Linear hidden transformations for adaptation of hybrid ANN/HMM models

Linear Hidden Transformations for Adaptation of Hybrid ANN/HMM Models[1]

Roberto Gemello[1], Franco Mana[1], Stefano Scanzio[2], Pietro Laface[a2] and Renato De Mori[3]

[a] Corresponding Author
Pietro Laface
[2] *Politecnico di Torino*
Corso Duca degli Abruzzi, 24
10129 Torino Italy
Email: {Pietro.Laface, Stefano.Scanzio}@polito.it
Phone: +39 011 564-7004, Fax: +39 011 564-7099

Coauthors
[1] *LOQUENDO*
Via Val della Torre, 4 A
10100 Torino Italy
Email: {Roberto.Gemello, Franco Mana}@loquendo.com
Phone: +39 011 291-3458, Fax: +39 011 291…..

[3] *LIA - University of Avignon*
339, Chemin des Meinajaries
Agroparc BP 1228
84911 AVIGNON Cedex 9 France
Email: Renato.Demori@lia.univ-avignon.fr,
Phone: +33 49 0 84 3515, Fax: +33 49 0 84 3501

Abstract

This paper focuses on the adaptation of Automatic Speech Recognition systems using Hybrid models combining Artificial Neural Networks with Hidden Markov Models.
A classical adaptation technique consists in adding a linear transformation network that acts as a pre-processor to the main network. We investigated the application of linear transformations not only to the input features, but also to the outputs of the internal layers. The motivation is that the outputs of an internal layer represent a projection of the input pattern into a space where it should be easier to learn the classification or transformation expected at the output of the network.

[1] This paper is an extended version of the paper "Adaptation of Hybrid ANN/HMM Models Using Linear Hidden Transformations and Conservative Training" accepted for publication at ICASSP 2006.

To compensate for the lack of adaptation samples for some phonetic units, a new solution, called Conservative Training, is also proposed that is a variation to the standard method of assigning the target values.
Supervised adaptation experiments with different corpora and for different adaptation types are described. The results show that the proposed approach always outperforms the use of transformations in the feature space and yields even better results when combined with linear input transformations.

Key words: Automatic Speech Recognition; Speaker Adaptation; Neural Network Adaptation; Catastrophic Forgetting

1 Introduction

The literature on speakers, environments, and applications adaptation is rich of techniques for refining Automatic Speech Recognition (ASR) systems by adapting the acoustic features and the parameters of stochastic models [1-5]. More recently, particular attention has been paid to discriminative training techniques and their application to the acoustic feature transformation [6,7].
Since the acoustic-phonetic Artificial Neural Networks (ANN) models are also trained with discriminative methods, it is worth exploring methods for adapting their features and model parameters.
Several solutions to this problem have been proposed. Some of these techniques for adapting neural networks are compared in [8,9]. A classical approach consists in adding a linear transformation network (LIN) that acts as a pre-processor to the main network, or simply adaptig all the weights of the original network.
A tied-posterior approach is proposed in [10] to combine Hidden Markov Models (HMM) with ANN adaptation strategies. The weights of a hybrid ANN/HMM system are adapted by optimizing the training set cross entropy. A sub-set of the hidden units is selected for this purpose. The adaptation data are propagated through the original ANN and the nodes exhibiting the highest variance are selected, since hidden nodes with a high variance transfer a larger amount of information to the output layer.

Recent adaptation techniques have been proposed with the useful properties of not requiring to store the previously used adaptation data and to be effective even with a small amount of adaptation data. Methods based on speaker space adaptation [2] and eigenvoices [3] are of this type and can be applied both to Gaussian Mixture HMMs as well as to the ANN inputs as proposed in [11]. The parameters of the transformations are seen as the components of a vector in a parameter adaptation space. Principal components can be found in this space to define a speaker space. Rapid adaptation consists in finding the values of the coordinates of a specific speaker point in the speaker space. If a limited number of adaptation data is available, then only fewer eigenvoices are used.

This paper explores a new possibility consisting in adapting ANN models with transformations of an entire set of internal model features. Values for these features are

collected at the output of a hidden layer for which the number of outputs is usually of the order of a few hundreds. These features are supposed to represent an internal structure of the input pattern. As for input feature transformation, a linear network can be used for hidden layer feature transformation. In both cases the estimation of the parameters of the adaptation networks can be done with error Back-Propagation by keeping unchanged the values of the parameters of the ANN. Internal transformations can also be obtained by linear combination of "eigenvoices".

