateral
	Frontal/Lateral	
7	Multiple/Simple	b7=1 multiple; b7=0 simple

Table 1. Phonetic Encoding Scheme being used.

This results in a 256-node graph, in which each node is connected to other 8 neighbours, forming an 8-dimensional hypercube. Every node will have an eight-bit code attached, but not all these code words will correspond to a real sound, as it may be easily inferred. This means that most of the nodes in the hypercube will be empty. The code is based in Hamming Distance, in the sense that two sounds will form a Minimum Distance Pair if the Hamming Distance of their associated code words is the unity. The set of code words associated with the sounds present in the Spanish language is given in Fig. 8.

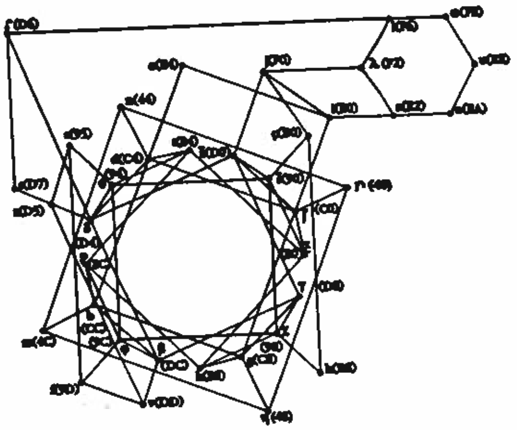


Fig. 8. Subgraph of the Encoding Scheme being proposed.

In it the phoneme and its corresponding hexadecimal code (between brackets) appear close to the corresponding node. The hexadecimal code is used as an abreviation of the binary code, taking b_0 as the most significant bit. It, may be seen that two sounds which form a Minimum Distance Pair, such as /t/ and /d/ are assigned code words (84) and (C4) which differ just in one bit, which in this case is b_1, corresponding to the feature voiced/unvoiced. It seems that this set of sounds is spontaneously organized in three different subsets. The first one is formed by 16 consonants, which include all the variations in the four bits coding voicing, articulation place and degree of closure, these being /p/t/c/k/b/d/j/g/ß/δ/z/γ/φ/θ/s/χ/. This subset is compact, in the sense that it may completely fill a 4-dimensional hypercube. The second subset is formed by other consonantal sounds related to the former, these being the subclusters /s/z/r/r/, /j/ç/h/, /m/n/n/η/ and /v/f/, which are colateral to the former ones. The third subset is formed by the cardinal vowels /e/i/a/o/u/ and other wowel-like consonants, as /w/l/λ/.

Results and Discussion

The structure being designed is rather complex to be trained and operated as a whole in a first approach, thus its different parts have been built and checked separately. In Fig. 9 a and b we present the results obtained when operating the AM Filter.

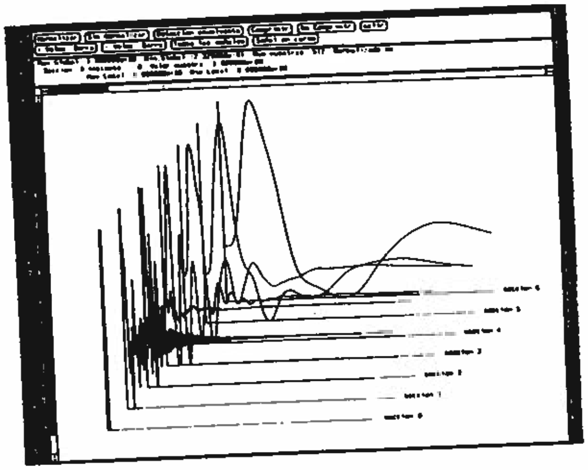


Fig. 9.a. Response of the AM Filter to an impulse excitation.

In Fig. 9.a, the response of this filter to a rectangular pulse 100 μsec.-wide is shown. Section 0 shows the excitation, and Sections 1-6 show the progression of the excitation through the model. The pulse broadens in time as we get deep into the Auditory Model, and the travel time, counted from section 0 to section 6 on the x axis relative to the peak maximum increases approximately exponentially with distance. The pulses are presented in relative amplitude (normalized to peak amplitude), the energy associated to the travelling wave becoming considerably smaller as the wave gets deeper (approx. -60 dB for Section 6 vs Section 0).

