ELSEVIER

Neurocomputing 37 (2001) 91–126

NEUROCOMPUTING

www.elsevier.com/locate/neucom

A survey of hybrid ANN/HMM models for automatic speech recognition

Edmondo Trentin[a,*], Marco Gori[b]

[a]*ITC-irst (Centro per la Ricerca Scientifica e Tecnologica), V. Sommarive, 18-Povo, Trento, Italy and Università di Firenze, V. S. Marta, 3 - Firenze, Italy*

[b]*Dipartimento di Ingegneria dell'Informazione, Università di Siena, V. Roma, 56 - Siena, Italy*

Received 27 April 1999; accepted 15 April 2000

Abstract

In spite of the advances accomplished throughout the last decades, automatic speech recognition (ASR) is still a challenging and difficult task. In particular, recognition systems based on hidden Markov models (HMMs) are effective under many circumstances, but do suffer from some major limitations that limit applicability of ASR technology in real-world environments. Attempts were made to overcome these limitations with the adoption of artificial neural networks (ANN) as an alternative paradigm for ASR, but ANN were unsuccessful in dealing with long time-sequences of speech signals. Between the end of the 1980s and the beginning of the 1990s, some researchers began exploring a new research area, by combining HMMs and ANNs within a single, hybrid architecture. The goal in hybrid systems for ASR is to take advantage from the properties of both HMMs and ANNs, improving flexibility and recognition performance. A variety of different architectures and novel training algorithms have been proposed in literature. This paper reviews a number of significant hybrid models for ASR, putting together approaches and techniques from a highly specialistic and non-homogeneous literature. Efforts concentrate on describing and referencing architectures and algorithms, their advantages and limitations, as well as on categorizing them into broad classes. Early attempts to emulate HMMs by ANNs are first described. Then we focus on ANNs to estimate posterior probabilities of the states of an HMM and on "global" optimization, where a single, overall training criterion is defined over the HMM and the ANNs. Connectionist vector quantization for discrete HMMs, and other more recent approaches are also reviewed. It is pointed out that, in addition to their theoretical interest, hybrid systems have been allowing for tangible improvements in recognition performance over the standard HMMs in difficult and significant benchmark tasks. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Speech recognition; Hidden Markov model; Neural network; Hybrid system

[*] Corresponding author. Tel.: + 39-0461-314592; fax: + 39-0461-314591.
E-mail address: trentin@itc.it (E. Trentin).

0925-2312/01/$ - see front matter © 2001 Elsevier Science B.V. All rights reserved.

PII: S0925-2312(00)00308-8

1. Introduction

Automatic speech recognition (ASR) [100,29,69,62,28] has been the most investigated topic in speech processing along the last few decades. Broadly speaking, it can be formulated as the problem of *recognizing* (not *understanding*) the words from a given dictionary uttered by a speaker, relying only on the information contained in the uttered *speech signal* and on prior knowledge on the problem domain. In the 1950s, when the research on ASR began, many researchers believed that the incoming new computer technologies would have made ASR rather an easy task. Unfortunately, a few decades later, we are realizing that the conjecture was false. ASR emerged to be a very hard problem, and nowadays many difficulties and open questions are far from being solved, in spite of the efforts of a number of long-termed research groups throughout the world. Difficulties are related to increasing dictionary size (e.g. more than 50,000 words), continuous speech recognition versus isolately uttered words, number and vocal characteristics of speakers in speaker-independent (SI) recognizers (versus simpler speaker-dependent (SD) systems, capable to recognize only speech uttered by the speaker whose acoustic material was used to train the recognizer itself), spontaneous speech phenomena (um's, ah's, false starts, out-of-vocabulary words, etc.), robustness to environmental conditions (noise and distortion over the channel, multiple microphones spread throughout the room to allow hands-free dictation, etc.), and so on. Most of these problems arise when moving from the laboratory, where simulation experiments often allow for excellent recognition performance, to real-world conditions, where a dramatic degradation in performance is far from being an exception.

Nonetheless, ASR has resulted highly effective in a variety of applicative scenarios: dictation, that is automatic generation of written text from the speech signal, access to databases, human–machine interface, access to remote automatic services on the telephone line, and control of machines.

The ASR problem can be formulated as a statistical *classification* problem, according to classical pattern recognition [30,40]. Once the *classes* have been defined as sequencies W of allowable words from a "closed" dictionary, a parametric representation of the speech signal has been chosen (e.g. a sequence of acoustic feature vectors X), and a *Maximum a Posteriori* (MAP) criterion has been adopted, the classification problem can then be stated as finding the sequence of words $\tilde{W}$ which maximizes the quantity $Pr(W \mid X)$. The latter is usually factorized using Bayes' theorem [30] as

$$Pr(W \mid X) = \frac{Pr(X \mid W)Pr(W)}{Pr(X)}. \tag{1}$$

Given an acoustic observation sequence X, the efforts on the maximization of $Pr(W|X)$ can be moved to the search for the class $\tilde{W}$ which maximizes the numerator of the right-hand side of Eq. (1), i.e. $Pr(X \mid W)Pr(W)$. The quantity $Pr(W)$, usually referred to as the *language model* (LM) [103] depends on high-level constraints and linguistic knowledge about allowed word strings for the specific task. The quantity $Pr(X \mid W)$ is known as the *acoustic model*. It describes the statistics of sequencies of parametrized acoustic observations in the feature space given the corresponding uttered words (e.g. certain phonemes). Hidden Markov models (HMM) [101,52] are the most popular (parametric) model at the acoustic level. A brief review of HMMs is presented in Section 2. Although HMMs are effective approaches to the problem of acoustic modeling in ASR, allowing for good recognition performance under many circumstances, they also suffer from some limitations. These limitations, discussed at the end of Section 2, are the rationale behind the research for different paradigms.

Starting from the late 1980s, many researchers began to use artificial neural networks (ANN) for ASR. Neural nets were expected to carry out the recognition task (e.g. classification of phonemes or words) when discriminatively trained on acoustic features. Milestones in this respect are [120–122,42,45,3,39,23,46,118,112,24,110, 10,111], among the others. Lippmann [75] wrote a comprehensive survey of the state of the art in connectionist speech recognition at the end of the Eighties. The topic was of crucial interest in classic conferences and workshops traditionally dedicated to speech recognition and even in specialized workshops.

To take the temporal dependencies typical of speech signals into account, two major classes of neural networks were proposed, namely *time-delay neural network*, and *recurrent neural networks*. Time-delay neural networks (TDNNs) [120,121,148], also known as *tapped delay lines*, represent an effective attempt to train a static multilayer perceptron (MLP) [114] for time-sequence processing, by converting the temporal sequence into a spatial sequence over corresponding units. The idea was applied in a variety of ASR applications, mostly for phoneme recognition [120,121,13]. An example of a TDNN is shown in Fig. 1. The input layer has been

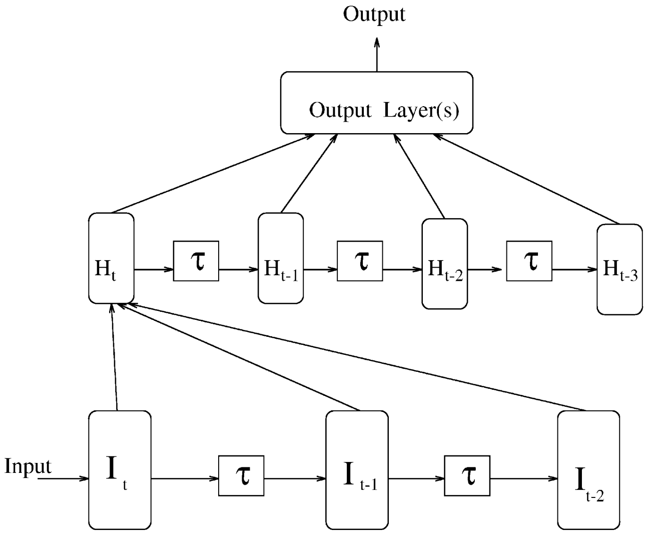

Fig. 1. Time-delay neural network. Input is fed into the leftmost set of input units (I_t) at time t. Previous inputs (I_{t-1}, I_{t-2}) are shifted to the right, with unit delays represented by boxes labeled with τ. A similar mechanism holds in the hidden layer ($H_t, \dots$). An integration over time of the input sequence is carried out by the leftmost set of hidden units, while the output layer of the net integrates over time the activations of the hidden units.

enlarged to accept as many input patterns as the (fixed) sequence length to be processed at each time step. Input vectors enter the network from the leftmost set of input units. At each time step, inputs are shifted to the right through the unit delay line that links each set of input units to the right-adjacent one, and the next input pattern is fed into the leftmost position. The same extension can also be applied to subsequent layers, introducing a tapped-delay mechanism between hidden units (e.g. only the first block of units in the tapped line actually receives input from the previous layer), giving the ability to deal with more complicated time dependencies. The *backpropagation* (BP) [123,114,67] algorithm can be used to train such a network.

Using TDNNs, Lang and Hinton [66] obtained a 7.8% error rate in multi talker classification of the isolated letters "B, D, E, V", using acoustic material collected among 100 male speakers. Waibel et al. [122] were able to recognize isolated consonants uttered by a Japanese speaker with a low error rate (4.1%), using a combination of specialized TDNNs. A significant 1.4% error rate in vowel recognition was obtained in the same experiments. As illustrated in the next Sections, TDNNs are often adopted instead of MLPs within hybrid paradigms.

Recurrent neural networks (RNN) provide a powerful extension of feed-forward connectionist models by allowing to introduce connections between arbitrary pairs of units, independently from their position within the topology of the network. Self-recurrent loops of a unit onto itself, as well as backward connections to previous layers, or lateral links between units belonging to the same layer are all allowed. An example of a generic RNN architecture is given in Fig. 2, to fix ideas.

RNNs behave like dynamical systems. Once fed with an input, the recurrent connections are responsible for an evolution in time of the internal state of the network. RNNs are particularly suited for sequence processing, due to their ability to keep an internal trace, or memory, of the past. This memory is combined with the current input to provide a context-dependent output. Several RNN architectures were

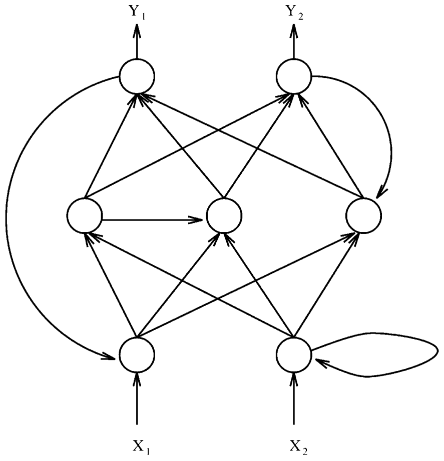

Fig. 2. Generic recurrent neural network.

proposed in literature [60,33,90], along with a variety of training algorithms, mostly based on gradient-descent techniques. Among the latter ones, particularly remarkable are recurrent back propagation [98], back-propagation for sequences (BPS) [42], real-time recurrent learning [127,126], Time-dependent recurrent back-propagation [124,95,115] and the most popular *back-propagation through time* [83,114].

As examples of application of RNNs to ASR problems, particularly remarkable are the results obtained for SI phoneme recognition by Robinson et al. [109,112,111], where RNNs are used as state-space machines, capable to compute an output and the next state, given the input and the current state. This results in a non-linear extension of linear control theory. A recurrent network for plosive recognition was successfully applied in [7,34]. In some cases, as described in the next sections, RNNs are combined with HMMs within hybrid architectures.

In spite of their ability to classify short-time acoustic-phonetic units, such as individual phonemes, ANNs failed as a general framework for ASR, especially with long sequencies of acoustic observations like those required in order to represent words from a dictionary or whole sentences. This is mainly due to the lack of ability to model long-term dependencies in ANNs, even when recurrent architectures are considered. The theoretical motivations underlying this problem were well analyzed by [11]. In the early 1990s this fact led to the idea of combining HMMs and ANNs within a single, novel model, broadly known as *hybrid* HMM/ANN [38,70,15, 87,93,44,118,8]. A recently published survey paper [125] investigates the "border" between ANNs and Markovian models (HMMs for ASR and *Markov random fields* for image processing), summarizing approaches where ANNs emulate Markov models and vice versa, and reporting hybrid architectures. The hybrid paradigm relies on maintaining an underlying HMM structure, capable of modeling long-term dependencies, with the integration of ANNs, which provide non-parametric universal approximation, probability estimation, discriminative training algorithms, fewer parameters to estimate than those usually required in standard HMMs, efficient computation of outputs at recognition time, and efficient hardware implementability.

Different hybrid architectures and training/decoding algorithms have been proposed in the literature, according to the nature of the ASR task, to the type of HMM, or to the specific role of the ANN within the system. The hybrid approach often allowed for significant improvements in performance with respect to standard approaches to difficult ASR tasks. This paper is a survey of hybrid HMM/ANN systems for ASR, putting together approaches and techniques from a highly specialistic and non-homogeneous literature. Efforts concentrate on describing and referencing architectures and algorithms, their advantages and limitations, as well as on categorizing them into broad classes.