A problem, however, occurs in distributed connectionist learning when a network, trained with a large set of patterns, has to learn new input patterns. This problem, called "catastrophic forgetting," [13] is particularly severe when a network is adapted with new data that do not adequately represent the knowledge included in the original training data. This effect is evident when adaptation data do not contain examples for a subset of the output classes.
A review of several approaches that has been proposed to solve this problem is presented in [13]. Among them, the use of a series of pseudo-patterns, i.e. random patterns, associated to the output values produced by the connectionist network before adaptation. These pseudo-patterns are added to the set of the new patterns to be learned [14] to try keeping stable the classification boundaries related to classes that have few or no samples in the new set of patterns. This effectively decreases catastrophic forgetting of the originally learned patterns. Since it seems difficult to generate these pseudo-patterns when the dimensionality of the input features is high, it has been proposed [15] to include in the adaptation set examples of the missing classes taken from the training set.

This paper proposes a solution to this problem introducing Conservative Training, a variation to the standard method of assigning the target values, which compensates for the lack of adaptation samples in some classes. Experimental results on the adaptation test for the Wall Street Journal task [16] using the proposed approach compare favorably with published results on the same task [10,16].

The paper is organized as follows: Section 2 gives a short overview of the acoustic-phonetic models of the ANN used by the ASR system, and presents the Linear Hidden Networks, which transform the features at the output of hidden layers. Section 3 is devoted to the illustration of the problem of catastrophic forgetting in connectionist learning, and proposes our Conservative Training approach as a possible solution and illustrates its benefits using an artificial classification task of 16 classes. Section 4 reports the experiments performed on several databases with the aim of clarifying the behavior of the new adaptation techniques with respect to the classical LIN approach. Finally the conclusions and future developments are presented in the last Section.

2 Feature transformations

The LOQUENDO-ASR decoder uses a 4-layer hybrid combination of Hidden Markov Models (HMM) and Multi Layer Perceptron (MLP), where each phonetic unit is described in terms of a single or double state left-to-right automaton with self-loops. The HMM transition probabilities are uniform and fixed, and the emission probabilities are computed by a MLP

[12]. The MLP has an input layer of 273 units (39 parameters of a 7 frame context), a first hidden layer of 315 units, a second hidden layer of 300 units and an output layer including a variable number of units, which is language dependent (600 to 1000). Using two hidden layers, rather than a larger single hidden layer, has the advantage of reducing the total number of connections. Moreover, it allows considering the activation values of each hidden layer as a progressively refined projection of the input pattern in a space of features more suitable for classification.
The acoustic models are based on a set of vocabulary and gender independent units including stationary context-independent phones and diphone-transition coarticulation models.
These models have been successfully used for the acoustic models of 15 languages released with the LOQUENDO-ASR recognizer, and are the seed models for adaptation experiments of Section 3, if not differently specified.

2.1 Input feature transformations

The simplest and more popular approach to speaker adaptation with ANNs is Linear Input Transformation [8,9]. The input space is rotated by a linear transformation to make the target conditions more consistent with the training conditions. The transformation is performed by a linear layer interface (referred to, in this paper, as linear input network or LIN) between the input observation vectors and the input layer of the trained ANN as shown in Figure 1. The LIN weights are initialized with an identity matrix, and they are trained by minimizing the error at the output of the ANN system keeping fixed the weights of the original ANN.
Using few training data, the performance of the combined architecture LIN/ANN is usually better than adapting the whole network, because it involves the estimation of a lower number of parameters.

