DATA-DRIVEN EXECUTION OF MULTI-LAYERED NETWORKS FOR AUTOMATIC SPEECH RECOGNITION

Renato DE MORI, Yoshua BENGIO and Régis CARDIN
Centre de Recherche en Informatique de Montréal (CRIM)
School of Computer Science, McGill University
805 Sherbrooke Street West,
MONTRÉAL, QUÉBEC, CANADA H3A 2K6

ABSTRACT

A set of Multi-Layered Networks (MLN) for Automatic Speech Recognition (ASR) is proposed. Such a set allows the integration of information extracted with variable resolution in the time and frequency domains and to keep the number of links between nodes of the networks small in order to allow significant generalization during learning with a reasonable training set size. Subsets of networks can be executed depending on preconditions based on descriptions of the time evolution of signal energies allowing spectral properties that are significant in different acoustic situations to be learned.

Preliminary experiments on speaker-independent recognition of the letters of the E-set are reported. Voices from 70 speakers were used for learning. Voices of 10 new speakers were used for test. An overall error rate of 9.5% was obtained in the test showing that results better than those previously reported can be achieved.

1. INTRODUCTION

Important efforts have been devoted in recent years to the coding of portions of the speech signal into **representations.**

Characterizing **Speech Units** (SU) in terms of speech properties or speech parameters requires a form of **learning** with a relevant **generalization** capability. Structural and stochastic methods have been proposed for this purpose [Jelinek, 1984; De Mori et al., 1987b].

Recently, a large number of scientists have investigated and applyied learning systems based on **Multi-Layered Networks** (MLN). Definitions of MLNs, motivations and algorithms for their use can be found in [Rumelhart et al., 1986; Plout and Hinton, 1987; Hinton and Sejnowski, 1986; Bourlard and Wellekens, 1987; Watrous and Shastri; 1987; Waibel et al., 1988]. Theoretical results have shown that MLNs can perform a variety of complex functions [Rumelhart et al., 1986]. Furthermore, they allow **competitive learning** with an algorithm based on well established mathematical properties.

Our interest in the use of MLNs is justified by previously published work. We have introduced a data-driven paradigm for extracting **acoustic properties** from continuous speech [De Mori et al., 1987a] and have investigated methods based on fuzzy or stochastic performance models for relating acoustic properties with SUs. MLNs appear to be good operators for automatically learning how to extract acoustic properties and relate them with phonetic features and words automating most of the activity which formely required a large amount of effort from a human expert. The human expert used knowledge acquired by generalizing observations of time-frequency-energy patterns. We will investigate in this paper how such learning can be performed by a set of MLNs whose execution is decided by a **data-driven strategy.**

By applying an input pattern to an MLN and clamping the output to the values corresponding to the code of the desired output, weights of connections between MLN nodes can be learned using error-back propagation [Plout and Hinton, 1987]. When a new input is applied to an MLN, its outputs may assume values between zero and one. If we interpret each output as representing a **phonetic property**, then the output value can be seen as a **degree of evidence** with which that property has been observed in the data [De Mori, 1983].

If phonemes are coded using a known set of **phonetic features**, the MLNs will learn how to detect evidence of each feature without being told all the details of the acoustic properties relevant for that feature.

Statistical distributions of feature evidences can be collected in performance models of SUs conceived as Hidden Markov Models (HMM). These models can be used to represent the time evolution of feature evidences for each SU or word. It is also possible to compute distances between time evolutions of real and desired degrees of evidences and to use such distances to rank word hypotheses, each word being characterized by a desired time evolution of degrees of evidences.

Experimental results obtained in the speaker-independent recognition of letters and digits ending with the phoneme /i/ will be reported. After a learning phase involving 70 speakers, a test was performed involving 10 new speakers and an error rate of 9.5% was found in the test.

2. ORGANIZATION OF MULTI-LAYERED NETWORKS

Figure 1 shows the general scheme of an MLN. The input layer is fed by a **Property Extractor** (PE), that acts as a window analyzing the data with variable time and frequency resolution. PEs may also extract data from the speech waveform.

The MLN in Figure 1 has two hidden layers and one output layer. Different MLNs may be used concurrently.

The following considerations motivate the use of different PE extractors and of different MLNs.

In the speech signal there are events characterized by abrupt transients. A plosive sound or the beginning of an utterance may produce events of that nature. In such situations, indicated as S_1, it is important to analyze possible bursts requiring a PE with high time resolution and not very high frequency resolution in high frequency bands.

