Speech Communication 9 (1990) 15–29
North-Holland

PHONETICALLY-BASED MULTI-LAYERED NEURAL NETWORKS FOR VOWEL CLASSIFICATION

Piero COSI
Centro di Studio per le Ricerche di Fonetica, C.N.R., Piazza Salvemini, 13, 35131 Padova, Italy

Yoshua BENGIO and Renato DE MORI
School of Computer Science, McGill University, 805 Sherbrooke Str. W., Montreal, Quebec, Canada H3A 2K6, and Centre de Recherche en Informatique de Montreal, 1550, de Maisonneuve Blvd. W., Montreal, Quebec, Canada H3G 1N2

Received 30 October 30 1989

Abstract. The vowel sub-component of a speaker-independent phoneme classification system will be described. The architecture of the vowel classifier is based on an ear model followed by a set of Multi-Layered Neural Networks (MLNN). MLNNs are trained to learn how to recognize articulatory features like the place of articulation and the manner of articulation related to tongue position.

Experiments are performed on 10 English vowels showing a recognition rate higher than 95% on new speakers. When features are used for recognition, comparable results are obtained for vowels and diphthongs not used for training and pronounced by new speakers. This suggests that MLNNs suitably fed by the data computed by an ear model have good generalization capabilities over new speakers and new sounds.

Zusammenfassung. Beschrieben wird eine Klassifizierungsstufe für Vokale als Teil eines sprecherunabhängigen Phonemklassifizierungssystems. Die Architektur dieses Vokalklassifikators basiert auf einem Ohrmodell, das von einem Satz mehrschichtiger neuronaler Netze gefolgt wird. Diese neuronalen Netze werden darauf trainiert, artikulatorische Merkmale, wie z.B. den Ort der Artikulation oder die Art der Artikulation – bezogen auf die Position der Zunge – zu erkennen.

Experimente mit 10 englischen Vokalen ergeben eine Erkennungsrate von mehr als 95% für neue, dem System bisher unbekannte Sprecher. Werden phonetische Merkmale für die Erkennung herangezogen, so lassen sich vergleichbare Resultate für solche Vokale und Diphthonge erreichen, die für das Training nicht verwendet oder von neuen Sprechern geäußert wurden. Dies legt nahe, daß mehrschichtige neuronale Netze, auf passende Weise mit den Ausgangsdaten eines Ohrmodells angesteuert, sich bei der Erweiterung dieser Aufgabe auf neue Sprecher oder neue Laute als gut geeignet erweisen.

Résumé. Nous présentons un système de classification de phonèmes indépendant du locuteur et appliqué aux voyelles. L'architecture du classificateur de voyelles est basée sur un modèle d'oreille suivi d'un ensemble de réseaux neuronaux à plusieurs couches (MLNN). Les MLNNs apprennent à reconnaître les traits articulatoires, par exemple le lieu et le mode d'articulation en relation avec la position de la langue.

Des expériences ont été effectuées sur 10 voyelles anglaises et montrent un taux de reconnaissance supérieur à 95% sur de nouveaux locuteurs. Lorsque les traits sont utilisés pour la reconnaissance, des résultats comparables sont obtenus pour des voyelles et des diphtongues qui n'ont pas été utilisées lors de l'apprentissage et prononcées par de nouveaux locuteurs. Ceci suggère que, pour des données calculées par un modèle d'oreille, les MLNNs présentent un bon pouvoir de généralisation pour de nouveaux locuteurs et de nouveaux sons.

Keywords. Speaker independent system, classification, recognition, multi-layered neural networks, articulatory features, vowels, ear model.

1. Introduction

In order to capture robust speech properties useful for coding speech for Automatic Speech Recognition (ASR) and characteristic of a large population of speakers, a Multi-Layered Neural Network (MLNN) system, specialized on phonetic feature hypothesisation, was proposed (Bengio et al., 1989).

The main motivation for such an approach was

that perceptually significant features, even if they possibly exhibit great differences in the space of acoustic parameters, exhibit small distortions in a perceptual space whose dimensions and metrics are not a-priori known and are difficult to specify.

Decision functions which have to describe the existence of phonetic properties in speech intervals can be learned by examples rather than being defined by properly conceived algorithms.

MLNNs are networks with an input layer of nodes, one or more hidden layers and an output layer whose nodes represent a coded version of the input. Nodes are connected by links. Weights are associated to links. All the links bringing a signal into a node contribute to the calculation of the excitation of that node. The excitation is the sum of the product of the weights of each link and the value of the output coming from the node from which the link carries its signal. The output of a node is a function of the node excitation. By choosing the link weights a large variety of classifiers can be designed having specific properties. Link weights can be obtained by a learning process. Learning can be supervised or unsupervised. When learning is supervised, the network input is fed by sets of patterns. Each set corresponds to a class of patterns that have to be coded with the same values appearing at the output nodes. The output nodes are clamped with the desired values and algorithms exist for computing the values of the link weights in such a way that the network codes the sets of input patterns as desired. These learning algorithms have a relevant generalization capability.

Recently, a large number of scientists are investigating and applying learning systems based on MLNNs. Definitions of MLNNs, motivations and algorithms for their use can be found in Rumelhart et al. (1986); Plaut and Hinton (1987); Hinton and Sejnowski (1986). Theoretical results have shown that MLNNs can perform a variety of complex functions (Plaut and Hinton, 1987). Applications have also shown that MLNNs have interesting generalization performances capable of capturing information related to pattern structures as well as characterization of parameter variation (Bourlard and Wellekens, 1987; Watrous and Shastri, 1987). Algorithms exist for MLNNs with proven mathematical properties that allow learning to be competitive and to focus on the properties that make different patterns belonging to different classes. Furthermore, in MLNNs the knowledge about a set of competing classes (in our case Speech Units or phonemes) is distributed in the weights associated to the links between nodes.

If we interpret each output of the classifier as representing a phonetic property, then an output value can be seen as a degree of evidence with which that property has been observed in the data.

Two important research problems can be studied with such an approach. The first problem investigates the possibility of learning the features of each phoneme only in some phonetic contexts and rely on the generalization capability of a network for generating correct hypotheses about phonemes in contexts that have not been used for learning. The second problem is similar to the first one but deals with the possibility of learning all the required features and using them for correctly hypothesizing phonemes that have not been used for learning. As for the second problem it is necessary to code the output with some features in order to learn features and to represent each class (phoneme or speech unit) as a combination of features.

In this paper we focused only on the speaker independent vowel classification problem, thus the system being described has to be viewed as the vowel-specific sub-component of the more complete architecture initially mentioned (Bengio et al., 1989), whose final aim was to globally solve the speaker independent phoneme recognition problem.