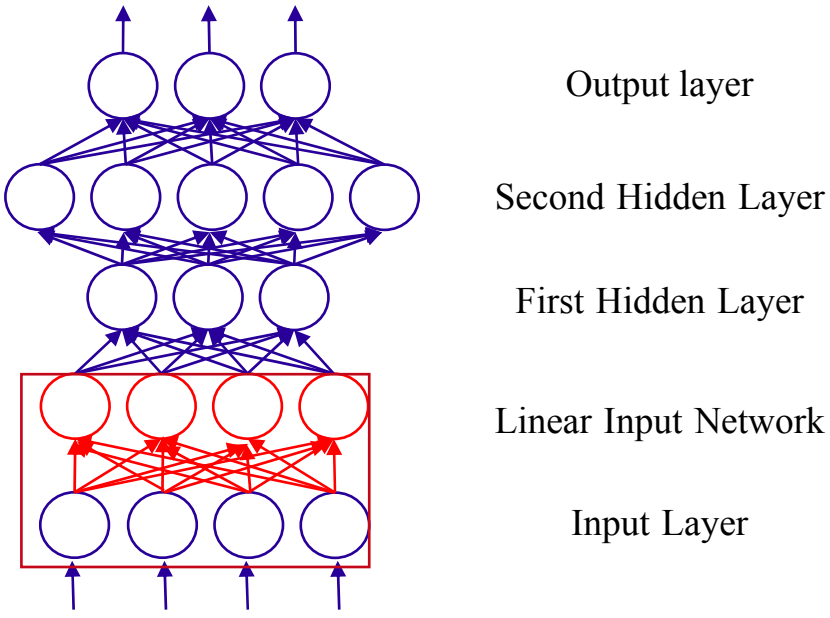

Fig. 1. Artificial Neural Network including a linear input layer

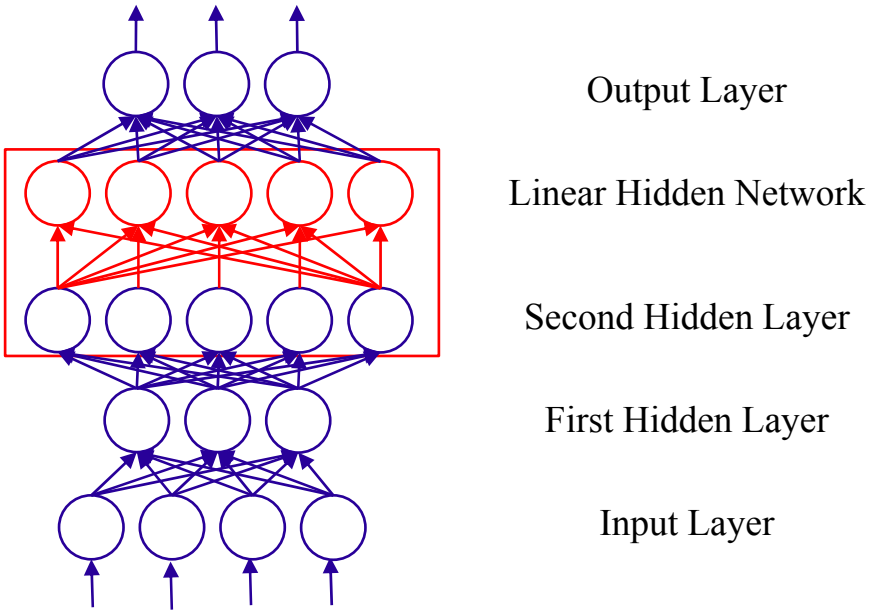

Fig. 2. Artificial Neural Network including a linear hidden layer

2.2 Hidden feature transformations

Assuming that the activation values of a hidden layer represent an internal structure of the input pattern in a space more suitable for classification, a linear transformation can be applied to the activations of the internal layers. Such a transformation is performed by a Linear Hidden Network (LHN). As for the LIN, the values of an identity matrix are used to initialize the weights of the LHN. The weights are trained using a standard Back-Propagation algorithm keeping frozen the weights of the original network. It is worth noting that, since the LHN performs a linear transformation, once the adaptation process is completed, the LHN can be removed combining LHN weights with the ones of the next layer using the following simple matrix operations:

$$\begin{aligned} W_a &= W_{LHN} \times W_{SI} \\ B_a &= B_{SI} + B_{LHN} \times W_{SI} \end{aligned} \qquad (1)$$

where W_a and B_a are the weights and the biases of the adapted layer, W_{SI} and B_{SI} are the weights and biases of the layer following the LHN in the original Speaker Independent network, and W_{LHN} and B_{LHN} are the adapted weights and the biases of the linear hidden network.
In our experiments the LHN has been applied to the last hidden layer.

3 Catastrophic Forgetting

It is well known that in connectionist learning, acquiring new information in the adaptation process, can damage previously learned information [13,14]. This effect must be taken into

account when adapting an ANN with limited amount of data, which do not include enough samples for all the acoustic-phonetic units. The problem is more severe in the ANN modeling framework than in the classical Gaussian Mixture HMMs. The reason is that an ANN uses discriminative training to estimate the posterior probability of each acoustic-phonetic unit. The minimization of the output error is performed by means of the Back-Propagation algorithm that penalizes the units with no observations in the adaptation set by setting to zero the target value of the their output units for *every* adaptation frame. This target assignment policy induces in the ANN a forgetting of its capability to classify the corresponding acoustic-phonetic units. Thus, while the Gaussian Mixture models with little or no observations remain un-adapted or share some adaptation transformations of their parameters with other similar acoustic models, the units with little or no observations in the ANN model loose their characterization rather than staying not adapted. Thus, adaptation may destroy the correct behavior of the network for the unseen units.
To mitigate the problem of loosing characterization of the units with little or no observations, it has been proposed [15] to include in the adaptation set examples of the missing classes taken from the training set. The disadvantage of this approach is that a substantial amount of the training set must be stored so that examples of the missing classes can be retrieved for each adaptation task. In [14], it has been proposed to approximate the real patterns with pseudo-patterns rather than using the training set. Pseudo-patterns consist of pairs of random input activations and the corresponding output. These pseudo-patterns are included in the set of the new patterns to be learned to prevent catastrophic forgetting of the original patterns. It seems difficult, however, to generate these pseudo-patterns when the dimensionality of the input features is high.

