On the Generalization Capability of Multi-Layered Networks in the Extraction of Speech Properties

Renato DE MORI*, Yoshua BENGIO*, and Piero COSI**

* School of Computer Science,
McGill University,
3480 University St.,
Montreal, Que., Canada H3A 2A7

** Centro di Studio per le Ricerche
di Fonetica,
C.N.R.,
Via G. Oberdan, 10,
35122 PADOVA, ITALY

Abstract

The paper describes a speech coding system based on an ear model followed by a set of Multi-Layer Networks (MLN). MLNs are trained to learn how to recognize articulatory features like the place and manner of articulation. Experiments are performed on 10 English vowels showing a recognition rate higher than 95% for 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 MLNs suitably fed by the data computed by an ear model have good generalization capabilities over new speakers and new sounds.

1. Introduction

Coding speech for Automatic Speech Recognition (ASR) can be performed with Multi-Layer Networks (MLN). This approach is interesting because it captures relevant speech properties useful for ASR at the stage of coding. A large number of scientists is currently investigating and applying learning systems based on MLNs [Rumelhart *et al.* 1986, Plout & Sejnowski 1987]. Applications have shown that MLNs have interesting generalization behaviour capable of capturing information related to pattern structures as well as characterization of parameter variation [Bengio *et al.* 1989, Bourlard & Wellekens 1987, Watrous & Shastri 1987]. Algorithms exist for MLNs with proven mathematical properties that allow learning to be discriminative and to focus on the properties that permit the separation of patterns belonging to different classes.

If we interpret each output of the coder 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. An important research problem can be studied with such an approach; it deals with the possibility of learning all the required features and their use in correctly hypothesizing phonemes that were not used for learning. As a first attempt at solving this problem, we have chosen to represent vowels and diphthongs with the place of articulation and the manner of articulation related to tongue position since 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 MLN for each feature and discretized each feature with five qualitative values, namely PL1 ,....PLi,...,PL5 for the place and MN1 ,...MNj,....MN5 for the manner.

Various tests have been performed, always with new speakers. The first test consists of 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 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 confusion with vowels not used for learning. The third experiment is an attempt to recognize new vowels pronounced by new speakers. This showed how the MLNs generalize to combinations of values of features not seen in the training set. This generalization capability was verified with 8 new sounds pronounced by 20 new speakers. Without any learning of the new sounds, but just using expectations based on phonetic knowledge on the composing features and their time evolution, an error rate of 7.5% was found.

2. Training of the MLNs

The Error Back Propagation Algorithm (EBPA) was used for training. EBPA was recently introduced [Rumelhart *et al.* 1986] for a class of non-linear MLNs. The networks used for the experiments described in this paper are feedforward (non-recurrent) and organized in layers. A weight is associated with the (unidirectional) connections between two nodes.

With EBPA the weights are computed iteratively in such a way that the network minimizes a cost C (the sum of the square of the differences between output unit values and target output values for the training examples). The EBPA uses gradient descent in the space of weights to minimize the error:

$$\Delta W <= \text{momentum}*\Delta W - \text{learning\_rate}*\partial C/\partial W \qquad (1$$

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 presented 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. When using on-line learning, one has to be careful in choosing the order of presentation of examples. We presented examples of each class one after the other, going through all the different classes. Batch learning provides an accurate measure of the performance of the network as well as of the gradient 3E/3W. 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 simplest one is to add a "momentum" term to the weight update rule [Rumelhart *et al.* 1986]. More interesting techniques involve adapting the learning rate as a function of 1) the evolution of the cost (deviation from target output), and 2) the evolution of the direction of the gradient. In other words, when the cost is improving sufficiently or when the gradient tends to point in the same direction from cycle to cycle, the learning rate should be increased. A further refinement consists of using (and adapting) a different learning rate for each connection. To improve learning time, a subset S of the training examples is used for training: those that produce errors. Once every few learning iterations on this subset, all the patterns are tested in order to decide which ones will go in S. Of course, with this technique the global cost and the global gradient are not evaluated at each iteration, so it is more suited to on-line learning. Another way to reduce the learning time is to divide the problem into subproblems which are as independent as possible and assign those subproblems to subnetworks: this is modularization. In our case we used separate networks for place of articulation and for manner of articulation. Outputs of small modules can be combined heuristically according to our knowledge of the functions they perform based on speech theory. They can also be combined to form a bigger network using Waibel's (1988) glue units. Another technique we used to train big networks was first to minimize the error using a simple architecture (e.g., no hidden units). Once the simple network has been trained, it can be augmented with more hidden units (and many more weights) in order to reduce the error significantly. This strategy provided significant gains in training time in some cases.