For the vowel classification problem we have chosen as main features the place of articulation and the manner of articulation related to tongue position. The reason is that these features are well characterized by physical parameters that can be measured or estimated. Phoneticians have characterized vowels and other sounds by discretizing place of articulation and manner of articulation related to tongue position, which are in nature continuous acoustic parameters. We have inferred an MLNN for each feature and we have discretized each feature with five qualitative values, namely PL_1, ..., PL_i, ..., PL_5 for the place and

MN_1, ..., MN_j, ..., MN_5 for the manner. We have used ten vowels pronounced by many speakers in a fixed context for training the two networks, each vowel being represented by one of the PL_i and one of the MN_j. In order to describe all the vowels of American English with enough redundancy, we have introduced another network with two outputs, namely T = tense and L = lax. We have also inferred the weights of a network with ten outputs, one for each vowel. The performances of this network have shown that it is possible to obtain an excellent generalization of the parameters when training is performed on a limited number of male and female speakers using data that make evident acoustic properties having little variance across speakers when the same vocalic sound is pronounced. The performances of this network have also been used as reference.

Tests have always been performed with new speakers. The first test consists in pronouncing the same vowels in the same context as in the data used for learning. This test is useful for comparing the results obtained with a mathematical model of the ear (Seneff, 1984; 1985; 1986; 1988) with those obtained with the more popular Fast-Fourier Transformation (FFT). This test is also useful for assessing the capabilities of the network learning method in generalizing knowledge about acoustic properties of speakers pronouncing vowels. The second test has the objective of recognizing vowels through features. This test has been useful for investigating the power of the networks with respect to possible confusions with vowels not used for learning. The third experiment consists in attempting to recognize new vowels pronounced by new speakers in order to investigate the capability of the networks to detect the same features used for learning, but integrated into sounds that were not used for learning. This generalization capability was verified with 8 new sounds pronounced by 20 new speakers. Without any learning on the new sounds, but just using expectations based on phonetic knowledge on the composing features and their time evolutions, an error rate of 7.5% was found.

Section 2 describes the details of the ear model. Section 3 describes the details of an MLNN trained for the speaker-independent recognition of 10 English vowels. Section 4 describes the details of the recognition experiments. Section 5 describes new networks trained to learn articulatory properties and their performances in the speaker-independent recognition of vowels and diphthongs not used for learning.

2. The ear model

Cochlear transformations of speech signals result in an auditory neural firing pattern significantly different from the spectrogram, a popular time-frequency-energy representation of speech.

In recent years basilar membrane, inner cell and nerve fiber behaviour have been extensively studied by auditory physiologists and knowledge about the human auditory pathway has become more accurate. A number of studies have been accomplished and a considerable amount of data has been gathered in order to characterize the responses of nerve fibers in the eighth nerve of the mammalian auditory system using tone, tone complexes and synthetic speech stimuli (Delgutte, 1980; Delgutte and Kiang, 1984a, b, c, d; Young and Sachs, 1979; Sachs and Young, 1980; Miller and Sachs, 1983; Sinex and Geisler, 1983; Kiang et al., 1965).

Phonetic features probably correspond in a rather straightforward manner to the neural discharge pattern with which speech is coded by the auditory nerve. For these reasons, even an ear model that is just an aproximation of physical reality appears to be a suitable system for identifying those aspects of the speech signal that are relevant for recognition.

The computational scheme proposed in this paper for modelling the human auditory system is derived from the one proposed by Seneff (1984; 1985; 1986; 1988). The overall system structure which is illustrated in Fig. 1 includes three blocks: the first two of them deal with peripheral transformations occurring in the early stages of the hearing process, while the third one attempts to extract information relevant to perception. The first two blocks represent the periphery of the hearing system. They are designed using knowledge of the rather well-known responses of the corresponding human auditory stages (Sinex and Geisler, 1983; Kiang et al., 1965). The third unit attempts to

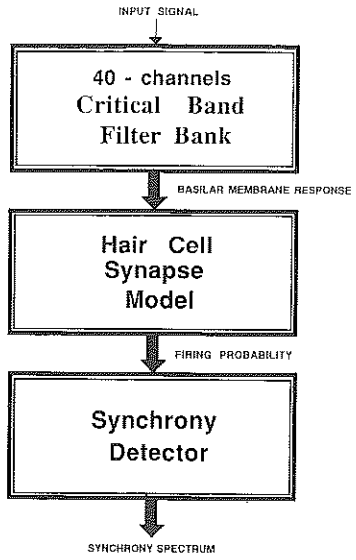

Fig. 1. Block-diagram of the ear model.

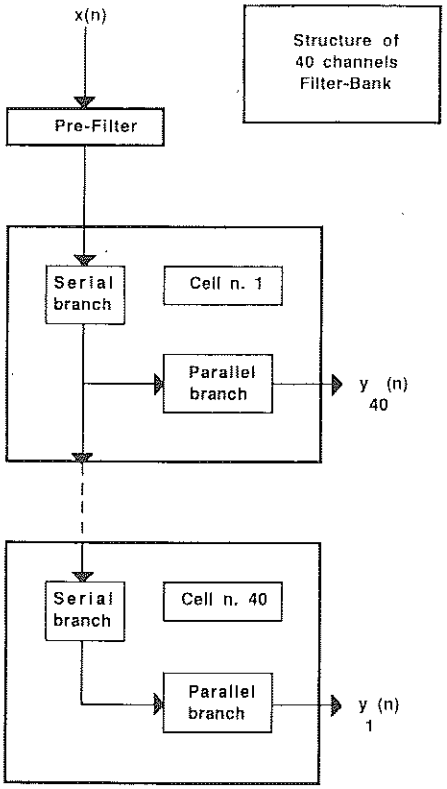

Fig. 2. Filter bank structure.

apply a useful processing strategy for the extraction of important speech properties like spectral lines related to formants (Seneff, 1984; 1985).

The speech signal, band-limited and sampled at 16 kHz, is first pre-filtered through a set of four complex zero pairs to eliminate the very high and very low frequency components. The signal is then analyzed by the first block, a 40-channel critical-band linear filter bank. Fig. 2 shows the block diagram of the filter bank which was implemented as a cascade of complex high frequency zero pairs with taps after each zero pair to individual tuned resonators. Filter resonators consist of a double complex pole pair corresponding to the filter center frequency (CF) and a double complex zero pair at half its CF. Filters, whose transfer functions are illustrated in Fig. 3, were designed in order to optimally fit physiological data like those observed by Kiang et al. (1965). Frequencies and bandwidths for zeros and poles of each filter were designed almost automatically by an interactive technique developed by Seneff and described in her thesis (1985).

The second block of the model, whose block diagram is shown in Fig. 4, is called the hair cell synapse model, it is nonlinear and is intended to capture prominent features of the transformation from basilar membrane vibration, represented by the outputs of the filter bank, to probabilistic response properties of auditory nerve fibers. The outputs of this stage, in accordance with Seneff (1988), represent the probability of firing as a function of time for a set of similar fibers acting as a group.