We here propose a solution, that we call Conservative Training (CT), to mitigate the forgetting problem.
Since the Back-Propagation technique used for MLP training is discriminative, the units for which no observations are available in the adaptation set will have zero as a target value for all the adaptation samples. Thus, during adaptation, the weights of the acoustic MLP will be biased to favor the output activations of the units with samples in the adaptation set and to weaken the other units, which will tend to always have a posterior probability close to zero.
Conservative Training does not set to zero the value of the targets of the missing units, using instead as target values the outputs computed by the original network.

Let F_p be the set of phonetic units included in the adaptation set (*p* indicates presence), and let F_m be the set of the missing units. In Conservative Training the target values are assigned as follows:

$$T(f_i \in F_m \mid O_t) = OUTPUT\_ORIGINAL\_NN(f_i \mid O_t)$$

$$T(f_i \in F_p \mid O_t \quad \& \quad correct(f_i \mid O_t)) = (1.0 - \sum_{j \in F_m} OUTPUT\_ORIGINAL\_NN(f_j \mid O_t)$$

$$T(f_i \in F_p \mid O_t \quad \& \quad !correct(f_i \mid O_t)) = 0.0$$

where $T(f_i \in F_p \mid O_t)$ is the target value associated to the input pattern O_t for a unit f_i that is present in the adaptation set, $T(f_i \in F_m \mid O_t)$ is a target value associated to the input pattern O_t for a unit not present in the adaptation set, $OUTPUT\_ORIGINAL\_NN(f_i \mid O_t)$ is the output of the original network (before adaptation) for the phonetic unit i given the input pattern O_t, and $correct(f_i \mid O_t)$ is a predicate which is true if the phonetic unit f_i is the correct class for the input pattern O_t.

Thus, a phonetic unit that is missing in the adaptation set, rather than obtaining a zero target value for each input pattern, will keep the value that it would have had with the original un-adapted network.

3.1 Experimental results on artificial data

An artificial two-dimensional classification task has been used to investigate the effectiveness of the Conservative Training technique. An MLP has been used to classify points belonging to 16 classes having the rectangular shapes shown by the green borders in Figure 3. The MLP has 2 input units, two 20 node hidden layers, and 16 output nodes. It has been trained using 2500 uniformly distributed patterns for each class.

Figure 3 shows the classification behavior of the MLP after training based on Back-Propagation. In particular, a dot has been plotted only if the score of the corresponding class was greater than 0.5. MLP outputs have also been plotted for test points belonging to regions that have not been trained, and outside the green rectangles: they are at the left and right sides of Figure 3. The average classification rate for all classes, and particularly for classes 6 and 7, is reported in the first row of Table 1.

Afterward, an adaptation set was defined to simulate an adaptation condition where only two of the 16 classes appear. The 5000 points in this set define a border between classes 6 and 7 shifted toward the left, as shown in Figure 4.

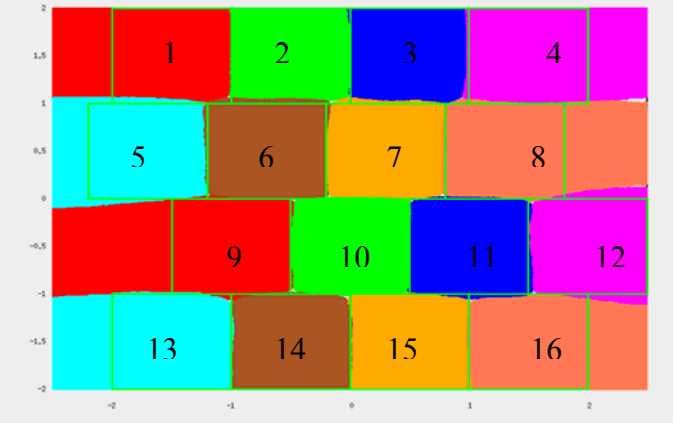

Fig. 3. Training 16 classes on a 4 layer network with 760 weights

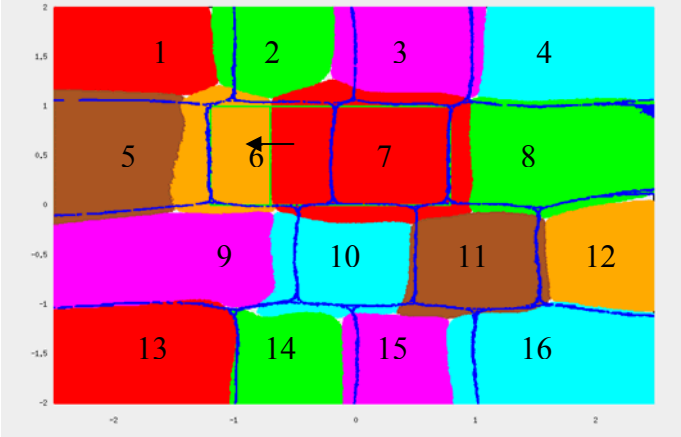

Fig. 4. Adaptation of all the network weights. The adaptation set includes examples of class 6 and class 7 only.