After a brief review of basic HMM concepts and typologies (Section 2) the paper reports the theoretical aspects and the performance obtained with different models (Section 3), which have been categorized according to their architecture (Sections 3.1–3.5). These categories are summarized in Table 1, with pointers to the respective sections. Experimental results are reported from literature in terms of recognition rate on specific test databases, which are usually public-domain. The most

Table 1
Categorization of hybrid models according to their architectures and nature, along with brief description of characterizing features and pointer to corresponding Section in the paper

Category	Brief description	Section
Early attempts	Approaches that relied on ANN architectures that attempted to emulate HMMs.	3.1
ANNs to estimate the HMM state-posterior probabilities	A probabilistic interpretation of the ANN outputs is given, e.g. ANNs perform an estimate of the posterior probability of CDHMM states given the acoustic observations.	3.2
Global optimization	Introduction of a training scheme aimed at the optimization of a "global" criterion function, defined at the whole-system (i.e., ANN and HMM simultaneously) level.	3.3
Networks as vector quantizers for discrete HMMs	Unsupervised ANNs are used to perform a *quantization* in the acoustic feature space for discrete HMMs.	3.4
Other approaches	Hybrid systems based on particular combination techniques between ANNs and HMMs, not belonging to any of the previous categories, and often focused on specific tasks.	3.5

common evaluation criterion in the case of connected or continuous speech recognition is the *word error rate* (WER), that is defined as

$$\mathrm{WER} = 100(Ins + Del + Sub)/N_{\mathrm{words}}, \tag{2}$$

where N_{words} is the total number of words in the uttered text, and the number of errors of the recognizer is expressed counting out word insertions (*Ins*), deletions (*Del*) or substitutions (*Sub*), respectively.

Although state-of-the-art hybrid models seem to yield a gain in performance with respect to standard HMM recognizers under many circumstances, several open problems in ASR still remain far from being solved. Some of them and a sketch of challenging future issues are briefly discussed in Section 4, where concluding remarks are also drawn.

2. Hidden Markov models

An HMM is a pair of stochastic processes: an *hidden* Markov chain and an *observable* process which is a probabilistic function of the states of the former. This means that observable events in the real world (e.g., acoustic observations) are modeled with (possibly continuous) probability distributions, that are the observable part of the model, associated with individual states of a discrete-time, first-order Markovian process. In general, the latter is not ergodic. The semantics of the model (conceptual correspondence with physical phenomena) is usually encapsulated in the hidden part: for instance, in ASR an HMM can be used to model a word in the task-dependent vocabulary, where each state of the hidden part represents a phoneme (or sub-phonetical unit), whereas the observable part accounts for the statistical characteristics of the corresponding acoustic events in a given feature space (e.g. sampled acoustic signal, represented in a proper way).

More precisely, an HMM is defined by:

(1) A set S of Q states, $S = \{S_1, \dots, S_Q\}$, which are the distinct values that the discrete, hidden stochastic process can take.
(2) An *initial state* probability distribution, i.e. $\pi = \{Pr(S_i \mid t = 0), S_i \in S\}$, where t is a discrete time index.
(3) A probability distribution that characterizes the allowed transitions between states, that is $\boldsymbol{a}_{ij} = \{Pr(S_j \text{ at time } t \mid S_i \text{ at time } t-1),\ S_i \in S,\ S_j \in S\}$ where the *transition probabilities* $\boldsymbol{a}_{ij}$ are assumed to be independent of time t.
(4) An *observation* or *feature* space F, which is a discrete or continuous universe of all possible observable events (usually a subset of $\boldsymbol{R}^d$, where d is the dimensionality of the observations).
(5) A set of probability distributions (referred to as *emission* or *output* probabilities) that describes the statistical properties of the observations for each state of the model: $\boldsymbol{b}_x = \{b_i(\boldsymbol{x}) = Pr(\boldsymbol{x} \mid S_i),\ S_i \in S,\ \boldsymbol{x} \in F\}$.

HMMs represent a learning paradigm, in the sense that examples of the event that is to be modeled can be collected and used in conjunction with a training algorithm in order to *learn* proper estimates of π, $\boldsymbol{a}$ and $\boldsymbol{b}_x$. The most popular of such algorithms are the *forward–backward* (or *Baum–Welch*) [101] and the Viterbi [101] algorithms. Whenever continuous emission probabilities are considered, both of them are based on the general maximum-likelihood (ML) criterion, i.e. they aim at maximizing the probability of the samples given the model at hand. In particular, the Viterbi algorithm concentrates only on the most alike path throughout all the possible sequences of states in the model.

These algorithms belong to the class of *unsupervised learning* techniques, since they perform unsupervised parameter estimation of the probability distributions without requiring any prior labeling of individual observations (within the sequences used for training) as belonging to specific states.

Once training has been accomplished, the HMM can be used for *decoding* or *recognition*. Two different instances occur according to the specific task. Whenever N different HMMs — corresponding to models of N different events or *classes* defined in the feature space — are used (e.g. in ASR), decoding (classification) means assigning each new sequence of observations to the most alike model. On the contrary, when a single HMM is used, decoding (recognition) means finding out the most alike path of states within the model and assigning each individual observation to a given state within the model. If the ML criterion is used, the forward–backward or the Viterbi algorithms are still suitable for the recognition task.

One major distinction has to be made between *discrete* HMMs and *continuous density* HMMs (CDHMMs) [52]. The former use discrete probability distributions to model the emission probabilities, i.e. they rely on the assumption of a finite alphabet of symbols in input or, equivalently, they require a *quantization* of a continuous input space, e.g. via any *clustering* [30] technique, into a finite size *codebook*. CDHMMs, on the other side, use continuous probability density functions (pdfs), usually referred to as *likelihoods*, to describe statistics of the acoustic features within the HMM states, and are usually best suited for very difficult ASR tasks (i.e, continuous speech dictation with large vocabularies), since they exhibit better modeling accuracy. Gaussian or mixtures of Gaussian components are the most popular and effective choices of pdf's for CDHMMs.

HMMs have been successfully applied in a variety of tasks, mainly in speech recognition. Unfortunately, standard HMMs in conjunction with the above-mentioned training and decoding algorithms suffer from some major intrinsic limitations. The combination of ANNs with HMMs is intended here as an attempt to overcome some of these limitations (see for instance [13]). Standard CDHMMs, trained with forward–backward or Viterbi algorithms, present poor discriminative power among different models, since they are based on the ML criterion, which is itself non-discriminative. The classical HMMs rely on strong assumptions on the statistical properties of the phenomenon at hand. For instance, the stochastic processes involved are modeled by first-order Markov chains, and the parametric form of the probability density functions that represent the emission probabilities associated with all states is heavily constraining. In addition, the number of parameters in HMMs do strongly limit their implementability in hardware. Given these limitations, the use of ANNs with their discriminative training, their capability to perform non-parametric estimation over whole sequences of patterns, and their limited number of parameters (which allows for effective development of neural microchips) appear definitely promising.

3. State of the art of hybrid ASR systems

Whereas standard HMMs have a consolidated and homogeneous theoretical framework, hybrid ANN/HMM systems are a recent research field with no unified formulation. A variety of different architectural and algorithmic solutions have been proposed in the literature. The following sections are organized according to four major categories:

(1) *Early attempts*. Early approaches (between the late 1980s and the beginning of the 1990s) relied on ANN architectures that attempted to emulate HMMs [16,93]. Some of them incorporated the dynamic programming (DP) algorithm [5] within the ANN itself [44,71]. These approaches strengthened the idea that ANNs could be effectively used for ASR, but the straight emulation of standard HMMs did not allow to overcome limitations of the latter (summarized in the previous section).

(2) *ANNs to estimate the HMM state-posterior probabilities*. Some ANN/HMM hybrids assume that the output of an ANN is sent to an HMM for ASR [70,38,44,118]. The architectures proposed by Bourlard et al. rely on a probabilistic interpretation of the ANN outputs [13,88]. Each output unit of the ANN is trained to perform a non-parametric estimate of the posterior probability of

a CDHMM state given the acoustic observations (and, possibly, given the previous state). This represents a fundamental class of hybrid models, which had a strong influence over a number of following approaches. Points of strength of this paradigm are the relative ease of implementation, due to a training scheme that rely on alternating between BP and a standard Viterbi alignment, as well as a discriminative training criterion that may allow for improved recognition performance. Weakness of the approach comes from the unavailability of an analytically motivated global optimization scheme defined at the whole-system level, and also from the requirement of huge ANNs architectures (possibly with millions of connection weights) to be trained for the most complicated ASR problems, e.g., SI continuous speech recognition with large vocabularies.

(3) *Global optimization*. The ANN and the HMM are often trained separately, but techniques have also been proposed in which a simultaneous training is accomplished. In [5,8], the ANN is used to transform a vector of acoustic features into more effective observations for CDHMMs, according to a "global optimization" of the parameters of the combined system. Apart from the feature extraction aspect, what is more relevant in this case is the introduction of an analytically motivated training scheme aimed at the optimization of a global criterion function, the latter being defined at the whole-system (i.e., recognition) level. Other researchers proposed different globally optimized hybrids in recent years. A brief comparison of hybrid architectures relying on global discriminative training algorithms can be found in [58].

(4) *Networks as vector quantizers for discrete HMM*. As mentioned in Section 2, discrete HMMs assume that a finite alphabet of input symbols has to be modeled. Since in ASR the assumption does not hold, a *quantization* in the acoustic feature space is required. Instead of using standard clustering algorithms, unsupervised ANNs can be effectively used as vector quantizers. This class of hybrids is characterized by distinct training steps for the ANN and for the HMM. Lack of a combined, global optimization scheme is compensated by the reduced complexity of the overall machine (mainly due to the ease of use of discrete HMMs versus CDHMMs) and by good recognition performance, the latter being close to that yielded by CDHMMs in some small-scale ASR tasks. Examples can be found in [64,55,106].

(5) *Other approaches*. Along with the architectures introduced so forth, during the Nineties several researchers proposed hybrid systems relying on particular combination techniques between ANNs and HMMs. Such techniques do not properly belong to any of the previous categories, and often focused on specific tasks, e.g. *rescoring* or *word-spotting*. In [5,44,118], the ANN outputs are interpreted as "scores" which are used within a DP algorithm to perform alignment and segmentation. As an alternative, the ANN can be used for re-scoring the *N*-best hypotheses produced with an HMM [130]. Although a uniform treatment of these approaches does not appear feasible, Section 3.5 provides a summary of major architectures, algorithms and experimental results reported from literature.

3.1. Early attempts (Viterbi Net, Alpha Net)

In 1987 Lippmann and Gold [76] proposed a recurrent neural network architecture, implemented in VLSI, able to mimic the decoding behavior of the continuous-density Viterbi algorithm, which was for the recognition of isolated words. It was called the Viterbi net. Although the recognition performance did not represent any improvement with respect to what could be achieved with standard HMMs, this connectionist architecture is remarkable for historical reasons.

The structure of the Viterbi net has as many input units as the dimensionality of the acoustic vectors. Acoustic observations are fed into the network in sequence, one at a time. Inputs are propagated forward through a single, fully connected layer, and summed up before being passed to each one of a set of state units representing the states of a corresponding left-to-right HMM. A Viterbi network is built for each word model of the HMM. The state units have a threshold-logic activation and a fixed delay on the output, and are laterally connected (each of them to the following one) in a way that resembles the topology of the left-to-right HMM. In addition to the summed inputs, each state unit also receives a feedback input from an associated subnetwork, which is able to select the maximum between the output from the state unit itself and the output from the adjacent state unit on the left.

There is no actual training procedure for the Viterbi net, and this is one of its major limitations. It is initialized using the parameters of the corresponding HMM, obtained using the Baum–Welch algorithm. After the initialization, when fed with a sequence of input vectors, the recurrent dynamics of the net result in a parallel version of Viterbi decoding, producing as the final output the logarithm of the likelihood of the input sequence (and most likely state path) given the model.

Experimental results on isolated word recognition tasks, performed using 12 *MEL* cepstral coefficients [27] and 13 differential MEL cepstra as features, were satisfactory and comparable with those obtained with state-of-the-art HMMs.

In 1990 Bridle [16] proposed a connectionist architecture which was able to behave like an HMM for ASR. The idea underlying his approach was to look at the forward and backward computation of the probabilities in HMMs, and to give them an interpretation in terms of a neural network. The model was called *Alpha Net*, because of the way its architecture and dynamics were calibrated to resemble the forward computation of the alphas in the Baum–Welch algorithm.

