Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2006) 398–408
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Inversion-based nonlinear adaptation of noisy acoustic parameters for a neural/HMM speech recognizer

Edmondo Trentin*, Marco Gori

Dipartimento di Ingegneria dell'Informazione, Università di Siena, V. Roma, 56 Siena, Italy

Received 25 July 2005; received in revised form 13 December 2005; accepted 28 December 2005
Communicated by G. Palm
Available online 27 June 2006

Abstract

Spoken human–machine interaction in real-world environments requires acoustic models that are robust to changes in acoustic conditions, e.g. presence of noise. Unfortunately, the popular hidden Markov models (HMM) are not noise tolerant. One way to increase recognition performance is to acquire a small adaptation set of noisy utterances, which is used to estimate a normalization mapping between noisy and clean features to be fed into the acoustic model. This paper proposes an unsupervised maximum-likelihood gradient-ascent training algorithm (instead of the usual least squares regression) for a neural feature adaptation module, properly combined with a hybrid connectionist/HMM speech recognizer. The algorithm is inspired by the so-called "inversion principle", that prescribes the optimization of the input features instead of the model parameters. Simulation results on a real-world speaker-independent continuous speech corpus of connected Italian digits, corrupted by noise, validate the approach. A small neural net (13 hidden neurons) trained over a single adaptation utterance for one iteration yields a 18.79% relative word error rate (WER) reduction over the bare hybrid, and a 65.10% relative WER reduction over the Gaussian-based HMM.

Keywords: Hybrid ANN/HMM; Model inversion; Feature adaptation; Speech recognition

1. Introduction

Spoken human–machine interaction in real-world environments often raises the problem of dealing with acoustic conditions that are far different from those characterizing the training stage of the acoustic model. In particular, automatic speech recognition (ASR) under noisy conditions has become a major topic in the speech processing community [31,22,14,21]. While, during the 1990s, the attention of researchers was mostly focused on large vocabulary systems under relatively quiet conditions (clean signals), at present time the challenge mainly concerns recognition of small vocabularies (e.g. the digits and/or a few dozen words) in noisy contexts such as a car, an office environment, or the telephone line. It is an established fact that hidden Markov models (HMM) [25], the most popular acoustic models, are extremely noise sensitive [32]. In addition, standard HMM-based ASR systems suffer from intrinsic limitations [5,29], in particular: (1) imposition of a specific parametric assumption for the emission probability density functions (*pdfs*); (2) lack of exploitation of possible correlations among input features; (3) limited *generalization* capabilities. We argue that this last drawback has a significant responsibility on the noise-sensitiveness of HMMs, given also the lack of suitable regularization techniques for HMMs. A survey of limitations of HMMs can be found in [29].

Approaches to tackle the problem of noise-tolerance usually rely on signal processing techniques (e.g., noise filtering), or on the adaptation of acoustic model parameters in order to increase the model robustness. Examples are noise reduction via spectral subtraction [23], blind source separation [1], parameter adaptation via input feature *normalization* (relying on linear transformations [10,19,12] or connectionist acoustic front-ends [28]), etc. Similar scenarios occur whenever other conditions change

*Corresponding author. Tel.: +39 0577 234636; fax: +39 0577 233602.
E-mail addresses: trentin@dii.unisi.it (E. Trentin), marco@dii.unisi.it (M. Gori).

0925-2312/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2005.12.130

from the training to the test setups: relevant instances concern the vocal tracts of the speaker, in the so-called "speaker normalization" problem [28], the communication channel (e.g., the telephone line) in the channel-compensation problem, or the transducer (i.e., the microphone, or a microphone array [13,22]) used to feed the ASR system. In the following, the term "noise" may thus refer to generic situations in which test acoustic conditions are not constrained to reflect the training conditions.

In [30] we introduced a novel ASR hybrid system that turned out to be intrinsically more robust to noise than standard HMMs, and that overcame the limitations listed above. This ANN/HMM hybrid is based on the combination of artificial neural networks (ANNs) and HMMs. It is related to the popular paradigms proposed by Bourlard and Morgan [4,5] and by Bengio [3,2]. Bourlard and Morgan's approach features an architecture where a multilayer perceptron (MLP) [27] estimates probabilistic quantities (the so-called *conditional transition probabilities*) associated with individual states of the underlying HMM. Unfortunately, in this framework the training of the MLP is heuristic. It is accomplished by applying standard backpropagation (BP) [27] over a synthetic supervised training set, which specifies artificial 0/1 targets for the ANN outputs. Bengio's approach uses instead an ANN as a feature extractor (featuring dimensionality reduction) that feeds a standard Gaussian-based HMM. It provides us with a formalism that forms the basis for the development of hybrid training algorithms aimed at the extremization of a global criterion function via joint optimization of the HMM parameters and of the ANN weights.

The hybrid we proposed in [30] relies on an HMM topology, including standard *initial* probabilities π and *transition* probabilities $\mathbf{a} = [a_{ij}]$ estimated by means of the *Baum–Welch* algorithm [25], while the *emission* probabilities $\mathbf{b}(\mathbf{y})$ [25] are estimated by an ANN. An output unit of the ANN holds for each of the states in the HMM, with the understanding that ith output value $o_i(t)$ represents the emission probability $b_{i,t}$ for the corresponding (i-th) state, evaluated over current acoustic observation $\mathbf{y}_t$. In the following, we refer to this ANN with the symbol ψ. Whereas recognition is accomplished applying the usual Viterbi algorithm [25,16], novel training techniques have been introduced. Learning rules for connection weights and for neuron biases are calculated according to gradient-ascent to maximize a global criterion function, namely the *likelihood* of the acoustic model given the acoustic observation sequences (ML criterion). In [30] we also pointed out that the performance of the system may be increased using a discriminative maximum likelihood-ratio criterion, called MAP, on the same architecture and relying on the same calculations.

In this paper we extend this ANN/HMM with the introduction of a connectionist adaptation module that realizes an acoustic feature normalization for the ANN/HMM itself. The adaptation is based on the "inversion principle", introduced in [20] for standard ANNs, and generalized to HMMs in [24]. The idea underlying the model inversion is the following. In the usual BP training algorithm, in order to learn the ANN connection weights, the gradient method is applied by backpropagating the partial derivatives of an error functional w.r.t. the weights themselves. In so doing, a reduction is obtained of the mismatch between the ANN output and a given target output associated with current (fixed) input. The ANN inversion exploits the duality between weights and inputs, by prescribing to carry out the backpropagation of partial derivatives of the error w.r.t. the inputs, while keeping the weights fixed. This is expected to lead to inputs that are more suitable to the desired ANN output. This method was successfully applied to different scenarios, including solution to inverse problems [15], or ANN query-learning [17,18]. In [24] this approach is extended to HMMs, aiming to transform acoustic feature vectors toward regions of the acoustic space that are "closer" to the recognition system, i.e. yielding higher likelihood of the HMM given the acoustic observations. A joint optimization of the acoustic parameters and of the model parameters emerges as a suitable extension of the paradigm. This algorithm is rather complex, due to the difficulties of computing partial derivatives of the input vectors within the Baum–Welch training scheme of the HMM. It also raises numerical stability problems, mostly due to the computation of derivatives over a training procedure that is not inherently gradient-based, as well as to the presence of Gaussian emission *pdfs* within the HMM (which basically tend to yield close-to-zero values over the whole feature space). In [24] an attempt to tackle these drawbacks is proposed in terms of an overall gradient-based re-estimation algorithm. Another problem with this approach is that the transformation of the input in order to increase the likelihood of the system does not necessarily lead to a projection of the acoustic vectors toward the correct state-sequence within the HMM (i.e., the sequence of phone-models that corresponds to the uttered signal), but, possibly, only to an overall increase of the likelihood of the whole HMM set, inducing the risk of a loss in terms of discriminative power.