Adaptation method	*Forgetting mitigation technique*	*Average classification rate (%)*	*Class 6 classification rate (%)*	*Class 7 classification rate (%)*
1. None	None	95.9	98.5	93.3
2. Whole network	None	83.1	100.0	98
3. Whole network	CT	89.8	97.8	94.8
4. LIN	None	42.6	100	95.7
5. LIN	CT	69.0	99.0	91.8
6. LHN	None	65.4	99.6	97.2
7. LHN	CT	86.7	98.0	93.3

Table 1
Correct classification rates on the artificial data task.

In the first adaptation experiment, all the 760 MLP weights and 56 biases of the network were adapted. The catastrophic forgetting behavior of the adapted network is evident in Figure 4, where a blue grid has been superimposed to indicate the original class boundaries learned by full training.
Classes 6 and 7 do actually show a relevant increase of their correct classification rate, but they have a tendency to invade the neighbor classes. Moreover, a marked shift toward the left affects the classification regions of all classes, even the ones that are distant from the adapted classes. This undesired shitf of the boundary surfaces induced by the adaptation process damages the overall average classification rate as shown in the second row of Table 1.

To mitigate the catastrophic forgetting problem, the adaptation of the network has been performed using Conservative Training. Figure 5 shows how the trend of classes 6 and 7 to invade neighbor classes is largely reduced, Class 6 and 7 fit well their true classification

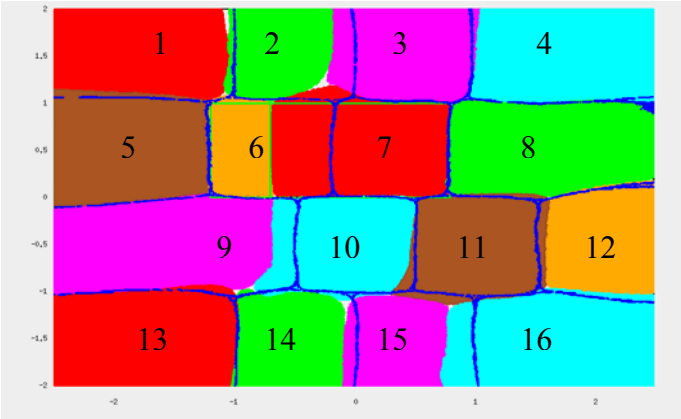

Fig. 5. Conservative Training adaptation of all the network weights.

regions, and although the left shift syndrome is still present, the adapted network performs better as shown by the average classification rate in the third row of Table 1.

Our artificial test-bed is not well suited to LIN adaptation because the classes cover rectangular regions: thus a linear transformation matrix that is able to perform a single *global* rotation of the input features is ineffective. Moreover the degree of freedom of this LIN is really poor: the LIN includes 4 weights and 2 biases only. These considerations are confirmed by the results reported in line 4 of Table 1. Classes 6 and 7 are well classified, but the average classification is very bad because the adaptation of the LIN weights to fit the boundary between class 6 and 7 has the catastrophic forgetting effect of enlarging the regions of all classes.
The mitigation of these effects introduced by Conservative Training is shown in Figure 6 and line 5 of Table 1. The drag toward left syndrome is still visible, but the horizontal boundary surfaces are correct.

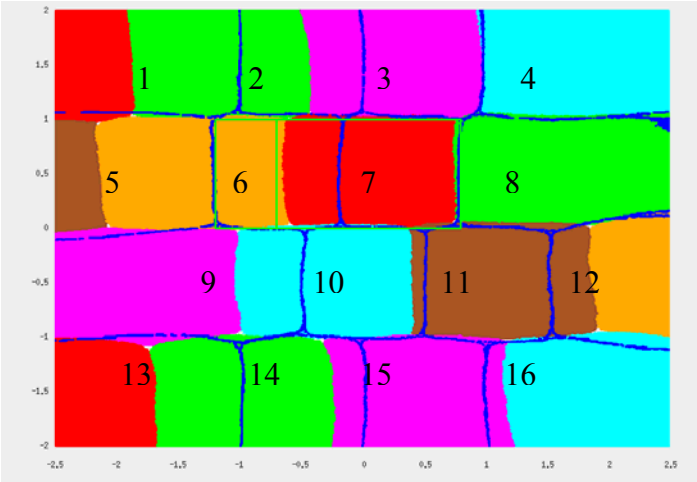

Fig. 6. Conservative Training LIN adaptation.