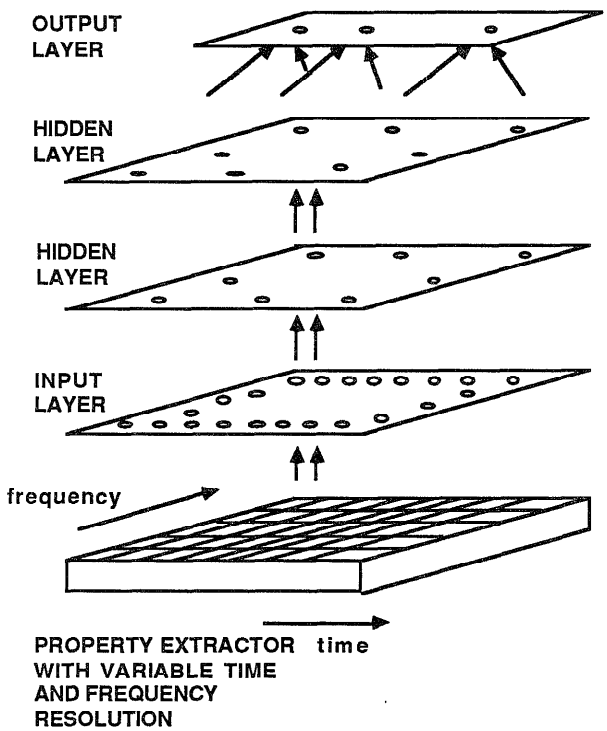

Figure 1 Multi-layered network with variable resolution Property extractor

It is also important to detect voicing with a PE spanning low frequencies for a relatively long time interval and to analyze possible formant transitions with PEs examining frequency bands between 0.3 and 3.5 kHz.

Recognition performance is improved by taking into account acoustic properties related to the morphology of time evolution of certain speech parameters following the approach proposed in [De Mori et al., 1987b].

The network in Figure 2 shows five PEs. Most of them are positionned on a speech spectrogram at the onset time after a silence, a buzz-bar or a frication noise interval.

The PEs are mostly rectangular windows subdivided into cells as shown in Figure 1. A vector of time resolutions (VT) and a vector of frequency resolutions (VF) describe the size of the cells in each PE (time values are in msecs, frequency values are in kHz). A symbol t* is inserted into VT to indicate the time reference for the position of the window.

The PEs introduced in Figure 2 have the following VTs and VFs:

PE11 : VT = {30,30,t*,10,10,10,10}
VF = {0.1,0.25,0.3,0.5}

meaning that two time intervals of 30 msecs each are analyzed before t* and four time intervals of 10 msecs each are analyzed after t*. The analysis is based on filters whose bands are delimited by two successive values of VF. There are 20 nodes on the first layer above PE11 and 10 nodes of the second layer.

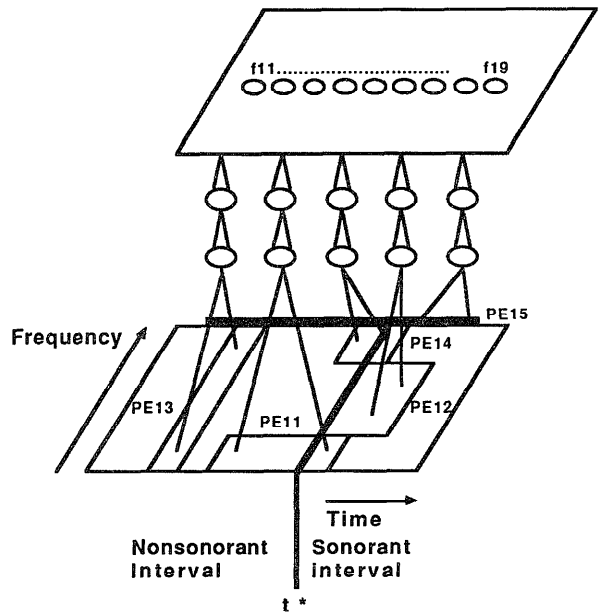

Figure 2 Property extractors of MLN1

PE12 has 39 filters each spanning three successive time intervals of 40 msecs. The filter bandwidth is 200 Hz, the position in frequency is decided based on spectral lines as defined in [Merlo et al., 1986], the first filter contains the spectral line that corresponds to the first formant in the last time interval. This allows speaker normalization by aligning filters with spectral lines. Default conditions are established in order to ensure that filters are positionned in prescribed frequency ranges even if spectral lines are not detected.

