MLP Based Hierarchical System for Task Adaptation in ASR

Joel Pinto [1,2], Mathew Magimai.-Doss [1], and Hervé Bourlard [1,2]

[1] Idiap Research Institute, Martigny, Switzerland.
[2] École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{joel.pinto, mathew, bourlard}@idiap.ch

***Abstract*—We investigate a multilayer perceptron (MLP) based hierarchical approach for task adaptation in automatic speech recognition. The system consists of two MLP classifiers in tandem. A well-trained MLP available off-the-shelf is used at the first stage of the hierarchy. A second MLP is trained on the posterior features estimated by the first, but with a long temporal context of around 130 ms. By using an MLP trained on 232 hours of conversational telephone speech, the hierarchical adaptation approach yields a word error rate of 1.8% on the 600-word Phonebook isolated word recognition task. This compares favorably to the error rate of 4% obtained by the conventional single MLP based system trained with the same amount of Phonebook data that is used for adaptation. The proposed adaptation scheme also benefits from the ability of the second MLP to model the temporal information in the posterior features.**

I. Introduction

Multilayer perceptron (MLP) classifiers are being extensively used for acoustic modeling in automatic speech recognition (ASR) [1][2][3][4][5]. The MLP is trained using acoustic features such as perceptual linear predictive (PLP) cepstral coefficients, and its output classes represent the subword units of speech such as phonemes. A well trained MLP can estimate the posterior probabilities of its output classes, conditioned on the input acoustic features [6][7].

The phonetic class conditional probabilities estimated by the MLP are typically used in hidden Markov model (HMM) based speech recognition as (a) local emission scores in the hybrid HMM/MLP [7] system or (b) features (after appropriate transformation) to an HMM/GMM system [8]. In this work, whenever the estimated phonetic class conditional probabilities are used as features to train another MLP classifier, we refer them to as *posterior features*.

In more recent works, we proposed an MLP based hierarchical system for estimating the phonetic class conditional probabilities [9][10][11]. The system consisted of two MLP classifiers connected in tandem. The first MLP was trained using PLP features with a temporal context of 90 ms. The second MLP was trained using the posterior features estimated by the first classifier, but with a relatively longer temporal context of 150-230 ms. Extensive experiments on recognition of phonemes on TIMIT as well as conversational telephone speech (CTS) showed that the hierarchical system is a better estimator of the phonetic class conditional probabilities.

There is growing interest in the research community in adapting the MLPs trained on large/different corpora to new tasks, where the amount of training data is limited. For instance, in [12] the MLP trained on a large subset of CTS data was adapted for ASR on meetings. Adaptation was achieved by performing three additional iterations of back-propagation training using the meeting data. The adapted MLP was used to extract posterior (and hidden activation) features on meetings, and this approach was shown to yield better performance. In [13], the weights connecting the last hidden layer and output layer were re-learned (regularized adaptation) using a small amount of adaptation data.

In this work, we investigate the MLP based hierarchical approach for task adaptation. A well trained MLP available off-the-shelf is used at the first stage of the hierarchical system. The second MLP is trained with a small amount of adaptation data specific to the target task. In this work, we use an MLP trained on 232 hours of CTS data for isolated word recognition on the Phonebook database. We compare the proposed approach to the standard hybrid system (single MLP trained on Phonebook data) and the hierarchical MLP approach, where both the MLPs are trained on the same Phonebook data.

We also study the performance of the hierarchical adaptation system with respect to: (a) the temporal context on the posterior features at the input of the second MLP (b) the quality of the posterior features estimated by the first MLP (c) the complexity of the second MLP in terms of the number of parameters, and (d) the amount of adaptation data used for adaptation. We ascertain if these results are consistent with the conclusions from our previous study [9], where both the MLPs in the hierarchical system were trained on the same task.

II. Background

In this section, we briefly discuss the motivation for the MLP based hierarchical system, and list some of the major findings of our previous study [9]. Fig. 1 shows the block schematic of the proposed system.

