TANDEM SYSTEM ADAPTATION USING MULTIPLE LINEAR FEATURE TRANSFORMS

Y.-Q. Wang and M. J. F. Gales

Engineering Department, Cambridge University
Trumpington St. Cambridge, CB2 1PZ, U.K.
{yw293, mjfg}@eng.cam.ac.uk

ABSTRACT

Adaptation to speaker and environment changes is an essential part of current automatic speech recognition (ASR) systems. In recent years the use of multi-layer percpetrons (MLPs) has become increasingly common in ASR systems. A standard approach to handling speaker differences when using MLPs is to apply a global speaker-specific constrained MLLR (CMLLR) transform to the features prior to training or using the MLP. This paper considers the situation when there are both speaker and channel, communication link, differences in the data. A more powerful transform, front-end CMLLR (FE-CMLLR), is applied to the inputs to the MLP to represent the channel differences. Though global, these FE-CMLLR transforms vary from time-instance to time-instance. Experiments on a channel distorted dialect Arabic conversational speech recognition task indicates the usefulness of adapting MLP features using both CMLLR and FE-CMLLR transforms.

Index Terms— MLP feature, acoustic model adaptation

1. INTRODUCTION

In recent years, the use of multi-layer perceptions (MLPs) for automatic speech recognition (ASR) has received considerable research interests [1, 2, 3]. An MLP usually takes several frames of short-term spectral-based feature vector (e.g., MFCC or PLP) as input to predict the center phone (or phone-state) identity. There are two broad approaches to using MLPs in the ASR systems. The first one, proposed in early 90's, replaces the Gaussian mixture model (GMM)-based emission probabilities by the class posterior probabilities estimated by MLPs [4]. This approach, usually referred to as hybrid artificial neural network-hidden Markov model (ANN-HMM), has recently become popular, since it is found training MLPs using context-dependent tied triphone state as target with more than 3 layers is able to deliver extremely good performance [5]. The other approach, usually referred to as probabilistic TANDEM approach, was first proposed in [6], combines the posterior obtained by MLPs with MFCC or PLP to form the TANDEM feature, which is modelled by the conventional GMM-based systems. An alternative form was proposed in [7], where a bottleneck layer is introduced in which neural nets are constrained to have a very narrow hidden layer, the bottleneck layer, in the middle and the linear output of that layer is taken as output instead of posteriors. The advantage of TANDEM approach is that almost all the techniques developed for the GMM-based system can be equally applied, e.g., adaptation, adaptive training and discriminative training. This work, considering using the existing adaptation techniques in GMM-based systems for MLPs, thus sits in this TANDEM framework.

It is well understood that speech signals are highly affected by various factors. Thus the ability to adapt ASR systems to new operating conditions, unseen in the training data, is important. One approach to adapting MLPs is to augment the neural nets with a linear transformation network connected to the input, e.g., the Linear Input Network (LIN) adaptation [8]. The transform matrix is then estimated by minimising the cross entropy between the supervision hypothesis and the model prediction. Due to this discriminative criterion, the estimation is sensitive to the error in supervision hypothesis. An alternative approach is to apply a global constrained maximum likelihood linear regression [9] (CMLLR) transform to the features prior to training or using the MLP (e.g., [10, 11]). This removes the need to estimate parameters discriminatively [12]. However the use of a single global transform, as the transform must be used for all classes, limits the ability to model the complexity of environment and channel distortions.

This work focuses extending adaptation approaches for MLPs by leveraging the existing adaptation techniques already developed for GMM-based systems. In particular, this work considers designing ASR systems to recognise speech transmitted though different communication channels (links). In the previous work [13], a front-end CMLLR (FE-CMLLR) [14] technique was used to normalise the impact of the communication channel while CMLLR was used to normalise the speaker. In this work, these schemes are also applied to nor-

This work was partially supported by Google research award and DARPA under RATS program. The paper does not necessarily reflect the position or the policy of US Government and no official endorsement should be inferred.

malise the input of a MLP with a bottleneck topology. FE-CMLLR is suitable for this task as, though the transform is applied globally, it varies from time-instance to time-instance. Effectively it yields a non-liner transform in the model space.

The rest of this paper is organised as follows. Section 2 briefly reviews the CMLLR and FE-CMLLR techniques developed for the GMM-based system. Section 3 discusses options to adapt TANDEM systems. Experiment and results are discussed in section 4 with the conclusions in section 5.