Each filter receives at its input a normalized value of the energy in its time-frequency window. For each filter there is a corresponding input that receives points of spectral lines detected inside the window corresponding to the time and frequency resolutions of the filter. There are 20 nodes in the first hidden layer above PE12 and 10 nodes on the second hidden layer.

PE13 is supposed to capture properties of frication noise and is characterized by the following vectors:

PE13: VT = {20,t*, 20}

VF = {1,2,3,4,5,6,7,8,9},

PE13 is executed every 20 msecs in the frication interval. It has 16 nodes in the first layer and 10 nodes in the second layer.

PE14 captures properties of the burst, when there is one, and is characterized as follows:

PE14: VT = {5,5,t*,5,5}
VF = {2,3,4,5,6,7}.

PE15 receives at its input normalized properties of the time evolution of energies in various bands as well as properties extracted from the speech waveform. This is a subset of the properties defined in [De Mori et al., 1987b] and contains those properties not included in what is extracted by the other PEs. There are 20 nodes in the first layer above PE14 and PE15 and 10 nodes in the second layer.

Let MLN1 be the network shown in Figure 2. It is executed when a situation characterized by the following rule is detected:

SITUATION S_1
((deep_dip)(t*)(peak)) **or**
((ns)(t*)(peak)) **or**
(deep_dip)(sonorant-head)(t*)(peak)) (1)
--> execute (MLN1 at t*)

(deep_dip), (peak), (ns) are symbols of the PAC alphabet representing respectively a deep dip, a peak in the time evolution of the signal energy and a segment with broad-band noise; t* is the time at which the first description ends, sonorant-head is a property defined in [De Mori et al., 1987a]. Similar preconditions and networks are established for nonsonorant segments at the end of an utterance.

The output in Figure 2 corresponds to the features defined in Table I.

TABLE I
Features corresponding to output code of MLN1:

output	feature
f11	voiced
f12	unvoiced
f13	sonorant
f14	plosive
f15	fricative
f16	labial
f17	alveolar
f18	palatal
f19	dental

For example, phoneme /b/ will be described by the following values : (f11=1 f12=0, f13=0, f14=1, f15=0, f16=1, f17=0, f18=0, f19 = 0). The code in Table I is redundant and can be modified. We have chosen it because MLNs give degrees of evidence for each output (feature) and it will be possible to compare the performance of MLN1, in which property extraction is based on automatically derived algorithms (i.e. learned with the weights) with past work done on property extraction performed by algorithms designed by a human expert [De Mori, 1983].

The features defined in Table I have been used for for recognizing letters of the E1 set defined as follows :

E1 : { b,c,d,e,g,k,p,v,3} (2)

In the learning phase the outputs were clamped according to the coding scheme of Table II.

TABLE II
Word coding for the E1-set

word	f11	f12	f13	f14	f15	f16	f17	f18	f19
b	1	0	0	1	0	1	0	0	0
c	0	1	0	0	1	0	1	0	0
d	1	0	0	1	0	0	1	0	0
e	0	0	1	0	0	0	0	0	0
g	1	0	0	0	1	0	0	1	0
k	0	1	0	1	0	0	0	1	0
p	0	1	0	1	0	1	0	0	0
t	0	1	0	1	0	0	1	0	0
v	1	0	0	0	1	1	0	0	0
3	0	1	0	0	1	0	0	0	1

By adding six other features to those in Table I, all the phonemes can be represented with phonetic features.

It was suggested that in theory the number of examples required for obtaining a good generalization in MLNs grows with the number of weights (links) [Plout and Hinton, 1986]. For this reason, a set of networks wherein each operates with a limited number of input nodes on a limited frequency interval can be better trained with a learning set of reasonable size. Furthermore, in spite of the theoretical complexity, a good generalization can be obtained with a reasonable number of experiments if the properties extracted from the inputs are chosen in such a way that a limited variation can be expected in their values if many speakers pronounce the same sound.

Experimental results on the recognition of the E1 set will be described in Section 3.

Sonorant segments can be extracted from continuous speech using a procedure described in [De Mori et al., 1985]. Sonorant segments are characterized by narrow-band resonances from which spectral lines as introduced in [Merlo et al., 1986] are extracted. For sonorant segments an MLN called MLN2 is used.