The posterior features are endowed with two important properties:

- **Linear separability:** The model parameters of the first MLP are optimized to minimize the cross entropy be-

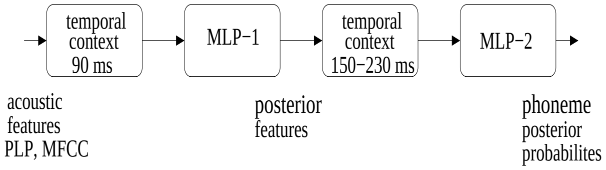

temporal context 90 ms
MLP−1
temporal context 150−230 ms
MLP−2
acoustic features PLP, MFCC
posterior features
phoneme posterior probabilites

Fig. 1. Estimation of posterior probabilities of phonemes using an hierarchy of two MLPs. The second MLP is trained using the posterior probabilities of phonemes estimated by the first MLP.

tween the estimated posterior probability vectors and the output target vectors, which are typically in the hard-target or one-hot format. In other words, the hard target vectors are at the simplex of the posterior feature space, which makes them linearly separable. Hence, a well trained model attempts to achieve linear separability in the estimated posterior features.

- **Lesser variabilities:** It has been shown that a well trained (large population of speakers, and different noise and channel conditions) MLP classifier can achieve invariance to speaker [3] as well as environmental [14] characteristics in the acoustic features.

In other words, posterior features represent soft-decisions on the underlying sequence of phonemes (the linguistic message), have lesser nonlinguistic variabilities when compared to acoustic features, and are simpler to classify. As a consequence, contextual information in the posterior features spanning longer temporal contexts can be effectively modeled by the second MLP. Contextual information in the posterior features manifests in the evolution of the trajectories of the estimated probabilities within a phoneme (sub-phonemic level) as well as in its transition to neighboring phonemes (sub-lexical level).

The major findings of our work are summarized below:

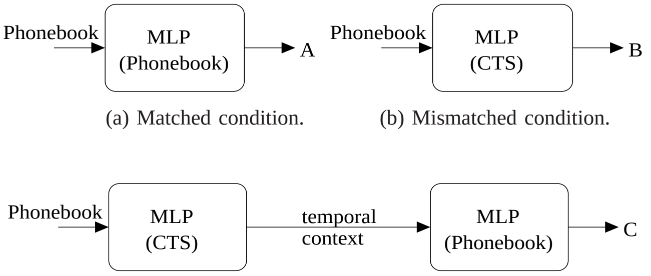
- The useful contextual information in the posterior features spans a temporal context of 150-230 ms.
- The second MLP learns the phonetic-temporal patterns in the posterior features, which include the phonetic confusions at the output of the first MLP classifier and to a certain extent the phonotactics of the language as observed in the training data.
- As the posterior features are trained to achieve linear separability, the second MLP could be simpler in terms of the number of parameters. Even a single layer perceptron, which is a linear classifier, yielded better performance in comparison with the single MLP system.
- As the posterior features have lesser nonlinguistic variabilities, the second MLP can be trained with a lesser amount of training data.

Motivated by the above findings, we investigate the possibility of using the hierarchical system for task adaptation. We use an off-the-shelf MLP trained on 232 hours of CTS data for recognition of isolated words on the Phonebook task. The second MLP in the hierarchical system is trained using the posterior features estimated by the first MLP, with a long temporal context.

III. Adaptation System

Fig. 2 illustrates the adaptation studies carried out in this paper. Fig. 2 (a) shows the single MLP based modeling, where the training and test conditions are matched. Fig. 2 (b) shows the mismatched condition where an MLP trained on CTS is directly used for the target Phonebook task. There can be a mismatch in the estimated posterior features at points A and B in the figure. The mismatch could arise due to the following:

- Differences in the pronunciation dictionary used in the development of the system, *i.e.*, for forced alignment to obtain the ground truth phonetic transcription for training the MLP and the one available for decoding in the target task. The differences could be in the number of phonemes used and/or in the pronunciation lexicon of the words in the dictionary.
- The acoustic mismatch due to different characteristics of speech used for training the MLP and in the target task.



(a) Matched condition.

(b) Mismatched condition.

(c) Mismatched condition plus adaptation.

Fig. 2. The adaptation scheme: (a) matched conditions, (b) mismatched conditions, and (c) the adaptation using the hierarchical system

Fig. 2 (c) is a block schematic of the proposed adaptation scheme. The parameters of the second MLP are learned using a small amount of data available on the target task. Based on our previous findings, we can expect the second MLP to compensate for any systematic perturbations in the posterior features, arising due to the mismatch in the training and test conditions of the first MLP classifier.

