Transfer learning for tandem ASR feature extraction

Joe Frankel[1,2]*, Özgür Çetin*[2]*, Nelson Morgan*[2]

1. University of Edinburgh, 2. International Computer Science Institute

joe@cstr.ed.ac.uk

Abstract

Tandem automatic speech recognition (ASR), in which one or an ensemble of multi-layer perceptrons (MLPs) is used to provide a non-linear transform of the acoustic parameters, has become a standard technique in a number of state-of-the-art systems. In this paper, we examine the question of how to transfer learning from out-of-domain data to new tasks.

Experiments in the meetings domain show that adapting MLPs originally trained on conversational telephone speech leads to lower word error rates than training MLPs solely on the target data with a random initialization. Multi-task learning, in which a single MLP is trained to perform a secondary task (in this case a speech enhancement mapping from farfield to nearfield signals) is also shown to be advantageous.

Recognition experiments on broadcast news data suggest that structure learned from English speech can be adapted to Mandarin Chinese. The performance of tandem MLPs trained on 440 hours of Mandarin speech with a random initialization was achieved by adapted MLPs using only 70 hours of data in the target language.

1. Introduction

This work is concerned with the use of multi-layer perceptrons (MLPs) to provide non-linear transformations of acoustic features for use in automatic speech recognition (ASR). This approach is known as tandem ASR [1], and has become a common addition to modern ASR systems. For example, several of the meeting ASR systems presented at the NIST Rich Transcription 2006 spring evaluation (RT06s) included the use of non-linear feature transforms using MLPs.

The process of producing tandem approach is sketched in Figure 1. Multiple frames of acoustic parameters are fed into one or an ensemble of MLPs. Rather than interpreting the outputs as phone class posteriors as in hybrid artificial neural network (ANN)/HMM modelling, they are subjected to a log transformation and dimensionality reduction, then treated as observations. Once computed, they are appended to standard acoustic parameters in a hidden Markov model (HMM) system with Gaussian mixture model (GMM) observation distributions. The power of tandem ASR is two-fold. Firstly, multiple frames of acoustic features are used as input to the MLP, which introduces longer-span contextual information. Secondly, the non-linear mapping is trained against phone targets, which has the effect that separation between phone classes is maximized in the output space. This separation leads to improved discrimination by the GMM which describes the output space associated with each HMM state. In addition, tandem features have been shown to exhibit some cross-task and cross-language portability [2]. Recent work [3] has also shown that the tandem feature extraction is complimentary to other discriminative feature (e.g.,fMPE) and estimation (e.g., MPE) methods.

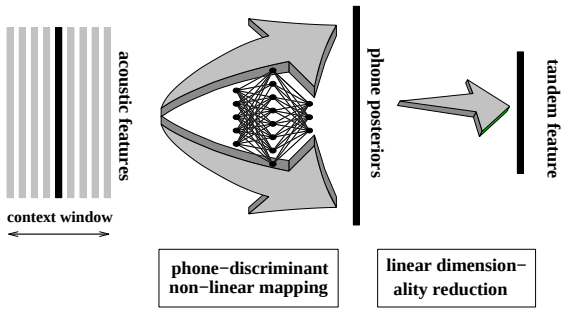

Figure 1: Schematic of tandem feature mappings.

Given the ever-increasing amounts of data on which systems are trained, and the considerable computational expense of training large MLPs, it is of interest to determine if learning can be transferred across domains and tasks. In this paper, we examine this question using experimental work within two scenarios. The first is recognition of farfield speech signals within the meetings domain. It was shown in [2] that adapting MLPs trained on out-of-domain data lead to better performance compared with unadapted MLPs, and that a slight improvement was given by Gaussian adaptation of HMMs onto the new features. In this paper we consider whether it is advantageous to adapt a previously-trained MLP or to retrain from scratch on the target data. Additionally, we present experiments on training MLPs using multi-task learning (MTL) [4], in which a single MLP is trained to perform a number of tasks. The second scenario we consider is the building a set of MLP-based features for a Mandarin broadcast news system.