MLN2 is executed in situation S_2 characterized by peaks and high energy valleys of the signal energy in which frication noise has not been detected. MLN2 has two PEs, namely, PE21 and PE22. PE21 receives normalized energies in cells characterized by the following vectors:

PE21 : VF = {0.2,0.2,0.2,0.2,0.4,0.4,0.4,0.4,0.4,0.4}
VT = {20}

PE21 is positionned in such a way that the lowest frequency window contains the spectral line corresponding to

the first formant. The values of VF represent the frequency interval of cells rather than breakpoints.

PE22 is positionned like PE21 and receives points of spectral lines according to the following vectors:

PE22 : VF = {0.2,0.2,0.2,0.2,0.2,0.4,0.4,0.2, 0.2,0.2,0.2,0.2,0.4}
VT = {30,t,30}.

PE22 takes into account summaries about time evolutions of relevant parameters in time. The values in VF have the same meaning as for PE21.

These property extractors are applied every 20 msecs in the sonorant region.

The approach just introduced can be generalized in a formal way. In general, MLNs may have PEs that act as running windows advancing on a spectrogram by fixed time intervals. For each time reference, the output of the invoked MLNs is represented by a vector **M** of degrees of evidence for each of the feature outputs:

$$\mathbf{M} = \{ \mu_1, \mu_2, \ldots \mu_j, \ldots \mu_J \} \qquad (3)$$

where μ_j is the degree of evidence of feature f_j. As time reference varies from the beginning to the end of the speech signal, if T is the sampling interval and nT is the n-th set of feature evidences, these evidences will be represented by a vector **M(n)**.

M(n) represents a code for a speech interval. Time evolutions of feature evidences can be used for building word models conceived as Markov sources.

3. EXPERIMENTAL RESULTS

In order to test the ideas proposed in this paper, the E1-set as defined in (2) was used.

The ten words of the E1 set were pronounced twice by 80 speakers (40 males and 40 females). Data acquisition was performed with an HP 9000-236 especially equipped for speech processing. Learning and recognition were performed on a VAX 8650. The data from 70 speakers were used as a training set while the data from the remaining 10 speakers (6 males and 4 females) were used for the test. A confusion matrix is shown in Figure 3.

For the error-back propagation algorithm, the learning rate was 0.1 and the momentum 0.3 (See definition of these parameteThe error on the test set decreased as the network iterated on the set of training examples for MLN1, stabilizing near 400 iterations, and increasing thereafter.

An overall error rate of 9.5% was obtained with a maximum error of 20% for the letter /d/. This result is much better than ones we obtained before which were published recently [De Mori et al., 1987b]. It compares with results recently obtained by [Bahl et al., 1988] working only with male speakers on nine letters using competitive learning based on cross entropy. An observation of the confusion matrix shows that most of the errors represent cases that appear to be difficult even in human perception. Such cases are confusions b->e and d->e representing a low evidence of burst and formant transitions in voiced plosives (this might also be due to our poor resolution in data acquisition), confusions b->v, v->b, d->b, p->t, t->p, t->k indicating wrong estimation of the place of articulation, and confusions d->t, p->b, e->b indicating errors in the characterization of voicing. The fact that the error rate in the training set could reach a very low level (less than 1%) makes us hope that recognition performances may improve if more data are used for training the MLNs.

A preliminary experiment was performed using a version of MLN2 for the recognition of the place of articulation of vowels and diphtongs. An error rate of 4.2% was found in an experiment whose details are described in [Bengio and De Mori, 1988].

PRONOUNCED

RECOGNIZED \	b	c	d	e	g	k	p	t	v	3
b	17		1	1			2		2	
c		20								
d			16							
e	1		2	19						
g					18					
k						20		1		
p							17	1		
t			1		2		1	18		2
v	2								18	
3										18

Total error rate 19/200 = 9.5%
test set: 6 male and 4 female speakers 2 pronounciations per word per speaker (20 tokens per word)

Figure 3 Confusion matrix for the recognition of the E1-set pronounced by ten speakers not used in the training set.

Speaker-independent recognition of 5 vowels was performed with a 3.4% error rate. Speaker-independent recognition of 10 vowels resulted in a 13% error rate.