The proposed hierarchical system can be considered as adaptation via feature transformation. The second MLP classifier can be viewed as a discriminatively trained nonlinear transformation from the posterior feature space corresponding to the CTS task to the posterior feature space corresponding to the target Phonebook task.

IV. Experimental setup

Experiments are performed on the Phonebook database [15]. In adaptation studies, off-the-shelf MLPs trained on conversational telephone speech (CTS) [16] are used. This section describes these databases and the experimental setup in detail.

A. Databases

The Phonebook task was designed for speaker-independent, isolated word recognition studies. There are no common words in the training, validation and test sets of the corpus. We use the same definition of training, validation, and test sets as discussed in [17]. The training set consists of 19421 isolated utterances from 243 speakers, amounting to 6.7 hours of speech.[1] The validation set consists of 7920 utterances (2.5 hours) from 106 speakers, and the test set consists of 6598 utterances from 96 speakers. The phonetic transcription required for training the MLP is obtained by forced alignment. For this, we use HMM/GMM acoustic models and the Phonebook pronunciation dictionary, containing 42 phonemes.

The test set in our Phonebook setup is made up of 8 subsets, each containing 75 unique words. Isolated word recognition is performed on the test set by following the two protocols defined in [17].

- **75-lexicon task:** Separate pronunciation dictionaries, each consisting of 75 words, are used for each of the eight subsets. The reported WER is the average across all the eight subsets.
- **600-lexicon task:** A common pronunciation dictionary, consisting of 600 words, is used across all the eight subsets in the test sets.

In adaptation studies, the first MLP is trained using a subset of the 277.7 hours of CTS *ctstrain04* data set, which was used in the development of the AMI RT05 system [16]. The output of the MLP represents 45 phonemes corresponding to the UNISYN pronunciation dictionary [18].

B. Feature Extraction and Modeling

The first 13 PLP cepstral coefficients are used as acoustic features in all the experiments. Speaker specific mean and variance normalization is applied on the base features. Dynamic cepstral coefficients (delta and delta-delta) are appended to the base features to obtain a 39 dimensional feature vector for every 10 ms of speech. These features are applied at the input of the MLP with a temporal context of 90 ms.

A three-layered MLP with a sigmoid nonlinear activation function at the hidden layer, and a softmax activation function at the output layer is used throughout the studies. The parameters of the MLP are trained using the minimum cross entropy error criterion. The input features to the MLP are normalized to zero mean and unit variance, and these statistics are typically estimated on the training data. However, in adaptation studies, where an MLP trained on CTS is used on Phonebook, mean/variances are reestimated on the Phonebook features.

C. Decoding

The HMM/MLP hybrid approach is used for decoding. Each phoneme is modeled by a three-state, strictly left-to-right HMM, thereby enforcing a minimum duration of 30 ms. The (scaled) emission likelihood in each of the three states is the same, and is obtained by normalizing the estimated phonetic class conditional probabilities by the respective class priors. The Viterbi algorithm is applied with a simple loop-of-words language model.

V. Results and Analysis

Table I shows the word error rates (WER) obtained on the Phonebook test set using 75-lexicon and 600-lexicon decoding protocols. The baseline system consists of an MLP ($351 \times 600 \times 42$), trained on PLP features from 6.7 hours of Phonebook speech. In the adaptation system, posterior features for the Phonebook speech are first estimated by using an MLP ($351 \times 5000 \times 45$) trained on 232 hours of CTS. The estimated posterior features are used to train a second MLP classifier with a temporal context of 130 ms.

TABLE I
Word error rates (WER) in percentage on the Phonebook test set obtained using 75-lexicon and 600-lexicon test protocols.

task	baseline (%)	adaptation (%)	rel. drop (%)
75-lexicon	1.2	0.5	58.3
600-lexicon	4.0	1.8	55.0

It can be seen from the table that we obtain an absolute reduction of 0.7% and 2.2% respectively on the 75-lexicon and 600-lexicon test protocols. The significance of these results was confirmed by performing the McNemar test [19], which yielded a p-value < 0.0001 on both the tasks.

It is interesting to note that the adaptation system outperforms the single MLP based system, even though the latter is trained and tested in matched conditions. This is due to a combination of two factors. Firstly, the first MLP in the adaptation system is well trained on a large amount of data. Secondly, there is a inherent advantage in using the hierarchical system as shown in our earlier study [9]. Later in this paper, we show that hierarchical system trained on the same Phonebook data also yields improvement, but not as much as the adaptation system.