2. GMM-BASED SYSTEMS ADAPTATION

A popular choice of adapting GMM-based systems is to use linear transform-based schemes, for example MLLR and CMLLR. One of the advantages of CMLLR is that it can be viewed as a transform acting on feature when a global class is used [9]. When this form of CMLLR is used for speaker adaptation, each speaker s is associated with one line transform $\mathbf{W}^{(s)} = [\mathbf{A}^{(s)}, \boldsymbol{b}^{(s)}]$ and the distribution of each component m for speaker s is:

$$p(\boldsymbol{y}_t^{(s)}|\mathcal{M}_{\mathtt{x}}, s, m) = |\mathbf{A}^{(s)}|\mathcal{N}(\mathbf{A}^{(s)}\boldsymbol{y}_t^{(s)} + \boldsymbol{b}^{(s)}; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}) \quad (1)$$

where $\boldsymbol{y}_t^{(s)}$ is the speech feature produced by speaker s, $\mathcal{M}_{\mathtt{x}} = \{\boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}\}$ the canonical model parameters. As CMLLR in this form is acting on the feature, it is very efficient to use in speaker adaptive training (SAT). The estimation of the canonical model $\mathcal{M}_{\mathtt{x}}$ and speaker transform $\mathbf{W}^{(s)}$ is done via iteratively maximising the likelihood function using EM. An efficient iterative row-by-row maximising method is used to estimate the transform [9].

To model complex acoustic conditions multiple linear transforms can be used. In [13], an alternative feature transformation, FE-CMLLR, is used to model the distortion of communication channel, in conjunction with the CMLLR used for modelling the speaker differences. For a communication channel (link) distorted speech vector $\boldsymbol{y}_t$, a FE-CMLLR is applied to yield $\hat{\boldsymbol{x}}_t$:

$$\hat{\boldsymbol{x}}_t = \sum_{c=1}^{C} \gamma_{\mathtt{c}t}^{(c)} (\mathbf{A}_{\mathtt{c}}^{(c)} \boldsymbol{y}_t + \boldsymbol{b}_{\mathtt{c}}^{(c)}) = \mathbf{A}_{\mathtt{c}t}\boldsymbol{y}_t + \boldsymbol{b}_{\mathtt{c}t} \quad (2)$$

where $\gamma_{\mathtt{c}t}^{(c)}$ is obtained from a front-end GMM and

$$\mathbf{A}_{\mathtt{c}t} = \sum_{c=1}^{C} \gamma_{\mathtt{c}t}^{(c)} \mathbf{A}_{\mathtt{c}}^{(c)} \; ; \quad \boldsymbol{b}_{\mathtt{c}t} = \sum_{c=1}^{C} \gamma_{\mathtt{c}t}^{(c)} \boldsymbol{b}_{\mathtt{c}}^{(c)} \quad (3)$$

Estimating the FE-CMLLR transform, $\mathcal{M}_c = \{\mathbf{A}_{\mathtt{c}}^{(c)}, \boldsymbol{b}_{\mathtt{c}}^{(c)}\}$, and the canonical model $\mathcal{M}_{\mathtt{x}}$ is also done via maximising the likelihood function using EM. An approximated method was used in [13] to perform the optimisation. It is obvious that FE-CMLLR, similar as global CMLLR, operates in the feature space. Different from global CMLLR, using multiple-component transforms allows FE-CMLLR to model complex distortions such as communication channels. FE-CMLLR also enables a consistent space for speaker adaptation [13].

To simultaneously model the effect of speaker and link, it is possible to combine FE-CMLLR with CMLLR. In [13], speaker transform $\mathbf{W}^{(s)}$ was applied in a link space defined by FE-CMLLR:

$$p(\boldsymbol{y}_t^{(s)}; s, m, \mathcal{M}_{\mathtt{x}}, \mathcal{M}_{\mathtt{c}}) = |\mathbf{A}^{(s)}||\mathbf{A}_{\mathtt{c}t}| \times \mathcal{N}(\mathbf{A}^{(s)}(\mathbf{A}_{\mathtt{c}t}\boldsymbol{y}_t^{(s)} + \boldsymbol{b}_{\mathtt{c}t}) + \boldsymbol{b}^{(s)}; \boldsymbol{\mu}_{\mathtt{x}}^{(m)}, \boldsymbol{\Sigma}_{\mathtt{x}}^{(m)}) \quad (4)$$

By transforming the feature using the speaker transform (CMLLR) and the link transform (FE-CMLLR), a link and speaker adaptive trained (LSAT) model can be built in the normalised feature space. The upper branch in Figure 1 shows the speaker and link adaptation of PLP feature.