The Alpha net is a recurrent neural network. As for the Viterbi net, its parameters are the same as those of the corresponding HMM, but a complete forward estimation of the likelihood of the observations given the models is accomplished, instead of a single best-path search as occurs with the Viterbi algorithm. Furthermore, a learning procedure, derived from the backward step of the Baum–Welch algorithm, is available. This procedure is based on a discriminant cost function (maximum mutual information) and backpropagation of its partial derivatives. A recurrent architecture is built for each unit (word) to be included in the model. The neurons are organized in order to represent the states of the HMM. For instance, in the most common case of left-to-right HMMs, each unit (neuron) is connected with a recurrent connection to itself, and with a forward connection to the unit representing the following adjacent state in the HMM. The weights of these connections are equal to the state transition probabilities between the corresponding pairs of states. The likelihoods of the emission probabilities are fed into the recurrent loops from another, distinct part of the network, and are multiplied instead of summed. The units are linear, with a unit delay, thus resulting in the computation of the product of the joint probability of transition and emission for that state at each time step. In so doing, the overall behavior is consistent with the probabilistic framework of the HMM. The separate network that is responsible for the computation of these likelihoods is supposed to rely, for example, on multipliers and exponentials in order to approximate as closely as possible the likelihoods generated by the Gaussian or mixture of Gaussians associated to the states of the corresponding HMM. The final output of such an architecture is the actual likelihood of the input acoustic sequence given the model, summed over all possible paths within the network.

3.2. ANNs to estimate the state posterior probabilities

Bourlard et al. [15,87,12,13] proposed HMM/ANN hybrids for continuous ASR in which a MLP was trained to estimate the posterior probabilities of HMM states, with the ultimate objective of maximizing the posterior probability of a given (left-to-right) Markov model M_i given an acoustic observation sequence X. Posterior probabilities can be written as

$$\begin{aligned} Pr(M_i \mid X) &= \sum_{q_1^L} Pr(q_1^L, M_i \mid X) \\ &= \sum_{q_1^L} Pr(q_1^L \mid X) Pr(M_i \mid q_1^L, X) \\ &= \sum_{q_1^L} Pr(q_1^L \mid X) Pr(M_i \mid q_1^L), \end{aligned} \tag{3}$$

where the model M_i is supposed to have Q states $S_1, \dots, S_Q$, and the acoustic observation sequence $X = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_L)$ is assumed to be of length L. In Eq. (3) the sums are extended over all possible sequences q_1^L of states.

The quantity $Pr(M_i \mid q_1^L)$ does not depend on the acoustics (observation sequence X), but only on higher-level choices made in the definition of the models and can thus be computed separately.

Repeatedly applying the properties of joint probabilities, Eq. (3) can be rewritten as

$$\begin{aligned} Pr(M_i \mid X) &= \sum_{q_1^L} Pr(q_1 \mid X) Pr(q_2 \mid X, q_1) \\ &\quad \dots Pr(q_L \mid X, q_1, \dots, q_{L-1}) Pr(M_i \mid q_1^L) \\ &= \sum_{q_1^L} \left\{ \prod_{\ell=1}^{L} Pr(q_\ell \mid X, q_1^{\ell-1}) \right\} Pr(M_i \mid q_1^L). \end{aligned} \tag{4}$$

Attempts to determine analytical developments for the present formulation, similar to those adopted in the forward–backward algorithm for maximization of the likelihood of the observations given the model, are not practicable. This is due to the constraint $\sum_i Pr(M_i \mid X) = 1$ that must be satisfied within the present *maximum* a posteriori discriminant framework. Bourlard and Morgan's idea was then to use feedforward neural networks as estimators of the posterior probabilities of states given the observations and the previous state sequence. In so doing, advantages are taken from ANNs discriminant training and from their capability to estimate Bayesian posterior probabilities when trained by BP on a mean-squared-error (MSE) criterion. The basic architecture is schematically represented in Fig. 3.

Actually, approximate versions of Eq. (4) are used, e.g., by taking

$$Pr(M_i \mid X) \approx \sum_{q_1^L} \left\{ \prod_{\ell=1}^{L} Pr(q_\ell \mid \boldsymbol{x}_{\ell-k}, \ldots, \boldsymbol{x}_{\ell+k}, q_{\ell-1}) \right\} Pr(M_i \mid q_1^L), \tag{5}$$

that is to say, the network is trained to estimate the state posterior, called *conditional transition probability*, $Pr(q_\ell \mid \boldsymbol{x}_{\ell-k}, \ldots, \boldsymbol{x}_{\ell+k}, q_{\ell-1})$ given a fixed number $2k+1$ of acoustic vectors $\boldsymbol{x}_{\ell-k}, \ldots, \boldsymbol{x}_{\ell+k}$ (a *window* or *context* of size k centered in the current acoustic observation $\boldsymbol{x}_\ell$) and the previous state. This is accomplished using the BP algorithm on a MLP, which has an output unit for each state, that represents the estimate of the state posterior probability. Other attempts were made to use the MLP to estimate the posteriors in different ways, for example as a function of the current observation and of the previous state only, or as a function also of previous MLP

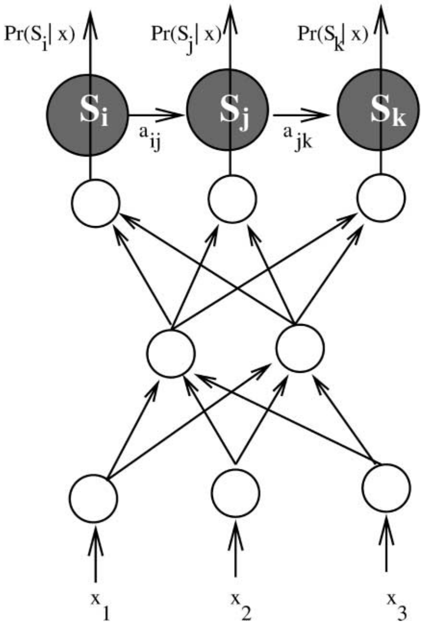

Fig. 3. Basic hybrid architecture where a 2-layer feedforward ANN estimates the posterior probabilities of states S_i, S_j, S_k of a left-to-right HMM given an hypothetic acoustic observation $x = (x_1, x_2, x_3)$.

outputs (previous states). In any case, the speech recognition performance of the resulting overall system was surprisingly poor in many experiments [14]. This was attributed to a mismatch between the priors estimated from relative frequencies over the training data, and priors implicitly constrained by the topology of the models. A step backward was then made by moving the system back to likelihoods. This issue was pursued by using a somewhat standard version of the HMM, in conjunction with neural networks. The latter were trained to perform exactly the same probability estimation as in Eq. (5), but with their outputs divided by the a priori probabilities of the corresponding states, in order to reduce probabilities to scaled likelihoods, normalized by the unconditional likelihood of each observation (using Bayes' theorem). Priors can be computed apart, from the training data or from statistical considerations on the constraints given by the specific task. This solution was effective and allowed the system to reach the recognition performance of state-of-the-art HMM recognizers on large vocabulary continuous speech tasks, but at the expense of the original theoretical framework.

One central point raised with this approach concerns the training procedure. Indeed, networks are trained by BP, which would require knowledge of target values for the outputs in order to compute the gradient of the cost function. With the exception of toy tasks, no supervised labeling of acoustic frames is actually available (labeling by hand in real-sized databases is not feasible). Bourlard et al. suggested an iterative training procedure that starts up with an initial segmentation of the acoustic observations, performs training of the networks according to that segmentation, then uses the Viterbi algorithm, in conjunction with the newly trained networks as estimators of the state-posterior probabilities, in order to produce a new and more reliable segmentation of the data, that in turn is used to train again the networks, and so on in an iterative fashion. The initial segmentation may be obtained using a standard HMM, or by dividing the observation sequence into equally sized segments; each segment is associated to the corresponding state in the correct HMM state-sequence. An analogous training scheme was proposed also in [38].

The approach by Bourlard et al. was successfully adopted in [105] for the DARPA resource management (RM) SI, continuous speech task, with a vocabulary size of 991. An HMM with 2 or 3 states left-to-right phone models was used, with parameter tying of the pdfs associated to the states of a given model.

The performance of the hybrid system was compared with that of Decipher, an HMM-based recognizer where emission probabilities are modeled with mixtures of Gaussian components. Decipher was also used to bootstrap the hybrid, i.e. to obtain the initial segmentation of acoustic observations used for the first training step of the MLP.

Experimental results obtained averaging on the three data sets used for testing, using 12 MEL plus energy and the corresponding first-order time-derivatives and a word-pair grammar (that reduced the perplexity to 60), yielded an average error of 8.7% for the context-independent hybrid, a 7.0% for the context-dependent HMM and a 5.5% for the interpolation of the two models. Moreover, results on one of the test sets (February91) showed considerably better performance of the context-independent hybrid (5.8%) with respect to the context-independent HMM (11.0%).

Singer and Lippmann [116] used radial basis function (RBF) [99] networks instead of MLPs as Bayesian probability estimators; the resulting hybrid was used in an isolated word recognition task.

Recently, Hennebert et al. [47] have proposed a reinforcement of the theoretical framework originally formulated by Bourlard, Morgan et al., by generalizing the local connectionist estimates of posterior probabilities to global posterior of models, formulating also a novel training algorithm within that framework.

An extension to the hybrid paradigm based on connectionist estimation of state posterior probabilities was proposed by Franco et al. [36], where the basic context-independent HMM scheme was replaced by a model capable to take into account acoustic contexts. The introduction of context dependency is accomplished using Bayes' theorem to obtain the following factorization:

$$p(X \mid q_j, c_k) = \frac{Pr(q_j \mid X, c_k) p(X \mid c_k)}{Pr(q_j \mid c_k)}, \tag{6}$$

where X is an acoustic observation, q_j is a state of the HMM, and c_k is a certain context ($k = 1, \dots, K$).

In Eq. (6) the quantity $Pr(q_j \mid X, c_k)$ is estimated by an MLP. Bourlard et al. used a single MLP to estimate all state posteriors $Pr(q_j \mid X)$, $q_j \in S$; [36] adopts K MLPs if K is the number of contexts to be modeled. The kth MLP is trained as an estimator for $Pr(q_j \mid X, c_k)$, for all $q_j \in S$, applying BP on the feature vectors embedded within kth context in the training sequencies.

The quantity $p(X \mid c_k)$ is expanded using Bayes' theorem again, writing a further factorization of the form

$$p(X \mid c_k) = \frac{Pr(c_k \mid X) p(X)}{Pr(c_k)}, \tag{7}$$

where $Pr(c_k \mid X)$ is estimated by an MLP in Bourlard's way. The other quantities involved in Eq. (7) are directly computable from statistical evaluations (analysis of relative frequencies) on the training data, otherwise they come out to be irrelevant in Viterbi search for the optimal path.

Franco et al. [36] used 2-layer, context-dependent MLPs with sigmoidal activations. The connection weights between input and hidden layer were shared among all the networks associated to a given phone class, independently of the context $c_k, k = 1, \dots, K$, and only the weights of connections between hidden and output layer were specialized on the specific context. In so doing, the complexity of the system (number of free parameters) was considerably reduced, thus limiting the effect of reduction of available training data for individual networks as the number K of contexts increased. A novel training procedure was introduced for this architecture.

Experimental results, carried out on a continuous speech, SI task (DARPA RM) using generalized biphone phonetic contexts, showed a 28.0% relative WER reduction with respect to the context-independent hybrid. When adopting a word-pair grammar and averaging over three different test sets, performance improved from

8.8% WER to 6.3% WER, whereas a standard context-dependent HMM scored a 7.0% WER.

Connectionist estimate of posterior probabilities is also adopted in [37], where the estimation is accomplished using a MLP which takes in input a time-window of previous acoustic observations, in addition to the current acoustic vector, so as to take the correlations among adjacent observations into account.

Robinson et al. [111,50,49] extended Bourlard's approach with the introduction of a recurrent network instead of a static MLP to estimate state-posteriors. Their system, called ABBOT, successfully participated to November 1993 ARPA Wall Street Journal Test. ABBOT is a continuous speech, SI recognition system for large vocabularies (more than 10,000 words). In [50] the basic system was improved by providing:

(1) *Combination of neural models*. Different recurrent nets are trained on different acoustic features, namely MEL and PLP coefficients. Furthermore, "backward recurrent" networks (i.e. models that start processing from the end of an input sequence back to the beginning) are introduced in addition to the conventional "forward recurrent" models, thus resulting in four parallel probability estimators. These models are then combined, either in a linear fashion (averaging on the corresponding probability estimates), or averaging in the log-domain (the latter choice allowing for a further gain in performance).

(2) *Introduction of a "posterior-directed path pruning"*. Pruning of the search space is accomplished relying, for each new input frame, on the connectionist probability estimates of individual phones. Paths containing phones, the posterior probability of which is below a given threshold, are immediately pruned.

Further improvements in performance for the ABBOT system are presented in [49], which discusses more sophisticated combination strategies for the recurrent models and improved techniques for phone-duration modeling.

Another approach which is conceptually analogous to that by Bourlard et al. is presented in [129], which does not use an HMM explicitly, but a Viterbi alignment strategy is rather applied on state scores computed as the output values of the second hidden layer of a feed-forward ANN.