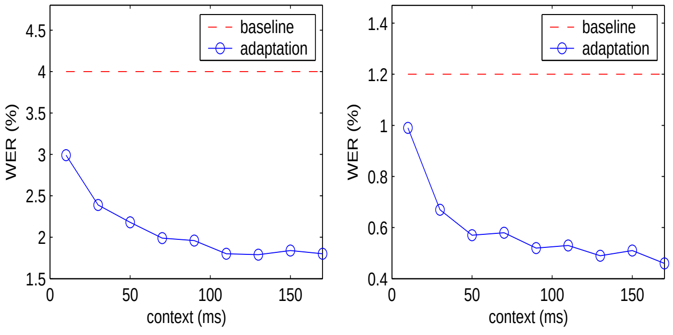
To the best of our knowledge, these are the lowest error rates to be reported on this particular Phonebook task. In the following sections, we study the performance of the system with respect to (a) the temporal context on the posterior features (b) goodness of the posterior features (c) complexity of the second MLP and (d) size of the adaptation data.

A. Role of Temporal Context

In Fig. 3, we plot the word error rate as a function of temporal context applied on the posterior features at the input of the second MLP. The posterior features are estimated by the MLP trained on 232 hours of CTS. The second MLP is trained using 6.7 hours of Phonebook speech. The horizontal dashed line indicates the word error rate obtained by the single MLP based baseline system, trained on 6.7 hours of Phonebook speech.

It can be seen that even without any temporal context, we obtain an absolute reduction of 1% in the error rate over the baseline system on the 600-lexicon task. In the hierarchical system without any temporal context, the second MLP can be viewed as a local mapping between the phonemes in the UNISYN dictionary to the Phonebook dictionary. The second MLP could be correcting any systematic perturbations in the

[1] 32% of the training data is silence.



(a) 600-lexicon

(b) 75-lexicon

Fig. 3. Word error rate as a function of the temporal context at the input of the second MLP. The first MLP is trained using 232 hours of CTS data. The results are for the (a) 75-lexicon task and (b) 600-lexicon task. The horizontal dashed line indicates the WER obtained from the baseline system trained on Phonebook.

estimated posterior probabilities due to the mismatch in the dictionaries. However, this aspect needs to be further analyzed.

As the temporal context is increased, the WER reduces further, and saturates for around 130-150 ms. With increase in temporal context, the second MLP classifier is also able to capture the contextual information in the posterior features. This trend was also observed in our previous studies, where both the MLPs in the hierarchical system were trained on the same task. It was shown that the phoneme error rate reduces with context, and saturates at around 150-230 ms of context [9]. A similar trend is observed in the case of lexicon-75 decoding as shown in Fig. 3 (b). In subsequent experiments, we fix the temporal context to 130 ms at the input of the second MLP classifier.

B. *Goodness of the Posterior Features*

The effectiveness of the proposed adaptation framework depends on the goodness of the posterior features estimated on the target task (here Phonebook). In Table II, we compare the posterior features estimated by two MLP classifiers, one trained on 6.7 hours of Phonebook speech (Phonebook-6.7 system), and the other trained on 232 hours of CTS speech (CTS-232 system). The comparison is done in two ways:

TABLE II
WORD ERROR RATES IN PERCENTAGE ON THE 600-LEXICON TEST PROTOCOL OBTAINED BY USING THE POSTERIOR FEATURES DIRECTLY (ROW-1) AND BY HIERARCHICAL SYSTEM (ROW-2).

system	Phonebook-6.7 (42 phns)	CTS-232 (45 phns)	CTS-232 (42 phns)
direct	4.0	-	5.5
hierarchy	3.3	1.8	1.9

1. Direct Decoding:
The estimated phonetic class conditional probabilities (posterior features) are used directly in the HMM/MLP hybrid decoding by using the Phonebook pronunciation dictionary, consisting of 42 phonemes. In this case, the training and testing conditions are matched, and we obtain a word error rate of 4.0%. In the case of the CTS-232 system, the estimated posterior probabilities of phonemes (conditioned on PLP features) have to be mapped to the Phonebook phoneme set. [2] As seen in the table, we obtain a word error rate of 5.5%.