2. Tandem ASR

Using MLP-based features in the tandem paradigm has been the subject of endeavour by a number of groups, including AMI [5], IDIAP [6] and the International Computer Science Institute at Berkeley (ICSI) [7]. The approach that is long-pursued at ICSI, and that taken in this work is to use a combination of the posteriors from two sets of MLPs.

The first of these is PLP-MLP, in which a 9 frame window of perceptual linear prediction (PLP) cepstra is used as input to a 3-layer MLP. The hidden units have sigmoid activation functions, and there is a softmax over the outputs.

The second is known as hidden activation temporal patterns (HATS) [8]. Under this configuration, separate small (e.g. 60 hidden unit) 3-layer MLPs are trained for phone classification with log critical band energies as input (one MLP per critical band with 50 frame windows). The hidden activations from each of these individual MLPs are then fed into a larger merger MLP, which is again trained on the task of phone classification.

The performance gains from using tandem and HATS feature independently are similar, though the additional information appears to be complementary, as the best results have been found using posterior combinations of the two [9].

Prior to the Fall 2004 DARPA speech-to-text (STT) evaluation, which focused on continuous telephone speech (CTS), significant quantities of data (close to 2000 hours) became available in the form of the Fisher corpus. Training of MLPs on this data is described by [9], a process which took approximately 6 weeks despite the use of code optimized to run on a multiple-core machine, and introducing a techniques specifically designed to reduce the training time. Training the MLPs on an order of magnitude more data than had been done previously led to improved performance. The focus of this paper is to explore methods by which the information encoded by these MLPs may be transferred to other domains.

3. Meetings domain

The meetings domain offers a particular set of challenges due to the nature of spontaneous multiparty speech in which speakers frequently overlap. Where the participants' speech is recorded using individual headset microphones (IHM), the high signal-to-noise ratio means that recognition has a word error rate (WER) of around 20%. However, it is not always possible or practical to have participants wearing individual microphones. In that case, table-top recording is required, which creates a new set of problems due to reduced signal-to-noise ratio and presence of effects such as reverberation. These lead to significantly higher word error rates, in the region 30-35%. The NIST RT06s meeting ASR evaluation specified three different farfield conditions:

- single distant microphone (SDM) - a single tabletop microphone source.
- multiple distant microphone (MDM) - tabletop microphone array with between 4 and 8 nodes.
- all distant microphones (ADM) - all channels used, which may include multiple microphone arrays.

In this work we use data from the MDM condition. The input waveforms were subject to delay-and-sum as described in [10].

All systems are gender-dependent, and employ many decoding stages including speaker adaptation, lattice generation, consensus decoding, n-best list rescoring, and cross-adaptation. For a full description, see [10].

3.1. Adaptation procedure

Since the targets against which the MLPs are trained are the English phone set as used to train the CTS MLPs, the adaptation procedure we adopt is to carry out a few epochs of further training from the CTS-trained nets.

CTS MLPs were trained on 8kHz data, and the original CTS front-end configurations were preserved when generating input features for the meeting data. Both tandem (3-layer 9-frame PLP input) and HATS (15 critical band MLPs with 51 frame input followed by merger on hidden activations) were adapted. For the HATS, only the merger MLP was adapted.

The time-aligned phone segmentation which is used to provide training targets was produced by segmenting against the nearfield signals, which have a higher signal-to-noise ratio, and therefore were assumed to produce a more accurate and consistent segmentation. Any regions of overlapped speech were removed, and targets were generated for the farfield signals by matching each frame against the nearfield frame closest in time.