In this paper we elaborate on the development and application of the inversion principle to the proposed ANN/HMM hybrid. As pointed out in Section 2, the inversion of a model that is intrinsically trained by means of a gradient-ascent method turns out to be computationally simple and natural. Moreover, inherent numerical stability follows from the homogeneity of quantities involved in both the ANN/HMM dynamics and in the inversion process. In addition, we take the bare inversion principle a step further, by using it as the basis for training an additional ANN model that can be used to transform the input features in a suitable manner. The ANN adapts acoustic parameters in order to increase the overall training criterion function that underlies the acoustic model, i.e. the likelihood of the ANN/HMM given the acoustic observations (see Section 2). Eventually, the gradient of the

likelihood function w.r.t. the ANN/HMM (obtained by backpropagating partial derivatives through the trellis [25] of the underlying HMM [30]) also yields the corresponding gradient through the novel ANN in an easy-to-compute manner.

The basic idea is the following. Let us start by considering a classic approach to the parameter adaptation problem for HMMs [10,28]. A small amount of data, called the *adaptation set* $\mathscr{A}$, is collected under the new (noisy) acoustic conditions, e.g. a few utterances from the user of the system. In practice, $\mathscr{A}$ is so small that no further training of the acoustic model may be accomplished without loss of its prior configuration (since too many free parameters should be determined from too little training data). The acoustic feature vectors in $\mathscr{A}$ are then individually matched against the corresponding vectors in the clean acoustic space. This requires that the latter ones must be available somehow, or they have to be synthetically generated; a pairwise alignment procedure is also needed, usually relying on the Viterbi algorithm. Finally, a least-squares regression model

$$\varphi : \mathscr{R}^d \rightarrow \mathscr{R}^d, \tag{1}$$

where d is the dimensionality of the feature space, is estimated from the aligned adaptation data. The ultimate goal is to apply φ to map noisy vectors onto the corresponding clean patterns, before feeding them into the HMM.

The idea of such a mapping φ from noisy to clean data appears to be a kind of ultimate solution to the problem of ASR under adverse conditions. Unfortunately, in spite of the huge potentialities of the approach, several practical drawbacks in finding out the right model for φ do impose severe limitations on its application. Formulation in terms of a least-squares regression over a roughly aligned set of (*input*, *target*) pairs is quite unnatural and approximated, and the method is known to be effective only to a limited extent. In particular, clean target values that correspond to the noisy features are generally not known in advance, and the regression problem itself has to be substantially posed in a heuristic manner.

Here, we train an ANN ϕ to learn the model φ for the same purpose: instead of a least-squares regression with standard BP, we develop an unsupervised gradient-ascent algorithm aimed at maximizing the natural criterion that underlies the acoustic model, namely the probability of the observations given the ANN/HMM hybrid model. It turns out that the algorithm includes a model inversion step, which resembles the usual ANN inversion scheme. The latter prescribes calculation of the gradient of the criterion function w.r.t. the ANN input features. Whenever the input features are the outputs of another ANN, say ϕ, the former gradient can be further backpropagated through ϕ, yielding a learning rule for the connection weights of ϕ itself. In addition, a *forward-propagation* scheme through the original ANN ψ is needed, as explained in Section 2.

A concise comparison with the technique described in [24] for Gaussian-based HMMs shows that: (i) the present approach is conceptually and architecturally different: indeed, it features an additional model ϕ which is trained once through an inversion-like scheme, and which does not require any further adaptation when presented with new data; (ii) the computation is quite simple, due to the availability of the gradient of the likelihood (and of its backpropagation through the model) in the underlying ANN/HMM paradigm; (iii) the homogeneity of quantities (sigmoids and their partial derivatives) over the whole system (HMM, ψ and ϕ) makes the algorithm highly stable from a numerical viewpoint (e.g., it does not require the computation of derivatives of close-to-zero Gaussian pdfs).

As a side effect, the application of the present inversion procedure along with the original training algorithm for the ANN/HMM results in a joint acoustic parameter/model parameter adaptation procedure (along the line of [24]), with global optimization over the ML criterion. In this case, the overall architecture/training of the system shares stronger similarities with Bengio's hybrid. Simulation results, described in Section 3, validate the method to a significant extent, showing dramatic improvement over the bare ANN/HMM, as well as over HMMs with/without linear parameter adaptation.

2. The inversion algorithm

In [30] we use a multilayer feed-forward network ψ (e.g. an MLP), the jth output of which (computed over tth input observation $\mathbf{y}_t$) is expected to be a nonparametric estimate of the emission probability $b_{j,t}$ [25] associated with jth state of the HMM at time t. This ANN/HMM is trained using a joint ANN-HMM parameter optimization over the maximum-likelihood criterion. An activation function $f_j(x_j(t))$, either linear or nonlinear, is attached to each unit j of ψ, where $x_j(t)$ denotes input to the unit itself at time t. The corresponding output $o_j(t)$ is given by $o_j(t) = f_j(x_j(t))$. The net is assumed to have ℓ layers $\mathscr{L}_0, \mathscr{L}_1, \ldots, \mathscr{L}_\ell$, where $\mathscr{L}_0$ is the input layer and is not counted, and $\mathscr{L}_\ell$ is the output layer. For notational convenience we write $i \in \mathscr{L}_k$ to denote the index of ith unit in layer $\mathscr{L}_k$.

Our goal is to derive a training algorithm for the connectionist inversion model ϕ that is aimed at the maximization of the overall criterion function that underlie the ANN/HMM recognizer. Gradient-ascent is applied in the following to realize a learning rule for the connection weights of the model ϕ. First, let us introduce the global criterion function to be maximized, namely the *likelihood* L of the model given the acoustic observations[1] [26,2]:

$$L = \sum_{i \in \mathscr{F}} \alpha_{i,T}. \tag{2}$$

[1] A standard notation is used in the following to refer to quantities involved in HMM training. See, for instance, [25].