Four different neural mechanisms are modeled in this nonlinear stage (Seneff, 1988). The transfer function of a transduction module which half-wave rectifies its input is shown in Fig. 5. The rectifier is applied to the signal in order to simulate the high level distinct directional sensitivity present in the inner hair cell current response. The short-term adaptation which seems due to the neurotransmitter release in the synaptic region between the inner hair cell and its connected nerve fibers is simulated by the so-called "membrane model", which was conceived following the work by Goldor (1985). The mathematical equations describing the mechanism which influences the evolution of the neurotransmitter concentra-

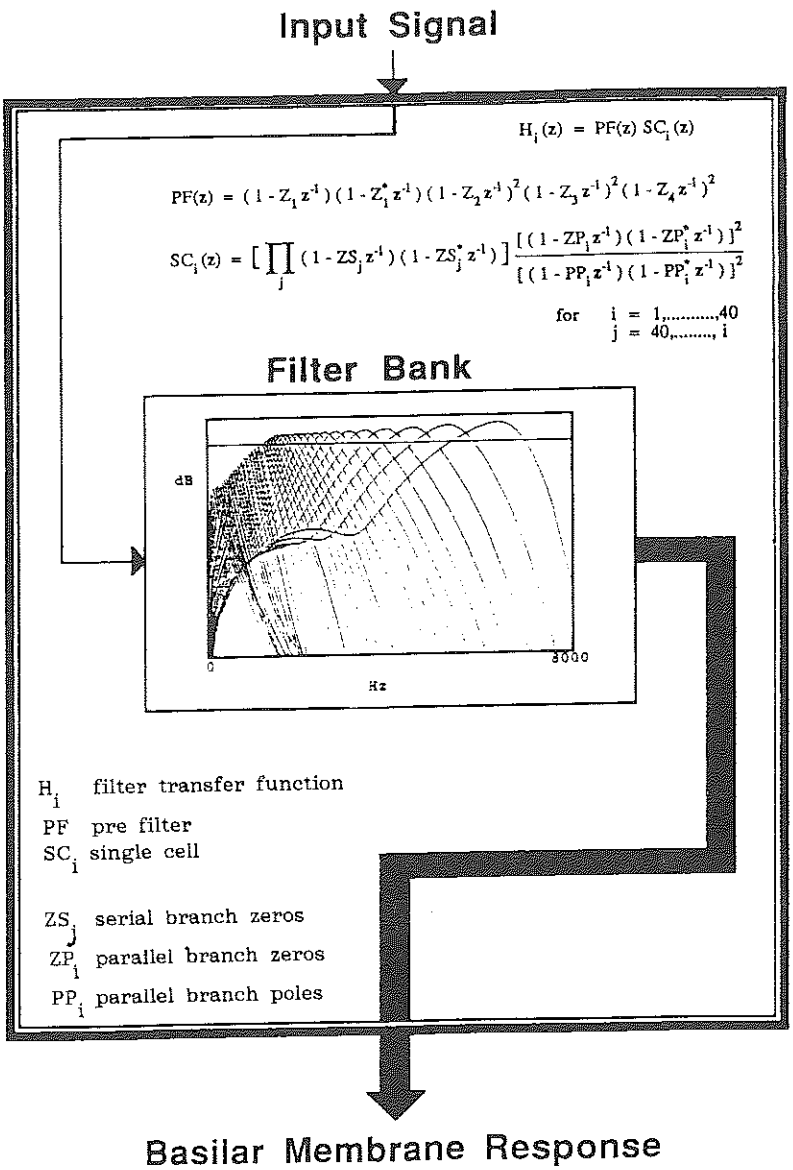

Fig. 3. Filter frequency responses.

tion inside the cell membrane are given in Fig. 6. The third unit represents the observed gradual loss of synchrony in nerve fiber behaviour as stimulus frequency is increased and it is implemented by a simple low-pass filter. The last unit is called "Rapid Adaptation". It performs "Automatic Gain Control" and implements a model of the refractory phenomenon of nerve fibers.

The third and last block of the ear model is the synchrony detector, which implements the known "phase locking" property of the nerve fibers. It enhances spectral peaks due to vocal tract resonances. Auditory nerve fibers tend to fire in a "phase-locked" way responding to low frequency periodic stimuli, which means that the intervals between nerve fibers tend to be integral multiples of the stimulus period. Consequently, if there is a "dominant periodicity" (a prominent peak in the frequency domain) in the signal, with the so

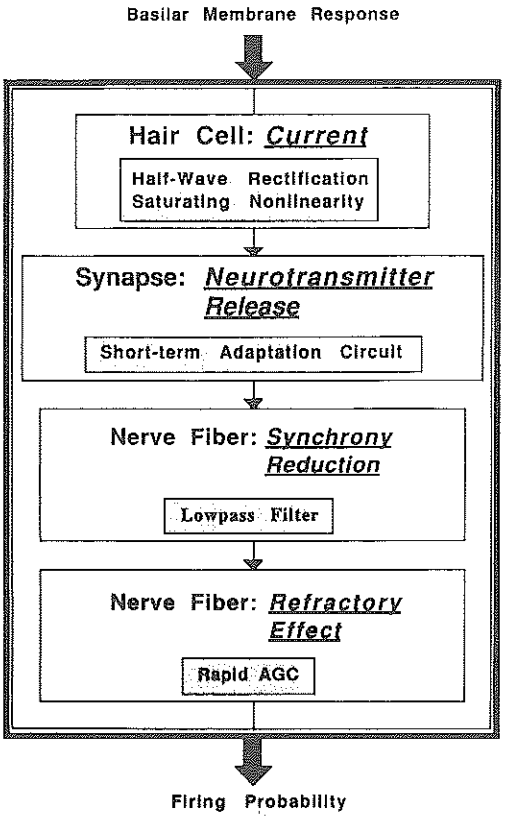

Fig. 4. Hair cell synapse module.

called Generalized Synchrony Detector (GSD) processing technique (Seneff, 1984; 1985), only those channels whose central frequencies are closest to that periodicity will have a more prominent response. The block diagram of the GSD, as applied to each channel is shown in Fig. 7.

3. Speaker-independent recognition of ten vowels in fixed contexts

A first experiment was performed for speaker-independent vowel recognition. The experimental environment is described in Fig. 8. The purpose was that of training an MLNN capable of discriminating among 10 different American-English vowels represented with the ARPABET by the following VSET:

VSET: {iy, ih, eh, ae, ah, uw, uh, ao, aa, er} (1)

The interest was to investigate the generalization capability of the network with respect to inter-speaker variability. Some vowels (ix, ax, ey, ay, oy, aw, ow) were not used in this experiment because we attempted to recognize them through features learned by using only VSET.

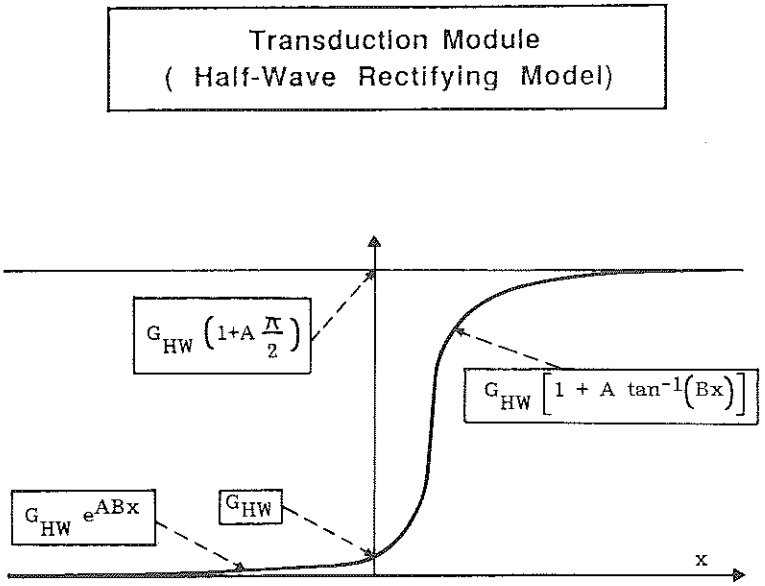

Fig. 5. Transduction module.