As both Phonebook as well as CTS is acquired over the telephone channel, we presume the channel mismatch to be minimal. Nonetheless, the higher error rate in the case of CTS net is not surprising as there could be a mismatch in the estimated posterior probabilities, due to the differences in (a) the speaking style - conversational speech for training versus read speech in testing and (b) the pronunciation dictionaries used in training and test conditions.

2. Hierarchical Modeling:
A second MLP classifier is trained on the posterior features estimated for the Phonebook speech by using the Phonebook-6.7 and CTS-232 systems. In the latter case, the hierarchical system is trained using posterior features of dimension 45 and 42 (after mapping to the Phonebook phoneme set). In these experiments, the second MLP is trained using 6.7 hours of posterior features, with a temporal context of 130 ms. The estimated class conditional probabilities are used in HMM/MLP hybrid decoding.

It can be seen from Table II that

- The hierarchical system using CTS-232 posterior features outperforms the baseline system (direct, Phonebook-6.7) as well as the hierarchical system on the same Phonebook task (hierarchy, Phonebook-6.7). In contrast, the performance of the CTS-232 system by direct decoding is below these systems. This demonstrates the effectiveness of the proposed adaptation scheme.
- The hierarchical system on the same Phonebook task also results in the reduction of the WER by 0.7% with respect to the baseline system. This improvement in performance confirms previous studies [10][9][11] on the effectiveness of the hierarchical system in ASR.
- The three additional phonemes in the case of the CTS system did not contribute significantly (0.1%) to the reduction in the error rates.

TABLE III
WORD ERROR RATES ON THE 75-LEXICON TEST PROTOCOL OBTAINED BY USING THE POSTERIOR FEATURES DIRECTLY (ROW-1) AND BY HIERARCHICAL SYSTEM (ROW-2).

system	Phonebook-6.7 (42 phns)	CTS-232 (45 phns)	CTS-232 (42 phns)
direct	1.2	-	1.7
hierarchy	0.9	0.5	0.6

[2] The MLP trained on CTS has 45 output classes corresponding to the phonemes in the UNISYN pronunciation dictionary, whereas the Phonebook dictionary consists of 42 phonemes. There exists a one-to-one mapping between the phoneme sets, except for three phonemes /el/, /em/, and /en/. The probability mass corresponding to these phonemes are added to phonemes /l/, /m/, and /n/ respectively.

Table III shows similar trends in results on the 75-lexicon decoding protocol. In subsequent studies, results are reported only for the 600-lexicon test protocol.

C. Complexity of the second MLP

As discussed earlier in Section II, the posterior features are trained to achieve linear separability in the posterior feature space. The degree to which they actually achieve linearly separability depends on the complexity of the task. An important consequence of this property is that the ensuing classifier could be simpler in terms of model capacity.

In Fig. 4, we plot the word error rate on the 600-lexicon task as a function of the size of the hidden layer at the second stage of the hierarchical system. The size of the hidden layer controls the amount of nonlinearity that the MLP can model.

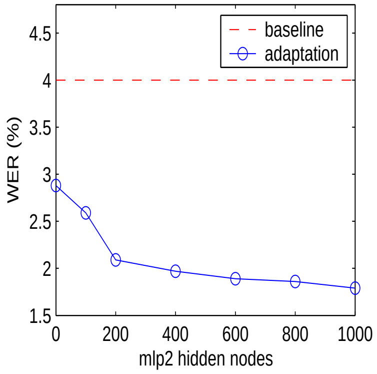


Fig. 4. The word error rate on the 600-lexicon task as a function of the size of the hidden layer of the MLP. The temporal context on the posterior features is fixed to 130 ms. The WER obtained by using a single layer perceptron is plotted as the number of hidden nodes equals zero.

It can be seen from the plot that the fall in the performance is minimal as the size of the hidden layer is reduced from 1000 to 200 units. As the size is reduced further, the performance drops more sharply. However, the adaptation system still outperforms the baseline system. As an extreme case, a single layer perceptron is used at the second stage of the hierarchy, and this is plotted as the number of hidden nodes equals zero in the figure. As seen in the figure, a simple linear classifier yields an absolute reduction of 1.1% in the error rate over the baseline single MLP system. This observation is consistent with our previous study [9], where lower phoneme error rates were obtained even when an SLP was used at the second stage of the hierarchy.