The sum is extended to the set $\mathscr{F}$ of all *final* states [2] within the HMM which corresponds to the current phonetic transcription. The transcription in terms of HMMs is supposed to involve Q states, and T is the length of the current observation sequence $Y = \mathbf{y}_1, \ldots, \mathbf{y}_T$. The *forward* terms $\alpha_{i,t} = Pr(q_{i,t}, \mathbf{y}_1, \ldots, \mathbf{y}_t)$ and the *backward* terms $\beta_{i,t} = Pr(\mathbf{y}_{t+1}, \ldots, \mathbf{y}_T | q_{i,t})$ for state i at time t can be computed recursively as follows [26]:

$$\alpha_{i,t} = b_{i,t} \sum_j a_{ji} \alpha_{j,t-1} \tag{3}$$

and

$$\beta_{i,t} = \sum_j b_{j,t+1} a_{ij} \beta_{j,t+1}, \tag{4}$$

where a_{ij} denotes the transition probability from ith state to jth state, $b_{i,t}$ denotes emission probability associated with ith state over tth observation $\mathbf{y}_t$ (normalized via ϕ), and the sums are extended to all possible states within the HMM. The initialization of the forward probabilities is accomplished as in HMMs [26], whereas the backward terms at time T are initialized in a slightly different manner, namely

$$\beta_{i,T} = \begin{cases} 1 & \text{if } i \in \mathscr{F}, \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

Given a generic weight w of the adaptation ANN ϕ, gradient-ascent over L prescribes a learning rule of the kind

$$\Delta w = \eta \frac{\partial L}{\partial w}, \tag{6}$$

where η is the *learning rate*. After [2], the following property can be easily shown to hold true by straightforwardly taking the partial derivatives of the left- and right-hand sides of Eq. (3) with respect to $b_{i,t}$

$$\frac{\partial \alpha_{i,t}}{\partial b_{i,t}} = \frac{\alpha_{i,t}}{b_{i,t}}. \tag{7}$$

In addition, borrowing the scheme proposed by [6], [2], the following theorem can be proved to hold true (see [30]): $\partial L / \partial \alpha_{i,t} = \beta_{i,t}$, for each $i = 1, \ldots, Q$ and for each $t = 1, \ldots, T$. Given this theorem and Eq. (7), by repeatedly applying the chain rule we can expand $\partial L / \partial w$ by writing

$$\begin{aligned} \frac{\partial L}{\partial w} &= \sum_{i=1}^{Q} \sum_{t=1}^{T} \sum_{j=1}^{d} \frac{\partial L}{\partial b_{i,t}} \frac{\partial b_{i,t}}{\partial y_{t,j}} \frac{\partial y_{t,j}}{\partial w} \\ &= \sum_{i=1}^{Q} \sum_{t=1}^{T} \sum_{j=1}^{d} \frac{\partial L}{\partial \alpha_{i,t}} \frac{\partial \alpha_{i,t}}{\partial b_{i,t}} \frac{\partial b_{i,t}}{\partial y_{t,j}} \frac{\partial y_{t,j}}{\partial w} \\ &= \sum_{i=1}^{Q} \sum_{t=1}^{T} \sum_{j=1}^{d} \beta_{i,t} \frac{\alpha_{i,t}}{b_{i,t}} \frac{\partial b_{i,t}}{\partial y_{t,j}} \frac{\partial y_{t,j}}{\partial w} \end{aligned} \tag{8}$$

where d is the dimensionality of the feature space, and $\mathbf{y}_t = (y_{t,1}, \ldots, y_{t,d})$ is the output vector from the network ϕ when fed with the acoustic feature vector at time t.

Eq. (8) is formally similar to the initial calculations that we used in [30] in order to derive the learning rule for the ANN/HMM hybrid itself. A double summation (over HMM states and time) was sufficient in [30], while a triple sum (over states, time, and the dimensionality of the feature space) is required in the present case. In addition, in spite of formal resemblances, some terms involved in the sums have different meanings, which reflects the differences in the nature of the corresponding ANNs outputs. As a matter of fact, the quantity $\partial y_{t,j} / \partial w$ in Eq. (8) is simply the partial derivative of jth output of ϕ w.r.t. its generic weight w, and can be computed by applying BP (this does not require any further developments in the remainder of this paper). The term $\partial b_{i,t} / \partial y_{t,j}$, on the contrary, is the core of the inversion principle. Since $b_{i,t}$ is the output of ψ at time t over its jth input $y_{t,j}$, the term $\partial b_{i,t} / \partial y_{t,j}$ is equivalent to the usual ANN inversion scheme. This means that, from a certain point of view, in the present framework the whole model inversion relies on a simple ANN inversion. This is the main reason why the present approach is mathematically and computationally simpler and more straightforward than the Gaussian-based HMM inversion proposed by [24].

Expansion of $\partial b_{i,t} / \partial y_{t,j}$ requires calculations over ψ. Note that the involvement of ψ in the calculations of the learning rule (6) for ϕ does not imply any changes to the weights in ψ: the inversion leaves the original ANN/HMM unchanged.

At this point, in order to use Eq. (8) within learning rule (6), we need an explicit expression for the quantity $\partial b_{i,t} / \partial y_{t,j}$. Given the above notation, it is immediately seen that an application of the chain rule yields

$$\frac{\partial b_{i,t}}{\partial y_{t,j}} = f_i'(x_i(t)) \frac{\partial x_i(t)}{\partial y_{t,j}}. \tag{9}$$

The term $\partial x_i(t) / \partial y_{t,j}$ can be computed by applying a forward-propagation (FP) strategy through ψ, opposite to BP through ϕ, in the following manner. The calculations proceed by induction on the number ℓ of layers of ψ. First of all, let $\ell = 1$ (e.g., ψ is a simple perceptron). We can write

$$\begin{aligned} \frac{\partial x_i(t)}{\partial y_{t,j}} &= \frac{\partial}{\partial y_{t,j}} \sum_{k \in \mathscr{L}_0} w_{ik} y_{t,k} \\ &= \sum_{k \in \mathscr{L}_0} w_{ik} \delta^{(1)}_{t,k,j}, \end{aligned} \tag{10}$$

where we have set

$$\delta^{(1)}_{t,k,j} = \begin{cases} 1 & \text{if } j = k, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

Eq. (10) represents the induction basis. Let us now pass to the inductive step by defining the quantity

$$\delta^{(n+1)}_{t,k,j} = f_k'(x_k(t)) \sum_{m \in \mathscr{L}_{n-2}} w_{km} \delta^{(n)}_{t,m,j} \tag{12}$$

and assuming that $\partial x_i(t) / \partial y_{t,j} = \sum_{k \in \mathscr{L}_{n-1}} w_{ik} \delta^{(n)}_{t,k,j}$ for $n = 2, 3, \ldots, \ell - 1$ and for $x_i(t)$ belonging to nth layer. We are going to show that, under the inductive assumption, this

property holds true also for $n = \ell$. If $x_i(t)$ is ith unit in layer ℓ, then

$$\begin{aligned}\frac{\partial x_i(t)}{\partial y_{t,j}} &= \frac{\partial}{\partial y_{t,j}} \sum_{k \in \mathscr{L}_{\ell-1}} w_{ik} f_k(x_k(t)) \\ &= \sum_{k \in \mathscr{L}_{\ell-1}} w_{ik} \frac{\partial f_k(x_k(t))}{\partial y_{t,j}} \\ &= \sum_{k \in \mathscr{L}_{\ell-1}} w_{ik} \frac{\partial f_k(x_k(t))}{\partial x_k(t)} \frac{\partial x_k(t)}{\partial y_{t,j}} \\ &= \sum_{k \in \mathscr{L}_{\ell-1}} w_{ik} f'_k(x_k(t)) \sum_{m \in \mathscr{L}_{\ell-2}} w_{km} \delta^{(n-1)}_{t,m,j}, \end{aligned} \tag{13}$$

where the inductive hypothesis was applied to obtain the last step. By applying definition (12), it is immediately seen that Eq. (13) can be rewritten as

$$\frac{\partial x_i(t)}{\partial y_{t,j}} = \sum_{k \in \mathscr{L}_{\ell-1}} w_{ik} \delta^{(n)}_{t,k,j}. \tag{14}$$