3. TANDEM SYSTEMS ADAPTATION

In the TANDEM system, MLP is used for feature extraction. Short-time spectral-based features (in this work, 13-dimensional PLP) with dynamic features and context frames are fed into MLP. Linear output of the bottleneck layer is decorrelated by principle component analysis (PCA) or SEMIT transform [15] and concatenated with PLP to form the TANDEM PLP+MLP feature. To allow simple concatenation, both PLP and MLP-based features are extracted using the same frame rate. The lower branch in Figure 1 shows the architecture for generating TANDEM feature.

There are two possible approaches to adapting a TANDEM system. First, the MLP input can be transformed to a normalised space. Usually the same linear transform is used for each context frame to reduce the number of adaptable parameters. The transform can be estimated by minimising a frame-level cross-entropy based criterion, as in LIN adaptation [8]. Alternatively, the feature transforms estimated in the GMM-based systems can be borrowed. This is shown in Figure 1 when the switch is in position 2. Second, as the TANDEM feature is again modeled by GMMs, the same adaptation techniques, such as CMLLR and FE-CMLLR, can be used. The dashed box in the lower branch in Figure 1 illustrates this.

It is interesting to compare these two forms of TANDEM system adaptation. Transforms which directly adapt the TANDEM feature are estimated by maximising the likelihood of TANDEM acoustic model. Using the transforms from the PLP system to modify the MLP input feature, which indirectly adapts the TANDEM feature, does not guarantee the increase in likelihood of TANDEM acoustic model or frame accuracy of MLPs. On the other hand, linearly transforming the MLP input yields a nonlinear transform of TANDEM feature, while directly adapts the TANDEM feature using CMLLR or FE-CMLLR is a linear (or piecewise linear) transform of the TANDEM feature. Given the differences between two approaches, it may be useful to combine them. In addition, these forms of feature transformation can be further combined with model-based adaptation, e.g., MLLR.

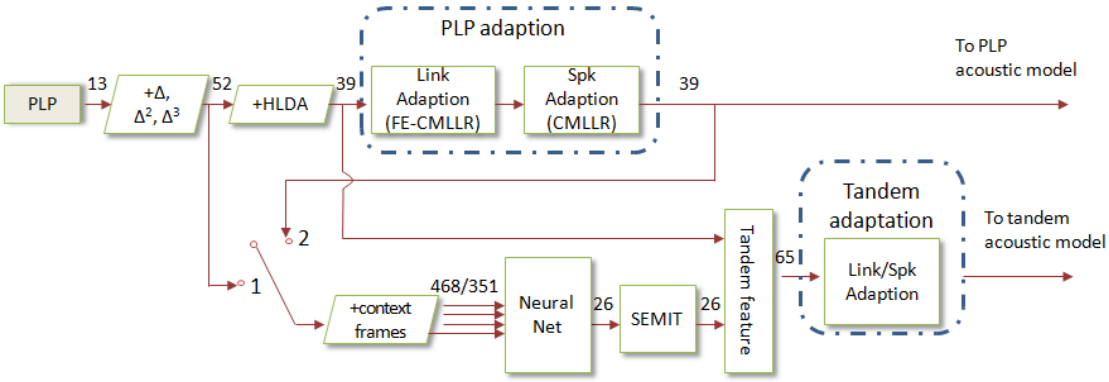

Fig. 1. Flowchart of PLP and TANDEM system adaptation.

4. EXPERIMENTS

Experiments were carried out on the training and test data provided from Robust Automatic Transcription (RATS) program for Arabic keyword spotting. The data was collected by retransmitting Levantine Arabic conversational telephone speech data over eight communication channels (links) which are labelled as A to H. A wide range of distortion are associated with these links. The training data include data from all eight channels plus the original clean speech. Part of the retransmitted data was held-out to form a test set, dev1. For each of the channels there was 2 to 2.5 hours test data, depending on how much of the retransmitting speech passed quality assurance tests. The clean Levantine Arabic transcriptions (excluding the dev1 test data), approximately 1.6 million words, were used to train a trigram language model.

The acoustic data was parameterised using 13-dimensional PLP, including C0. Delta, delta-deltas and triples were appended followed by an HLDA projection from 52 dimensions to 39. Speaker (side) based cepstral mean normalisation was applied. Word-based graphemic systems, incorporating word boundary information were build. Cross-word decision-tree state-clustered triphone models were then trained using MPE criterion. There are about 3K distinct states with an average of 36 components per state. In addition to the above speaker (and link) independent (SI) system, SAT system was build using global and full CMLLR transforms at the speaker level. For link representation, a 128-component FE-CMLLR was used for each link. A single front-end GMM was used for all links. SAT was also built in the FE-CMLLR normalised space, yielding the LSAT system.