D. Amount of Adaptation Data

As discussed in Section II, the posterior features have lesser non-linguistic variabilities such as speaker and noise characteristics. As a consequence, the second MLP classifier can be trained using a lesser amount of data. In Fig. 5, we plot the word error rate obtained on the 600-lexicon task as a function of the amount of Phonebook data used for training. The hierarchical systems are trained with a temporal context of 130 ms. The plots in the figure correspond to the following four systems.

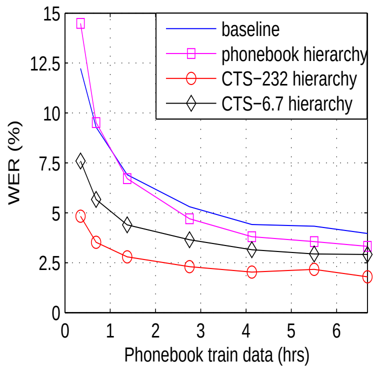


Fig. 5. The word error rate as a function of the amount of adaptation data (Phonebook) used. A temporal context of 130 ms is considered in the case of the hierarchical system.

Baseline system: Here, a conventional single MLP based system is trained on the PLP features obtained on the Phonebook speech. It can be seen that the performance of the system falls sharply with the reduction of training data. With 20 minutes of training data, we obtain a word error rate of 12%.

Phonebook hierarchy: An hierarchical system is trained with the posterior feature estimated by the baseline system. The second MLP is trained with the same Phonebook data that was used to train the baseline system. It can be seen that the hierarchical system yields lower error rates when compared to the baseline system. However, as the training data is further reduced, the hierarchical system ceases to show improvements over the baseline system.

CTS-232 hierarchy: In this adaptation system, the posterior features are estimated on the Phonebook task using an MLP trained on 232 hours of CTS. It can be seen that with just 30 minutes of adaptation, the system yields the same performance as the baseline system trained with 6.7 hours of task specific Phonebook speech.

CTS-6.7 hierarchy: In this adaptation system, the first MLP classifier is trained using 6.7 hours of CTS. In this case, we need 2 hours of adaptation data to obtain the same word error rate as that on the baseline system.

To briefly summarize, if the first MLP in the hierarchical system is trained using a larger amount of data, then smaller amount of adaptation data is sufficient. Furthermore, the difference between the word error rates obtained from CTS-232 system and CTS-6.7 system is larger when the adaptation data is limited, and this gap reduces with the increase in the amount of adaptation data.

VI. Discussion

The second MLP classifier can be viewed as a mapping of a trajectory in the posterior feature space corresponding to CTS phonemes to a point in the posterior feature space corresponding to the Phonebook phonemes. The following two factors contribute to the effectiveness of the hierarchical adaptation system.

- The second MLP can compensate for any systematic perturbations in the posterior features arising out of any mismatch in the training and test conditions for the first MLP classifier. This is evident from the fact that in Fig. 3, even without any temporal context, the hierarchical adaptation system yields lower word error rates.
- With increase in temporal context, the second MLP is able to learn the contextual information in the posterior features. This is reflected in the reduction of the word error rate with increase in the temporal context as shown in Fig. 3. This trend was also observed in our previous study [9], where both the MLPs in the hierarchical system were trained on the same task. Analysis of the trained parameters of the second MLP revealed that it learns the phonetic confusions at the output of the first MLP classifier, and to a certain extent the phonotactics of the language as observed in the training data. Future work will include such an analysis of the adaptation system.

The extent to which each of the above factors contributes to the overall decrease in the word error rate is difficult to ascertain as the second MLP jointly learns the phonetic-temporal patterns to minimize the cross entropy error criterion. However, this aspect can be better understood through carefully designed experiments. For example, the first MLP could be trained on CTS data, but with the phonetic transcription obtained by using the source pronunciation dictionaries used to derive the Phonebook pronunciation dictionary.

The main objective of this work was to investigate the feasibility of using the MLP based hierarchical system for task adaptation. This objective is clearly met as reflected in the experimental results. In fact, we obtain better results with the adaptation system in comparison with the baseline system. This is because of the ability of the second MLP to model the contextual information in the posterior features.

The present work also confirms that the conclusions drawn in our previous study (both MLPs trained on the same data) hold even in the case of task adaptation. To this end, we demonstrated that (a) the optimal temporal context for this particular task is around 130 ms (b) a simpler classifier is sufficient at the second stage of the hierarchy, and (c) the second MLP classifier can be trained using a lesser amount of adaptation data.