A variant on hybrid systems where the ANNS are used as state-posterior probability estimators is described in [20]. The output of the networks is interpreted as a discriminant function (for instance, applying Bayes' theorem and assuming that all the states of the HMM share the same prior probability) capable to discriminate among the states of the model, i.e. a classical classification task. In this perspective, the approach of [20] consists in training MLPs as "state recognizers", relying on the discriminative training capabilities of feedforward nets, combining then the MLPs into a sequence (sequential multilayer perceptron: SMLP) with the aim of modeling and recognizing isolated words, within a DP scheme.

The system models the words in the vocabulary with such a sequence of MLPs, using a left-to-right chain of as many nets as the number of states in the corresponding underlying HMM. Each net has an output unit for each of the classes (words) to be recognized. The input sequence of acoustic features is fed through the networks, producing sequences of scores for each class. A decoding engine based on Viterbi algorithm is run on the scores obtained this way.

The training is in two steps:

(1) Each MLP is roughly trained by BP on pre-labeled sequences (e.g. the label for a given acoustic observation may represent the corresponding phoneme).
(2) Fine-tuning of MLPs parameters is accomplished using an optimization scheme based on Generalized Probabilistic Descent (GPD) [63,18] and BP of partial derivatives throughout the layers of the net.

Another variant to Bourlard's approach can be found in [104]. In that paper RBF networks, instead of MLPs, are used to estimate state-posterior probabilities. The training takes place again in two steps:

(1) RBF models are trained by BP over a training set which was previously labeled by a standard HMM.
(2) A global discriminative training of the hybrid, based on GPD with the minimum classification error criterion [61], is accomplished.

Also in [128] ANNs are used to estimate state posteriors, but a novelty is proposed in the training scheme. Instead of considering only the strict values "0" and "1" as targets for output units, according to Viterbi segmentation of training sequencies, the probability targets in the continuous range (0,1) generated within the forward–backward mechanism are considered. The training technique was evaluated on a recognition task of continuous digits collected over the telephone channel, where the novel hybrid scored a 4.9% WER, whereas the hybrid with standard training yielded a 6.0% WER, and a standard CDHMM scored a 5.0% WER.

Table 2 summarizes the main hybrid models of the present category, giving references, brief description and experimental results, in order to allow for a concise comparison.

3.3. The "global optimization" approach

The hybrid model described in [4,5,8] is based on the simultaneous estimation of all the parameters for both the HMM and the ANN according to the joint optimization of a single criterion, defined at the sequence level (e.g., word or sentence). The approach was originally inspired from the Alpha net (see Section 3.1).

The ANN is trained as a *feature extractor* for a CDHMM, with the goal to transform input acoustic representations into compact, but significant, low-dimensional representations that are more suitable to be modeled by the emission probabilities of the HMM than standard acoustic parameters. In so doing, an increased recognition rate is expected.

The ANN maps a raw input sequence $\boldsymbol{u}_1^L$ onto a low-dimensional observation sequence $\boldsymbol{x}_1^L$. An assumption is made that the optimization criterion C used to train the HMM is a continuous and differentiable function of the observations $\boldsymbol{x}_t$, $t = 1, \dots, L$, so that a gradient descent optimization method can be applied. The partial derivatives $\partial C / \partial x_t$ (here x_t denotes a generic component of tth observation $\boldsymbol{x}_t$)

Table 2
Summary of main hybrid architectures based on a connectionist estimate of posterior probability of HMM states (Section 3.2). First column identifies the model, giving the bibliographic reference. Second column provides a brief description. Last column reports experimental results from literature

Model	Brief description	Performance
Reference model [12,13,105]	A MLP is trained with BP over a window of acoustic frames to estimate the posterior probabilities of CDHMM states given an acoustic observation sequence. A discriminative maximum a posteriori criterion is considered	DARPA resource management (RM) "February91" test set, SI, continuous speech task, vocabulary size of 991: 5.8% WER (standard CDHMM: 11.0% WER)
Replacing MLP with RNN [111,50,49]	Extension of the approach relying on RNNs, instead of static MLPs, to estimate posterior probabilities	The system, called ABBOT, successfully participated in November 1993 ARPA Wall Street Journal Test. In 1994 a 8.8% WER on Nov93 *hub 2* test set, using a trigram LM, was reported
Explicit context modeling [36]	Acoustic context is explicitly considered: the criterion is the likelihood of observations given the HMM state *and* the context. The latter is computed from the posterior probability of states given the observations and the context. MLPs (one for each context) are trained to estimate such probability, using a variant of BP with sharing of input-to-hidden weights among the nets	DARPA RM, SI, continuous speech task, averaging over three test sets: 6.3% WER (context-independent hybrid: 8.8% WER; standard CDHMM: 7.0% WER)
Continuous target outputs [128]	Variant in the training scheme: instead of the strict targets "0" and "1", probability targets in the continuous range (0,1) (generated within the forward–backward mechanism) are considered	Recognition task of connected digits collected over the telephone channel: 4.9% WER (hybrid with Bourlard et al. training scheme: 6.0% WER; standard CDHMM: 5.0% WER)

can then be computed and used (along with the *chain rule*) to train the weights w of the ANN. The training scheme is sketched in Fig. 4.

Interestingly, the computation of $\partial C/\partial x_t$ is very simple, since the forward–backward algorithm implicitly computes the derivative of the HMM likelihood $Pr(\boldsymbol{x}_1^L \mid Y)$ with respect to the emission probabilities $b_{i,t} = Pr(\boldsymbol{x}_t \mid q_t = S_i)$.[1]

Beering in mind the definitions and the notation introduced in Section 2 and referring to [101] for details, it can be stated that in the forward pass the Baum–Welch

[1] For notational convenience, we write $b_{i,t}$ instead of $b_i(\boldsymbol{x}_t)$ to denote emission probability of state S_i at time t, keeping in mind that the observation sequence $\boldsymbol{x}_1^L$ is under consideration.

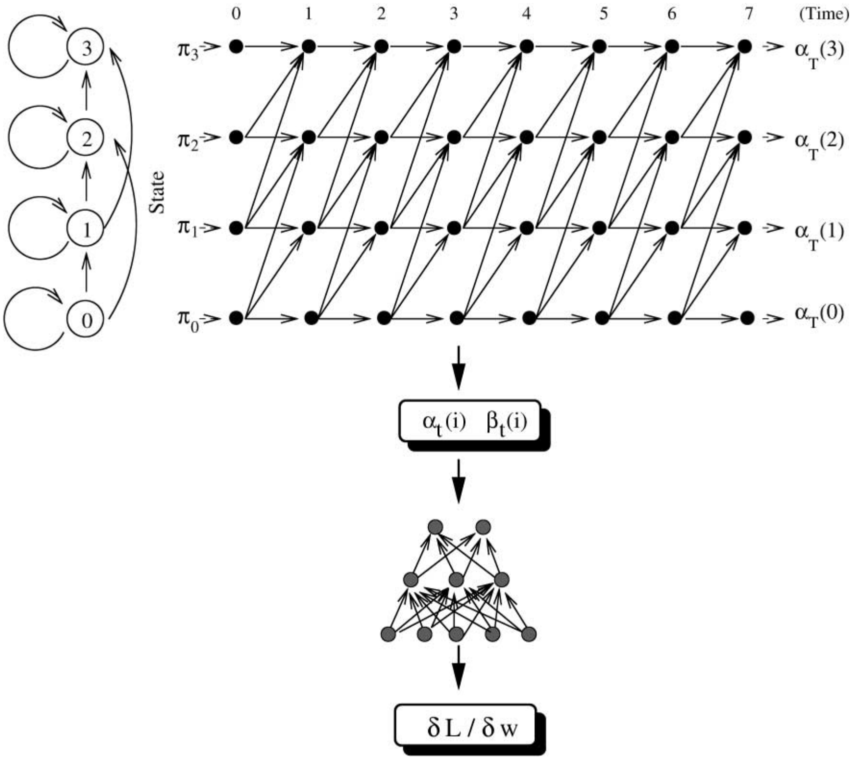

Fig. 4. Training scheme of Bengio's "global optimization" approach. A gradient-ascent learning rule for weight w of the ANN (which transforms raw input acoustic representations into low-dimensional, significant features) is obtained by taking partial derivatives of the likelihood L of the acoustic observations given the model. Quantities involved in the calculations, namely α's and β's at each time $t = 1, \ldots, T$ and for each HMM state i (a 4-state left-to-right HMM is shown), are *backpropgated* from the usual *forward–backward* algorithm applied to the standard *trellis* structure shown in the upper part of the figure.

algorithm computes

$$\alpha_{i,t} = b_{i,t} \sum_{j} a_{ij}\alpha_{j,t-1}, \tag{8}$$

where $\alpha_{i,t} = Pr(q_t = S_i, \boldsymbol{x}_1^t)$. It can be shown [101] that the likelihood $Pr(\boldsymbol{x}_1^L)$ can be obtained from the α's at the last time step, for the set $\mathscr{F}$ of "final" states of the HMM, as follows:

$$Pr(\boldsymbol{x}_1^L) = \sum_{i \in \mathscr{F}} \alpha_{i,L}. \tag{9}$$

The backward pass is initialized as follows:

$$\beta_{i,L} = 1 \text{ if } i \in \mathscr{F}, \quad 0 \text{ otherwise}, \tag{10}$$

where $\beta_{i,t} = Pr(\boldsymbol{x}_{t+1}^{L} \mid q_t = S_i)$. These are computed recursively as

$$\beta_{i,t} = \sum_{j} a_{ji} \beta_{j,t+1} b_{j,t+1}. \tag{11}$$

Note that from Eqs. (9) and (10) $\partial Pr(\boldsymbol{x}_1^L)/\partial \alpha_{i,L} = \beta_{i,L}$. From Eqs. (11) and (8) and proceeding by induction (applying also the chain rule), it is possible to write

$$\frac{\partial Pr(\boldsymbol{x}_1^L)}{\partial \alpha_{i,t}} = \beta_{i,t}. \tag{12}$$

Derivatives with respect to the emission probabilities can now be obtained from Eq. (8):

$$\frac{\partial Pr(\boldsymbol{x}_1^L)}{\partial b_{i,t}} = \frac{\alpha_{i,t}\beta_{i,t}}{b_{i,t}}. \tag{13}$$

When considering the case in which the training criterion C is the likelihood of the observations, and using an HMM constrained by the correct word sequence Y, the gradient of the criterion C with respect to the emission probabilities can be immediately obtained from Eq. (13). To compute the gradient with respect to the ANN outputs and thus, via chain rule, with respect to the ANN weights, the computation of the partial derivatives of the emission density function with respect to the observations is needed. Bengio et al. [5,8] give an example in the case of Gaussians or Gaussian mixtures emission pdfs.

It should be noted that in the above scheme an infinite value of the likelihood can be obtained whenever the ANN produces a constant output, since in this case the Gaussian mixtures can converge to zero variance, and the emission probabilities increase towards infinity, as well as the likelihood $Pr(\boldsymbol{x}_1^L)$. This troublesome behavior is the same that occurs when training standard Gaussian mixtures to maximize the likelihood when a Gaussian concentrates on one point. The problem did not occur in the experiments discussed in [5,8]. However, the problem can be definitely tackled by adopting discriminant training criteria, such as maximum mutual information or maximum a posteriori [5].

The basic model described above was further developed with the combination of multiple, specialized ANNs, trained on specific acoustic classes (e.g. plosives) with a class-dependent set of acoustic features. A *gathing* network that combined the outputs of specialized ANNs was initialized to perform principal component analysis. The whole HMM/multiple ANNs hybrid was then fine-trained with the global optimization technique.

Experimental results showed that training the ANN jointly with the HMM in the proposed hybrid architecture improved recognition performance over the standard CDHMM [5,8], lowering the error rate on a plosive recognition task from 19% to 14%.

In [59,57] the global optimization technique was used to train an MLP as an extractor of "adjoint" input features for an HMM relying on the Maximal Mutual Information criterion.

Recently, Riis and Krogh [108] proposed a more general framework for the global optimization approach, which extends the application of ANNs from feature extraction to the estimation of HMM parameters, obtaining encouraging experimental results in recognition of five broad phonetic classes when compared to a discrete HMM and a discriminative conditional ML training criterion was adopted.

As a final consideration, it should be noted that also the approaches where a joint optimization of the HMM and of the ANN is performed according to the GPD method (and that are described throughout the other Sections of this Survey) are, in general, instances of "globally optimized" architectures.

3.4. Networks as vector quantizers for discrete HMM

Starting from the late 1980s, a number of researchers began to apply ANNs to the problem of generating codebooks for discrete HMMs (see for instance [64,55]). Most of those works relied on Kohonen's learning vector quantization (LVQ) [65,79] as an effective neural alternative to standard clustering algorithms.

