ADAPTATION OF HYBRID ANN/HMM MODELS USING LINEAR HIDDEN TRANSFORMATIONS AND CONSERVATIVE TRAINING

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

[1] LOQUENDO, Torino – Italy
roberto.gemello@loquendo.com
franco.mana@loquendo.com

[2] Politecnico di Torino, Italy
pietro.laface@polito.it
stefano.scanzio@polito.it

[3] LIA . Univ. of Avignon, France
renato.demori@lia.univ-avignon.fr

ABSTRACT

A technique is proposed for the adaptation of automatic speech recognition systems using Hybrid models combining Artificial Neural Networks with Hidden Markov Models.
The application of linear transformations not only to the input features, but also to the outputs of the internal layers is investigated. 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.
A new solution, called Conservative Training, is proposed that compensates for the lack of adaptation samples in certain classes.
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.

1. INTRODUCTION

A large number of papers have described techniques for refining Automatic Speech Recognition (ASR) systems by adapting the acoustic features and the parameters of stochastic models to environments, applications and speakers [1-5]. More recently, particular attention has been paid to discriminative training techniques and their application to the acoustic feature transformation [6,7].
Since Artificial Neural Networks (ANN), used as acoustic models, are also trained with discriminative methods, it is worth exploring methods for adapting their features and model parameters. Some solutions to this problem have been proposed. In [8,9], different techniques for adapting neural networks are compared. These techniques include adding a linear transformation network that acts as a pre-processor to the main network or adapting all 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 optimising 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".
The risk of catastrophic forgetting [13] is particularly high when a distributed connectionist 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. If the outputs of the missing classes are forced to a value close to zero for all the adaptation samples, there is a risk that the network becomes less sensitive to input data belonging to these classes. This paper proposes a solution to this problem introducing Conservative Training, a variant 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 favourably 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 Loquendo 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. Section 4 reports the experiments performed on several databases with the aim of clarifying the behavior of the new adaptation

technique with respect to the classical LIN approach. Finally the conclusions and future developments are presented in the last Section.

2. FEATURE TRANSFORMATIONS

2.1 The ANN architecture

The Loquendo-ASR decoder uses a 4-layer hybrid HMM-MLP model where each phonetic unit is described in terms of a single or double state left-to-right automaton with self-loops. The models are based on a set of vocabulary and gender independent units including stationary context-independent phones and diphone-transition coarticulation models. The HMM transition probabilities are uniform and fixed [12].

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. We typically use 273 feature for the input layer (39 parameters of a 7 frame context), 315 nodes for the first hidden layer, and 300 for the second hidden layer.

These models have been successfully used for the 15 languages released with the Loquendo ASR recognizer, and are the starting models for adaptation experiments of section 4, if not differently specified.

2.2 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. 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.

2.3 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:

$$W_a = W_{LHN} \times W_{SI}$$
$$B_a = B_{SI} + B_{LHN} \times W_{SI} \quad (1)$$

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.

3. CATASTROPHIC FORGETTING

It is well known that in connectionist learning, acquiring new information in the adaptation process, can damage previously learned information [13]. 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 by assigning to them a zero target value for every adaptation frame. That induces in the ANN a forgetting of the 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 acoustic similar 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 of no observations, it has been proposed [14] 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 [15], 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.

A solution called Conservative Training (CT) is proposed here to mitigate the forgetting problem. Since ANN training is discriminative, the units for which no observations are available will have zero as a target for all the adaptation samples. Thus, during adaptation, the weights of the acoustic ANN 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 avoids to always associating the value zero to the target of the missing units, using instead as target 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 other missed ones. In Conservative Trainining 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 that is missing 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.

4. EXPERIMENTAL RESULTS

4.1 Test on different adaptation 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 types that have been considered are listed below.

Application adaptation: Directory Assistance
The adaptation to a *Directory Assistance* application has been tested. The corpus includes spontaneous utterances of the 9325 Italian town names. The adaptation set is made of 53713 utterances; the test set includes 3917 utterances.

Vocabulary adaptation: Command words
The lists A1-2-3 of SpeechDat-2 Italian, 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).

Speaker: WSJ0
The standard adaptation and test sets of WSJ0 (8 speakers, 40 utterances per speaker) have been used after 8 kHz down-sampling. The down-sampling was performed because the original ANN model is trained with the LDC Macrophone telephone speech corpus. Standard bigram language model is employed.
The results on these tests, reported in Table 1, 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 whole net is feasible only if many adaptation data are available, and is less effective than LHN.

Table 1. Adaptation results on different tasks with different methods (WER %). The adaptation starts from the standard Loquendo telephone models.

Adaptation type	***Application***	***Vocabulary***	***Channel-Environ.***	***Speaker***
Test case:	***Directory Assistance***	***Command Words***	***Aurora3 Ch1***	***WSJ0 (8 kHz)***
no adaptation	**14.6**	**3.8**	**24.0**	**16.4**
whole net	**10.5**	**3.2**	**10.7**	**15.3**
LIN+CT	**12.4**	**3.4**	**15.3**	**13.9**
LHN+CT	**10.1**	**2.3**	**10.4**	**12.1**

4.2 Speaker Adaptation (WSJ0)

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

- the default 8kHz telephone speech model (trained with LDC Macrophone – 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, for each type of models two architectures are tested: a 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.

The test set is the standard SI 5K read NVP Senneheiser microphone (si_et_05, 8 speakers x ~40 utterances) with bigram or trigram standard LM provided by Lincoln Labs.
The results, reported in Tables 2 and 3, show that also in these cases LHN is always better 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 WSJ0 because the adaptation set has a good phonetic coverage and the problem of unseen phonetic classes is not dramatic. Nevertheless, its use improves performances (see LIN standard vs. LIN+CT), because it avoids the adaptation of

prior probabilities of the phonetic classes on the (poor) prior statistics of the adaptation set.

Table 2. Speaker Adaptation results – WSJ0 8 kHz

Train Set	*Net type*	*Adaptation method*	*Bg LM*	*Tg LM*
MCRP	STD	NO adaptation	**16.4**	**13.6**
MCRP	STD	LIN standard	**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	LIN standard	**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	LIN standard	**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 3. Speaker Adaptation results – WSJ0 16 kHz

Train Set	*Net type*	*Adaptation method*	*Bg LM*	*Tg LM*
WSJ0	STD	NO adaptation	**10.5**	**8.4**
WSJ0	STD	LIN standard	**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	LIN standard	**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**

6. CONCLUSIONS[1]

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.

[1] This work was partially supported by the EU FP-6 IST Projects HIWIRE and DIVINES

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

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

[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 modelling of coarticulation and pronunciation variation", Proc. Interspeech-05, pp. 1353-1356, Lisbon, 2005.