Since the targets were generated using nearfield alignments, the nearfield cuttings can be considered as clean versions of the farfield data. Previous experiments showed that adapting MLPs to nearfield data improved farfield WER, so a single epoch of adaptation to nearfield data was carried out first. This was followed by 3 epochs of adaptation to the farfield or combined farfield and nearfield signals. The MLPs from the epoch which gave the highest cross-validation (CV) accuracy during training were used for experimentation.

For the farfield MLPs, a single channel was selected at random to provide the data for each segment, though input normalizations were calculated over all segments for any given speaker/channel combination. The starting learn rates were equal to those in the last epoch of training of the CTS MLPs.

3.2. Multi-task learning

The meeting data provides both nearfield and farfield signals, which gives the opportunity for a particular application of transfer learning known as multi-task learning (MTL) [4], in which a single MLP is trained to perform multiple related tasks. The rationale is that by using a shared representation, related tasks can act as a prompts for each other. Additionally, given that local minima of the error function are unlikely to fall at the same location for multiple tasks, the risk of over-training is reduced.

The idea of combining phone posterior estimation with speech enhancement was introduced in [11], and isolated digit classification experiments showed the benefits of this method. We evaluate this idea in the context of a state-of-the-art tandem system.

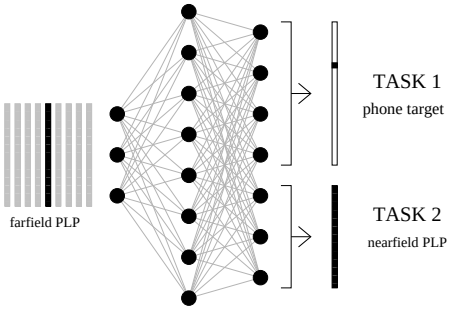

Figure 2: Multi-task MLP for tandem ASR. The MLP learns a speech enhancement mapping from farfield to nearfield PLP features in addition to the usual phone posterior estimation.

Figure 2 shows the multitask PLP-MLP configuration. Farfield PLPs provide the inputs to a fully-connected hidden layer. The output layer is divided into two regions. The first is a group of 46 units with a softmax activation function, trained to perform phone posterior estimation. The second is a set of 39 sigmoid units which map to the frame of nearfield PLPs matching that from the centre frame of the input window. The target PLPs are scaled to be in the range [0.0, 1.0] in order to match the output range of sigmoid units. In the case of HATS, it is the merger MLP which is trained to be multi-task.

3.3. Experiments

Once the MLPs were adapted, features were generated to match the training data, and HMM Gaussian adaptation was performed. Unlike for MLP training, the data from all channels

for each farfield segment was used as adaptation material. Once adapted, features were generated to correspond to the test data, and a multi-pass decoding pass was initiated, as described in [10]. The MLP features are used in the first stage of decoding, and are appended to Mel frequency cepstral coefficients.

The results presented below are WERs for the NIST RT05s evaluation data.

	MLP train data	word error rate (%)	
		one-pass	multi-pass
no MLP	N/A	50.2	36.7
random initialization	farfield	44.0	35.2
	farfield MTL	44.1	34.3
adapted from CTS	nearfield	52.7	41.2
	farfield	43.3	33.2
	nearfield+farfield	42.7	33.0

Table 1: *Results (WER) on the NIST RT05s MDM evaluation data for a number of different adapted MLPs.*

The results are shown in table 1. The WER is shown for each MLP type after the first stage of decoding (one-pass), and at completion of the final multi-pass sweep. A baseline system in which no MLP features were used gives WERs of 50.2% and 36.7% for one-pass and multi-pass decoding respectively. Using the features from randomly-initialized MLPs, trained only on the farfield data, there is a substantial reduction in the WER after the first pass, from 50.2% to 44.0%, and at completion of all passes, the gain is reduced, though still evident with the WER reducing from 36.7% to 35.2%. For the MLPs which were trained in a multi-task setting, there is no benefit evident for the one-pass decoding, though at completion, the WER is reduced to 34.3%. These results suggest that guiding learning with the addition of a speech enhancement task is of benefit.

