ADAPTATION OF HYBRID ANN/HMM USING WEIGHTS INTERPOLATION

Stefano Scanzio[1], Pietro Laface[1], Roberto Gemello[2], Franco Mana[2]

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

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

ABSTRACT

Many techniques for speaker or channel adaptation have been successfully applied to automatic speech recognition. Most of these techniques have been proposed for the adaptation of Hidden Markov Models (HMMs). Far less proposals have been made for the adaptation of the Artificial Neural Networks (ANNs) used in the hybrid HMM-ANN approach.

This paper presents an adaptation technique for ANNs that, similar to the framework of MAP estimation, tries to exploit in the adaptation process prior information that is particularly useful to deal with the problem of sparse training data.

We show that the integration of a priori information can be simply achieved by linear interpolation of the weights of an "a priori" network and of a speaker specific network.

Good improvements with respect to the baseline results are reported evaluating this technique on the Wall Street Journal WSJ0 and WSJ1 databases and on TIMIT corpus using different amounts of adaptation data.

1. INTRODUCTION

The recognition accuracy of speaker-independent recognition systems heavily depends on speaker variability. Significant performance degradations are experienced with outlier or non-native speakers. Environment, channel, and microphone variability is another important source of errors.

Several techniques have been proposed to deal with these variations in the framework of Gaussian Mixture HMMs. The Speaker Independent (SI) model parameters are adapted using a limited amount of new data characterising the new speaker, environment, channel, or microphone.

As far as speaker variability is concerned, speaker adaptation can be performed exploiting prior information included in the SI models [1], or mapping the input speech of new speakers to a SI system [2-4].

In comparison, far less proposals have been made for the adaptation of the Artificial Neural Networks used in the hybrid HMM-ANN approaches [5-7].

In [8] the eigenvoice approach proposed in [9] has been originally applied to the HMM-ANN approach to reduce the need of adaptation data.

The simplest and more popular approach to speaker adaptation with ANNs is Linear Input Transformation [5,6]. The input space is rotated – and shifted – by a linear transformation to make the target conditions more consistent with the speaker independent 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 SI ANN. The LIN weights 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 SI network, because it involves the estimation of a lower number of parameters.

Often, however, the available data do not include enough samples to allow accurate adaptation of all the acoustic-phonetic units. This is, of course, the main problem for every adaptation task. The problem is more severe in the ANN modeling framework than in the classical Gaussian Mixture HMMs. The reason is that an ANN estimates the posterior probability of each acoustic-phonetic unit state using discriminative training. 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. The result is that the posterior probability of the unseen units is inappropriately reduced.

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 behaviour of the network for the unseen units.

The LIN approach reduces these effects because a single linear transformation of the input parameters is performed. The learned transformation, however, may heavily depend on the phonetic content when few sentences are available for adaptation.

To mitigate the problem of unseen units, this paper proposes an adaptation technique that, similar to MAP estimation in the Gaussian Mixture framework, tries to exploit in the adaptation process prior information that is particularly useful to deal with the problems of sparse training data.

We show that the integration of the a priori information can be simply achieved by linear interpolation of the weights of an "a priori" network and of a speaker specific network.

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. Section 3 presents our interpolation technique that combines speaker adapted models and a priori information derived, for example, from the environment/channel information. Section 4 reports the experiments performed on three databases using different amounts of adaptation data. Finally the conclusions and

future developments are presented in the last Section.

2. NEURAL NETWORK 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 HMM 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 [10].
These models have been successfully used for the 15 languages provided by the Loquendo ASR recognizer.
45 stationary units and 485 transition units have been defined for the US-English language, for a total of 949 output states.

3. NETWORK WEIGHTS INTERPOLATION

To mitigate the problem of loosing characterization of the units with little of no observations, discussed in the introduction, it has been proposed [12] 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.
The idea developed in this work is to exploit in the adaptation process prior information, similar to MAP estimation in the Gaussian Mixture framework.
For example, the information gathered adapting an SI ANN to an environment/channel can be used as a priori knowledge to condition the adaptation of the SI network to a new speaker using small amounts of her/his data.
We illustrate our approach assuming that a SI model has to be adapted to a rather large set of speakers. Thus, we have enough data to adapt a SI network to a possibly new environment/channel, but a small amount of speaker specific data. We exploit, as a priori knowledge, the parameters estimated from the adaptation of the SI network with the data of *all the speakers*, i.e. the parameters of the environment/channel adapted network. The idea is to constrain the gradients of the weights of the Speaker Adapted network to take into account the gradients of the environment/channel adapted weights
An ANN is completely defined by its topology and by the set of its weights (including the biases). If we use the same topology for the seed network and for the adapted one, we can describe a network by the set of its *n* weights.
We will denote W^{SI}, W^{EA}, and W^{SA} the set of weights of a Speaker Independent, Environment/Channel Adapted, and Speaker Adapted networks respectively

