Improving LVCSR System Combination Using Neural Network Language Model Cross Adaptation

X. Liu, M. J. F. Gales & P. C. Woodland

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {*xl207,mjfg,pcw*}*@eng.cam.ac.uk*

Abstract

State-of-the-art large vocabulary continuous speech recognition (LVCSR) systems often combine outputs from multiple sub-systems developed at different sites. Cross system adaptation can be used as an alternative to direct hypothesis level combination schemes such as ROVER. The standard approach involves only cross adapting acoustic models. To fully exploit the complimentary features among sub-systems, language model (LM) cross adaptation techniques can be used. Previous research on multi-level n-gram LM cross adaptation is extended to further include the cross adaptation of neural network LMs in this paper. Using this improved LM cross adaptation framework, significant error rate gains of 4.0%-7.1% relative were obtained over acoustic model only cross adaptation when combining a range of Chinese LVCSR sub-systems used in the 2010 and 2011 DARPA GALE evaluations.

1. Introduction

State-of-the-art large vocabulary continuous speech recognition (LVCSR) systems often use system combination techniques. The diversity and complimentary features among multiple systems can be exploited to improve recognition performance [3, 10]. Two major categories of techniques are often used for system combination: hypothesis level combination and cross system adaptation. The former exploits the consensus among component systems using voting as well as confidence measures, such as ROVER [7] and confusion network combination (CNC) [5]. Alternatively the second category uses acoustic model (AM) cross adaptation [19, 16, 20, 17]. Cross adaptation is only applied to acoustic models. Previous research has shown that LVCSR systems' diversity may also be further exploited in language models (LM). Context dependent cross adaptation multi-level n-gram LMs was proposed to improve system combination performance [10]. However, when only limited amounts of text data is available in adaptation, the generalization ability of this approach remain limited. For any unseen context that only occurs in the test data, component n-gram models will back-off to lower order distributions. This leads to a reduced modelling resolution in the combined LM, irrespective of the nature of interpolation weights being used. Hence, techniques that can represent the target domain in continuous space [1, 18, 4] are preferred for LM adaptation. Along this line, a cascaded network based NNLM adaptation method was investigated in [15].

This work was in part supported by DARPA under the GALE program via a subcontract to BBN Technologies. The paper does not necessarily reflect the position or the policy of the US Government and no official endorsement should be inferred.

This paper investigates cross adaptation of NNLMs to improve LVCSR system combination. The rest of the paper is organized as follows. A generic WFST based LM combination scheme is reviewed in section 2. Context dependent cross adaptation of multi-level n-gram LMs is presented in section 3.1. A cascaded network based NNLM cross adaptation scheme is proposed in section 3.2. In section 4 LM cross adaptation schemes are evaluated on a range of Chinese LVCSR systems used in the DARPA GALE phase 4 and 5 evaluations.

2. Language Model Combination

In state-of-the-art LVCSR systems language model (LMs) are often constructed by combining different LMs. The precise nature of component LMs determines the appropriate form of combination to use. For example, when building word level LMs, in order to improve context coverage and generalization, a linear interpolation between component n-gram, class or neural network LMs [18, 4, 15] can be used. When introducing additional sub-word level linguistic constraints to increase discrimination, word and syllable level LMs can be log-linearly combined as multi-level n-gram LMs [11]. In order to capture local variation of modelling resolution, generalization, topics and styles among component LMs, context dependent LM interpolation and adaptation can be used [9]. As the combination configuration becomes more complex, these techniques require increasingly more extensive software changes. An alternative approach to combine and adapt LMs is to use *semi-ring* based weighted finite state transducers (WFSTs) [13, 11]. They provide a generic and well-defined framework to represent different types of LMs and their combined forms.