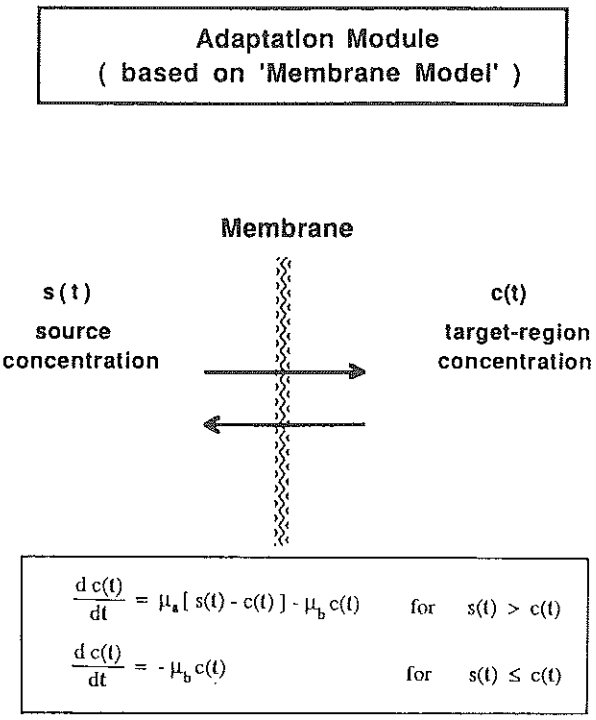

Fig. 6. Adaptation module.

Speech material consisted of 5 pronunciations of ten monosyllabic words containing the vowels of VSET. The words used are those belonging to the WSET defined in the following:

WSET: {BEEP, PIT, BED, BAT, BUT, BOOT, PUT, SAW, FAR, FUR} (2)

The signal processing method used for this experiment is the one described in the previous Section. The output of the Generalised Synchrony Detector (GSD) was collected every 5 ms and represented by a 40-coefficients vector. This type of output is supposed to retain most of the relevant speech spectral information.

The GSD output of the vocalic part of the signal was sent to an MLNN. The performances of an MLNN depend on its architecture, on the method used for learning and for producing an output but also on the type of input and the way the output is coded. In order to capture the essential information of each vowel it was decided to use 10 equally-spaced frames per vowel for a total of 400 network input nodes. A single hidden layer was used with a total of 20 nodes. Ten output

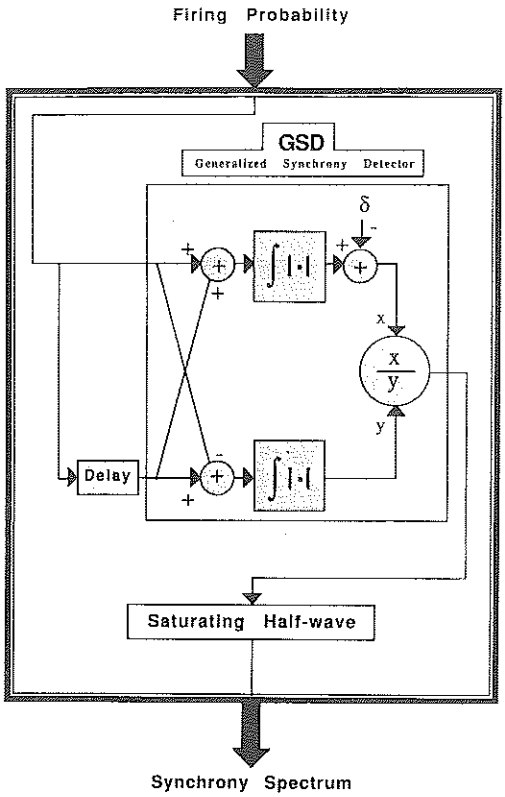

Fig. 7. Generalized Synchrony Detector (GSD) module.

nodes were introduced, one for each vowel as shown in Fig. 9.

Vowels were automatically singled out by an algorithm proposed in De Mori et al. (1985), and a linear interpolation procedure was used to reduce to 10 the variable number of frames per vowel (the first and the last 20 ms of the vowel segment were not considered in the interpolation procedure). The resulting 400 (40 spectral coefficients per frame × 10 frames) spectral coefficients became the inputs of the MLNN.

The Error Back Propagation Algorithm (EBPA) was used for training. EBPA was recently introduced (Rumelhart et al., 1986) for a class of non-linear MLNNs. These networks are made of connected units. The networks used for the experiments described in this paper are feed-forward (non-recurrent) and organized in layers. A weight is associated to each of the (unidirectional) connection between two nodes. Input nodes are on layer 0 and have no input connec-

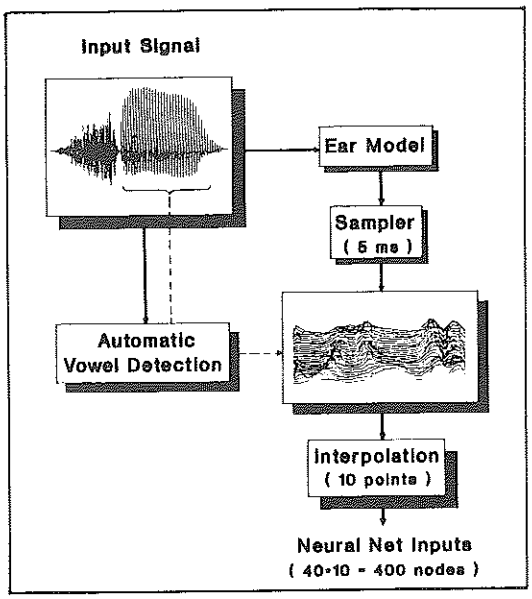

Fig. 8. Experimental environment for vowel recognition.

tions. Output nodes have no output connections and are on the last layer. Nodes which are neither input nor output units are called hidden units.

The network computes a non-linear function from the input units to the output units. The architecture of the network determines which functions it can compute. A typical architecture used in the experiments described in this paper is shown in Fig. 9. The nodes of the network compute a sigmoid function of the weighted sum of their inputs. Any output value takes values between 0 and 1 according to the following function:

$$Y_i = f\left(\sum_{j=1}^{J} Y_j W_{ij}\right), \tag{3}$$

with:

$$f(x) = 1 / (1 + \exp(-x)). \tag{4}$$

The sum in (3) is over the J units with an outgoing connection to unit i, the output value of this unit is Y_i. The weight W_{ij} is associated with the link between the output of unit i and the input of unit j.