$$W^{SI} = \{w_1^{SI}, w_2^{SI}, ..., w_n^{SI}\} \quad (1)$$

$$W^{EA} = \{w_1^{EA}, w_2^{EA}, ..., w_n^{EA}\} \quad (2)$$

$$W^{SA} = \{w_1^{SA}, w_2^{SA}, ..., w_n^{SA}\} \quad (3)$$

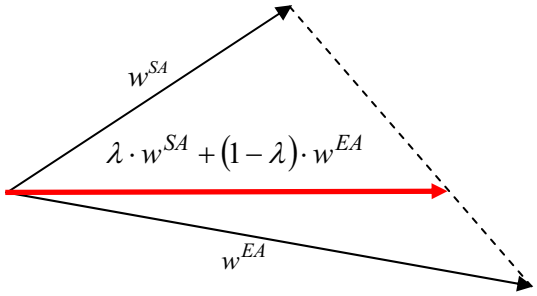

Figure 1: Interpolation of the weight gradients

The Environment/Channel adaptation will produce a new set of weights w^{EA}, and a gradient Δw^{EA} with respect to the original weights w^{SI}.

$$\Delta w_i^{EA} = w_i^{EA} - w_i^{SI} \qquad (i = 1,...n) \quad (4)$$

Adapting the SI network to a specific speaker produces, instead, another set of Speaker Adapted weights and a gradient Δw^{SA}.

$$\Delta w_i^{SA} = w_i^{SA} - w_i^{SI} \qquad (i = 1,...n) \quad (5)$$

Depending on the amount of adaptation data, the speaker adapted model can be unreliable.
Thus, to account for the a priori information given by the environment adaptation, we propose to linearly interpolate the gradient Δw^{SA} with the gradient Δw^{EA}, to get the new set of speaker adapted weights w^{AD}.

$$w_i^{AD} = w_i^{SI} + (1-\lambda) \cdot \Delta w_i^{EA} + \lambda \cdot \Delta w_i^{SA} \quad (6)$$

The idea is illustrated in Figure 1.

Since $\Delta w^{EA} = w^{EA} - w^{SI}$ and $\Delta w^{SA} = w^{SA} - w^{SI}$, it turns out that interpolating the weight gradients corresponds to the linear interpolation of the network weights

$$w_i^{AD} = (1-\lambda) \cdot w_i^{EA} + \lambda \cdot w_i^{SA} \quad (7)$$

Thus, even an SI network can be used as an "a priori" network to limits the effects due to the scarcity of data.

4. EXPERIMENTAL RESULTS

The adaptation corpora of the Wall Street Journal WSJ0 and WSJ1 databases [11] have been used for testing the adaptation-by-interpolation approach.
The WSJ0 test used is the standard 5K test set, including 8 speakers and about 40 sentences per speaker. The adaptation set consists of approximately 40 sentences pronounced by the same 8 speakers. Only the Senneheiser component has been used both for adaptation and for testing.
The WSJ1 test used here is the *spoke3*, which is a similar test containing 10 non-native speakers with 40 adaptation sentences and approximately 40 test sentences per speaker.
Since in this work we were mainly interested in testing the relative improvements due to the interpolation approach, the experimental test-bed was set as follows:

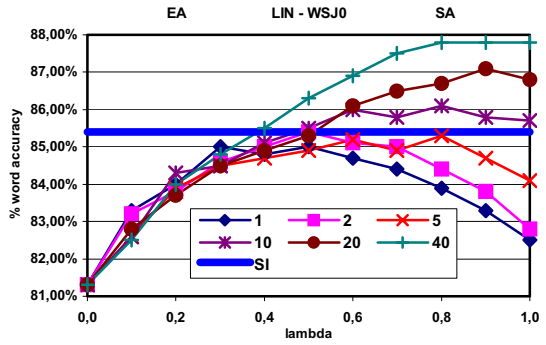

Figure 2: Word accuracy on WSJ0 as a function of the number of adaptation sentences (show in key). Interpolation of environment and speaker adapted LIN weights.