Linear Model Combination: is a *union* of all the individual experts. It helps overcome the sparsity issue when training individual component models and thus improves generalization. Let w_i denote the i^{th} word of a L word long sequence $\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$. The LM log-probability for the complete word sequence is given by

$$\ln P(\mathcal{W}) = \sum_{i=1}^{L} \ln \left(\sum_{m=1}^{M} \lambda_m P_m(w_i | h_i^{n-1}) \right) \quad (1)$$

where h_i^{n-1} represents the i^{th} word's history of $n-1$ words maximum, $< w_{i-n+1}, ..., w_{i-1} >$, and λ_m is the global, context free weight for the m^{th} component under a positive and sum-to-one subject. The WFST representation of the linearly combined LM can be derived using a component level *composition* between the n-gram and interpolation weight transducers prior to a final *log semi-ring* based n-gram level *union*.

Log-Linear Model Combination: provides an *intersection* of individual experts and yields a high likelihood only when all

component models agree. For the example above, the log-linearly interpolated LM probability, with the probability normalization term ignored, is

$$\ln P(\mathcal{W}) \quad = \quad \sum_{i=1}^{L} \sum_{m=1}^{M} \lambda_m \ln P_m(w_i|h_i^{n-1}) \qquad (2)$$

where λ_m is the context free log-linear weight for the m^{th} component. They are no longer subject to a positive and sum-to-one constraint and are fixed as equal in this paper. Using WFSTs, a log-linear model combination may be implemented using a sequence of *composition* operations between component n-gram model transducers after an *arithmetic scaling* of arc costs by their respective log-linear weights.

WFST based LM combination is highly flexible and can be used for a wide range of combination configurations. It not only supports the use of global, context free weights in LM combination, but also a more general case when context dependent weights are employed. In previous research this method was used for a multi-level n-gram LM derived by intersecting context dependently adapted LMs built at word and syllable level [11, 10]. In this paper this approach is extended to further include NNLMs in a context free, linear combination,

$$P(w_i|h_i^{n-1}) = \lambda P_{\text{NG}}(w_i|h_i^{n-1}) + (1-\lambda)\tilde{P}_{\text{NN}}(w_i|h_i^{n-1}) \quad (3)$$

where $\tilde{P}_{\text{NN}}(w_i|h_i^{n-1})$ is the normalized NNLM probabilities to ensure a sum-to-one constraint for the combined probabilities. λ is the context free interpolation weight assigned to the multi-level n-gram distribution $P_{\text{NG}}(\cdot)$. It is fixed as $\lambda = 0.5$ in all experiments of this paper. The above combination requires NNLMs to be converted on-the-fly to fully expanded n-gram WFSTs without a back-off chain during decoding.

3. Language Model Cross Adaptation

3.1. Multi-level LM Cross Adaptation

In order to improve robustness to varying styles or tasks, unsupervised LM adaptation to a particular broadcast show, for example, may be used [2, 6]. As directly adapting n-gram probabilities is impractical on limited amounts of data, the standard adaptation schemes only involve updating the context free, linear interpolation weights in equation (1). The aim is to optimize the LM log-probability of the supervision by re-estimating λ_m, the global, context free weight for the m^{th} component LM.

However, this approach can only adapt LMs to a particular genre, epoch or other higher level attributes. Local factors that determine the "usefulness" of sources on a context dependent basis, such as modelling resolution, generalization, topics and styles, are poorly modelled. To handle this issue, context dependent LM interpolation and adaptation can be used [9]. Thus equation (1) is extended to

$$\ln \check{P}(\mathcal{W}) = \sum_{i=1}^{L} c_i \ln \left(\sum_{m=1}^{M} \boldsymbol{\phi}_m(h_i^{n-1}) P_m(w_i|h_i^{n-1}) \right) \quad (4)$$

where $\boldsymbol{\phi}_m(h_i^{n-1})$ is the m^{th} component weight for context h_i^{n-1}. MAP and discriminative schemes are available to robustly estimate these weight parameters [9]. c_i is the confidence score for word w_i used to further improve robustness to the supervision quality. A count cut-off for different histories, for example, the average word level confidence score computed over the supervision hypotheses, is introduced. Contexts which do not have sufficient counts above such threshold will be pruned in weight estimation.