With EBPA the weights are computed iteratively in such a way that the network minimizes a square error measure defined over a set of training input/output examples. These examples be-

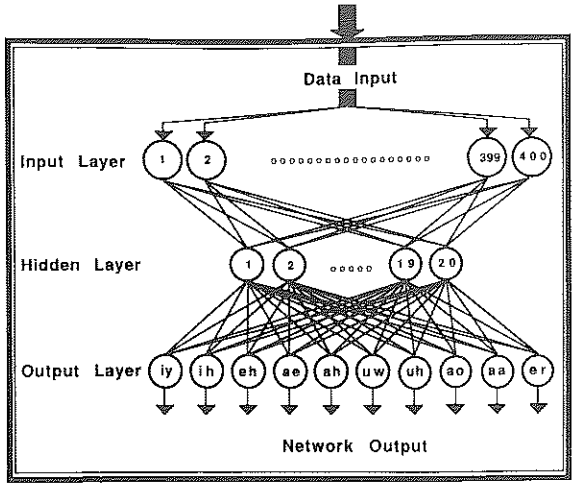

Fig. 9. Structure of the neural network used for vowel recognition.

long to a training set defined as follows:

$$\text{training set} = \left\{ \bigcup_{k=1}^{K} (\text{IN}_k, \text{OUT}_k) \right\}, \tag{5}$$

where IN_k is an input pattern and OUT_k is a desired output pattern that can be represented by the following vector of values: (OUT_{k1}, OUT_{k2}, ..., OUT_{kM}). The minimized square error measure is:

$$\text{E} = 0.5 * \left(\sum_{k=1}^{K} \sum_{m=1}^{M} (\text{OUT}_{km} - Y_m(\text{IN}_k))^2 \right), \tag{6}$$

where k varies over the training set of examples and m varies over the M nodes on the output layer. $Y_m(\text{IN}_k)$ is the value of the mth output node computed by the MLNN when IN_k is applied at the input layer.

EBPA uses gradient descent in the space of weights to minimize E. The basic rule for updating link weights is:

$$\Delta W = - \text{learning\_rate} * \partial E / \partial W, \tag{7}$$

where $\partial E/\partial W$ can be computed by back-propagating the error from the output units as described in Rumelhart et al. (1986).

In order to reduce the training time and accelerate learning, various techniques can be used. The classical gradient descent procedure modifies the weights after all the examples have been pre-

sented to the network. This is called batch learning. However, it was experimentally found, at least for pattern recognition applications, that it is much more convenient to perform on-line learning, i.e. updating the weights after the presentation of each example. Batch learning provides an accurate measure of the performance of the network as well as of the gradient $\partial E/\partial W$. These two parameters can be used to adapt the learning rate during training in order to minimize the number of training iterations. In our experiments we used various types of acceleration techniques. The most effective one consisted in switching from on-line learning to batch learning and vice-versa, depending on the behaviour of the gradient and the evolution of performances.

In contrast with classical Hidden Markov Models, where output probabilities are maximized independently for each class, MLNNs can learn from presentations of examples from all the classes that have to be recognized with the possibility of emphasizing what makes classes different and different examples of the same class similar.

The voices of 13 speakers (7 male, 6 female) were used for learning with 5 samples per vowel per speaker. The voices of seven new speakers (3 male, 4 female) were used for recognition with 5 samples per vowel per speaker. Speech material was recorded in a quiet office room and data acquisition was performed with a 12 bits A/D converter at 16 kHz sampling frequency. Learning was accomplished with about 60 training cycles with zero error rate on the training set. As for the test set the network produces degrees of evidence varying between zero and one, candidate hypotheses could be ranked according to the corresponding degree of evidence.

The confusion matrix represented in Table 1 was obtained. In 95.7% of the cases, correct hypotheses were generated with the highest evidence, in 98.5% of the cases correct hypotheses were found in the top two candidates and in 99.4% of the cases in the top three candidates. Similar results were obtained by repeating the experiments with different starting random weight configurations. These results were compared with those obtained by performing the same experiment, in the same experimental conditions as described in Figs. 8 and 9, but using a more classical acoustic front-end. By using 40 coefficients produced by an FFT-based mel-scale 40-channel filter bank instead of using data from the ear model, the recognition rate was about 87% showing a 9% significant reduction. In this particular task the use of the ear model coefficients showed better recognition performance than the use of classical FFT-based coefficients. Moreover further evidence needs to be provided (e.g. using other classical acoustic parameters like LPC coefficients or mel-scale cepstrum coefficients) in order to conclusively prove that the proposed perception-based auditory analysis in general performs better than other acoustic production-based front-end for automatic speech recognition. The use of the ear model allowed to produce spectra with a limited number of well-defined spectral lines. This

Table 1
Performance of the vowel recognition system using the entire vocalic segment

	BAT /ae/	BED /eh/	BEEP /iy/	BOOT /uw/	BUT /ah/	FAR /aa/	FUR /er/	PIT /ih/	PUT /uh/	SAW /ao/	THE /ax/
BAT /ae/	34	0	0	0	0	1	0	0	0	0	0
BED /eh/	0	35	0	0	0	0	0	0	0	0	0
BEEP /iy/	0	0	35	0	0	0	0	0	0	0	0
BOOT /uw/	0	0	0	34	0	0	0	0	1	0	0
BUT /aw/	2	0	0	0	33	0	0	0	0	0	0
FAR /aa/	0	0	0	0	1	34	0	0	0	0	0
FUR /er/	1	0	0	0	1	0	33	0	0	0	0
PIT /ih/	0	0	0	0	0	0	0	35	0	0	0
PUT /uh/	0	0	0	0	1	0	1	0	32	0	1
SAW /ao/	0	0	0	0	0	4	0	0	0	31	0

Table 2
Vowel representation using phonetic features

	Place of articulation					Manner of articulation						
	Back		Central	Front		Low		Mid	High		Law	Tense
	PL_1	PL_2	PL_3	PL_4	PL_5	MN_1	MN_2	MN_3	MN_4	MN_5	L	T
/ac/ BAT	0	0	0	1	0	1	0	0	0	0	0	1
/eh/ BED	0	0	0	1	0	0	1	0	0	0	1	0
/iy/ BEEP	0	0	0	0	1	0	0	0	0	1	0	1
/uw/ BOOT	0	1	0	0	0	0	0	0	0	1	0	1
/ah/ BUT	0	0	1	0	0	0	1	0	0	0	1	0
/aa/ FAR	1	0	0	0	0	1	0	0	0	0	0	1
/er/ FUR	0	0	1	0	0	0	0	1	0	0	0	1
/ih/ PIT	0	0	0	1	0	0	0	0	1	0	1	0
/uh/ PUT	0	1	0	0	0	0	0	0	1	0	1	0
/ao/ SAW	1	0	0	0	0	1	0	0	0	0	1	0
/ax/ THE	0	0	1	0	0	0	0	1	0	0	1	0

represents a good use of speech knowledge according to which formants are vowel parameters with low variance. The use of male and female voices allowed the network to perform an excellent generalization with samples from a limited number of speakers.

Encouraged by the results of this first experiment, other problems appeared worth investigating with the proposed approach. The problems are all related to the possibilities of extending what has been learned for ten vowels to recognize new vowels. An appealing generalization possibility relies on the recognition of vowel features. By learning a set of features in a set of vowels, new vowels can be characterized just by different combination of the learned features. Features like the place of articulation and the manner of articulation related to tongue position are good descriptors of the vowel generation system. It can be expected that their values have low variance when different speakers pronounce the same vowel.