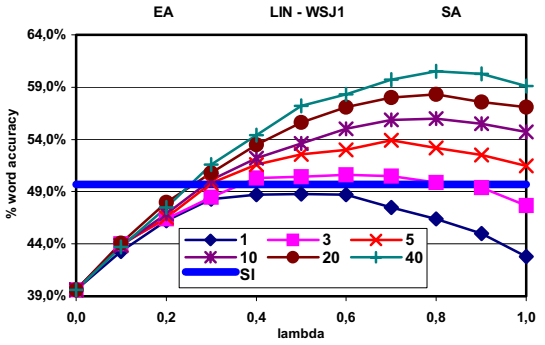

Figure 4: Word accuracy on WSJ1. Interpolation of environment and speaker adapted LIN network weights.

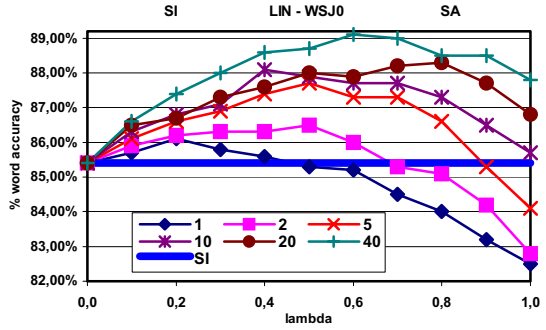

Figure 3: Word accuracy on WSJ0. Interpolation of speaker independent and speaker adapted LIN network weights.

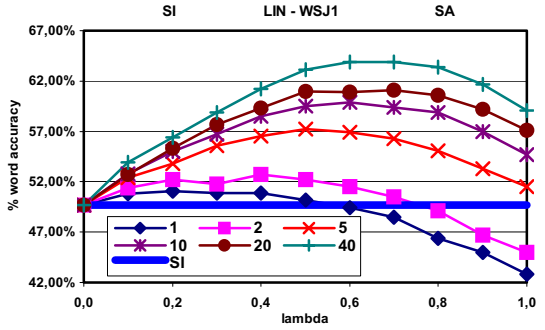

Figure 5: Word accuracy on WSJ1. Interpolation of speaker independent and speaker adapted LIN network weights.

– The speaker independent US-English models provided by the Loquendo ASR have been used for the baseline systems. Neither retraining nor adaptation of these *telephone* speech models (8 KHz) to the WSJ *microphone* environment or channel has been done with the WSJ0 *training* data.
– The supplied 5000 words, closed-vocabulary bigram and trigram models of the WSJ0 have been used as language models.
– Due to the often limited number of sentences available per speaker, for each speaker, a LIN network (SA) has been estimated adapting the SI model with an increasing number of speaker sentences, while all the sentences of the other speakers have been used to estimate the Environment/Channel adapted LIN network (EA).

Figure 2 shows the average word accuracy for the 8 speakers of the WSJ0 corpus, using the standard trigram language model, as a function of the interpolation factor and of the number of adaptation sentences. In the figure, the interpolation factor λ=0 corresponds to performing the tests with the EA model, while λ=1 gives the average performance of the speaker adapted models (SA).
The average performance of the EA models (81.3% word accuracy) is worse that the original SI models (85.4%) because they are adapted with the sentences of 7 speakers only and, moreover, those speakers may not have acoustic-phonetic characteristics in common with the test speaker.

The test confirms what was expected: the linear interpolation requires at least 10 sentences per speaker to prevail over the SI network, and the contribution of the interpolation is minimal because the a priori information is weak.
Better results, shown in Figure 4, have been obtained on the non-native speaker adaptation task of the WSJ1 corpus. With at least 5 adaptation sentences, the Speaker Adapted as well as the interpolated models outperform the SI model. The interpolation factor that gives the best performance increases with the amount of speaker specific data, but λ=0.7 is still a good setting.
Since the EA models are weak because too few sentences and speakers contribute to the adaptation, better performance is obtained interpolating the SI and the speaker adapted model, as reported in Figures 3 and 5 for the WSJ0 and WSJ1 respectively.
Using an improved SI model, a baseline SI WER of 6.6% has been obtained on the WSJ0 test with the standard trigram LM, and without cross-word specific acoustic models. The model was trained with the WSJ0 train set (16 kHz), a wider input window modeling a time context of 250 ms [13], and a third hidden layer.
Adapting to the speaker a combination of a LIN network and of another linear layer between the last hidden layer and the output layer gives a 5.0% WER.
As shown in Figure 6, the interpolation approach is still able to improve the performance of the SA model to 4.9% WER keeping fixed the parameter λ to 0.7.