This completes the induction and provides us with a technique to compute the desired learning algorithm, after using Eq. (14) to expand Eq. (9), and the latter to expand Eq. (8). As anticipated in Section 1, a discriminative maximum a posteriori variant of the algorithm, called MAP, may be used [30]. It basically relies on a *likelihood ratio* criterion, the partial derivatives of which can be computed as above, similarly to [30].

3. Experiments

A speaker-independent, continuous speech recognition task with a small vocabulary (namely, Italian digits from 0 to 9) is considered. Training of the acoustic models is accomplished on a clean training set, while the test set is corrupted by noise, i.e. a mismatch in acoustic conditions holds between training and test environments. The acoustic parameter adaptation models φ are estimated from a small noisy adaptation set $\mathscr{A}$, drawn from a few utterances in the training set. The following sections describe in some detail the dataset, the task (including the noise model), the topology of the acoustic models, and the performance evaluation criteria. Baseline results on the clean test set are reported. Finally, a comparative experimental analysis of the proposed inversion technique is accomplished, and its behavior is also investigated as a function of both the noise level and of the amount of the available adaptation material.

3.1. Dataset and recognition task

The SPK database[2] was used in the experiments. SPK is a *clean* corpus, i.e. it was collected under laboratory conditions, using a close-mouth microphone. The *cdigits* chapter of SPK contains continuous speech, and is divided into a training part and a test part. The training set includes 500 utterances of Italian connected 8-digit strings (4000 words in total), collected among 20 speakers (10 male and 10 female) and sampled at a rate of 16 kHz. The feature space was defined by using 20 ms Hamming windows, with an overlap of 10 ms, and applying standard spectral analysis to obtain 8 *Mel frequency scaled cepstral coefficients* (MFSCCs) [7] and the log-energy of the signal. The log-energy of the signal was computed on a frame-by-frame basis, after applying a standard pre-emphasis filter in the form $1 - \alpha z^{-1}$ and after extraction of the Hamming windows. Using 8 MFSCCs is one of the possible, typical choices in ASR, and it allows for a direct comparison with the experiments presented in [30] on the same dataset. Furthermore, the adoption of more than 8 MFSCCs during a preliminary experimental stage turned out to improve performance of the models on the clean data to a very limited extent, but worsened slightly the recognition rate on noisy signals. This is likely to occur because of the reduced generalization capabilities of over-complex models (indeed, basic results from machine learning theory show that the Vapnik–Chervonenkis (VC) dimension of a learning machine increases with the dimensionality of the feature space [33]).

Normalization of such a feature space was carried out to ensure that ANNs input values are uniformly distributed over the $[0, 1]$ interval. This was accomplished by transforming individual components of each input pattern into the corresponding value of the cumulative distribution function (*cdf*) of the inputs, estimated on a feature-by-feature basis. A mixture of Logistics[3] was assumed as a model of the *cdf* for a given component of the feature space. Maximum-likelihood estimation of the mixture parameters was performed from the available samples [9]. Although the inversion algorithm is suitable for any generic normalization techniques, the present approach turns out to be particularly effective. Moreover, it leads to (potentially) sample-invariant ANN topologies/learning parameters, since different data sets are reduced to the same distribution.

The portion of the SPK corpus used here for testing purposes is the test part of *cdigits*, the same used in [30] to evaluate the ANN/HMM under clean conditions. It consists of other 500 clean utterances (4000 words in total) from 20 speakers (11 male and 9 female) not involved in the training process. Real noise collected in an office environment (mostly due to the presence of noisy workstations) was added to the clean speech signal, assuming an additive model for the overall signal $s(t)$ at time t in the form

$$s(t) = x(t) + n(t), \tag{15}$$

where $x(t)$ is the clean signal and $n(t)$ is the additive noise.

[2] SPK, originally developed at ITC-irst (Trento, Italy), is available from the European Language Resources Association (ELRA).

[3] A mixture of standard sigmoids may be used as well. In doing so, the normalization step can be encapsulated within the ANN, in a suitable and straightforward manner, by adding an extra (pre)input layer with sigmoid activation functions.

Performance of the acoustic models was evaluated, throughout several experiments, for different values of SNR. The technique described in [13] was applied in order to compute the noisy signal (15). The adaptation set $\mathscr{A}$ is obtained by taking again a few utterances from speakers in the training set and corrupting them with noise according to the above additive model.

3.2. Topology of the acoustic models

The same HMM and ANN/HMM topologies used in [30] are adopted (allowing for a direct comparison of the results), namely 11 left-to-right word models (one per digit, plus one for the background noise, or "silence" model), with a number of states equal to the length of the phonetic transcription of each Italian digit according to the *SAMPA* (speech assessment methods phonetic alphabet)[4] acoustic–phonetic units. The HMM contains 8 Gaussian *pdfs* per state, an assumption on the form of the emission probabilities which is popular in ASR. More than 8 component densities do not yield significant improvement in terms of recognition rate on clean data, and tend to worsen the performance on noisy utterances. This phenomenon is similar to the situation described above when an attempt was made to use higher-dimensionality feature spaces: the increased complexity of the learning machine reduces its generalization capabilities over samples that differ substantially from the training data. The segmental k-means initialization [25], Baum–Welch training and Viterbi decoding algorithms are used. The ANN ψ to be applied within the hybrid framework was chosen according to the optimal configuration discussed in [30]: it is a 2-layer MLP with a 93-sigmoids hidden layer and a 40-sigmoids output layer. In so doing, the HMM and the ANN/HMM have a comparable number of free parameters to be determined from the data. The MLP is initialized according to the Bourlard and Morgan-like iterative BP/Viterbi procedure [30] and trained with the on-line version of the MAP hybrid algorithm in the linear domain [30]. Training of connection weights is accomplished in parallel with the re-estimation of initial and transition probabilities in the underlying HMM (via Baum–Welch).

3.3. Evaluation criteria and baseline results

Prior to discussing the experimental results under noisy conditions, we report on the performance of the acoustic models considered here (HMM, Bourlard and Morgan's hybrid, and the proposed ANN/HMM) on the clean test set, as summarized in Table 1 (see [30]). The evaluation criteria are the *word recognition rate* (WRR) and the *string recognition rate* (SRR). The former is defined as $WRR = 100\{1-(Ins+Del+Sub)/N_{\text{words}}\}\%$, where N_{words} is the total number of words in the uttered text, and the number of errors is expressed counting out word insertions (*Ins*), deletions (*Del*), and substitutions (*Sub*), respectively. The SRR is the fraction of test sequences that are recognized without any errors. Since an objective comparison between the *acoustic* models is sought, the language model [8] was tuned—from time to time—in order to have the same number of insertions and deletions for any given acoustic model. It is seen that the connectionist approaches do compare quite favorably with the Gaussian-based HMM.