4. The recognition of phonetic features

The same procedure introduced in the previous section was used for learning three networks, namely $MLNNV_1$, $MLNNV_2$ and $MLNNV_3$. These networks have the same structure as the one introduced in the previous section with the only difference that they have more outputs. $MLNNV_1$ has five additional outputs corresponding to the five places of articulation PL_1, ..., PL_i, ..., PL_5. $MLNNV_2$ has five new outputs, namely MN_1, ..., MN_j, ..., MN_5. $MLNNV_3$ has two additional outputs, namely T = tense and L = lax. The ten vowels used for this experiment have the features defined in Table 2.

After having learned the weights of the three networks, the outputs corresponding to the individual vowels were ignored and confusion matrices were derived only for the outputs corresponding to the phonetic features. An error corresponds to the fact that an output has a degree of evidence higher than the degree of the output corresponding to the feature possessed by the vowel whose pattern has been applied at the input.

The confusion matrix for the features is shown in Table 3. The overall error rates are 4.6%, 5.7% and 5.4% respectively for the three sets of features. As in the case of speaker-independent recognition of ten vowels in fixed contexts previously described, similar results were obtained repeating the present experiments with different starting random weight configurations. Error rates were always zero after a number of training cycles (between 60 and 70) of the three networks. Several rules can be conceived for recognizing vowels

Table 3
Performances in the recognition of features

Pronounced features	Place of articulation					Manner of articulation						
	Back		Central	Front		Low		Mid	High		Lax	Tense
	PL_1	PL_2	PL_3	PL_4	PL_5	MN_1	MN_2	MN_3	MN_4	MN_5	L	T
BAT	1	0	0	34	0	35	0	0	0	0	0	35
BED	0	0	0	35	0	0	35	0	0	0	35	0
BEEP	0	0	0	0	35	0	0	0	0	35	0	35
BOOT	0	32	0	3	0	0	0	0	1	34	2	33
BUT	0	0	30	5	0	8	27	0	0	0	32	3
FAR	34	0	1	0	0	35	0	0	0	0	8	27
FUR	0	2	29	4	0	1	1	31	2	0	0	35
PIT	0	0	0	35	0	0	0	0	35	0	35	0
PUT	0	35	0	0	0	0	2	0	33	0	33	2
SAW	35	0	0	0	0	5	30	0	0	0	31	4

through their features. The most severe rule is that a vowel is recognized if all the three features have been scored with the highest evidence. With such a rule, 313 out of 350 vowels are correctly recognized corresponding to 89.4% recognition rate.

In 28 cases, combinations of features having the highest score did not correspond to any vowel, so a decision criterion had to be introduced in order to generate the best vocalic hypothesis. It is important to consider as an error the case in which the features of a vowel not contained in the set defined by (1) receive the highest score. Considering these vowels as well as the vowels in (1) an error rate of 2.6% was found. This leads to the conclusions that an error rate between 2.6% and 10.6% can be obtained depending on the decision criterion used for those cases for which the set of features having the highest membership in each network do not correspond to any vowel.

An appealing criterion consists in computing the centers of gravities of the place and manner of articulation using the following relation:

$$CG = \left(\sum_{i=1}^{5} i\mu(i-1) \right) / \sum_{i=1}^{5} \mu(i-1) \ . \qquad (8)$$

Let CGP and CGM be respectively the center of gravity of the place and manner of articulation. A degree of "tenseness" has been computed by dividing the membership of "tense" by the sum of the memberships of "tense" and "lax". Each sample can now be represented as a point in a three-dimensional space having CGP, CGM and the degree of tenseness as dimensions. Euclidean distances are computed for those sets of features not corresponding to any vowel with respect to the points representing theoretical values for each vowel. With centers of gravity and Euclidean distance an error rate of 7.2% was obtained.

Another interesting criterion consists in introducing a subjective probability for a feature defined as the ratio of the feature membership over the sum of the memberships of the other features. For example for feature PL_i a probability π_i is defined as follows:

$$\pi_i = \mu(PL_i) / \left(\sum_{k=1}^{5} \mu(PL_k) \right) . \qquad (9)$$

The probability of a vowel is then defined as the product of the subjective probabilities of the features of the vowel. As the denominator of the probability of a vowel is the same for all the vowels, the vowel with the highest probability is the one with the highest product of the evidences of its features.

By smoothing each membership with its neighbours and multiplying the memberships of the features of each vowel an error rate of 8.8% was obtained. The error rate obtained with gravity centers is not far from the one obtained in the

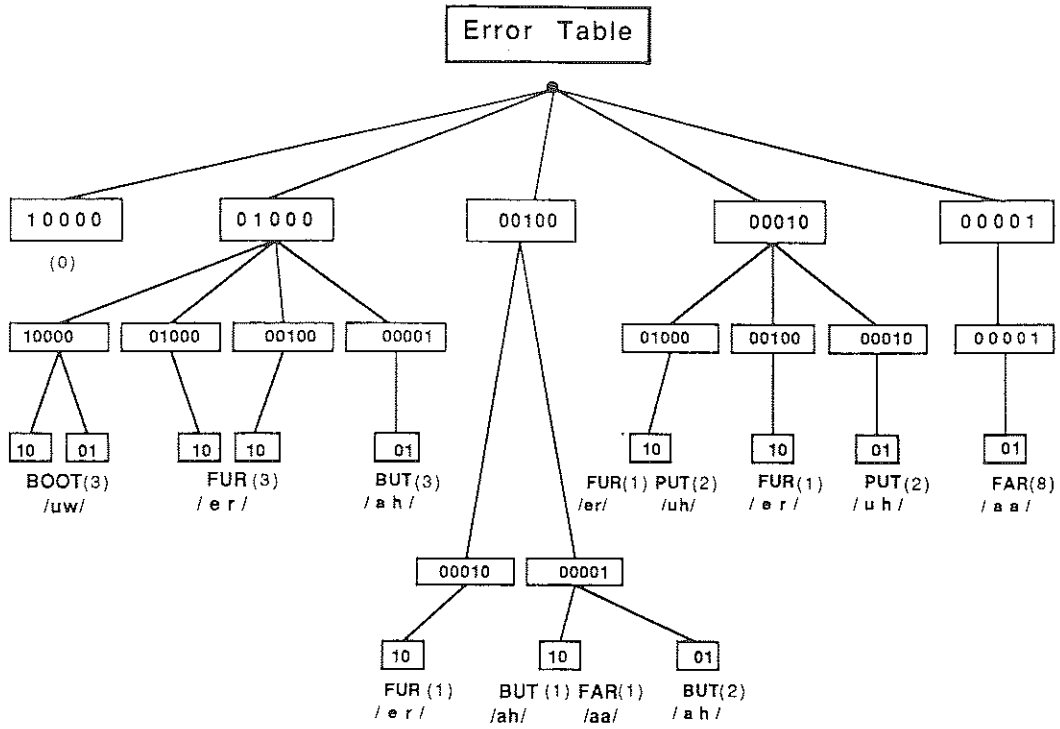

Fig. 10. Error tree for the vowels classified with a code that does not correspond to any vowel.

previous section with ten vowels. In this case the possibility of error was higher because the system was allowed to recognize feature combinations for all the vowels of American English.