A good example of the principles and motivations underlying the present approach can be found in [106]. The latter proposed a new neural architecture to perform vector quantization (VQ) on the acoustic features for a discrete HMM. The novelty of the approach concerned the training of the ANN. The latter was a feedforward net trained with an unsupervised algorithm based on the maximal mutual information (MMI) criterion, having defined the mutual information (MI) as

$$\mathrm{MI}(Y, W) = H(Y) - H(Y \mid W), \tag{14}$$

where $H(\cdot)$ denotes entropy, Y is the sequence of labels (*codewords*) produced by the vector quantizer on the input feature stream, and W is the corresponding words string.

The proposed network topology is basically a 1-layer MLP with activations:

$$z_j = \|\boldsymbol{w}_j - \boldsymbol{x}\| = \sum_{i=1}^{d} (w_{ji} - x_i)^2, \tag{15}$$

which somehow express a "distance" between input vector $\boldsymbol{x} = (x_1, \dots, x_d)$ and the jth "codeword" (or prototype) of the codebook, represented by the connection weights for jth output unit (see also competitive neural networks [48] used for clustering).

The training algorithm iteratively modifies the weights of the winner unit in order to increase quantity (14). Ref. [106] extends also the procedure to the cases of multilayer nets and multiple input features.

Experimental results obtained by initializing the weights of the ANN with the *k-means* (or *isodata*) [30] clustering algorithm, in a SD task of 18 isolated Japanese phonemes from the ATR speech database, using a codebook size of 100, showed a 25.0% relative error reduction over the VQ based on k-means, averaging over 5 different speakers and using the same discrete HMM. Further improvements to the system were presented in [107,92,113], where evolutions of ANN architecture and training were discussed, along with experimental results that compare well with those obtained with a CDHMM.

Jang and Un [56] present an hybrid system for SI isolated words recognition. The system is constituted by a connectionist fuzzy-vector quantizer (FVQ), the output of which is fed into a discrete HMM. The proposed FVQ is a modular combination of TDNNs, where the nets are trained by BP on a training set labeled by hand. The TDNNs are used as phoneme classifiers. Two additional nets are trained to discriminate between vowels and consonants.

Instead of feeding the outputs of the networks into the HMM, the latter takes in input the internal representations (activations of the second hidden layer) of the acoustic features that the TDNNs develop. The internal representation is thought of as the (nonlinear) projection of the feature space over a transformed space, replacing the standard vector quantization process. The HMM is separately trained on the transformed space, using standard HMM-training algorithms.

In order to improve performance of the system and to allow training of the system with few data, a smoothing technique [102] is used. Emission probabilities are smoothed relying on a “smoothing matrix” computed from the average values of the activation vectors of the second hidden layer of the TDNNs.

A 44.9% relative WER reduction with respect to the discrete HMM with a floor-smoothing technique was gained in a SI (5 males for training, 2 males for test), isolated Korean words (75 words dictionary) recognition task, when 12 LPC MEL-cepstral coefficients were used.

Strictly related to the idea of using ANNs as vector quantizers for discrete HMMs is the concept of neural *labelers* [77,17]. In its basic version, this system is based on an MLP (or TDNN), fed with input acoustic features, with an output unit per each phonetic class. The actual output value can be 0 or 1, according to the membership of the input observation to the class or not. The output from the MLP is passed on to a standard discrete HMM that uses it as a class label within a Viterbi decoding paradigm. Training is supervised, relying on BP over a phonetically balanced training set that is pre-labeled by means of a standard discrete HMM.

The system turns out to be useful mainly for small vocabularies, where a model for each word is used (instead of phoneme models). The main advantage over approaches based on the estimation of state posterior probabilities is that no implicit requirement is made of reaching the global minimum of the error functional to keep consistency with the theoretical framework.

The basic model was then extended [17] with the introduction of *N-Top*, that is to say, the *N*-best (highest) MLP outputs are considered (not only the winner label) and passed on to the HMM. This is of particular interest in the regions of the feature space in proximity of separation surfaces between pairs of adjacent acoustic classes, where misclassification is more alike. Other improvements concerned the use of parallel MLPs trained on different acoustic features (e.g. bare coefficients, their first- and second order derivatives, energy) and the introduction of a *fuzzy vector quantization* scheme.

The system was evaluated on a SI, isolated Flemish digits task over the telephone line. Results reported in [17] are difficult to interpret, since they miss a direct comparison with a standard CDHMM; nonetheless, using MLP as labelers performed better than vector quantization based on standard clustering techniques, and yielded recognition rates comparable to those obtained with an hybrid HMM/ANN system with connectionist probability estimation.

Main hybrid architectures based on a connectionist vector quantization of acoustic feature space for discrete HMMs are summarized in Table 3.

3.5. Other approaches

One important topic is related to the concept of connectionist *rescoring*. The idea, proposed in [130], is developed in [132], where a connectionist approach is applied to

Table 3
Summary of main hybrid architectures based on a connectionist vector quantization of acoustic feature space for discrete HMMs (Section 3.4)

Model	Brief description	Performance
VQ based on maximal mutual information (MMI) [106,107,92]	A feedforward net (1- or multi-layer Perceptron) is trained to perform VQ for a discrete HMM with a novel unsupervised algorithm based on the MMI criterion. Output activation functions somehow express a "distance" between acoustic input vector and codewords of the discretized codebook of input symbols	ATR speech database, SD task of 18 isolated Japanese phonemes, codebook size of 100, averaging over five different speakers: 25.0% relative error reduction over the VQ based on *k*-means (using the same discrete HMM)
Connectionist fuzzy-vector quantizer (FVQ) [56]	A FVQ is used. It is a modular combination of TDNNs, where the nets are trained as phoneme classifiers by BP on a training set labeled by hand. TDNNs internal representations (activations of the second hidden layer) of the acoustic features are used as inputs for a discrete HMM, the latter being separately trained using standard algorithms	SI (five males for training, two males for test), isolated Korean words (75 words dictionary) recognition task: 44.9% relative WER reduction with respect to the standard discrete HMM
Neural *labelers* [77,17]	An MLP (or TDNN) with an output unit per each phonetic class is trained to yield output values equal to 0 or 1, according to the membership of the input observation to the class or not. The output from the MLP is passed on to a standard discrete HMM that uses it as a class label within a Viterbi decoding paradigm. Training is supervised, relying on BP over a pre-labeled training set	SI, isolated Flemish digits task (over the telephone line); using MLP as labelers performed better than VQ based on standard clustering techniques, and yielded recognition rates comparable to those obtained with an hybrid HMM/ANN system with connectionist probability estimation

the problem of rescoring the hypothesis generated by an HMM which uses an N-best strategy. In this case the network does not compute scores on individual acoustic frames, but on whole *segments* (sub-sequencies) of frames, corresponding to phonemes. In this way, correlations between frames that are close in time — and that correspond to the same phoneme — are exploited, thus overcoming the usual limitations following the assumption of independence made in standard HMMs.

The connectionist model used for this purpose is called segmental neural network (SNN) [1]. It takes a given acoustic segment in input, and produces an estimation of the posterior probability of a certain frame given that segment. The overall score, which is actually used in the rescoring process of a whole sentence is obtained by multiplying the scores of the segments in which the sentence has been segmented by the HMM. The score obtained this way is then linearly combined with that computed by the HMM, yielding a total score for the sentence given a specific segmentation. The weights for the linear combination are empirically determined by cross-validation on a development set.

Since the input layer of the networks has fixed size, while the segments are made up of a variable number of frames, two alternative techniques of "normalization" of the length of the input window are proposed. The former is a semilinear temporal sampling, which drops some of the original frames if the segment is longer than the input dimensionality, or replicates some of them if the segment is shorter. The alternative is a discrete cosine transform (DCT) applied to the whole segment, retaining as many parameters as it is necessary to fill the input layer.

Training of the system is accomplished using a segmentation performed by the HMM, possibly considering the first N-best hypothesis of segmentation — where segmentations from the 2nd best down to the Nth best are used as negative examples. Networks are trained on a relative entropy criterion; the outputs are normalized in order to fit the probabilistic framework.

Different neural architectures were tried (1-layer perceptron, MLP, HyperBF) [132] without significant fluctuations in performance. They were also combined altogether to obtain a more robust rescoring process.

The system was evaluated on the DARPA RM, SI, continuous speech task, with a 1000 words vocabulary and a bigrams grammar (perplexity 60), using also duration modeling, one left context (at the segment level) and a regularization technique to improve performance. The rescoring approach obtained combining HMM, SNN based on MLP and HyperBF, improved the WER of the bare HMM from 3.4% to 2.7% on the test set RM Feb91, and from 6.0% to 5.5% averaging on the two test sets RM Sep92.

Another application of ANNs for rescoring of a standard HMM with N-best strategy can be found in [89].

Connectionist *wordspotting* was discussed, for instance, in [86,72,74].

Somewhat similar to connectionist rescoring of HMM hypothesis is the technique by [78]. In this case, a standard HMM with N-best is used for connected speech recognition with confusable words (e.g. acoustically similar words). The neural nets are applied after the HMM has generated the N-best word-sequence hypothesis, to "correct" (or confirm) those individual words that belong to specific "confusable" classes. For each such a class, an MLP is trained to produce output scores for all the words of the given class. The MLP is fed with an input vector of acoustic features (MEL frequency scaled cepstral coefficients) extracted in an ad hoc manner, according to a prior knowledge concerning which position within an acoustic segment is more likely to allow for better discrimination among words of a certain class. Training of the MLPs is accomplished using BP. The proposed technique allowed for a significant gain in performance for some classes of confusable letters (e.g., “B”, “D”, and “V”), from 491 spellings of names collected over the telephone channel.

The approach proposed by [31] is not truly a hybrid, since it uses a standard HMM (trained with Viterbi algorithm on the ML criterion), which is used in parallel with connectionist models by combining the estimates of the emission probabilities (likelihoods) provided by the HMM with the normalized scores obtained with the ANNs. The linear combination scheme is the following:

$$\log P = \mu \log P_{\mathrm{HMM}} + (1-\mu) Q_{\mathrm{net}} \tag{16}$$

where $\log P$ is used as the combined emission log-likelihood estimate for a given state of the HMM, P_{HMM} is the usual emission likelihood yielded by the standard HMM, Q_{net} is the properly normalized output of the ANNs (a score assigned to the same state of the HMM) and μ is a tuned weighting linear coefficient in the range (0.0,1.0).

The combined value obtained from Eq. (16) is used within a Viterbi decoding strategy.

The ANNs are a hierarchical mixture of TDNNs, with a gather (root of the one-level tree which defines the hierarchy) which is trained, along with the experts of the mixture (the leaves of the hierarchy), to assign each acoustic frame to the proper TDNN. Each TDNN expert is specialized on subsets of phonemes, in order to reduce the complexity of training on whole corpora of continuous-speech data.

The networks are trained with the usual MSE criterion, instead of ML, and the approach is expected to improve the baseline Markovian recognizer thanks to the combination of the different training criteria.

In fact, the hierarchical TDNNs mixture, trained and tested on the DARPA RM SD corpus significantly worsened the performance of the HMM (test was carried out by averaging on three different reference speakers), while the results obtained with the TDNN/HMM combination (1.7% average WER) were slightly better then those provided by the HMM alone (2.0% average WER), that is a 15% relative WER reduction at the expense of only 10% increase in the number of free parameters to be estimated during training.

In [43] the classical framework of an HMM-based recognizer where ANNs are used as preprocessors (feature extractors, VQs, etc.), or probability estimators, is inverted. The HMM is used as a preprocessor for an ANN, the latter being the classifier. The HMM maps sequencies of acoustic features into subsequencies (the normalized means of the Gaussian emission probabilities of the states associated to the observation sequence according to Viterbi alignment) of fixed length (the number of states within the corresponding left-to-right model). The subsequence obtained this way is then passed in input to the ANN which, in turn, carries out the classification.

In [53] MLPs are trained by BP to estimate the class-dependent probability density functions associated to the states of a SCHMM.

A different approach is discussed in [22]. In the latter, ANNs are used for *weighting* the emission probabilities (likelihoods) associated to the states of a CDHMM for continuous speech recognition. The system relies on a CDHMM with mixtures of Gaussian components. The HMM is trained and used for decoding of speech signals using conventional algorithms (e.g. Viterbi), but an MLP (with an input window of three left contexts and three right contexts, i.e. a basic TDNN) is associated to each state of the HMM in order to estimate a state "weight" (or credit), given the current acoustic observations. Performance improvement is expected from the combination of HMM capabilities to classify long sequencies, with discriminative training properties of ANNs.

In practice, [22] used a pair of MLPs for each phoneme, one for the acoustic features (LPC cepstral coefficients) and the other for the corresponding time derivatives; each net was "split" over the three adjacent states of 3-state, left-to-right HMMs that modeled phonemes.

The training algorithm is a three-step procedure:

(1) Training of the CDHMM, using a conventional algorithm.
(2) Training of the MLPs as phoneme classifiers, using BP.
(3) Re-training of the MLPs with an iterative scheme in which the models obtained at steps 1 and 2 above are used (the re-trained MLPs are then used during the successive iterations) to classify the training data, evaluating a differentiable, discriminative criterion function aimed at minimizing the error rate (GPD). The partial derivatives of the criterion with respect to the weights of the MLPs are then backpropagated within the networks to obtain new weight values, and the next iteration is started.