Table 1
String recognition rate (SRR) and word recognition rate (WRR) on the clean SPK test set

Architecture/algorithm	SRR (%)	WRR (%)
HMM with 8-Gaussian mixtures	46.60	90.03
Bourlard and Morgan's hybrid	47.40	90.20
Present ANN/HMM trained via MAP	68.60	94.65

3.4. Detailed analysis of results under intermediate noisy conditions

First of all, the inversion technique is evaluated under intermediate conditions of noise, and a detailed comparison with other models is accomplished. Section 3.5 presents results as a function of the SNR, ranging from 30 dB (light noise) to 10 dB (severe noise). Although it will be seen that the effect of the present normalization approach has a stronger impact on the overall recognition performance whenever the signal is more noisy, in this section we assume an intermediate reference noise level of 20 dB.

The adaptation set used here is constituted of 20 utterances. This amount of adaptation speech material may be reasonable in several practical scenarios, and it is aligned with experimental setups concerning parameter adaptation described in the literature [11,28]. Section 3.6 reports results obtained with different sizes of the adaptation set, ranging from 1 to 20 training utterances: it will be seen that the most interesting, and useful, performance is gained relying on a single adaptation utterance, while no improvement is obtained when the adaptation set is increased from 15 to 20 utterances. As a consequence, the present experimental conditions (20 dB, 20 adaptation utterances) are not optimal for the present algorithm.

Table 2 summarizes the results obtained with the different models/algorithms on the noisy task. The first row highlights the dramatic loss encountered by standard Gaussian-based HMMs once a mismatch between training and test acoustic conditions is encountered. Particularly relevant is the result on the second row of the table. It shows the effect of refining HMM parameters using Baum–Welch over $\mathscr{A}$, starting from the HMM parameters that were originally estimated during regular training on the clean training set. In the following, we will refer to this

[4]SAMPA is a machine-readable phonetic alphabet, originally developed under the E.U. ESPRIT project *SAM* (speech assessment methods).

Table 2
String recognition rate (SRR) and word recognition rate (WRR) on the noisy SPK test set at a SNR level of 20 dB (adaptation set size: 20 utterances).

Architecture/algorithm	SRR (%)	WRR (%)
HMM with 8-Gaussian mixtures	10.20	**69.28**
HMM with retraining	0.60	41.27
HMM with linear parameter adaptation	17.60	77.42
Bourlard and Morgan's hybrid	24.40	79.38
ANN/HMM hybrid trained via MAP	41.40	**86.80**
ANN/HMM with MAP retraining	40.00	84.78
ANN/HMM with neural spectral mapping	41.60	87.42
ANN/HMM with the proposed inversion model	49.80	**90.15**

scheme as to "model retraining". The result refers to the application of a single Baum–Welch iteration, since further re-estimation steps even worsen the performance. The phenomenon is analytically illustrated in Fig. 1, which represents the learning and generalization curves for the HMM during 30 iterations of the retraining process. The learning curve is expressed in terms of values of the training criterion function, i.e. the log-likelihood $\log P(\mathscr{A}|\mathscr{M})$ of the model $\mathscr{M}$ given the adaptation data $\mathscr{A}$. The generalization curve is given in terms of WRR on the test set. The values of the curves at the beginning (iteration 0) refer to the original (trained) HMM before retraining takes place. It is seen that while Baum–Welch actually maximizes $P(\mathscr{A}|\mathscr{M})$, the model tends to "forget" the previous (more general and representative) parameter estimates learned over the whole training set. As expected, the size of $\mathscr{A}$ is too small to allow for re-estimation of all the HMM parameters, namely about 5800 free parameters.

A standard parameter adaptation technique was then applied to the (original, nonretrained) HMM. The third row of Table 2 shows the results yielded by a linear transformation of the acoustic parameters, aimed at the compensation for the mismatch in acoustic conditions w.r.t. the training data. The linear transformation was realized according to the scheme proposed in [10,28], and it turned out to be effective. Estimation of the linear regression model was accomplished using the Widrow–Hoff algorithm [9]. This represents a significant benchmark, since HMM with linear parameter adaptation is a popular state-of-the-art solution to the problem of mismatch in acoustic conditions.

Bourlard and Morgan's hybrid was then tested, which turned out to be intrinsically more robust to noise than the HMM with acoustic parameter adaptation. This aspect is further emphasized by the next row of the table. Here, the ANN/HMM introduced in [30] is evaluated in a straightforward manner, i.e., without any adaptation on $\mathscr{A}$. These experiments confirm the advantages of having a nonparametric estimation, that does not impose any a priori assumptions on the form of the distribution of the data. Gaussian-based HMMs tend to "memorize" the training samples, placing precise Gaussian *pdfs* over specific locations of the feature space. On the contrary, ANNs tend to learn a general law, i.e. they do generalize better to novel data whenever conditions are changed (e.g. when the position of samples in the feature space is moved away as a consequence of the presence of noise). Just like in the HMM, retraining of the ANN/HMM on $\mathscr{A}$ was carried out (starting from the original model parameters learned from the clean training set). Again, performance of the overall recognition system was worsened, and the same considerations outlined in the case of the HMM still hold. In particular, $\mathscr{A}$ is too small and statistically little representative to permit learning of suitable model parameters (a total of nearly 5000 free parameters) without compromising the original configuration of the overall model.

Adaptation of the acoustic features was then realized, keeping the ANN/HMM parameters fixed (last two rows of the table). The adaptation relies on a feature transformation model φ, according to Eq. (1). First of all, a neural regression model is applied, following the approaches described in [28]. It is trained with standard BP in order to achieve a nonlinear transformation that minimizes the least squares criterion between noisy and clean acoustic vectors. In the present setup, where an additive noise model is used, the alignment between noisy input patterns and the corresponding (target) clean vectors is straightforward. This gives a potential advantage to this regression scheme, an advantage that is hardly realized under general conditions, where no exact relationship between noisy patterns and clean targets is known a priori (it is worth underlining that, by contrast, the inversion algorithm presented in this paper never requires any alignments at all). Indeed, a much more sophisticated and less precise alignment technique has to be adopted in the general case (see, for instance, [28]). The ANN used here has a simple architecture, namely 13 hidden sigmoid units and 9 linear outputs, for a total of 247 free parameters (i.e., connection weights and adaptive bias of sigmoids). This is the same topology that is used for the evaluation of the inversion technique, in order to allow for a fair comparison. The results show that the approach is effective, yielding a significant improvement over the bare ANN/HMM.