For those cases for which the features which reached the maximum evidence did not define a set corresponding to any vowel of American English an error analysis was made. The conclusions of these analyses are shown by the error tree in Fig. 10 (number of errors are indicated between parentheses). They suggest that most of the errors were systematic (PL_2 confused with PL_4 and MN_2 confused with MN_4).

Based on the tree in Fig. 10, features for maximum evidence can be used as a code for describing an unknown vowel. When this code does not correspond to any acceptable vowel, it can be mapped into the right one corresponding to the true features of the vowel when the wrong code always corresponded to the same vowel. When the wrong code corresponds to more than one vowel a procedure is executed that computes Euclidean distances on gravity centers. With this criterion, which is derived from the test data, an error rate of 3.2% can be obtained. This error rate cannot be used for establishing the performances of the feature networks because it corrects some errors by recoding the memberships using a function that has been learned by analyzing the test data. Nevertheless, it suggests that feature-based MLNNs may outperform a straightforward phoneme-based MLNN if successive refinements are performed using more than one training set. In fact, after a few experiments, interpretations for the codes PL = 00001, MN = 00001 and PL = 01000, MN = 10000 can be inferred and applied to successive experiments leading to a correct recognition rate close to 96%.

5. Recognition of new vowels and diphthongs

In order to test the generalization power of the networks for feature hypothesization a new experiment was performed involving 20 new speakers from 6 different mother tongues (English,

French, Spanish, Italian, German and Vietnamese) pronouncing letters in English.

According to other experimental works on vowel recognition (Leung and Zue, 1988), there are 13 vowels in American English and 3 diphthongs. The vowels and diphthongs that were not used in the previous experiments belong to the NSET:

{/ax/(the), /ey/(A), /ay/(I), /oy/(boy), /aw/(bough), /ow/(O)}. (10)

The vowel /ax/ does not exhibit transitions in time of the parameters CGM and CGP so its recognition was based on the recognition of the expected features as defined in Table 2. The other five elements of NSET exhibit evolutions of CGP and CGM in the time domain. For this reason, it was decided to use such evolutions as basis for recognition. Furthermore, the sequences /yu/ and /way/ (corresponding to the pronunciation of letters U and Y) were added to NSET in order to have a larger set of classes for testing the generalization capabilities of the system.

Although Hidden Markov Models could be and will be conceived for modeling time evolution of centers of gravities, a crude classification criterion was applied in this experiment.

Recognition was purely based on time evolutions of place and manner of articulation according to descriptions predictable from theory or past experience and not learned by actual examples. The centers of gravities CGP and CGM were computed every 5 ms and vector-quantized using five symbols for CGP according to the following alphabet:

$$\Sigma_1 = \{F, f, C, b, B\}. \quad (11)$$

F represents "strong front". Analogously, the following alphabet was used for quantizing the manner of articulation:

$$\Sigma_2 = \{H, h, M, I, L\}. \quad (12)$$

M represents "strong high".

Coding of CGP and CGM is based on values computed on the data of the ten vowels used for training the network.

Transitions of CGP and CGM were simply identified by sequences of pairs of symbols from Σ_1 and Σ_2. Fig. 11 gives an example of the time evolutions of CGP and CGM for letters A (ey) and Y (way) together with their codes.

The following regular expressions were used to characterize the words containing the new vowels and diphthongs:

A: (f, h)* (F, H)*
I: (b + C, l)* (f + F, h + H)*
O: (b + B, l)* (b + B, h + H)*
/oy/: (B, l)* (f + F, h + H)* (13)
/aw/: (C, l)* (b + B, h + H)*
U: (f + F, h + H)* (b + B, h + H)*
Y: (b + B, h + H)* (C, l + L)* (f + F, h + H)*.

The asterisk means in theory "any repetition" but in our case a minimum of two repetitions was required. The symbol "+" here means logical disjunction while a concatenation of terms between parentheses means a sequence in time. A short sequence with intermediate symbols was tolerated in transitions B-F, L-H, and vice versa.

For each new word, twenty samples were available based on the idea that speaker-independent recognition has to be tested with data from new speakers and repetition of data from the same speaker is not essential.

The errors observed were quite systematic. For /ax/, 1 case was confused with /ah/. For /ey/ (letter A), three errors were observed, all corresponding to a sequence (f, h)*, meaning that the transition from /eh/ was not detected. For /ow/ (letter O), three errors were observed corresponding to the sequence (b, l)*, meaning that the transition from /oh/ was not detected which may correspond to an intention of the speaker. Three errors were found for /oy/ confused with /ay/ and two errors for /aw/ confused with /ow/. For the other transitions, the expectations were always met. The repeatability of the describing strings was remarkable. A total of 12 errors out of 160 test data was found corresponding to an error rate of 7.5%. This provides evidence that a system made of an ear model followed by MLNNs trained to recognize normalized values of the place and manner of articulation reliably generates feature hypotheses about vowels and diphthongs not used for training.

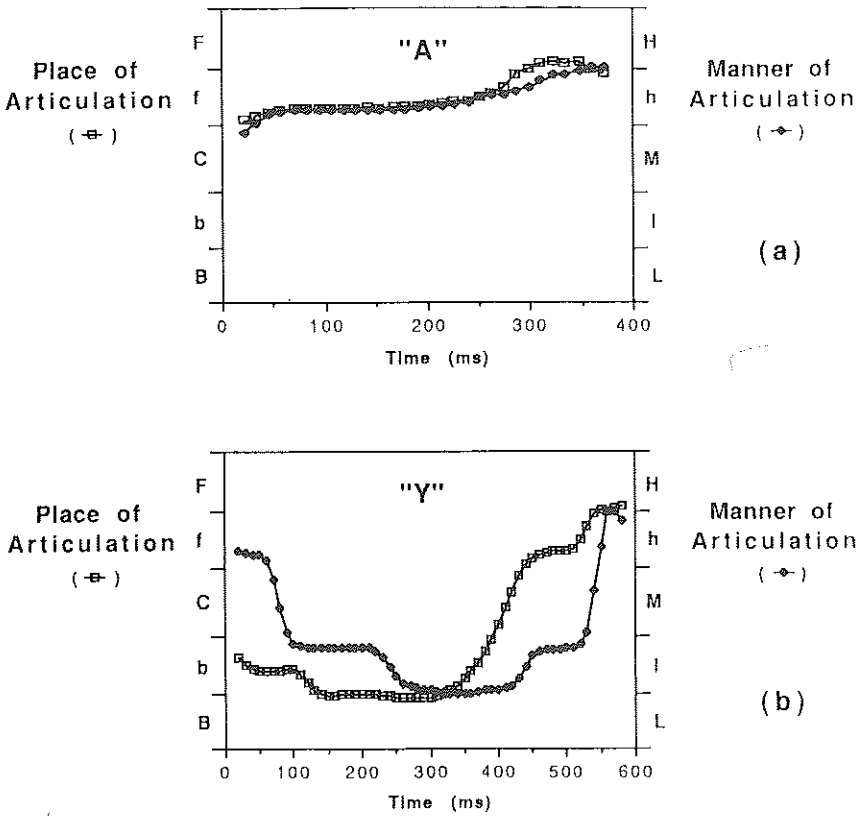

Fig. 11. Time evolution of CGM and CGP.