The system was experimentally evaluated over a continuous speech, SI (16 speakers for training, 10 for testing) task, with vocabulary of 102 Korean words. The best results were obtained using mixtures of four Gaussian components: WER on the test set was reduced from 15.9% (HMM) to 15.1% (HMM with non-retrained MLPs), and finally down to a significant 11.1% (HMM with retrained MLPs).

In [21] another approach is presented, which extends the idea of neural prediction model (NPM) [54] by integrating non-linear neural predictors within an hybrid framework. Each predictor is a 2-layer MLP, with a hidden layer of sigmoids and a linear output layer, trained by BP. The MLP is fed with acoustic features, and produces in output the prediction for the acoustic frame at time t, given a time window of the acoustic observations which immediately precede and follow the tth frame. Such predictors are associated to the states of a CDHMM, where mixtures of Gaussian components are used to model the distributions of the prediction errors.

The major advantages of this system are the capability of the networks to take into account the temporal correlations between adjacent acoustic observations and the global optimization of the system. The latter relies on a GPD-based algorithm that simultaneously trains the HMM and the ANNs (the latter ones acting as feature extractors for the HMM).

Two training techniques are proposed, based on ML and on a discriminative criterion (minimum classification error), respectively. Experiments accomplished on a database consisting of 102 Korean words, in a continuous speech, SI (16 speakers for training and 27 speakers for testing) task, using 12 LPC plus energy and their first-order time-derivatives, showed a 22.8% WER for the hybrid architecture trained with ML criterion, and a 11.1% WER when the discriminative training criterion was used, while the standard HMM resulted in a 16.0% WER.

In [96,97] a novel neural model, called OWE (orthogonal weight estimator), is used in order to face the problem of context dependence. The OWE is applied after a step of standard context-dependent HMM (or, alternatively, to feed a second-order HMM with transformed input frames) in recognition tasks from the TIMIT database.

Recently, [84] have proposed a novel approach based on the joint optimization of models and feature space in order to improve robustness to noise in a standard HMM. The approach is based on the principle of *inversion* of a neural network [73], although no strict integration between HMM and ANN is adopted.

The inversion principle refers to the duality which holds between the weights of an ANN and the values of the inputs in the computation of the gradient of the criterion (error) function. Indeed, analogously to the calculation of the delta rule [94] for weight updating in the BP algorithm [114], it is possible to obtain an updating rule for the current d-dimensional input $\boldsymbol{x} = (x_1, \ldots, x_d)$ of an ANN by writing

$$x'_i = x_i - \eta \frac{\partial C}{\partial x_i} \tag{17}$$

for $i = 1, \ldots, d$, where $\boldsymbol{x}' = (x'_1, \ldots, x'_d)$ is the updated feature vector, i.e. a transformed input which is made more suitable to the ANN (the parameters of the latter are kept fixed) in order to extremize the training criterion, and C is the criterion itself (e.g. MSE). The effective calculation of the partial derivatives in Eq. (17) is accomplished by repeatedly applying the chain rule and backpropagating them throughout the layers of the network, like in the standard BP case. [84] apply the same approach for inversion of an HMM, writing

$$x'_{it} = x_{it} + \eta \frac{\partial C_{\text{HMM}}}{\partial x_{it}} \tag{18}$$

where x_{it} is ith coefficient of tth acoustic frame in the input sequence $\boldsymbol{x}$, x'_{it} is the same coefficient after the transformation has been applied, and C_{HMM} is the criterion to be optimized to increase performance of the HMM. For instance, if a ML training criterion is used, then $C_{\text{HMM}} = \log Pr(\boldsymbol{x}|\lambda) = \log \alpha_{N,L}$ (log-likelihood), where λ is the vector of parameters of the HMM and $\alpha_{N,L}$ is the usual *forward probability* term [101] used in the forward–backward algorithm, for the final state S_N, assuming an input sequence of length L. It should be noticed that in the ML case a maximization is to be carried out, instead of a minimization of a cost (error) function as in ANNs, and this motivates the presence of the sign " + " in Eq. (18).

In addition to the application of the gradient method, Moon and Hwang [84] propose also an inversion technique directly based on the Baum–Welch algorithm, as well as a combined scheme where the HMM inversion is combined with a MINIMAX-based [82] model parameter adaptation [68]. Parameter adaptation can be accomplished in a batch manner, as well as alternating it with an inversion step, in an iterative fashion. Experimental results on isolated digits with noise of the TI database with 16 speakers, using CDHMMs with seven states (a model for each digit) and mixtures of four Gaussian components defined over the 12-dimensional LPC space, showed that the combined scheme yields an improvement over the standard HMM, increasing the recognition rate from 25.73% to 54.90% with signal-to-noise ratio (SNR) of 5 dB, and from 76.63% to 89.27% with SNR of 20 dB.

[81] and [80] use time-space neural network (TSNN), a variant of TDNN, for probability estimation of phonemes given the acoustic observations. Different techniques for the integration within an HMM are presented, featuring the adoption of fuzzy-probabilistic information in a continuous speech, SD task.

Finally, other approaches are those by [41], where a different kind of acoustic units (named stationary–transitional units) is successfully used within an hybrid with connectionist estimates of the emission probabilities, by [32], and by [9,6]; the latter introduces a novel paradigm, namely the *input–output HMM*, which extends the concept of HMM by allowing for generation of sequences of *output values* in front of sequences in input.

Table 4 summarizes the main approaches discussed throughout the present section.

4. Some future issues

Hybrid speech recognizers allowed for significant gain in performance with respect to standard HMMs in a variety of situations. In addition to the scientific interest related to their theoretical framework, hybrid systems appear to be a flexible and efficient alternative paradigm, to be taken into full consideration by the speech community in the next years to face open ASR problems.

This paper surveyed different combination techniques that have been proposed in literature, highlighting a number of opportunities to take advantage from both HMMs and ANNs. One future issue concerns merging some of these techniques together, e.g. globally optimized connectionist feature extraction (as in Bengio's approach) for an hybrid system where ANNs perform probability estimation (as in Bourlard and Morgan's paradigm). ANNs properties can be exploited by combining neural models at different levels within the ASR system, increasing robustness of the latter, its acoustic modeling capabilities and supporting integration with the LM. Increase in robustness to acoustic variability is expected as a consequence of the generalization ability of ANNs, as well as from selective (adaptive) extraction of relevant acoustic features. Accuracy in acoustic modeling is expected due to the universal approximation property of ANNS [25], potentially suitable to model any form of the emission pdfs associated with HMM states. A different perspective [12] is to consider ANNs as optimal non-parametric estimators of Bayesian classifiers. Finally, integration with the LM is sought whenever connectionist approaches to language modeling [85], in addition to acoustic modeling, are used: under these

Table 4
Summary of main "other approaches" discussed in Section 3.5

Model	Brief description	Performance
Connectionist *rescoring* [130,132]	A *segmental neural network* computes scores on whole *segments* of frames, corresponding to phonemes, for re-scoring the segmentation hypothesis yielded by an HMM which uses an N-best strategy.	DARPA RM, SI, continuous speech task (vocabulary size of 1000): 2.7% WER on Feb91 test set (standard CDHMM; 3.4% WER); 5.5% WER averaging on Sep92 test sets (standard CDHMM; 6.0% WER).
Rescoring of "confusable words" [78]	MLPs are applied after the HMM has generated the N-best word-sequence hypothesis, to "correct" (or confirm) those individual words that belong to specific "confusable" classes.	A significant gain in performance for some classes of confusable letters (e.g. "B", "D", and "V"), from 491 spellings of names collected over the telephone channel.
Parallel ANN/HMM state-probability estiamtes [31]	HMM estimates of the emission probabilities are linearly combined with scores obtained with a hierarchical mixture of TDNNs.	DARPA RM SD task, averaging on three different reference speakers: 1.7% WER (standard HMM alone: 2.0% WER).
Connectionist *weighting* of state-mission probabilities [22]	ANNs are associated with HMM states to estimate a state "weight" (or credit), given the current acoustic observations. Training combines BP and GPD.	Continuous speech, SI (16 speakers for training, 10 for testing) task, vocabulary of 102 Korean words: 11.1% WER (standard CDHMM: 15.9% WER).
Hybrid HMM/neural prediction model (NPM) [21]	MLPs are fed with a time window of acoustic observations centered in the tth frame and predict the acoustic frame at time t. The pdfs associated with the states of a CDHMM are used to model the distributions of the prediction errors.	Continuous speech, SI (16 speakers for training and 27 speakers for testing) task, 102 Korean words: 11.1% WER (standard HMM: 16.0% WER).
HMM *inversion* [84]	Joint optimization of models and feature space via "model inversion": an updating rule for the current input vector is developed, and the transformed input is expected to be more suitable to the model.	TI database, isolated digits task with noise, 16 speakers: increase in recognition rate from 76.63% (standard CDHMM) to 89.27% (applying inversion) with SNR of 20 dB.

circumstances, a scenario where a global backpropagation learning scheme is defined within a homogeneous (neural) combined system can be hypothesized. Some issues that are strictly related to the open problems introduced so far follow.

An important topic in state-of-the-art ASR systems is the *model adaptation* [131,91], e.g. *speaker normalization* [51,119] or *channel adaptation* [26,117]. Speaker normalization deals with the problem of recognizing speech signal acquired from a new speaker, not included in the training population used to estimate parameters for the ASR system, whose voice may significantly differ from those used for training.

Differences in vocal tracts dramatically affect the recognition performance. Similarly, in channel adaptation the diversity in acoustic conditions is introduced by a novel acquisition channel, e.g., the telephone line. Related problems, which have to deal with system robustness, concern the electronic transducer, namely the microphone(s) used to capture the voice signal: the microphone used for testing may differ from that used to collect the training data, or it may no longer be close-mouth but placed at a certain distance from the talker(s). A *microphone array* [35,19] is sometimes needed to allow hands-free recognition of people moving within a room, and so on. Model adaptation techniques usually aim at building an acoustic front-end to a pre-trained HMM-based recognizer, or to tune the parameters of the latter, in order to increase recognition performance whenever acoustic conditions are changed with respect to those which held during the training, using a limited amount of acoustic material collected in the new conditions and avoiding a full re-training of the whole system. The nature of ANNs, e.g., their adaptivity and generalization ability, should be exploited to yield effective approaches to the problem of model adaptation. In particular, *on-line* learning [5] given a global training criterion defined at the recognizer level (e.g., ML or minimum WER) appears promising in this respect.

Another hot issue concerns ASR in noisy environments, such as a room with many people talking or loud background noises (e.g., noisy electrical devices), a car or the telephone line. Many noise-reduction approaches, based also on ANNs, have been investigated by the speech community, such as extraction of robust acoustic features or blind separation of sources [2]. Future hybrid systems should allow for adaptive and robust connectionist reduction, or filtering, of noise. As mentioned above, a globally optimized robust feature extraction process is sought.

A final remark to VLSI implementation of ASR systems is due: one major limitation of state-of-the-art ASR systems based on standard HMMs is the difficulty to implement them in hardware (at least when non-trivial tasks are considered). ANNs, as pointed out in Section 1, require much less parameters than HMMs and are easily implemented in VLSI. This could be a considerable advantage of many hybrid systems, since availability of compact and efficient hardware devices would increase applicability and diffusion of ASR to a large extent.

Acknowledgements

The authors wish to acknowledge Renato De Mori and Yoshua Bengio for their unvaluable contributions, Fabio Brugnara for his trellis artwork, and two anonymous reviewers of *Neurocomputing*, whose comments allowed for a significant improvement of the paper.

References

[1] S. Austin, G. Zavaliagkos, J. Makhoul, R. Schwartz, Speech recognition using segmental neural nets, International Conference on Acoustics, Speech and Signal Processing, San Franscisco, March 1992, pp. I-625–628.

[2] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.

[3] Y. Bengio, Radial basis functions for speech recognition, in Speech Recognition and Understanding: Recent Advances, Trends and Applications, NATO Advanced Study Institute Series F: Computer and Systems Sciences, 1990, pp. 293–298.

[4] Y. Bengio, A connectionist approach to speech recognition, Int. J. Pattern Recognition Artif. Intell. 7 (4) (1993) 647–667.

[5] Y. Bengio, Neural Networks for Speech and Sequence Recognition, International Thomson Computer Press, London, UK, 1996.

[6] S. Bengio, Y. Bengio, An EM algorithm for asynchronous input/output hidden Markov models, in: L. Xu (Ed.), International Conference on Neural Information Processing, Hong-Kong, 1996.

[7] Y. Bengio, R. De Mori, G. Flammia, R. Kompe, Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks, Proceedings of EuroSpeech'91, 1991.