The next set of results correspond to MLPs which have been adapted from the CTS MLPs which were originally trained on 2000 hours of speech. Using the MLPs which were adapted on nearfield signals leads to increases in WER compared with a system with no MLP features. This can be attributed to mismatch of the nearfield and farfield signals. The MLPs adapted on only farfield data lead to WERs of 43.3% and 33.2% after one and multiple decoding passes respectively. These results show substantial improvement over the baseline non-MLP system, and further reductions over the MTL-trained MLPs. Finally, adapting MLPs on pooled nearfield and farfield data gives the best results overall, with WERs of 42.7% and 33.0% for one-pass and multi-pass decoding.

One possible explanation for the superior performance of adapted MLPs is the size of the MLPs. Each of the CTS MLPs has 8 million parameters, which translates to 20,800 hidden units for each of the male and female tandem nets. By contrast, the male and female tandem nets which were trained from scratch had 7125 and 1825 hidden units respectively. These were determined by setting the total number of free parameters to equal 15% of the number of training frames.

4. Mandarin broadcast news

In this section, we consider the problem of deriving tandem MLPs for use in a Mandarin Chinese system. Approximately 440 hours of Chinese broadcast news (BN) data was available, 70 of which had careful transcripts, and the remainder transcripts derived from a forced alignment of closed captions. Framewise labels were in terms of a set of 71 tonemes. In addition to training tandem and HATS MLPs from random initialization for the Mandarin data, the English CTS MLPs were used as a starting point for cross-lingual adaptation. Two strategies were explored: in the first (3-layer), the English CTS input-hidden layer was used in conjunction with a randomly initialized hidden-output layer. In the second (4-layer), an extra layer was added to the CTS MLPs. In both cases, training then proceeded with all weights being updated at each epoch. For HATS, it was the merger MLP which was adapted, and the critical band MLPs used unchanged.

initialization	Cross-validation accuracy	
	tandem	HATS
random 5% free params	74.1%	75.5%
random 10% free params	74.8%	76.2%
adapted, 3-layer	76.8%	77.2%
adapted, 4-layer	75.5%	76.5%

Table 2: *Cross validation accuracies on Mandarin broadcast news data for tandem and HATS MLPs both with random initialization, and adapting from English CTS. Accuracies are shown for randomly-initialized MLPs with the number of free parameters set to be 5% and 10% of the total training frames.*

Table 2 shows cross-validation (CV) accuracies for the Mandarin MLPs both with random initialization, and for the 3-layer and 4-layer MLPs which are initialized from English CTS MLPs. In all cases the training data is the 70-hour subset. For the randomly-initialized MLPs, accuracies are shown for the number of free parameters set to be 5% and 10% of the total training frames. We find that for both tandem and HATS MLPs, it is the adapted versions which give higher CV accuracy. Additionally, it was found that these MLPs converged much more rapidly, achieving close to highest CV accuracy within the first epoch of training.

Despite the higher CV accuracy of the 10% version, it is the 5% MLPs which lead to the lowest WERs, and are used for the recognition results presented below in Table 3.

	word error rate (%)	
	dev-04	eval-04
no MLP	9.5	19.5
Chinese (70 hours)	8.2	18.2
Chinese (440 hours)	8.0	17.9
adapted (70 hours)	7.7	18.0

Table 3: *Results (WER) on the dev-04 and eval-04 data sets.*

Table 3 gives recognition WERs for the 2004 development and evaluation sets for the GALE Mandarin Broadcast News recognition task. Performance for the system with no MLPs is 9.5% and 19.5% WER for the dev and eval sets respectively. The results using MLPs trained on 70 and 440 hours of data from a random initialization are given, with those trained on 440 hours giving a slightly lower error rate of 8.0% and 17.9% on the dev and eval sets respectively. The adapted MLPs give the lowest error of any on the development set, and all results are very similar on the evaluation data.