3. Experimental results

3.1 Speaker-Independent recognition of ten vowels in fixed context

A first experiment was performed for speaker-independent vowel recognition. The purpose was that of training an MLN 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} (2)

Our interest was in investigating the generalization capability of the network with respect to inter-speaker variability. Some vowels and diphthongs (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.

Speech material consisted of 5 pronunciations by 19 speakers of 10 monosyllabic words containing the vowels of VSET. The tokens from 12 of the speakers (6 males, 6 female) were used for training (600 tokens) and and the remaining ones (from 3 males, 4 females) were used for tests (350 tokens). Data acquisition was performed with a 12 bit A/D converter at a 16 kHz sampling frequency. The words used are those belonging to the WSET defined in the following:

WSET: (3)
{BEEP,PIT,BED,BAT,BUT,BOOT,PUT,SAW,FAR,FUR}

Two signal processing methods were used for this experiment. One was based on 128 point FFT spectra reduced to energy values in 40 bands, the other used the output of the Generalized Synchrony Detector (GSD), represented by a 40-coefficient vector. In both cases spectra were sampled every 5 ms. Spectral values were normalized to lie in the range 0 to 1. 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. Best results were obtained with a single hidden layer with a total of 20 nodes. Ten output nodes were introduced, one for each vowel.

Vowels were automatically singled out by an algorithm proposed in [De Mori *et al* 1985] and a linear interpolation procedure was used to obtain 10 equally-spaced 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 x 10 frames) spectral coefficients became the inputs of the MLN.

Training was stopped when the MLN made 0 errors on the training set. For the test set, the network produces degrees of evidence varying between zero and one, hence candidate hypotheses can be ranked according to the corresponding degree of evidence.

The error rates on the test set were 4.3% with the ear model and 13.0% with the FFT. The reason for such a difference is probably due to the fact that the use of the ear model allowed us to produce spectra with a limited number of well defined spectral lines. This represents a good use of speech knowledge according to which formants are vowel parameters with low variance.

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.

3.2. Recognition of phonetic features

The same procedure introduced in the previous section was used for learning in three networks, namely MLNV1, MLNV2 and MLNV3. These networks have the same structure as the one introduced previously, the only difference being that they have more outputs. MLNV1 has five additional outputs corresponding to the five places of articulation PL1,...,PLi.....PL5. MLNV2 has five new outputs, namely MN1,...,MNj,...MN5. MLNV3 has two additional outputs, namely T=tense and U=lax. The ten vowels used for this experiment have the features defined in Table 1. Training the first 10 outputs to correspond to the 10 vowels improved generalization over nets with only feature outputs. This might be explained as follows. A network with features as target outputs was slower to train and did not generalize as well on new speakers than a network with vowels as target outputs. This counterintuitive result might be explained by the possibility that the regions in the input space defined by the feature values are not as easily drawn (e.g., including several disjoint regions) as the regions in the input space defined by the vowel discrimination. Note that the acoustic definition of these features is imposed on the network based on speech production theory and might not represent the best choice of representation. Hence using 10 additional outputs representing the vowels forced the creation of hidden units that were useful to perform the vowel discrimination. These hidden units in turn could be used to produce the target feature values. The resulting network still does not generalize as well as the vowel discrimination network on new speakers, but it does generalize on new vowels.

		Place of Articulation					Manner of Articulation						
		Back		Central		Front	Low		Mid		High	Lax	Tense
ARPABET		PL1	PL2	PL3	PL4	PL5	MN1	MN2	MN3	MN4	MN5	L	T
/ae/	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	0	1	0	0	0	1	0
/ax/	THE	0	0	1	0	0	0	0	1	0	0	1	0

Features

Table 1: Vowel representation using phonetic features.

After having learned the weights of the three networks with the same methodology as for the first experiment, confusion matrices were derived only for the outputs corresponding to the phonetic features. An error was determined by comparing the feature value with the highest degree of evidence with the correct feature.

The overall error rates on the test sets were 4.57%, 5.71% and 5.43% respectively for the three sets of features. Error rates on the training set 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 through their features. The most severe rule is that a vowel is recognized if all three features have been scored with the highest evidence. With such a rule, 313 out of 350 vowels are correctly recognized corresponding to 10.5% error 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. In 2.57% of the examples, the three features corresponded to a wrong vocalic hypothesis. This leads to the conclusion that an error rate between 2.57% and 10.57% can be obtained depending on the decision criterion used for those cases in which the set of features having the highest membership in each network do not correspond to any vowel.