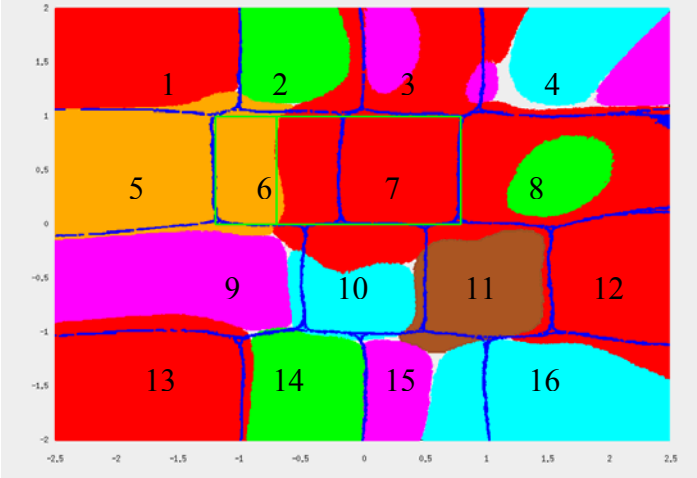

Fig. 7. LHN Adaptation.

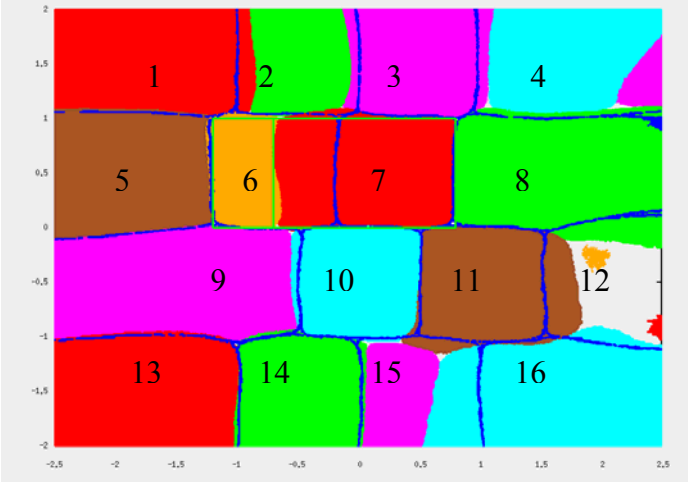

Fig. 8 Conservative Training LHN adaptation.

If we add, instead, a LHN between last hidden layer and the output layer, and we adapt its 420 weights plus biases only, we obtain better results than LIN adaptation (see line 6 of Table 1). However, as Figure 7 shows, the class separation surfaces are ugly. Class 6, and especially class 7 are spread out, class 3 is split, and thus the average classification rate is unacceptable.
Conservative Training does again a very good job, as shown in Figure 8 and in last line of Table 1, even if class 12 does not present high scores.

5 Experimental results on speech recognition tasks

Adaptation to a specific application may involve the speakers, the channel, the environmental noise and the vocabulary, especially if the application uses specific list of terms. The proposed techniques have been tested on a variety of cases requiring different types of adaptation. The adaptation tasks that have been considered are listed in sub-session 4.1

below. The LOQUENDO default speaker independent Italian models were the seed models for the adaptation.

The results of our experiments show that the problem of forgetting is dramatic especially when the adaptation set is not characterized by a good coverage of the phonemes of the language. The use of Conservative Training mitigates the forgetting problem, allowing adaptation with a limited performance decrease of the model on other tasks (some performance reductions are inevitable because the ANN is adapted to a specific condition and thus it is less general).

4.1 Tests on various adaptation tasks

Application adaptation: Directory Assistance

We tested the performance of models adapted to a Directory Assistance application. The corpus includes spontaneous utterances of the 9325 Italian city names. The adaptation set has 53713 utterances; the test set includes 3917 utterances.

Vocabulary adaptation: Command words

The lists A1-2-3 of the SpeechDat-2 Italian corpus, containing 30 command words, have been used. The adaptation and the test sets include 6189 and 3094 utterances respectively.

Channel-Environment adaptation: Aurora-3

The benchmark is the standard Aurora3 Italian corpus. The Well-Matched train set has been used for adaptation (2951 utterances), while the results on Well-Matched test set (the noisy channel, ch1) are reported (654 utterances).