[8] Y. Bengio, R. De Mori, G. Flammia, R. Kompe, Global optimization of a neural network-hidden Markov model hybrid, IEEE Trans. Neural Networks 3 (2) (1992) 252–259.

[9] Y. Bengio, P. Frasconi, Input/Output HMMs for sequence processing, IEEE Trans. Neural Networks 7 (5) (1996) 1231–1249.

[10] Y. Bengio, M. Gori, R. De Mori, Learning the dynamic nature of speech with back-propagation for sequences, Pattern Recognition Lett. 13 (5) (1992) 375–386. (Special issue on Artificial Neural Networks).

[11] Y. Bengio, P. Simard, P. Frasconi, Learning long-term dependencies with gradient descent is difficult, IEEE Trans. Neural Networks 5 (2) (1994) 157–166. (Special Issue on Recurrent Neural Networks, March 94).

[12] H. Bourlard, N. Morgan, Continuous speech recognition by connectionist statistical methods, IEEE Trans. Neural Networks 4 (6) (1993) 893–909.

[13] H. Bourlard, N. Morgan, Connectionist Speech Recognition. A Hybrid Approach, The Kluwer International Series in Engineering and Computer Science, Vol. 247, Kluwer Academic Publishers, Boston, 1994.

[14] H. Bourlard, N. Morgan, Connectionist Speech Recognition. A Hybrid Approach, Kluwer Academic Publishers, Boston, 1994, p. 117.

[15] H. Bourlard, C. Wellekens, Links between hidden Markov models and multilayer perceptrons, IEEE Trans. Pattern Anal. Mach. Intell. 12 (1990) 1167–1178.

[16] J.S. Bridle, Alphanets: a recurrent 'neural' network architecture with a hidden Markov model interpretation, Speech Commun. 9 (1) (1990) 83–92.

[17] P. Le Cerf, W. Ma, D. Van Compernolle, Multilayer perceptrons as labelers for hidden Markov models, IEEE Trans. Speech Audio Process. 2 (1) (1994) 185–193.

[18] P.C. Chang, B.H. Juang, Discriminative training of dynamic programming based speech recognizers, IEEE Trans. Speech Audio Process. 1 (1993) 135–143.

[19] C. Che, Q. Lin, J. Pearson, B. de Vries, J.L. Flanagan, Microphone arrays and neural networks for robust speech recognition, Proceedings of ARPA Human Language Technology (HLT), 1994, pp. 342–348.

[20] W.Y. Chen, S.H. Chen, C.J. Lin, A speech recognition method based on the sequential multi-layer perceptrons, Neural Networks 9 (4) (1996) 655–669.

[21] Y.J. Chung, C.K. Un, An MLP/HMM hybrid model using nonlinear predictors, Speech Commun. 19 (4) (1996) 307–316.

[22] Y.J. Chung, C.K. Un, Multilayer perceptrons for state-dependent weighting of HMM likelihoods, Speech Commun. 18 (1996) 79–89.

[23] P. Cosi, Y. Bengio, R. De Mori, Phonetically-based multi-layered networks for acoustic property extraction and automatic speech recognition, Speech Commun. 9 (1) (1990) 15–30.

[24] P. Cosi, P. Frasconi, M. Gori, N. Griggio, Phonetic recognition experiments with recurrent neural networks, Proceedings of the International Conference on Spoken Language, Banff, Canada, October 1992, pp. 1335–1338.

[25] G. Cybenko, Approximations by superpositions of a sigmoidal function, Math. Control Signals Systems 2 (1989) 303–314.

[26] S. Das, A. Nádas, D. Nahamoo, M. Pichney, Adaptation techniques for ambience and microphone compensation in the IBM Tangora speech recognition system, International Conference on Acoustics, Speech and Signal Processing, Adelaide, April 1994, pp. I-21–24.

[27] S.B. Davis, P. Mermelstein, Comparison of parametric representations of monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. 28 (4) (1980) 357–366.

[28] R. De Mori, Spoken Dialogues with Computers, Academic Press, London, UK, 1998.

[29] J.R. Deller Jr., J.G. Proakis, J.H. Hansen, Discrete Time Processing of Speech Signals, Macmillan Publishing Company, New York, USA, 1993.

[30] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

[31] C. Dugast, L. Devillers, X. Aubert, Combining TDNN and HMM in a hybrid system for improved continuous-speech recognition, IEEE Trans. Speech Audio Process. 2 (1) (1994) 217–223.

[32] S. Dupont, C. Ris, O. Deroo, V. Fontaine, J.M. Boite, L. Zanoni, Context-independent and context-dependent hybrid HMM/ANN systems for vocabulary independent tasks, Proceedings of EUROSPEECH, Vol. 4, Rhodi, 1997, pp. 1947–1950.

[33] J.L. Elman, Finding structure in time, Cognitive Sci. 14 (1990) 179–211.

[34] G. Flammia, Speaker independent consonant recognition in continuous speech with distinctive phonetic features, Master's Thesis, McGill University, School of Computer Science, 1991.

[35] J.L. Flanagan, D.A. Berkley, G.W. Elko, J.E. West, M.M. Sondhi, Autodirective microphone systems, Acustica 75 (1991) 58–71.

[36] H. Franco, M. Cohen, N. Moran, D. Rumelhart, V. Abrash, Context-dependent connectionist probability estimation in a hybrid hidden Markov model-neural net speech recognition system, Comput. Speech Language 8 (1994) 211–222.

[37] H. Franco, V. Digalakis, Temporal correlation modeling in a hybrid neural network/hidden Markov model speech recognizer, Proceedings of EUROSPEECH, Madrid, 1995, pp. 1681–1684.

[38] M.A. Franzini, K.F. Lee, A. Waibel, Connectionist Viterbi training: a new hybrid method for continuous speech recognition, International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM, 1990, pp. 425–428.

[39] P. Frasconi, M. Gori, G. Soda, Recurrent networks for continuous speech recognition, Computational Intelligence 90, Milano, Italy, Elsevier, Amsterdam, 1990.

[40] K. Fukunaga, Statistical Pattern Recognition, 2nd Edition, Academic Press, San Diego, 1990.

[41] R. Gemello, D. Albesano, F. Mana, Continuous speech recognition with neural networks and stationary-transitional acoustic units, ICNN, Houston, TX, USA, 1997, pp. 2107–2111.

[42] M. Gori, Y. Bengio, R. De Mori, BPS: a learning algorithm for capturing the dynamical nature of speech, Proceedings of the International Joint Conference on Neural Networks, Washington, DC, IEEE, New York, 1989, pp. 643–644.

[43] F.S. Gurgen, J.M. Song, R.W. King, A continuous HMM based preprocessor for modular speech recognition neural networks, Proceedings of ICSLP, Yokohama, 1994, pp. 1507–1510.

[44] P. Haffner, M. Franzini, A. Waibel, Integrating time alignment and neural networks for high performance continuous speech recognition, International Conference on Acoustics, Speech and Signal Processing, Toronto, 1991, pp. 105–108.

[45] P. Haffner, A. Waibel, K. Shikano, Fast back-propagation learning methods for large phonemic neural networks, Proceedings of Eurospeech'89, 1989.

[46] J.B. Hampshire, A.H. Waibel, A novel objective function for improved phoneme recognition using time-delay neural networks, IEEE Trans. Neural Networks 1 (2) (1990) 216–228.

[47] J. Hennebert, C. Ris, H. Bourlard, S. Renals, N. Morgan, Estimation of global posteriors and forward-backward training of hybrid HMM/ANN systems, Proceedings of EUROSPEECH, Vol. 4, Rhodi, 1997, pp. 1951–1954.

[48] J. Hertz, A. Krogh, R. Palmer, Introduction to the Theory of Neural Computation, Addision-Wesley, Reading, MA, 1991.

[49] M.M. Hochberg, S.J. Renals, A.J. Robinson, G.D. Cook, Recent improvements to the ABBOT large vocabulary csr system, International Conference on Acoustics, Speech and Signal Processing, Detroit, 1995, pp. 69–72.

[50] M.M. Hochberg, S.J. Renals, A.J. Robinson, D.J. Kershaw, Large vocabulary continuous speech recognition using a hybrid connectionist-HMM system, Proceedings of CSLP, Yokohama, 1994, pp. 1499–1502.

[51] X.D. Huang, Speaker normalization for speech recognition, International Conference on Acoustics, Speech and Signal Processing, San Franscisco, March 1992, pp. I-465–468.

[52] X.D. Huang, Y. Ariki, M. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh, 1990.

[53] H.-P. Hutter, Comparison of a new hybrid connectionist-SCHMM approach with other hybrid approaches for speech recognition, International Conference on Acoustics, Speech and Signal Processing, Detroit, 1995, pp. 3311–3314.

[54] K. Iso, T. Watanabe, Speaker-independent word recognition using a neural prediction model, International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM, 1990, pp. 441–444.

[55] H. Iwamida, S. Katagiri, E. McDermott, Speaker-independent large vocabulary word recognition using an LVQ/HMM hybrid algorithm, International Conference on Acoustics, Speech and Signal Processing, Toronto, 1991, pp. 553–556.

[56] C.S. Jang, C.K. Un, A new parameter smoothing method in the hybrid TDNN/HMM architecture for speech recognition, Speech Commun. 19 (4) (1996) 317–324.

[57] F.T. Johansen, Global optimisation of HMM input transformations, Proceedings of ICSLP, Vol. 1, Yokohama, 1994, pp. 239–242.

[58] F.T. Johansen, A comparison of hybrid HMM-architectures using global discriminative training, Proceedings of ICSLP, Philadelphia, 1996, 498–501.

[59] F.T. Johansen, M.H. Johnsen, Non-linear input transformations for discriminative HMMs, International Conference on Acoustics, Speech and Signal Processing, Vol. 1, Adelaide, 1994, pp. 225–228.

[60] M.I. Jordan, Serial order: a parallel, distributed processing approach, in: J.L. Elman, D.E. Rumelhart (Eds.), Advances in Connectionist Theory: Speech, Lawrence Erlbaum, Hillsdale, 1989.

[61] B.H. Juang, S. Katagiri, Discriminative learning for minimum error classification, IEEE Trans. Signal Process. 40 (12) (1992) 3043–3054.

[62] J.C. Junqua, J.P. Haton, Robustness in Automatic Speech Recognition: Fundamentals and Applications, Kluwer Academic Publishers, Boston, USA, 1996.

[63] S. Katagiri, C.-H. Lee, B.H. Juang, New discriminative training algorithms based on the generalized probabilistic descent method, Proceedings of the IEEE Workshop on Neural Networks for Signal Processing, 1991, pp. 299–308.

[64] D. Kimber, M.A. Bush, G.N. Tajchman, Speaker-independent vowel classification using hidden Markov models and LVQ2, International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM, 1990, pp. 497–500.

[65] T. Kohonen, Learning vector quantization for pattern recognition, Report TKK-F-A601, Helsinki University of Technology, Espoo, Finland, 1986.

[66] K.J. Lang, G.E. Hinton, The development of the time-delay neural network architecture for speech recognition, Technical Report CMU-CS-88-152, Carnegie-Mellon University, 1988.

[67] Y. LeCun, Learning processes in a asymmetric threshold network, in: F. Fogelman Soulie, E. Bienenstock, G. Weisbuch (Eds.), Disordered Systems and Biological Organization, Springer, Les Houches, France, 1986, pp. 233–240.

[68] C.-H. Lee, C.-H. Lin, B.-H. Juang, A study on speaker adaptation of the parameters of continuous density hidden Markov models, IEEE Trans. Signal Process. 39 (4) (1991) 806–814.

[69] C.H. Lee, F.K. Soong, K.K. Paliwal (Eds.), Automatic Speech and Speaker Recognition: Advanced Topics, Kluwer Academic Publishers, Boston, USA, 1996.

[70] E. Levin, Word recognition using hidden control neural architecture, International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM, 1990, pp. 433–436.

[71] E. Levin, R. Pieraccini, E. Bocchieri, Time-warping network: a hybrid framework for speech recognition, in: J.E. Moody, S.J. Hanson, R.P. Lippmann (Eds.), Advances in Neural Information Processing Systems 4, Denver, CO, 1992, pp. 151–158.

[72] K.P. Li, J.A. Naylor, A whole word recurrent neural network for keyword spotting, International Conference on Acoustics, Speech and Signal Processing, San Franscisco, March 1992, pp. II-81–84.

[73] A. Linden, J. Kindermann, Inversion of multilayer nets, International Joint Conference on Neural Networks, Washington, DC, June 1989, pp. 425–430.

[74] R. Lippmann, E. Singer, Hybrid neural-network/HMM approaches to wordspotting, International Conference on Acoustics, Speech and Signal Processing, 1993, pp. I–565–568.

[75] R.P. Lippmann, Review of neural networks for speech recognition, Neural Comput. 1 (1989) 1–38.

[76] R.P. Lippmann, B. Gold, Neural classifiers useful for speech recognition, IEEE Proceedings of First International Conference on Neural Networks, Vol. IV, San Diego, CA, 1987, pp. 417–422.