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

$$CG = \left(\sum_{i=1}^{5} i \, \mu(i-1) \right) / \left(\sum_{i=1}^{5} \mu(i-1) \right) \qquad (4),$$

where ji(i) is the degree of evidence for feature level i obtained by the MLNs. 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 from choices of feature values not corresponding to any vowel to the points representing theoretical values for each vowel. With centers of gravity and Euclidean distance an error rate of 7.24% was obtained. The error rate obtained with gravity centers is not far from that obtained with ten vowels but is higher because the system was allowed to recognize feature combinations for all the vowels of American English.

3.3. Recognition of new phonemes

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 isolated letters and words in English.

The MLNs, trained as described in section 3, have as input 10 frames of 40 parameters each. During training these frames were chosen so as to span the length of a stable vowel segment. In the test experiments described below, the 10 frames are 10 consecutive frames each representing 5 ms of speech. The MLN thus has an input window of 50 ms which scans the input speech data with a 5 ms step.

According to other experimental work on vowel recognition [Leung & 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:

NSET : {/ax/(the), /ey/(A), /ay/(I), /oy/(boy),
/aw/(bough), /ow/(O)} (5)

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 1. The other five elements of NSET exhibit evolution of CGP and CGM in the time domain. For this reason, it was decided to use such evolutions as the basis for recognition.

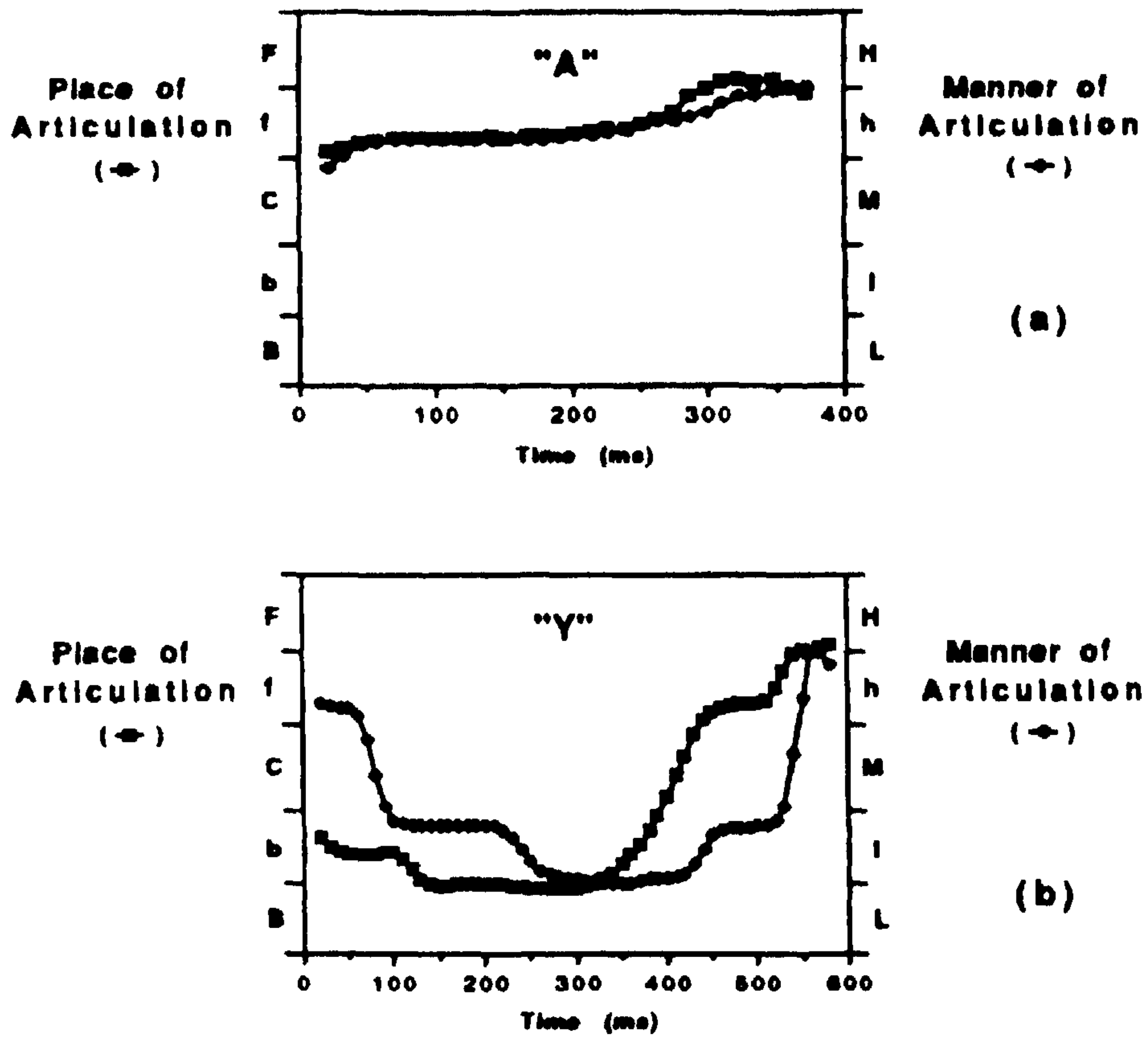

Figure 1: Time evolution of CGM and CGP for pronunciations of the letters "a" and "y" by speakers from the test set.