The learning and recognition algorithms based on error-back propagation (EBP) have been executed on a SUN 4/280 and on a VAX-8650. For an MLN of about 10,000 links, the time was 115 CPU msecs for the recognition of a spoken letter and 317 msecs for the learning of a spoken letter on the SUN 4/280. A 20% reduction was obtained on the VAX 8650.

The theorical complexity of the recognition algorithm is linear with the number of links.

The major contributions of this paper are the following.

First, new evidence is provided that speech sounds can be coded by features related to the place and manner of articulation. The presence of these features can be represented by degrees of evidence based on which stochastic performance models can be derived.

Second, it is shown how data-driven property extractors based on knowledge about acoustic correlates of phonetic features can provide useful information for training a multi-layered network with a reasonable number of data. Under these conditions significant performance has been achieved.

Third, new evidence is provided that information about frames of short duration combined with information about intervals of longer duration results in a better characterization of important speech sounds in view of automatic recognition.

Fourth, a remarkable gain in recognition accuracy (in accordance with [Bahl et al., 1988]) can be achieved if learning is competitive. Furthermore, our use of MLNs seems to perform equally well a normalization across male and female speakers.

Under the conditions described in this paper, MLNs can be trained effectively with a reasonable number of data to generate a single model for several sounds which can be incrementally trained.

ACKNOWLEDGEMENTS

This work was supported by the Natural Science and Engineering Council of Canada (NSERC) and Fond Concerté d'Aide à la Recherche (FCAR) of the province of Québec. The Centre de Recherche en Informatique de Montréal (CRIM) kindly provided computing time on its facilities.

REFERENCES

[Bahl et al., 1988] L.R. Bahl, P.F. Brown, P.U. De Souza and R.L. Mercer, Speech recognition with continuous-parameter Hidden Markov Models, In *Proceeding of ICASSP-88*, pages 40-43, International Conference on Acoustic, Speech and Signal Processing, April 1988.

[Bengio and De Mori, 1988] Y. Bengio and R. De Mori, Speaker normalization and automatic speech recognition using spectral lines and neural networks, In *Proceedings of CSCSI-88*, Canadian Conference on Artificial Intelligence, May 1988.

[Bourlard and Wellekens, 1987] H. Bourlard and C.J. Wellekens, Multilayer perceptron and automatic speech recognition, In *proceeding of ICNN-87*, pages IV407-IV406, International Conference on Neural Networks, June 1987.

[De Mori, 1983] R. De Mori, *Computer Models of Speech Using Fuzzy Algorithms*, Plenum Press, New-York, New-York, 1983.

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

[De Mori et al., 1987a] R. De Mori, E. Merlo, M. Palakal and J. Rouat, Use of procedural knowledge for automatic speech recognition, In *Proceedings IJCA-87*, pages 840-844, International Joint Conference on Artificial Intelligence, August 1987.

[De Mori et al., 1987b] R. De Mori, L. Lam and M. Gilloux, Learning and plan refinement in a knowledge-based system for automatic speech recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-9(2):289-305, March 1987.

[Hinton and Sejnowski] G.E. Hinton and T.J. Sejnowski, Learning and relearning in Boltzmann machines, In *Parallel Distributed Processing : Exploration in the Microstructure of Cognition*, volume 1, pages 282-317, MIT Press, Cambridge, Massachusetts, 1986.

[Jelinek, 1984] F. Jelinek, The development of an experimental discrete dictation recognizer, *IEEE Proceedings*, pages.1616-1624, November 1984.

[Merlo et al., 1986] E. Merlo, R. De Mori, M. Palakal and G. Mercier, A continuous parameter and frequency domain based Markov model, In *proceedings ICASSP-86*, pages 1597-1600, International Conference on Acoustics, Speech, Signal Processing, April 1986.

[Plout and Hinton, 1987] D.C. Plout & G.E. Hinton, Learning sets of filters using back propagation, Computer *Speech and Language*, 2(2):35-61, July 1987.

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

[Waibel et al., 1988] A. Waibel, T. Hanazawa, K. Shikano, Phoneme recognition : neural networks vs hidden Markov models, In *proceedings ICASSP-88*, pages 107-110, International Conference on Acoustics, Speech and Signal Processing, April 1988.

[Watrous and Shastri, 1987] R.L. Watrous and L. Shastri, Learning phonetic features using connectionist networks, In *proceedings IJCAI-87*, pages 851-854, International Joint Conference on Artificial Intelligence, August 1987.