[77] W. Ma, D. Van Compernolle, TDNN labeling for a HMM recognizer, International Conference on Acoustics, Speech and Signal Processing, 1990.

[78] J.-F. Mari, D. Fohr, Y. Anglade, J-C Junqua, Hidden Markov models and selectively trained neural networks for connected confusable word recognition, Proceedings of ICSLP, Yokohama, 1994, pp. 1519–1522.

[79] E. McDermott, S. Katagiri, LVQ-based shift-tolerant phoneme recognition, IEEE Trans. Signal Process. 39 (6) (1991) 1398–1411.

[80] X. Menendez-Pidal, R. de Cordoba, J. Ferreiros, J.M. Pardo, Incorporating fuzzy modelling in a hybrid HMM-ANNs system for CSR tasks, Proceedings of EUROSPEECH, Madrid, 1995, pp. 1689–1692.

[81] X. Menendez-Pidal, J. Ferreiros, R. de Cordoba, J.M. Pardo, Recent work in hybrid neural networks and HMM systems in CSR tasks, Proceedings of ICSLP, Yokohama, 1994, pp. 1515–1518.

[82] N. Merhav, C.H. Lee, A minimax classification approach with application to robust speech recognition, IEEE Trans. Speech Audio Process. 1 (1993) 90–100.

[83] M.L. Minsky, S.A. Papert, Perceptrons, MIT Press, Cambridge, 1969.

[84] S. Moon, J.N. Hwang, Robust speech recognition based on joint model and feature space optimization of hidden Markov models, IEEE Trans. Neural Networks 8 (2) (1997) 194–204.

[85] D.P. Morgan, C.L. Scofield, Neural Networks and Speech Processing, Kluwer, Norwell, MA, 1991.

[86] D.P. Morgan, C.L. Scofield, J.E. Adcock, Multiple neural network topologies applied to keyword spotting, International Conference on Acoustics, Speech and Signal Processing, Toronto, 1991, pp. 313–316.

[87] N. Morgan, H. Bourland, Continuous speech recognition using multilayer perceptrons with hidden Markov models, International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM, 1990, pp. 413–416.

[88] N. Morgan, Y. Konig, S.L. Wu, H. Bourlard, Transition-based statistical training for ASR, Proceedings of IEEE Automatic Speech Recognition Workshop (Snowbird), 1995, pp. 133–134.

[89] T. Moudenc, R. Sokol, G. Mercier, Segmental phonetic features recognition by means of neural-fuzzy networks and integration in an N-best solutions post-processing, Proceedings of ICSLP, Philadelphia, 1996, pp. 338–341.

[90] M.C. Mozer, Neural net architectures for temporal sequence processing, in: A. Weigend, N. Gershenfeld (Eds.), Predicting the future and understanding the past, Addison-Wesley, Redwood City, CA, 1993, pp. 243–264.

[91] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, T. Robinson, Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system, Proceedings of EUROSPEECH, Madrid, September 1995, pp. 2171–2174.

[92] C. Neukirchen, G. Rigoll, Advanced training methods and new network topologies for hybrid MMI-connectionist/HMM speech recognition systems, International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, 1997, pp. 3257–3260.

[93] L.T. Niles, H.F. Silverman, Combining hidden Markov models and neural network classifiers, International Conference on Acoustics, Speech and Signal Processing, Albuquerque, NM, 1990, pp. 417–420.

[94] Y.H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison-Wesley, Reading, MA, 1989.

[95] B.A. Pearlmutter, Learning state space trajectories in recurrent neural networks, Neural Comput. 1 (1989) 263–269.

[96] N. Pican, D. Fohr, J-F. Mari, HMMs and OWE neural network for continuous speech recognition, Proceedings of ICSLP, Philadelphia, 1996, pp. 1309–1312.

[97] N. Pican, J.-F. Mari, D. Fohr, Continuous speech recognition using a context sensitive ANN and HMMs, Proceedings of EUROSPEECH, Vol. 1, Rhodi, 1997, pp. 95–98.

[98] F.J. Pineda, Recurrent back-propagation and the dynamical approach to adaptive neural computation, Neural. Comput. 1 (1989) 161–172.

[99] M.J.D. Powell, Radial basis functions for multivariable interpolation: a review, in: J.C. Mason, M.G. Cox (Eds.), Algorithms for Approximation: IMA-1985 Conference, Clarendon Press, Oxford, 1987.

[100] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice-Hall, Englewood Cliffs, NJ, USA, 1993.

[101] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proc. IEEE 77 (2) (1989) 257–286.

[102] L.R. Rabiner, B.H. Juang, An introduction to hidden Markov models, IEEE ASSP Mag. 77 (1986) 257–285.

[103] L.R. Rabiner, B.H. Juang, C.H. Lee, An overview of automatic speech recognition, in: C.H. Lee, F.K. Soong, K.K. Paliwal (Eds.), Automatic Speech and Speaker Recognition: Advanced Topics, Kluwer Academic Publishers, Boston, USA, 1996.

[104] W. Reichl, G. Ruske, A hybird RBF-HMM system for continuous speech recognition, International Conference on Acoustics, Speech and Signal Processing, Detroit, 1995, pp. 3335–3338.

[105] S. Renals, N. Morgan, H. Bourlard, M. Cohen, H. Franco, Connectionist probability estimators in HMM speech recognition, IEEE Trans. Speech Audio Process. 2 (1) (1994) 161–1174.

[106] G. Rigoll, Maximum mutual information neural networks for hybrid connectionist-HMM speech recognition systems, IEEE Trans. Speech Audio Process. 2 (1) (1994) 175–1184.

[107] G. Rigoll, Ch. Neukirchen, J. Rottland, Large vocabulary speaker-independent continuous speech recognition with a new hybrid system based on MMI-neural networks, Proceedings of EUROSPEECH, Madrid, 1995, pp. 1659–1662.

[108] S.K. Riis, A. Krogh, Hidden neural networks: a framework for HMM/NN hybrids, International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, 1997, pp. 3233–3236.

[109] A.J. Robinson, F. Fallside, Static and dynamic error propagation networks with application to speech coding, in: D.Z. Anderson (Ed.), Neural Information Processing Systems, American Institute of Physics, New York, Denver, CO, 1988, pp. 632–641.

[110] T. Robinson, A real-time recurrent error propagation network word recognition system, International Conference on Acoustics, Speech and Signal Processing, Vol. I, 1992, pp. 617–620.

[111] T. Robinson, An application of recurrent nets to phone probability estimation, IEEE Trans. Neural Networks 5 (2) (1994) 298–305.

[112] T. Robinson, F. Fallside, A recurrent error propagation network speech recognition system, Comput. Speech Language 5 (3) (1991) 259–274.

[113] J. Rottland, Ch. Neukirchen, D. Willett, G. Rigoll, Large vocabulary speech recognition with context dependent MMI-connectionist/HMM systems using the WSJ database 1, Proceedings of EUROSPEECH, Vol. 1, Rhodes, 1997, pp. 79–82.

[114] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in: D.E. Rumelhart, J.L. McClelland (Eds.), editors, Parallel Distributed Processing, Vol. 1, MIT Press, Cambridge, 1986, pp. 318–362 (Chapter 8).

[115] M. Sato, A real time learning algorithm for recurrent analog neural networks, Biol. Cybernet. 62 (1990) 237–241.

[116] E. Singer, R.P. Lippmann, A speech recognizer using radial basis function neural networks in an HMM framework, International Conference on Acoustics, Speech and Signal Processing, Vol. 1, San Franscisco, March 1992, pp. 629–632.

[117] J. Takahashi, S. Sagayama, Telephone line characteristic adaptation using vector field smoothing technique, Proceedings of ICSLP, Yokohama, September 1994, pp. 991–994.

[118] J. Tebelskis, A. Waibel, B. Petek, O. Schmidbauer, Continuous speech recognition using linked predictive networks, in: R.P. Lippman, R. Moody, D.S. Touretzky (Eds.), Advances in Neural Information Processing Systems 3, Morgan Kaufmann, San Mateo, Denver, CO, 1991, pp. 199–205.

[119] E. Trentin, D. Giuliani, Speaker normalization with a mixture of recurrent networks, Proceedings of ESANN97, European Symposium on Artificial Neural Networks, Bruges, Belgium, April 1997.

[120] A. Waibel, Modular construction of time-delay neural networks for speech recognition, Neural Comput. 1 (1989) 39–46.

[121] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang, Phoneme recognition using time-delay neural networks, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 328–339.

[122] A. Waibel, H. Sawai, K. Shikano, Modularity and scaling in large phonemic neural networks, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 1888–1898.

[123] P. Werbos, Beyond regression: new tools for prediction and analysis in the behavioral sciences, Ph.D. Thesis, Harvard University, 1974.

[124] P.J. Werbos, Generalization of backpropagation with application to a recurrent gas market model, Neural Networks 1 (1988) 339–356.

[125] P. Wilinski, B. Solaiman, A. Hillion, W. Czamecki, Toward the border between neural and markovian paradigms, IEEE Trans. Systems Man Cybernet. 28 (2) (1998) 146–159.

[126] R.J. Williams, D. Zipser, Experimental analysis of the real-time recurrent learning algorithm, Connection Sci. 1 (1989) 87–111.

[127] R.J. Williams, D. Zipser, A learning algorithm for continually running fully recurrent neural networks, Neural Comput. 1 (1989) 270–280.

[128] Y. Yan, M. Fanty, R. Cole, Speech recognition using neural networks with forward–backward probability generated targets, International Conference on Acoustics, Speech and Signal Processing, Munich, Germany, 1997, pp. 3241–3244.

[129] D. Yu, T. Huang, D.W. Chen, A multi-stage NN/HMM hybrid method for high performance speech recognition, Proceedings of ICSLP, Yokohama, 1994, pp. 1503–1506.

[130] G. Zavaliagkos, S. Austin, J. Makhoul, R. Schwartz, A hybrid continuous speech recognition system using segmental neural nets with hidden Markov models, Int. J. Pattern Recognition Artif. Intell. (1993) 305–319. (Special Issue on Applications of Neural Networks to Pattern Recognition (I. Guyon Ed.)).

[131] G. Zavaliagkos, R. Schwartz. J. Makhoul, Batch, incremental and instantaneous adaptation techniques for speech recognition, International Conference on Acoustics, Speech and Signal Processing, Detroit, May 1995, pp. I-676–679.

[132] G. Zavaliagkos, Y. Zhao, R. Schwartz, J. Makhoul, A hybrid segmental neural net/hidden Markov model system for continuous speech recognition, IEEE Trans. Speech Audio Process. 2 (1) (1994) 151–160.

Edmondo Trentin received his Laurea "Cum Laude" in Computer Science from the Università di Milano, Italy, in 1990, with a research thesis in the area of hypertext systems. From 1990 to the end of 1993 he worked as a Project Leader in SGS-Elsag. Since January 1994 he has been working as a researcher at ITC-irst (Trento, Italy), in the area of Interactive Sensory Systems. His research interests include learning, artificial neural networks (ANN), statistical pattern recognition, robotics, hidden Markov models (HMM), and hybrid ANN/HMM systems. He is involved in projects of scientific and technological relevance in the field of speech processing. He is the author of more than 25 scientific publications. At present, he is also a PhD fellow at the Università di Firenze, Italy. Edmondo Trentin is a member of INNS (International Neural Network Society), SIREN (Italian Neural Networks Society) and IAPR-IC (International Association for Pattern Recognition, Italian Chapter).

Marco Gori received the Laurea in electronic engineering from Università di Firenze, Italy, in 1984, and the Ph.D. degree in 1990 from Università di Bologna, Italy. From October 1988 to June 1989 he was a visiting student at the School of Computer Science (McGill University, Montreal). In 1992, he became an Associate Professor of Computer Science at Università di Firenze and, in November 1995, he joint the University of Siena, where he is currently full professor. His main research interests are in pattern recognition (especially document processing) and neural networks. Dr. Gori was the general chairman of the Second Workshop of Neural Networks for Speech Processing held in Firenze in 1992, organized the NIPS'96 post-conference workshop on "Artificial Neural Networks and Continuous Optimization: Local Minima and Computational Complexity," and co-organized the Caianiello Summer School on "Adapting Processing of Sequences" held in Salerno on September 1997. He co-edited the volume *Topics in Artificial Intelligence* (Springer-Verlag, 1995) which collects the contributions of the 1995 Italian Congress of Artificial Intelligence. Dr. Gori serves as a Program Committee member of several workshops and conferences mainly in the area of Neural Networks and acted as Guest Co-Editor of the Neurocomputing Journal for the special issue on recurrent neural networks (July 1997). He is an Associate Editor of several journals in the area of expertise, including the IEEE Trans. Neural Networks, Neurocomputing, and Pattern Recognition. He is the Italian chairman of the IEEE Neural Network Council (R.I.G.) and is a member of the IAPR, SIREN, and AI*IA Societies. He is also a senior member of the IEEE.