The results on these tests, reported in Table 2, show that a linear transform on hidden units (LHN) always outperforms a linear transform on the input space (LIN). This indicates that the hidden units represent a projection of the input pattern in a space where it is easier to learn or adapt the classification expected at the output of the MLP. The adaptation of the

Adaptation task / Adaptation method	Application *Directory Assistance*	Vocabulary *Command Words*	Channel-Environment *Aurora3 Ch 1*
No adaptation	14.6	3.8	24.0
Whole network	10.5	3.2	10.7
LIN	11.2	3.4	11.0
LIN + CT	12.4	3.4	15.3
LHN	9.6	2.1	9.8
LHN + CT	10.1	2.3	10.4

Table 2
Adaptation results (WER %) on different tasks using various adaptation methods. The seed adaptation models are the standard LOQUENDO telephone models.

Adaptation method	*Directory Assistance Adapted Models*	*Command Words Adapted Models*	*Aurora3 Ch1 Adapted Models*
Whole network	36.3	*63.9*	*126.6*
LIN	36.3	42.7	*108.6*
LIN + CT	36.5	35.2	42.1
LHN	40.6	*63.7*	*152.1*
LHN + CT	40.7	45.3	44.2
No adaptation	29.3		

Table 3
Evaluation of the forgetting problem: recognition results (WER%) on Italian continuous speech with various adapted models.

whole net is feasible only if many adaptation data are available, and is less effective than LHN.
As expected, CT slightly reduces the performance, but the CT adapted models have greater generalization capabilities. This claim has been assessed by testing on a generic common task (continuous speech with a large vocabulary) the models adapted on a specific condition. The reference word error rate achieved using un-adapted acoustic models on the same task is given in the last line of the Table 3.
Because the adapted models have been specialized to a specific condition, some performance reductions are justified. But, the interest of the results shown in Table 3 is that they highlight the effects of catastrophic forgetting, which takes place when the vocabulary of the adaptation set is small and it has a poor phonetic coverage. This is particularly evident for the *Command words* and *Aurora 3* adapted models whose results on the generic continuous speech recognition task are emphasized in italics in Table 3. Conservative Training mitigates the problem, preserving an acceptable performance of the adapted model on the task for which the original network was trained (open vocabulary speech recognition).

4.2 Speaker Adaptation

Further experiments have been performed on the WSJ0 speaker adaptation test in several conditions. Three baseline models have been used:

- the default LOQUENDO 8kHz telephone speech model (trained with LDC MACROPHONE [18] – referred as MCRP in the Tables);
- a model trained with the WSJ0 train set (SI-84), 16 kHz.
- a model trained with the WSJ0 train set (SI-84), down-sampled to 8 kHz.

Furthermore, we tested two architectures for each type of models: the standard one (STD), described in sub-section 2.1 and an improved one (IMP), characterized by a wider input window modeling a time context of 250 ms [17], and by the presence a third 300 units hidden layer.
The adaptation set is the standard adaptation set of WSJ0 (si_et_ad, 8 speakers, 40 utterances per speaker), down-sampled to 8 kHz when necessary.

Train Set	*Net type*	*Adaptation method*	*Bigram LM*	*Trigram LM*
MCRP	STD	NO adaptation	16.4	13.6
MCRP	STD	Standard LIN	14.6	11.6
MCRP	STD	LIN+CT	13.9	11.3
MCRP	STD	LHN+CT	12.1	9.9
MCRP	STD	LIN+LHN+CT	11.2	9.0
WSJ0	STD	NO adaptation	13.4	10.8
WSJ0	STD	Standard LIN	14.2	11.6
WSJ0	STD	LIN+CT	11.8	9.7
WSJ0	STD	LHN+CT	10.4	8.3
WSJ0	STD	LIN+LHN+CT	9.7	7.9
WSJ0	IMP	NO adaptation	10.8	8.8
WSJ0	IMP	Standard LIN	9.8	7.6
WSJ0	IMP	LIN + CT	9.8	7.7
WSJ0	IMP	LHN + CT	8.5	6.6
WSJ0	IMP	LIN+LHN+CT	8.3	6.3

Table 4
Speaker Adaptation results – WSJ0 8 kHz.

The test set is the standard SI 5K read NVP Senneheiser microphone (si_et_05, 8 speakers x ~40 utterances), and the bigram or trigram standard Language Models provided by Lincoln Labs have been used.
The results, reported in Tables 4 and 5, show that also in these tests LHN always achieves better performance that LIN. The combination of LIN and LHN (trained simultaneously) is usually better than the use of LHN alone. Conservative training (CT) effects are of minor importance in these tests because the adaptation set has a good phonetic coverage and the problem of unseen phonetic classes is not dramatic.
Nevertheless, its use improves the performance (compare Standard LIN and LIN+CT), because it avoids the adaptation of prior probabilities of the phonetic classes on the (poor) prior statistics of the adaptation set.