5. Discussion

Mismatch of training and testing data frequently has a significant impact on the performance of ASR systems. However, when porting a system from one domain to another, it is usual to take advantage of previously trained models and adapt the Gaussian mixture models (GMMs) associated with each state to the new domain. In this work we have shown that similarly, it is advantageous to utilize a previously-trained MLP and adapt to the new domain. For example, at completion of all decoding passes, the WER using MLPs trained from a random initialization was 35.2%, compared to 33.2% when adapting from CTS. In addition, the MTL MLPs gave a lower word error rate, from 35.2% to 34.3%. This suggests that strategies which combine adaption with multi-task learning may prove useful. For example, extra output and (possibly hidden) units could be added to the CTS MLPs to provide a speech enhancement mapping prior to training on the target data.

The results were less conclusive on the cross-lingual adaptation to Mandarin Chinese, though there is some evidence that given 70 hours of target data, it is preferable to adapt from English rather than train MLPs with random initialization. This suggests that there may be potential for sharing hidden representations between languages in order to increase available data.

Additionally, for the MLPs trained on the Mandarin data from a random initialization, it was found that smaller MLPs (free parameters set to be 5% rather than 10% of the total training frames) gave the best performance. This results in an MLP with a hidden layer of 4232 units, compared with 20800 in the adapted version. In this type of adaptation, we are considering the input-hidden layer as a general speech pattern classifier, and the hidden-output as a mapping to the particular set of outputs. Initializing with a ready-trained input-hidden layer makes it possible to train many more free parameters.

The results presented in this paper show that whilst mismatch in training and testing domains leads to performance degradations (e.g. nearfield MLPs on farfield data), there are sufficient commonalities to find a benefit from transfer learning.

6. References

[1] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature stream extraction for conventional hmm systems," in *Proc ICASSP*, vol. III, Istanbul, Turkey, 2000, pp. 1635–1638.

[2] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, Toulouse, France, May 2006.

[3] J. Zheng, O. Çetin, M.-Y. Hwang, X. Lei, A. Stolcke, and N. Morgan, "Combining discriminative feature, transform, and model training for large vocabulary speech recognition." in *Proc. ICASSP*, 2007.

[4] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997. [Online]. Available: citeseer.ist.psu.edu/caruana97multitask.html

[5] T. Hain, L. Burget, J. Dines, G. Garau, M. Karafiat, M. Lincoln, J. Vepa, and V. Wan, "The AMI meeting transcription system : Progress and performance," in *NIST RT'06 Workshop*, 2006.

[6] H. Hermansky, "TRAP-TANDEM: Data-driven extraction of temporal features from speech," IDIAP, Martigny, Switzerland, IDIAP-RR 50, 2003.

[7] N. Morgan, Q. Zhu, A. Stolcke, K. Sonmez, S. Sivadas, T. Shinozaki, M. Ostendorf, J. Jain, H. Hermansky, D. Ellis, G. Doddington, B. Chen, O. cCetin, H. Bourlard, and M. Athineos, "Pushing the Envelope - Aside," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 81–88, September 2005.

[8] B. Chen, Q. Zhu, and N. Morgan, "Learning long term temporal feature in LVCSR using neural networks." in *Proc. ICSLP*, 2004, pp. 612–615.

[9] Q. Zhu, A. Stolcke, B. Y. Chen, and N. Morgan, "Using MLP features in SRI's conversational speech recognition system," in *Proc. Eurospeech*, Lisbon, Portugal, 2005.

[10] A. Janin, A. Stolcke, X. Anguera, K. Boakye, O. cCetin, J. Frankel, and J. Zheng, "The ICSI-SRI spring 2006 meeting recognition system," in *Proc. MLMI*, Washington DC, USA, 2006.

[11] S. Parveen and P. D. Green, "Multitask learning in connectionist asr using recurrent neural networks." in *Proc. Eurospeech*, Geneva, Switzerland, 2003.