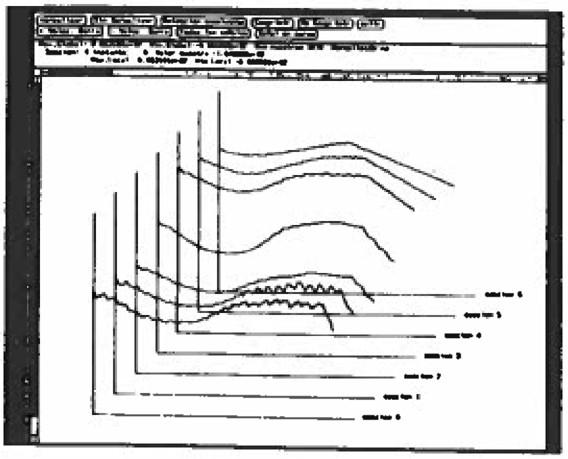


Fig. 9.b. Response of the AM filter to a speech trace. Envelope detection.

On the other hand, Fig. 9.b shows the effect of the filter on a speech trace, after filtering and envelope detection. As before, the x-axis represents time, and the y-axis relative amplitude. The distance along the cochlea is measured from section 0 (apex) to section 6 (belicotrema). The original utterance for which these traces were obtained corresponded to /ba/. Finally, in Table 2 the results for the training of the TDBPNN using two different structures are presented. The test subset was composed by structures VCV or CV using the first subset of 16-consonant sounds. The use of this subset allows to test the performance of the network under the worst conditions (minimum distance between sounds, thence minimum separability). Each speech trace was passed through the AM filter. The 6 output channels were envelope detected, re-sampled at a rate of one sample every 2.5 msec., and used to form the input vector x_{in}. Three of these vectors composed an 18-element vector y_{in} to be fed to the TDBPNN. The code word for training was fixed using a standard procedure for speech fragmentation [10]. The Network was trained using 935 training steps after which, a given speech trace was presented to the Network, and the Network's response was compared to the original code word. The first set of results, corresponding to a 18:9:8 network, showed a good performance, except in some fragments corresponding to /ka/ and /ta/. In the case of /ka/ shown in Table 2.a, the network anticipated the insertion point of the vowel.

Bit #	Output Value	Objective Value	Agreement
0	.9994	1	yes
1	.9957	0	no
2	.9999	0	no
3	.0000	0	yes
4	.0009	1	no
5	.0010	0	yes
6	.9999	0	no
7	.0006	0	yes

Table 2.a. The network anticipates the insertion of /a/.

Bit #	Output Value	Objective Value	Agreement
0	.9997	1	yes
1	.0224	0	yes
2	.0009	0	yes
3	.0000	0	yes
4	.0249	0	yes
5	.0000	1	no
6	.0008	0	yes
7	.0002	0	yes

Table 2.b. The network mistakes /t/ and /c/.

In other two cases, the network coded /ta/ as /ca/ (Table 2.b). Finally, after some more trials, the structure of the network was augmented to 18:11:8, and in this case, the match obtained was perfect using only 564 iteration steps with the same subset of sounds.

Acknowledgements

The present research is being carried out under Grants Nos. MIC88-0398-E and MIC90-1219-E from the Programa Nacional de Microelectrónica, Acción Concertada UPM A91-0020-02-78, Plan Regional de Investigación de la Comunidad Autónoma de Madrid C072/90, and Grant No. 0395/87 from the Nato Sci. Affairs Division.

References

[1] J. B. Allen, "Cochlear Modeling", IEEE ASSP Magazine, pp. 3-28, January 1985.

[2] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano and K. Lang, "Phoneme Recognition using time-delay Neural Networks", IEEE Trans. on ASSP, Vol. 37, No. 3, pp. 328-339, March 1989.

[3] T. Parsons, Voice and Speech Processing, McGraw-Hill, N. Y. 1987.

[4] V. Rodellar, P. Gómez, M. Hermida, A. Díaz and R. W. Newcomb, "A VLSI Architecture for the support of an Auditory Model for Hearing and Speech Processing", Proc. of the 33rd. Midwest Symp. on Circuits and Systems, Calgary, Alberta, Canada, pp. 787-790, August, 1990.

[5] P. Gómez, V. Rodellar, M. Hermida, A. Díaz and R. W. Newcomb, "VLSI Implementation of a Digital Filter for Hearing Aids", Proc. of the Digital Signal Processing-87 Conference, V. Capellini, Ed., North-Holland, Florence, Italy, pp. 341-345, September, 1987.