Train Set	*Net type*	*Adaptation method*	*Bigram LM*	*Trigram LM*
WSJ0	STD	NO adaptation	10.5	8.4
WSJ0	STD	Standard LIN	9.9	7.9
WSJ0	STD	LIN+CT	9.4	7.1
WSJ0	STD	LHN+CT	8.4	6.6
WSJ0	STD	LIN+LHN+CT	8.6	6.3
WSJ0	IMP	NO adaptation	8.5	6.5
WSJ0	IMP	Standard LIN	7.2	5.6
WSJ0	IMP	LIN+CT	7.1	5.7
WSJ0	IMP	LHN+CT	7.0	5.6
WSJ0	IMP	LIN+LHN+CT	6.5	5.0

Table 5
Speaker Adaptation results – WSJ0 16 kHz.

5 Conclusions

A method has been proposed for adapting all the outputs of the hidden layer of ANN acoustic models and for reducing the effects of catastrophic forgetting when the adaptation set does not contains examples for some classes. Experiments for the adaptation of an existing ANN to a new application, a new vocabulary, a new noisy environment and new speakers have been performed. They all show the benefits of CT, and also that LHN outperforms LIN. Furthermore, experiments on speaker adaptation show that further improvements are obtained by the simultaneous use of LHN and LIN showing that linear transformations at different levels produce different positive effects that can be effectively combined.
An overall WER of 5% after adaptation on WSJ0 using the standard trigram LM and without across word specific acoustic models compares favorably with published results.
Future work will explore unsupervised adaptation and the use of eigenvoices.

References

[1] J. L. Gauvain, C. H. Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains", IEEE Trans. on Speech and Audio Processing, Vol. 2, n. 2, pp. 291-298, 1994.

[2] M.J.F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", Computer Speech and Language, Vol. 12, pp. 75-98, 1998.

[3] R. Kuhn, J.-C. Junqua, P. Nguyen, N. Niedzielski. "Rapid Speaker Adaptation in Eigenvoice Space", IEEE Trans. on Speech and Audio Processing, Vol. 8, no. 4, pp. 695–707, Nov 2000.

[4] S. Sagayama, K. Shinoda, M. Nakai, and H. Shimodaira, "Analytic methods for acoustic model adaptation: A review", in Proc. Adaptation Methods for Speech Recognition, ISCA ITR-Workshop, France, 2001, pp. 67–76.

[5] C.-H. Lee and Q. Huo, "On adaptive decision rules and decision parameter adaptation for automatic speech recognition", Proc. IEEE, vol. 88, no. 8, pp. 1241–1269, Aug. 2000.

[6] R Hsiao and B. Mak, "Discriminative feature transformation by guided discriminative training", Proc. ICASSP-04, Montreal, pp. 897-900, 2004.

[7] X. Liu and M.J.F. Gales, "Model complexity control and compression using discriminative growth functions", Proc. ICASSP-04, Montreal, pp. 797-800, 2004.

[8] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," Proc. EUROSPEECH 1995, pp. 2183–2186, 1995.

[9] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, T. Robinson, "Speaker-adaptation for Hybrid HMM-ANN Continuous Speech Recognition System," Proc. EUROSPEECH 1995, pp. 2171–2174, 1995.

[10] J. Stadermann, G. Rigoll, "Two-stage speaker adaptation of hybrid tied-posterior acoustic models". Proc. ICASSP-05, Philadelphia, pp. I-997,1000, 2005.

[11] S. Dupont, L. Cheboub. "Fast speaker adaptation of artificial neural networks for automatic speech recognition", Proc. ICASSP 2000, pp. 1795-1798, 2000.

[12] D. Albesano, R. Gemello, F. Mana, "Hybrid HMM-NN Modeling of Stationary-Transitional Units for Continuous Speech Recognition", Int. Conf. On Neural Information Processing, pp. 1112–1115, 1997.

[13] M. French, "Catastrophic Forgetting in Connectionist Networks: Causes, Consequences and Solutions", in *Trends in Cognitive Sciences*, 3(4), pp. 128-135.

[14] A. Robins, "Catastrophic forgetting, rehearsal, and pseudo-rehearsal". *Connection Science*, 7, 123 – 146, 1995.

[15] M.F. BenZeghiba and H. Bourlard, **"**Hybrid HMM/ANN and GMM Combination for User-Customized Password Speaker Verification**,"** ICASSP-03, pp. 225-228, 2003.

[16] D. S. Pallett, J. G. Fiscus, W. M. Fisher, J. S. Garofolo, B. A. Lund, and M. A. Przybocki, “1993 Benchmark Tests for the ARPA Spoken Language Program”, In Proc. of the Human Language Technology Workshop, pp. 49–74, Plainsboro, 1994.
[17] S. Dupont, C. Ris, L. Couvreur and J. M. Boite. “A study of implicit and explicit modeling of coarticulation and pronunciation variation”, Proc. Interspeech-05, pp. 1353-1356, Lisbon, 2005.
[18] Available at http://www.ldc.upenn.edu