In order to incorporate richer linguistic constraints, it is possible to train and combine LMs that model different unit sequences, for example, syllables and words [11]. Context dependently interpolated LMs built at word and syllable level are intersected using equation (2) as a multi-level LM. This LM leverages from both linear and log-linear forms of model combination and aims to achieve a good balance between generalization and discrimination. Its hierarchical nature may also be used to exploit the additional, non-acoustic system diversity at word and syllable level to improve system combination.

3.2. Neural Network LM Cross Adaptation

As discussed in section 1, the use of continuous space representation of words in NNLMs provides a stronger generalization ability than n-gram LMs. This advantage can also be exploited when adapting NNLMs to a given target domain using limited amounts of supervision data. A cascaded network based NNLM cross adaptation scheme is considered in this paper [15]. An additional adaptation layer is added between the projection and hidden layers. It acts as a linear input transformation to the hidden layer of a task independent NNLM. The precise location of such adaptation layer is determined by two factors. First, in conventional NNLMs much fewer nodes are often used for projection and hidden layers than input and output layers [18, 4]. Hence, given very limited amounts of data, adopting a compact structure of the adaptation layer is important. Second, non-linear activation functions are used in hidden and output layers of standard NNLMs. It is also preferable to retain the same discriminative power and non-linearity during NNLM adaptation. This scheme provides a direct adaptation of NNLMs via a non-linear, discriminative transformation to a new domain.

To reduce computational cost, conventional NNLMs only model the probabilities of a small and more frequent subset of the whole vocabulary, commonly referred to as the *short-list*. The output layer only contains nodes for in-shortlist words. Two issues arise when using this conventional NNLM architecture. First, NNLM parameters are trained only using the statistics of in-shortlist words thus introduces an undue bias to them. Secondly, as there is no explicit modelling of probabilities of *out-of-shortlist* (OOS) words in the output layer, statistics associated with them are also discarded in NNLM training. To handle these issues, an improved NNLM architecture with an additional output node explicitly modelling the probability mass of OOS words is used [15]. This ensures that all training data are used in NNLM training, and the probabilities of in-shortlist words are smoothed by the OOS probability mass, thus obtaining a more robust parameter estimation. The architecture of an adapted NNLM of this form is illustrated in figure 1. During adaptation, only the part of network connecting the adaptation layer and projection layer are updated (shown in dashed lines in figure 1), while other parameters are fixed.

As discussed in section 2, a linear interpolation between adapted multi-level n-gram LMs and NNLMs using equation (3) can be used to obtain a good coverage of contexts and strong generalization ability. This requires the probability mass computed from the OOS output layer node re-distributed among all OOS words [15]. Such normalization can be very expensive for LVCSR tasks. To improve efficiency, an approximated form of normalization can be used, assuming that the OOS probability

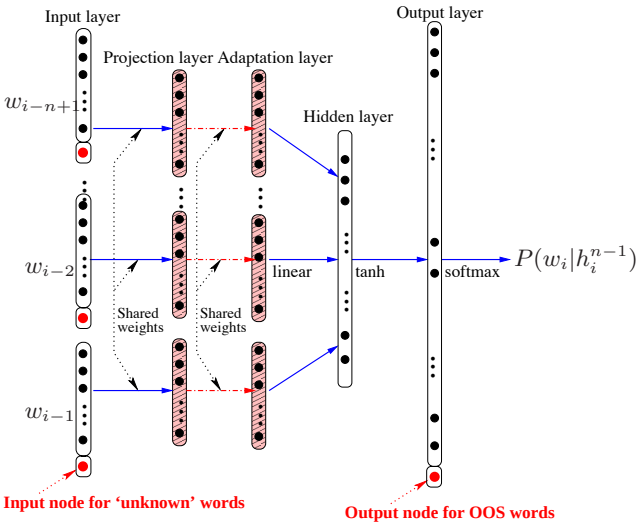

Figure 1: Architecture of a NNLM with adaptation layer.