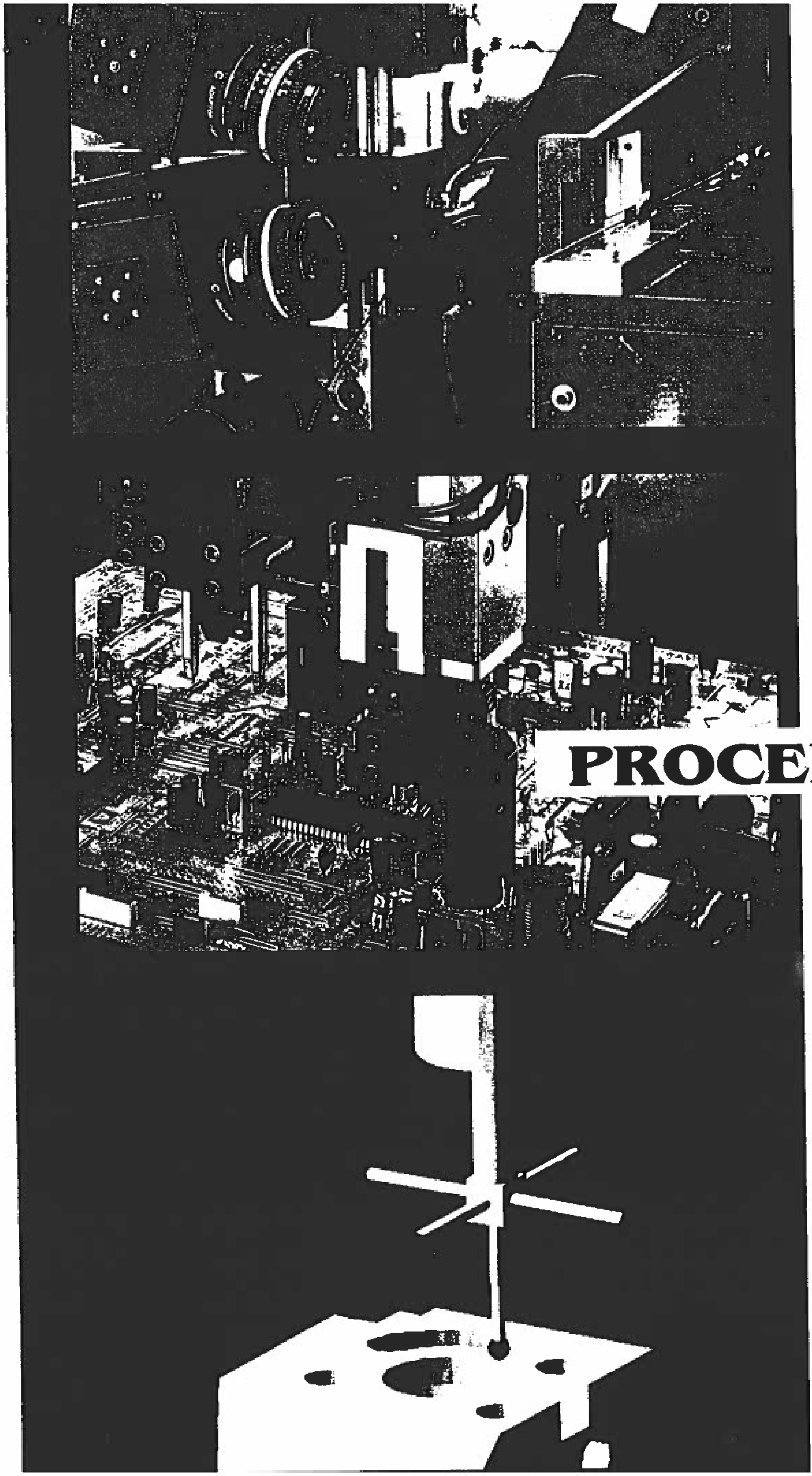
[6] P. Gómez, V. Rodellar, A. Díaz and M. Hermida, "Métodos Numéricos en el Desarrollo de Filtros Espacio-Temporales en Percepción Auditiva", Memorias del I Congreso en Métodos Numéricos en Ingeniería, Las Palmas, Canary Islands, Spain, pp. 249-256, June 1990.

[7] A. V. Oppenheim and R. W. Schafer, Digital Signal Processing, Prentice-Hall, Englewood Cliffs, N. J., 1975.

[8] N. K. Bose, Digital Filters: Theory and Applications, North-Holland, New York, 1985.

[9] T. Khanna, Foundations of Neural Networks, Addison-Wesley, Reading, Massachusetts, 1990.

[10] V. Rodellar, P. Gómez, C. García, M. Pérez and C. Gonzalo, "Performance of a Phonetic Encoding Scheme for Speech Recognition using Neural Networks", accepted for its presentation at PANEL'92 Conference, to be held at Las Palmas, Canary Islands, Spain, September 1992.



ICARCV '92

■

SECOND
INTERNATIONAL
CONFERENCE ON
AUTOMATION

■

ROBOTICS AND

■

COMPUTER
VISION

PROCEEDINGS

VOLUME 3 OF 3

ORGANISED BY

School of Electrical and
Electronic Engineering
Nanyang Technological University
Singapore

The Institution of Engineers,
Singapore