TANDEM SI, SAT and LSAT systems were built using the "fast" system build method detailed in [3]. The TANDEM features for this work were 26-dimensional with decorrelating transform constructed in the same fashion as in [3]. Initially 52-dimensional PLPs (static, delta, delta-deltas and triples) were used for each frame. If MLP input adaptation is switched on, the HLDA estimated in PLP system was used to project the 52-dimension PLPs down to 39 dimensions, followed by CMLLR and/or FE-CMLLR transforms. 9 context frames were used. The inputs to MLP were also mean and variance normalised on the side level. In the initial investigation, a 4-layer MLP topology was used in which the first hidden layer has 3500 hidden nodes while the second layer, bottleneck layer, has 26 nodes. The neural net was trained using back-propagation in mini-batch (800 frames) mode. Ten percents of the training data (randomly chosen at side level) was used as the cross validation set.

For the SI systems, a one-pass unadapted decoding was performed using the trigram language model. For SAT and LSAT, the PLP SI system was first used to generate the supervision hypothesis, which was then used to estimate the speaker transforms. In this work, a CMLLR and a MLLR mean transform were used, both were global and full transforms. After the speaker adaptation, a second pass decoding was performed using the adaptively trained models (SAT or LSAT). During the test, it was assumed the segmentation and link identity of each utterance were known. Three representative links in terms of distortions were given in the first Table: link A (high), C (medium); and G (low). All results are based on confusion network (CN) decoding.

Initially, PLP acoustic models were built. As this task is known to be very challenging, the overall performances are as expected, quite poor. Decoding using a PLP SI model gave a WER (averaging on all links) of 68.4%. Using SAT/LSAT for PLP-based systems yields average WERs of 63.9% and 63.3% respectively, while most of the gains coming from the high distortion link such as link A. The initial TANDEM systems were build without MLP input adaptation. Using TANDEM feature alone yield considerable gains over SI PLP systems (64.2% vs 68.4%), while adaptively trained SAT/LSAT TANDEM systems gave further gains, as shown in the first rows of block 2 and block 3 in table 1. Preliminary investigation on MLP input adaptation showed transforming PLPs using only HLDA does not give any significant gains, which correlates to the findings in [12] for a deep neural net used

# layers	TANDEM Systems	MLP input adapt. Speaker	MLP input adapt. Link	Link A	B	C	D	E	F	G	H	Avg
4	SI	–	–	71.9	73.6	67.9	62.8	76.0	65.7	57.4	71.8	68.0
	SAT	✓	–	70.4	70.5	63.7	58.0	73.2	63.4	53.8	69.2	64.9
	LSAT	✓	✓	70.2	70.7	63.6	57.2	72.3	62.5	52.9	67.9	64.2
5	SAT	–	–	68.8	69.6	63.4	57.2	71.7	61.1	53.0	67.4	63.6
	SAT	✓	✓	69.2	69.4	62.4	55.7	71.1	61.4	51.7	67.4	63.1
	LSAT	✓	✓	69.3	69.5	62.3	55.4	70.8	61.3	51.4	67.1	62.9
7	SAT	–	–	68.8	69.0	63.4	57.1	71.6	61.3	53.0	67.5	63.4
		✓	✓	69.4	69.4	61.5	55.2	71.4	61.4	51.7	67.2	62.9

Table 2. Performance contrast by adapting TANDEM system to speaker and/or link using different number of layers of bottleneck neural nets.

in the hybrid architecture. As the supervision hypothesis had such a high WER, it is suspected LIN adaptation, estimated by the discriminative criterion, will not give any gains either. However, MLP input adaptation using CMLLR and FE-CMLLR does give gains, as shown in table 1. The first block of table 1 shows the contrast on SI systems. Compared with the performance of the TANDEM SI system without input adaptation, 0.7% absolute gains can be achieved by link adaptation, while about 1.5% gains can be obtained by adapting the MLP input to speaker or speaker/link. Note that adapting the MLP input to link does not require a supervision, therefore can be used in the initial decoding. The second and third block of table 1 show the gains by combining MLP input adaptation with adaptively trained TANDEM models. On the most advanced systems (LSAT), using CMLLR and FE-CMLLR as MLP input feature normalisation, there is about 0.6% performance gains (61.2% vs. 60.6%). The first block of table 2 shows the overall adaptation gains on all links by using speaker or speaker/link information. In total 3.1%-3.8% gains can be obtained on this difficult task. This shows adaptation of TANDEM systems is helpful.