Finally, estimation of φ via the inversion algorithm was evaluated. The ANN ϕ is initialized via BP on a small, synthetic dataset of input-target pairs in the form $(\mathbf{x}, \mathbf{x})$, uniformly scattered over the feature space. This initialization forces the network to realize the identical transformation, i.e. $\varphi(\mathbf{x}) = \mathbf{x}$. In so doing, it is immediately seen that feeding the ANN/HMM with the outputs from ϕ yields a baseline result which is identical to the performance of the bare ANN/HMM. Next, 10 training epochs of the inversion algorithm, namely Eq. (6), are applied. The resulting WRR (90.15%) represents a significant 25.38% relative word error rate (WER) reduction w.r.t. the ANN/HMM without parameter adaptation, scoring even higher

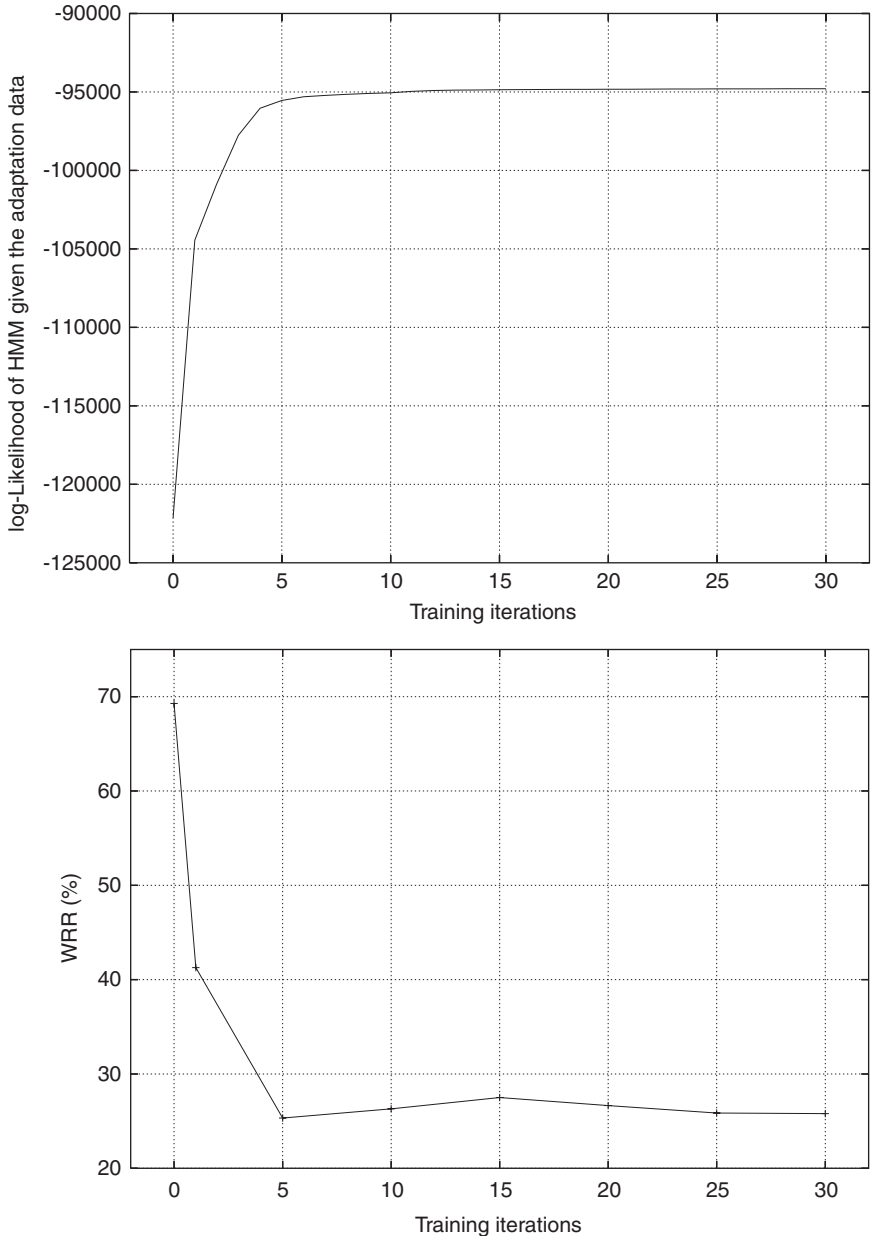

Fig. 1. Learning curve (top) in terms of log-likelihood of the model given the adaptation data, and generalization curve (bottom) in terms of WRR on the noisy test set, as functions of the number of Baum–Welch retraining iterations over the adaptation set for the standard Gaussian-based HMM.

than the 90.03% WRR yielded by the Gaussian-based HMM over the clean signals. This fact has a relevance also in an application-oriented perspective.

3.5. Recognition results under different noisy conditions

The relevance of the inversion technique to the overall recognition performance depends on the amount of noise present in the input acoustic features. In the limit case, whenever no noise is present at all, acoustic parameter adaptation is useless and performance cannot be improved. Fig. 2 shows the results obtained with the three major paradigms, namely Gaussian-based HMM, ANN/HMM with and without model inversion, represented as WRR curves in function of the SNR. The SNR ranges from 30 dB (light noise) to the left, to 10 dB (severe noisy conditions) to the right. The WRR is evaluated at 5 dB steps in this range. It is seen that the adaptation ϕ of acoustic features turns out to be more effective (upper curve) as the noise level increases over the signal. The HMM drops quickly (bear in mind that it yielded a 90.03% WRR over the clean speech), and it is increasingly far below the performance provided by the ANN/HMM, with or without inversion.

At a SNR of 30 dB the inversion algorithm scores a 4.72% relative WER reduction over the ANN/HMM without parameter adaptation. The gap dramatically

increases to a maximum reached at a SNR level of 10 dB, where the inversion model scores 65.08% WRR versus 52.12% WRR yielded by the ANN/HMM, being 40.92% WRR the performance of the HMM. It is worth underlining that similar breakdowns in Gaussian-based HMMs under severely adverse conditions are well-known from the literature [14,32,22].

3.6. Results for different sizes of the adaptation set

In the previous experiments we used 20 training utterances—taken from the SPK training set and corrupted by noise—to form the adaptation set $\mathscr{A}$. In this section we investigate the behavior of the inversion algorithm whenever the amount of adaptation data varies. More precisely,