Although Hidden Markov Models could be and will be conceived for modelling the time evolution of

feature values, a crude classification criterion was applied in this experiment. Recognition was based purely on time evolution of place and manner of articulation according to descriptions predictable from theory or past experience and not learned by actual examples. The centers of gravity CGP and CGM were computed every 5 ms and vector-quantized using five symbols for CGP according to the following alphabet:

$$\Sigma 1 = \{F,f,C,b,B\} \qquad (6),$$

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

$$\Sigma 2 = \{H,h,M,l,L\} \qquad (7),$$

where H represents "strong high". Coding of CGP and CGM is based on values computed from 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 Σ1 and Σ2. Figure 1 shows definitions of Σ1 and X2 and gives an example of the time evolution 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:

$$\begin{array}{ll} A: & (f,h)^*(F,H)^* \\ I: & (b+C,l)^*(f+F,h+H)^* \\ O: & (b+B,l)^*(b+B,h+H)^* \\ /oy/: & (B,l)^*(f+F,h+H)^* \\ /aw/: & (C,l)^*(b+B,h+H)^* \end{array} \qquad (8)$$

In theory the asterisk means "any repetition", but in our case a minimum of two repetitions was required. The symbol V 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, as well as in initial and final transients.

For each vowel and diphthong, 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 token 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/. The repeatability of the describing strings was remarkable. Performance can be improved with a more rigorous word recognition algorithm.

4 Conclusions

The work reported in this paper shows that a combination of an ear model and multi-layer networks results in 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 MLNs for ASR [Leung & Zue 1988, Waibel *et al,* 1988].

Furthermore, training a set of MLNs with a sma'l number of training speakers on a number of well distinguishable vowels resulted in a very good generalization on new speakers (with a variety of accents) as well as 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.

The error-back propagation algorithm seems to be a suitable one for learning weights of internode links in MLNs. 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 MLNs, their architecture, the coding of their input and output and the learning strategy 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 informatique de Montr6al (CRIM) kindly provided computing time on its facilities.

References

[Bengio et al. 1989] Bengio, Y., Cardin, R., De Mori, R., Merlo, E., "Programmable Execution of Multi-Layered Networks for Automatic Speech Recognition", Communications of the Association for Computing Machinery, February 1989, vol. 32, no. 2, pp. 195-199.

[Bourlard & Wellekens 1987] H. Bourlard and C.J. Wellekens, "Multilayer perceptron and automatic speech recognition", IEEE first International Conference on Neural Networks, San Diego, June 1987, pp. IV407-IV416.

[De Mori *et al.* 1985] R. De Mori, P. Laface and Y. Mong Y.f "Parallel algorithms for syllable recognition in continuous speech", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. PAMI-7, N. 1,1985, pp. 56-69.

[Leung & Zue 1988] H. C. Leung and V. W. Zue, "Some phonetic recognition experiments using artificial neural nets". Proc. International Conference on Acoustics, Speech and Signal Processing, New York, N.Y., 1988, pp. 422-425.

[Plout & Hinton 1987] D.C. Plout & G.E. Hinton, "Learning sets of filters using back propagation", Computer Speech and Language, 1987, vol. 2, pp.35-61.

[Rumelhart *et al.* 1986] D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning internal representation by error propagation", Parallel Distributed Processing: Exploration in the Microstructure of Cognition, vol. 1, MIT Press, 1986, pp.318-362.

[Seneff 1988] S. Seneff, "A joint synchrony/mean-rate model of auditory speech processing", Journal of Phonetics, January 1988, pp. 55-76.

[Waibel *et al.* 1988] A. Waibel, T. Hanazawa, K. Shikano, "Phoneme recognition: neural networks vs hidden Markov models", Proc. International Conference on Acoustics, Speech and Signal Processing 1988, New York, N.Y., paper 8.S3.3.

[Watrous & Shastri 1987] R.L. Watrous and L. Shastri, "Learning phonetic features using connectionist networks", Proceedings of the 10th International Joint Conference on Artificial Intelligence, 1987, pp.851-854.