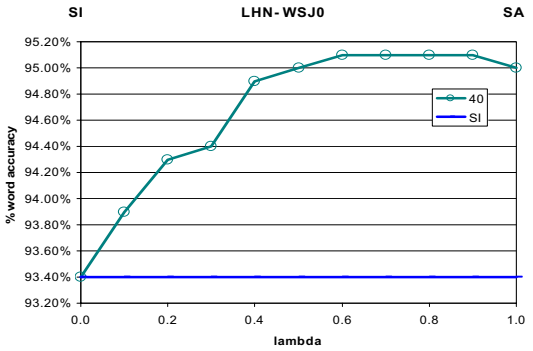

Figure 6: Word accuracy on WSJ0. Interpolation of speaker independent and speaker adapted LIN network and linear hidden layer weights.

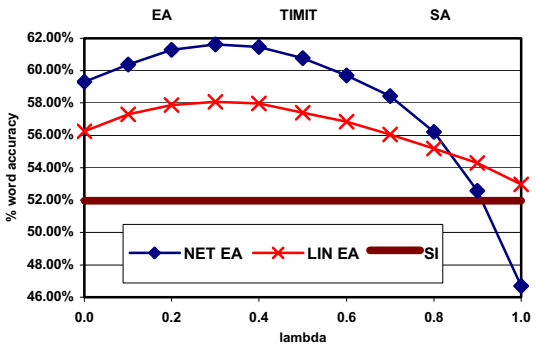

Figure 7: Word accuracy on TIMIT. Interpolation of the environment and speaker adapted LIN or net weights.

Finally, the interpolation approach has also been tested on the TIMIT corpus to test its behavior on a task where a fairly large amount of environment/channel adaptation data is available.

The task is phone recognition with a vocabulary of the 45 phones defined in our standard US-English model (i.e. not the ones appearing in the supplied manual transcriptions of the TIMIT corpus), with 462 training speakers, 4620 training sentences, and 168 speakers 1344 test sentences. Again, the seed model is the standard SI model. The results have been obtained with a *single* adaptation sentence per speaker, and without using a phone bigram language model. The two curves of Figure 7 show the results of the interpolation of the LIN weights only, and those referring to the interpolation of *all* the weights of the EA and SA networks. Here the EA network is far better than the SI one, but the LIN interpolation is still able to improve the phone accuracy of 2% absolute. A reduced interpolation factor λ=0.3 must be used in this case because the EA network is well trained, whereas the SA adaptation is performed with a single utterance only. The same adaptation property is offered by the interpolation of all the network weights. Although, as expected, the network SA model performance is very poor, worse than the one obtained with the SI network weights or with the LIN SA models, the interpolated model outperforms the EA model, and achieves a 10% better score than the speaker independent model.

6. CONCLUSIONS[1]

An adaptation technique for ANNs has been presented that, similar to the MAP estimation, exploits prior information to reduce the effects of data scarcity interpolating the weights of two networks.

Good improvements have been obtained in the WSJ0 adaptation task and with the adaptation of non-native speakers of the WSJ1. Moreover, fast adaptation has been demonstrated on the TIMIT corpus.

Work is in progress to apply this approach to other tasks, such as parallel network training, and to further mitigate the problem of units with scarce or no observations.

[1] This work was supported by the EU FP-6 IST Project DIVINES – Diagnostic and Intrinsic Variabilities in Natural Speech

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] C. J. Leggetter, P. C. Woodland. Maximum likelihood linear regression for speaker adaptation. Computer Speech and Language, Vol. 9, pp.171-185, 1995.

[3] V. Digalakis, D. Rtischev and L. Neumeyer, "Speaker Adaptation Using Constrained Estimation of Gaussian Mixtures", IEEE Trans. on Speech and Audio Processing, Vol. 3, no. 5, pp. 357-366, 1995.

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

[5] H.Bourlard, and N. Morgan, "Connectionist Speech Recognition – A Hybrid Approach", Kluwer Academic Press, 1994.

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

[7] 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.

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

[9] 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.

[10] 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.

[11] 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.

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

[13] 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, Lisbon, 2005.