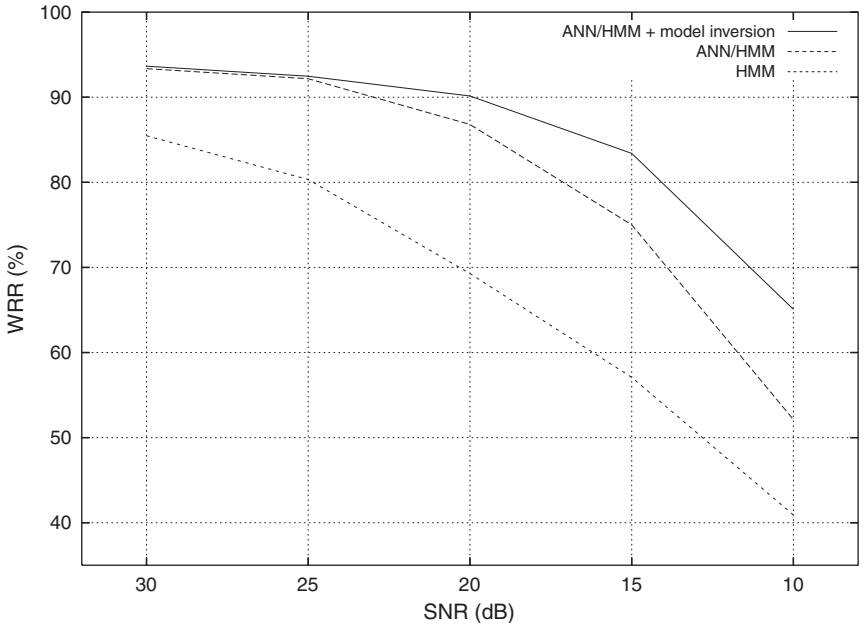

Fig. 2. WRR curves as functions of the SNR in the three major cases: Gaussian-based HMM (lower curve); ANN/HMM hybrid without inversion (middle curve); ANN/HMM hybrid with inversion-based feature adaptation (upper curve).

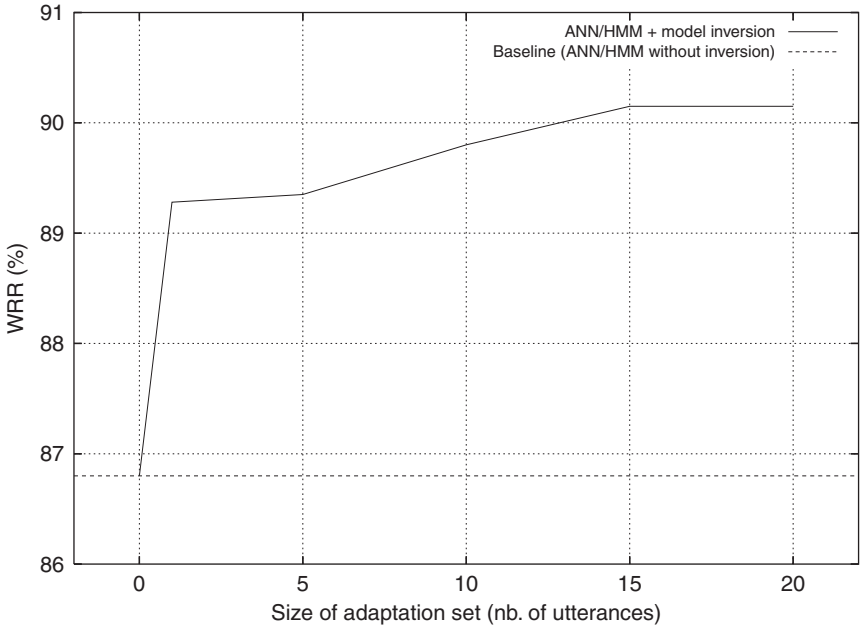

Fig. 3. WRR yielded by the proposed inversion-based feature adaptation scheme, at a reference SNR level of 20 dB, plotted as a function of the adaptation set size (number of adaptation utterances). The curve is compared with the baseline represented by the WRR yielded by the ANN/HMM without inversion.

we consider experimental setups where the cardinality of $\mathscr{A}$ ranges from 1 (a single adaptation utterance) to a maximum of 20. As above, the adaptation utterances are corrupted by additive noise at a reference SNR level of 20 dB. If n is the cardinality of $\mathscr{A}$, then n speakers from the training part of SPK are considered, and a single utterance (taken at random) from each of them is inserted in the adaptation set. From an application-oriented point of view, it is crucial to keep the amount of adaptation material as small as possible. As a matter of fact, it is essential to evaluate the proposed technique whenever only a few utterances are at the system's disposal to accomplish acoustic adaptation. Fig. 3 plots the WRR curve yielded by the inversion algorithm as a function of the number of adaptation utterances. The ANN topology and its training parameters (e.g., 10 training epochs) are the same as before. The curve is compared with the (constant) baseline of 86.80% WRR yielded by the ANN/HMM without parameter adaptation. The figure shows that, in spite of the fact that we used 20 utterances in the comparative analysis of Section 3.4, no gain in performance is actually obtained when the cardinality of $\mathscr{A}$ is increased over 15. On the contrary, only a few adaptation utterances are sufficient to reach a significant improvement in terms of WRR: using a unique utterance yields a 89.28% WRR. Moreover, training ϕ for a single epoch of the inversion algorithm using again one utterance also scores 89.28% WRR. It is our conviction that this result should be taken into consideration in a practical perspective, in the light of the 69.28% WRR provided by state-of-the-art Gaussian-based HMMs at the same SNR level.

4. Conclusions

This paper introduced an inversion scheme for an ANN/HMM hybrid system where a feed-forward ANN ψ carries out estimation of emission probabilities associated with the states of an underlying HMM topology. The bare model inversion principle was extended to form the basis of a training algorithm for a neural parameter-adaptation model ϕ. A joint optimization scheme of acoustic parameters and of model parameters is obtained once the ANNs ϕ and ψ are trained in parallel. It was shown that:

1. Neural-based acoustic models, whenever trained over proper probabilistic criteria, score higher than Gaussian-based HMMs, being much more noise-tolerant. This is due, both to the generalization capabilities of ANNs and to their nonparametric universal approximation property.
2. The proposed algorithm allows for simple normalization models (13 sigmoid neurons) that can learn effective normalization mappings from a few adaptation utterances. It outperforms standard linear, as well as neural, parameter adaptation techniques.
3. Due to its reduced complexity and its limited demand of adaptation data, the model may be particularly suitable for hardware implementation on small devices for real-word applications.

Although the simulations were carried out on data corrupted by noise, the approach is suitable for generic situations featuring an underlying mismatch in acoustic conditions between the training and the test environments, e.g., speaker normalization, channel or transducer compensation. Furthermore, the approach may be applied to large-vocabulary ASR systems, upon availability of a suitable language model, and provided that phone-based HMMs are used instead of word models.

The technique, albeit evaluated in real-world ASR tasks, may be suitable for broader classes of sequence processing problems, i.e. offline handwritten character recognition, cheminformatics, or bioinformatics. Finally, an alternative application of the inversion algorithm is also viable: once trained within the ANN/HMM framework, the ANN ϕ may be used as a nonlinear acoustic parameter adaptation model φ to feed a standard, Gaussian-based HMM. This may turn out to be useful whenever a particularly stable, performing and large-scale HMM-based ASR system has already been developed, and the adoption of the ANN/HMM hybrid from scratch is not sought.

Acknowledgments

The authors are grateful to Marco Matassoni, who helped us in setting up the experimental environments. The invaluable inspiration and contributions from our friends Renato De Mori and Yoshua Bengio are gratefully acknowledged. Special thanks to Stefania Biscetti, who helped us improving the quality of the paper. The development of the present research was possible only upon agreement with ITC-irst (Povo, Italy), in the person of Gianni Lazzari.