VII. Conclusions

We investigated an MLP based hierarchical approach for task adaptation in ASR. This adaptation scheme facilitates the reusability of well-trained MLP classifiers available off-the-shelf to new scenarios. The proposed hierarchical system can achieve adaptation as well as exploit the contextual information in the posterior features, leading to significant reduction in the word error rates. Experimental studies also revealed that the second stage of the hierarchical system can be simpler in terms of number of parameters, and can be trained using a lesser amount of training data.

Acknowledgment

This work was supported by the Swiss National Science Foundation (SNSF) under the Indo-Swiss joint research program on keyword spotting (KEYSPOT) and the Swiss National Center for Competence in Research (NCCR) under the Interactive Multimodal Information Management (IM2) project. More information on these projects can be found at http://indo-swiss.epfl.ch and http://www.im2.ch.

References

[1] N. Morgan *et al.*, "Pushing the Envelope - Aside," *IEEE Signal Process. Magazine*, vol. 22, no. 5, pp. 81–88, 2005.

[2] A. Stolcke *et al.*, "Recent Innovations in Speech-to-Text Transcription at SRI-ICSI-UW," *IEEE Trans. Audio. Speech. Language. Process.*, vol. 14, no. 5, pp. 1729–1744, 2006.

[3] Q. Zhu, B. Chen, N. Morgan, and A. Stolcke, "On Using MLP Features in LVCSR," *Proc. of Interspeech*, pp. 921–924, 2004.

[4] P. Fousek, L. Lamel, and J.-L. Gauvain, "Transcribing Broadcast Data using MLP Features," *Proceedings of Interspeech*, 2008.

[5] J. Park, F. Diehl, M. Gales, M. Tomalin, and P. Woodland, "Training and Adapting MLP Features for Arabic Speech Recognition," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, 2009.

[6] M. Richard and R. Lippmann, "Neural Network Classifiers Estimate Bayesian *a posteriori* Probabilities," *Neural Computation*, vol. 3, pp. 461–483, 1991.

[7] H. Bourlard and N. Morgan, *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[8] H. Hermansky, D. Ellis, and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1635–1638, 2000.

[9] J. Pinto, G. Sivaram, M. Magimai.-Doss, H. Hermansky, and H. Bourlard, "Analysis of MLP based Hierarchical Phoneme Posterior Probability Estimator," Idiap Research Institute, Tech. Rep., 2009.

[10] J. Pinto, B. Yegnanarayana, H. Hermansky, and M. Magimai.-Doss, "Exploiting Contextual Information for Improved Phoneme Recognition," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4449–4452, 2008.

[11] H. Ketabdar and H. Bourlard, "Hierarchical Integration of Phonetic and Lexical Knowledge in Phone Posterior Estimation," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 4065–4068, 2008.

[12] A. Stolcke *et al.*, "Further Progress in Meeting Recognition: The ICSI-SRI Spring 2005 Speech-to-text Evaluation System," *Second Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, pp. 463–475, 2005.

[13] X. Li and J. Bilmes, "Regularized Adaptation of Discriminative Classifiers," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, 2006.

[14] S. Ikbal, "Nonlinear Feature Transformations for Noise Robust Speech Recognition," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne (EPFL), Switzerland, 2004.

[15] J. Pitrelli *et al.*, "PhoneBook: A Phonetically-rich Isolated-word Telephone Speech Database," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 101–104, 1995.

[16] T. Hain *et al.*, "The Development of AMI System for Transcription of Speech in Meetings," in *Machine learning for Multimodal Interaction: 2nd International Workshop, Revised Selected Papers*, S. Renals and S. Bengio, Eds. Springer-Verlag, 2005, no. 3869, pp. 344–356.

[17] S. Dupont, H. Bourlard, O. Deroo, V. Fontaine, and J.-M. Boite, "Hybrid HMM/ANN Systems for Training Independent Tasks: Experiments on 'PhoneBook' and Related Improvements," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 1767–1770, 1997.

[18] S. Fitt, "Documentation and User Guide to UNISYN Lexicon and Post-lexical Rules," Center for Speech Technology Research, University of Edinburgh, Tech. Rep., 2000.

[19] L. Gillick and S. Cox, "Some Statistical Issues in the Comparison of Speech Recognition Algorithms," *Proc. of IEEE Conf. Acoust. Speech Signal Process. (ICASSP)*, pp. 532–535, 1989.