6. Conclusions

The work reported in this paper shows that a combination of an ear model and multi-layer neural networks allows us to obtain an effective generalization among speakers in coding vowels. The results obtained in the speaker-independent recognition of ten vowels add a contribution that justifies the interest in the investigation of the use of MLNNs for ASR (Leung and Zue, 1988; Waibel et al., 1988).

Furthermore, training a set of MLNNs on a number of well distinguishable vowels allows us to obtain a very good generalization on new vowels and diphthongs if recognition is based on features.

By learning how to assign degrees of evidence to articulatory features it is possible to estimate normalized values for the place and manner of articulation which appear to be highly consistent with qualitative expectations based on speech knowledge.

Speech coders that produce degrees of evidence of phonetic features can be used for fast lexical access, for word spotting, for recognizing phonemes in new languages with limited training or for constraining the search for the interpretation of a sentence.

Effective learning and good generalizations can be obtained using a limited number of speakers in analogy with what humans do.

Performance models of the time evolutions of evidences or derived parameters like CGP and CGM can be made using Hidden Markov Models. Degrees of evidences can be used as "pseudo-probabilities", parameters and evidences can be vector-quantized or their continuous densities can be estimated for the models.

The error-back propagation algorithm seems to be a suitable one for learning weights of inter-

node links in MLNNs. A better understanding of the problems related to its convergence is a key factor for the success of an application. The choice of the number of MLNNs, their architecture, the coding of their input and output are also of great importance, especially for generalization.

The computation time of the system proposed in this paper is about 150 times real-time on a SUN 4/280. The system structure is suitable for parallelization with special purpose architectures and accelerator chips. It is not unrealistic to expect that with a suitable architecture, such a system could operate in real-time.

Acknowledgements

This work was supported by the Natural Science and Engineering Council of Canada (NSERC). The Centre de Recherche en Informatique de Montreal (CRIM) kindly provided computing time on its facilities.

References

Y. Bengio, R. Cardin, P. Cosi, R. De Mori and E. Merlo (1989), "Speech coding with multilayer networks", in *Neurocomputers* (NATO ASI Series), ed. by F. Fogelman (Springer Verlag, Berlin)

D.E. Rumelhart, G.E. Hinton and R.J. Williams (1986), "Learning internal representation by error propagation", in *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, Vol. 1 (Cambridge, MA: MIT Press) pp. 318–362.

D.C. Plaut and G.E. Hinton (1987), "Learning sets of filters using back propagation", *Computer Speech and Language*, Vol. 2, pp. 35–61.

G.E. Hinton and T.J. Sejnowski (1986), "Learning and relearning in Boltzmann machines", in *Parallel Distributed Processing: Exploration in the Microstructure of Cognition*, Vol. 1 (Cambridge, MA: MIT Press) pp. 282–317.

H. Bourlard and C.J. Wellekens (1987), "Multilayer perceptron and automatic speech recognition", *Proc. IEEE first Intern. Conf. Neural Networks (ICNN), San Diego, June 1987*, pp. IV407–IV416.

R.L. Watrous and L. Shastri (1987), "Learning phonetic features using connectionist networks", *Proc. 10th Intern. Joint Conf. on Artificial Intelligence (IJCAI), 1987*, pp. 851–854.

S. Seneff (1984), "Pitch and spectral estimation of speech based on an auditory synchrony model", *Proc. IEEE Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), San Diego, CA, 1984.*

S. Seneff (1985), "Pitch and spectral analysis of speech based on an auditory synchrony model", *RLE Technical Report*, No. 504 (Cambridge, MA: MIT Press).

S. Seneff (1986), "A computational model for the peripheral auditory system: Application to speech recognition research", *Proc. IEEE Intern. Conf. Acoust. Speech, Signal Proc. (ICASSP), Tokyo, 1986*, pp. 37.8.1–37.8.4.

S. Seneff (1988), "A joint synchrony/mean-rate model of auditory speech processing", *J. of Phonetics, Vol. 16, No. 1, pp. 55–76.*

B. Delgutte (1980), "Representation of speech-like sounds in the discharge patterns of auditory-nerve fibers", J. Acoust. Soc. Am., Vol. 68, No. 3, pp. 843–857.

B. Delgutte and N.Y.S. Kiang (1984a), "Speech coding in the auditory nerve: I. Vowel-like sounds", *J. Acoust. Soc. Am.*, Vol. 75, No. 3, pp. 866–878.

B. Delgutte and N.Y.S. Kiang (1984b), "Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds", *J. Acoust. Soc. Am.*, Vol. 75, No. 3, pp. 897–907.

B. Delgutte and N.Y.S. Kiang (1984c), "Speech coding in the auditory nerve: III. Voiceless fricative consonants", *J. Acoust. Soc. Am.*, Vol. 75, No. 3, pp. 887–896.

B. Delgutte and N.Y.S. Kiang (1984d), "Speech coding in the auditory nerve: IV. Sounds with consonant-like dynamic characteristics", *J. Acoust. Soc. Am.*, Vol. 75, No. 3, pp. 897–907.

E.D. Young and M.B. Sachs (1979), "Representation of steady-state vowels in the temporal aspects of the discharge pattern of populations of auditory nerve fibers", *J. Acoust. Soc. Am.*, Vol. 66, No. 5, pp. 1381–1403.

M.B. Sachs and E.D. Young (1980), "Effects of nonlinearities on speech encoding in the auditory nerve", *J. Acoust. Soc. Am.*, Vol. 68, No. 3, pp. 858–875.

M.I. Miller and M.B. Sachs (1983), "Representation of stop consonants in the discharge patterns of auditory-nerve fibers", *J. Acoust. Soc. Am.*, Vol. 74, No. 2, pp. 502–517.

D.G. Sinex and C.D. Geisler (1983), "Responses of auditory-nerve fibers to consonant-vowel syllables", *J. Acoust. Soc. Am.*, Vol. 73, No. 2, pp. 602–615.

N.Y.S. Kiang, T. Watanabe, E.C. Thomas and L.F. Clark (1965), *Discharge Patterns of Single Fibers in the Cat's Auditory-Nerve Fibers* (Cambridge, MA: MIT Press).

R.S. Goldhor (1985), "Representation of consonants in the peripheral auditory system: A modeling study of the correspondence between response properties and phonetic features", *RLE Technical Report*, No. 505 (Cambridge, MA: MIT Press).

R. De Mori, P. Laface and Y. Mong (1985), "Parallel algorithms for syllable recognition in continous speech", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. PAMI-7, No. 1, 1985, pp. 56–69.

H.C. Leung and V.W. Zue (1988), "Some phonetic recognition experiments using artificial neural nets", *Proc. Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), New York, 1988*, pp. 422–425.

A. Waibel, T. Hanazawa and K. Shikano (1988), "Phoneme recognition: Neural networks vs. hidden Markov models", *Proc. Intern. Conf. Acoust., Speech, Signal Process. (ICASSP), New York, 1988*, paper 8.S3.3.