References

[1] A.J. Bell, T.J. Sejnowski, An information-maximization approach to blind separation and blind deconvolution, Neural Comput. 7 (1995) 1129–1159.
[2] Y. Bengio, Neural Networks for Speech and Sequence Recognition, International Thomson Computer Press, London, UK, 1996.
[3] 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.
[4] H. Bourlard, N. Morgan, Continuous speech recognition by connectionist statistical methods, IEEE Trans. Neural Networks 4 (6) (1993) 893–909.
[5] 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.
[6] J.S. Bridle, Alphanets: a recurrent 'neural' network architecture with a hidden Markov model interpretation, Speech Commun. 9 (1) (1990) 83–92.
[7] 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.
[8] R. De Mori, Spoken Dialogues with Computers, Academic Press, London, UK, 1998.
[9] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.
[10] M.-W. Feng, Improved speaker adaptation using text dependent spectral mappings, in: Proceedings of ICASSP, 1988, pp. I-131–134.
[11] C. Furlanello, D. Giuliani, E. Trentin, S. Merler, Speaker normalization and model selection of combined neural nets, Connection Sci. 9 (1) (1997) 31–50 Special Issue on Combining Neural Nets.
[12] M.J.F. Gales, Maximum likelihood linear transformations for hmm-based speech recognition, Comput. Speech Lang. 12 (1998).
[13] D. Giuliani, M. Matassoni, M. Omologo, P. Svaizer, Training of hmm with filtered speech material for hands-free recognition, in: International Conference on Acoustics, Speech and Signal Processing, Phoenix, 1999.
[14] H.G. Hirsch, D. Pearce, The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions, Proc. ISCA (2000).
[15] D. Hoskins, J.N. Hwang, J. Vagners, Iterative inversion of neural networks and its application to adaptive control, IEEE Trans. Neural Networks 3 (1992) 292–301.
[16] X.D. Huang, Y. Ariki, M. Jack, Hidden Markov Models for Speech Recognition, Edinburgh University Press, Edinburgh, 1990.
[17] J.N. Hwang, J.J. Choi, S. Oh, R.J. Marks II, Query learning applied to partially trained multilayer perceptrons, IEEE Trans. Neural Networks 2 (1991) 131–136.
[18] J.N. Hwang, H. Li, Interactive query learning for isolated speech recognition, in: Proceedings of IEEE International Workshop on Neural Networks and Signal Processing, Helsinger, September 1992, pp. 93–101.
[19] C.J. Leggetter, P.C. Woodland, Speaker adaptation of continuous density hmms using miltivariate linear regression, in: Proceedings of ICSLP, Yokohama, 1994, pp. 451–454.
[20] A. Linden, J. Kindermann, Inversion of multilayer nets, in: Proceedings of the International Joint Conference on Neural Networks, Washington, DC, 1989, pp. 425–430.
[21] D. Macho, L. Mauuary, B. Noe, Y.M. Cheng, D. Ealey, D. Jouvet, H. Kelleher, D. Pearce, F. Saadoun, Evaluation of a noise-robust DSR front end on Aurora databases, Proc. ICSLP (2002).
[22] M. Matassoni, M. Omologo, L. Cristoforetti, D. Giuliani, P. Svaizer, E. Trentin, E. Zovato, Some results on the development of a hands-free speech recognizer for car-environment, in: Proceedings of the 1999 international workshop on Automatic Speech Recognition and Understanding (ASRU), Keystone, Colorado, USA, December 12–15 1999.
[23] C. Mokbel, D. Jouvet, J. Monné, R. De Mori, Robust speech recognition, in: R. De Mori (Ed.), Spoken Dialogues with Computers, Academic Press, London, UK, 1998, pp. 435–439.
[24] 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.
[25] L.R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77(2), 1989, pp. 257–286.
[26] L. Rabiner, B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, NJ, USA, 1993.
[27] D.E. Rumelhart, G.E. Hinton, R.J. Williams, Learning internal representations by error propagation, in: D.E. Rumelhart, J.L. McClelland (Eds.), Parallel Distributed Processing, vol. 1, MIT Press, Cambridge, 1986, pp. 318–362.
[28] E. Trentin, D. Giuliani, A mixture of recurrent neural networks for speaker normalization, Neural Comput. Appl. 10 (2001) 120–135.
[29] E. Trentin, M. Gori, A survey of hybrid ANN/HMM models for automatic speech recognition, Neurocomputing 37 (1–4) (2001) 91–126.
[30] E. Trentin, M. Gori, Robust combination of neural networks and hidden Markov models for speech recognition, IEEE Trans. Neural Networks 14 (6) (2003).
[31] E. Trentin, M. Matassoni, Robust segmental-connectionist learning for recognition of noisy speech, in: Proceedings of the First Workshop on Robust Methods for Speech Recognition in Adverse Conditions, Tampere, Finland, May 25–26, 1999, pp. 159–162.
[32] E. Trentin, M. Matassoni, The regularized SNN-TA model for recognition of noisy speech, in: Proceedings of IJCNN2000 (International Joint Conference on Neural Networks), Como, Italy, 24–27 July, 2000, pp. V 97–102.
[33] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.

Edmondo Trentin received his Laurea "Cum Laude" in Computer Science from the Università di Milano (Italy) in 1990, and his Ph.D. in Automation and Information Engineering from the Università di Firenze (Italy) in 2001. From 1990 to the end of 1993 he worked as a Project Leader at SGS-Elsag. From 1994 to 2000 he was a researcher at ITC-irst (Trento, Italy), in the area of Interactive Sensory Systems. Since December 2000 he has been carrying out his research and teaching activities at the Dipartimento di Ingegneria dell'Informazione, Università di Siena, Italy, where he is an Assistant Professor.

His research interests include learning, artificial neural networks (ANN), statistical pattern recognition, hidden Markov models (HMM), and hybrid ANN/HMM systems. He has been involved in projects of scientific and technological relevance in the fields of speech processing and bioinformatics.

Dr. Trentin is the author of more than 30 scientific publications. He 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 Universitdi Firenze, Italy, in 1984, and the Ph.D. degree in 1990 from Universitdi 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 Universitdi Firenze and, in November 1995, he joined the Universitdi Siena, where he is currently full professor of computer science.

His main interests are in machine learning, with applications to pattern recognition, Web mining, and game playing. He has leaded a number of research projects on these themes with either national or international partners, and has been involved in the organization of many scientific events, including the IEEE–INNS International Joint Conference on Neural Networks (Como, July 2000) and the NATO-ARW on Limitations and Future Trends in Neural Computation (Siena, October 2001).

Dr. Gori serves as an Associate Editor of a number of technical journals related to his areas of expertise, including Pattern Recognition, Neural Networks, Neurocomputing, Pattern Analysis and Application, the International Journal of Document Analysis and Recognition, and the International Journal on Pattern Recognition and Artificial Intelligence. He has been the recipient of best paper awards and keynote speakers in many international conferences.

He is the Chairman of the Italian Chapter of the IEEE Computational Intelligence Society and he has been the President of the Italian Association for Artificial Intelligence. Dr. Gori is a fellow of the IEEE.