Systems	MLP input adaptation Speaker	MLP input adaptation Link	Link A	C	G	Avg
SI	–	–	71.9	67.9	57.4	64.2
	CMLLR	–	71.3	65.6	56.0	62.8
	–	FE-CMLLR	71.1	66.4	57.2	63.5
	CMLLR	FE-CMLLR	71.1	65.6	56.0	62.7
SAT	–	–	70.4	64.9	54.7	61.8
	CMLLR	–	70.4	63.7	53.8	61.1
	–	FE-CMLLR	70.8	63.5	53.5	61.0
	CMLLR	FE-CMLLR	70.4	63.4	52.9	60.6
LSAT	–	–	70.4	64.1	53.8	61.2
	CMLLR	–	70.6	63.6	53.4	60.9
	–	FE-CMLLR	70.9	63.2	53.1	60.7
	CMLLR	FE-CMLLR	70.2	63.6	52.9	60.6

Table 1. MLP input speaker and/or link adaptation. The bottleneck neural net had 4 layers.

Finally, the effectiveness of these MLP adaptation techniques were examined on two more complex neural nets: a 5-layer and a 7-layer bottleneck neural net. The 5-layer neural net had 2 hidden layer each with 2K nodes, while the 7-layer neural net had 4 hidden layer each with 1K nodes. Other layers were kept the same. To get the best perform, the 7-layer neural net were discriminatively pre-trained and then fine-tuned as in [12]. Adding additional hidden layers yields gains, as shown in the second and third blocks of table 2. The average WERs of TANDEM SAT system was 63.1% and 63.4% for the 5-layer and 7-layer neural nets respectively, which compared to 64.9% WER of the system using 4-layer neural nets. On the other hand, it seems that linearly transforming the MLP input is still helpful when combining with SAT TANDEM system, achieving 0.5% gains. Future work will exam the trend of these gains when deeper MLPs and context-dependent targets are used.

5. CONCLUSIONS

This paper has discussed approaches to TANDEM system adaptation in degraded communication channels. Multiple linear transforms were constructed to normalise the MLP input: a global CMLLR was used to normalise the speaker differences, and a more powerful FE-CMLLR was employed for channel difference normalisation. Different from the global CMLLR, which only allows a single transform for each speaker, FE-CMLLR varies from time-instance to time-instance. This gives FE-CMLLR a flexibility to normalise more complicated, channel, distortions. By combing these linear transforms, distortions caused by multiple acoustic factors (speaker and channel differences in this work) can be better normalised. These transforms were estimated in the GMM-based system using maximum likelihood criterion. Although used in a different system, they are shown to be useful. MLP input adaptation is also combined with adaptive trained TANDEM models. Experiments on the channel distorted dialect Arabic conversational speech recognition task demonstrated the benefits of TANDEM system adaptation using multiple linear transforms.

6. REFERENCES

[1] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sinath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 2, no. November, pp. 82–97, 2012.

[2] Z. Tüske, M. Sundermeyer, R. Schlüter, and H. Ney, "Context-dependent MLPs for LVCSR: TANDEM, hybrid or both?," in *Interspeech*, 2012.

[3] J. Park, F. Diehl, M. J. F. Gales, M. Tomalin, and P. C. Woodland, "The efficient incorporation of MLP features into automatic speech recognition systems," *Computer Speech & Language*, vol. 25, no. 3, pp. 519–534, July 2011.

[4] N. Morgan and H.A. Bourlard, "Neural networks for statistical recognition of continuous speech," *Proceedings of the IEEE*, vol. 83, no. 5, 1995.

[5] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Interspeech*, 2011, pp. 437–440.

[6] H. Hermansky, D. P. W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *ICASSP*, 2000.

[7] F. Grézl, M. Karafiát, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *ICASSP*, 2007.

[8] J. P. Neto, C. Martins, and L. B. Almeida, "Speaker-adaptation in a hybrid HMM-MLP recognizer," in *ICASSP*, 1996, pp. 3382–3385.

[9] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, Apr. 1998.

[10] A.-R. Mohamed, T. N. Sainath, G. Dahl, B. Ramabhadran, G. E. Hinton, and M. A. Picheny, "Deep belief networks using discriminative features for phone recognition," in *Proc. ICASSP*, 2011, pp. 5060–5063.

[11] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in *Proc. ICASSP*, 2012.

[12] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in Context-Dependent Deep Neural Networks for conversational speech transcription," in *ASRU-2011*. Dec. 2011, pp. 24–29, Ieee.

[13] M. J. F. Gales and F. Flego, "Model-based approaches for degraded channel modelling in Robust ASR," in *Interspeech*, 2012.

[14] H. Liao and M. J. F. Gales, "Joint uncertainty decoding for noise robust speech recognition," in *Interspeech*, 2005.

[15] M. J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *IEEE Transaction on speech and audio processing*, vol. 7, no. 3, 1999.