mass assigned by the NNLM and n-gram LM are equal,

$$\tilde{P}_{\mathsf{NN}}(w_i|h_i^{n-1}) \approx \begin{cases} P_{\mathsf{NN}}(w_i|h_i^{n-1}) & w_i \text{ in shortlist} \\ P_{\mathsf{NG}}(w_i|h_i^{n-1}) & \text{otherwise.} \end{cases} \quad (5)$$

4. Experiments and Results

In this section various cross adaptation configurations are evaluated on the AGILE Chinese LVCSR systems used in the 2010 and 2011 DARPA GALE evaluations.

The 2009 CU Chinese LVCSR system was trained on on 1960 hours of broadcast speech data. A total of 5.6 billion characters from 28 text sources were used in LM training. These account for 3.7 billion words after a longest first based character to word segmentation. A 63k word list was used. A word level 4-gram NNLM with an OOS output layer node was trained using 23 million words from acoustic transcription sources only. The size of NNLM input and output vocabularies are 63k and 20k words respectively. The system uses a multi-pass recognition and system combination framework. In the initial lattice generation stage, an interpolated 4-gram word level baseline LM and adapted gender dependent cross-word triphone MPE acoustic models with HLDA projected PLP and pitch features were used in decoding. The lattices generated were then rescored using a context dependently adapted multi-level LM, which models both 4-gram word and 6-gram character sequences. It can be optionally further combined with the 4-gram NNLM at word level. The resulting lattices were then used in a "P3" acoustic re-adaptation and lattice rescoring stage, before a final 4-way CNC combination [10].

Five GALE Chinese speech test sets of mixed broadcast news (BN) and conversation (BC) genre: 2.6 hour **d07**, 1 hour **d08**, 3 hour **d09s**, 2.6 hour **p2ns** and 1.5 hour **p3ns** were used. Performance of CU systems using various LM configurations are shown in table 1. Both context dependent multi-level n-gram LM adaptation and NNLMs gave consistent character error rate (CER) reductions and their gains also largely additive. For example, CER gains of 0.1%-0.4% absolute were obtained using the "SA CD/Base" system over the comparable baseline "Base/Base", which used an unadapted n-gram LM and NNLM. Using an NNLM this "SA CD/Base" system gave 0.3%-0.5% absolute CER gains over the comparable "SA CD/-" baseline. NNLM adaptation gave small CER gains of 0.1% on **d07** and **p2ns** for the "Base/SA Proj" system over the "Base/Base" baseline, both of which used an unadapted n-gram LM. However, no performance improvement was obtained over the "SA CD/Base" system, which used an adapted n-gram LM. These trends suggest that NNLM adaptation may not capture additional useful information when combined with n-gram LM adaptation.

NGLM	NNLM	**d07**	**d08**	**d09s**	**p2ns**	**p3ns**
Base	-	8.2	7.4	8.7	7.8	10.9
Base	Base	8.0	6.9	8.4	7.5	10.5
Base	SA Proj	7.9	6.9	8.4	7.4	10.5
SA CD	-	7.9	7.3	8.4	7.6	10.6
SA CD	Base	7.6	6.8	7.9	7.3	10.1
SA CD	SA Proj	7.6	6.8	8.0	7.4	10.2

Table 1: CER Performance of 2009 CU systems. "Base" stands for no LM adaptation, "SA" for self-adaptation. "CD" for context dependent. "Proj" for projection layer NNLM adaptation.

The 2009 AGILE Chinese LVCSR system was built by combining a range of systems separately developed at Cambridge University, BBN Technologies and LIMSI-CNRS using cross site adaptation. The BBN and LIMSI systems were trained on the same amount of speech and text data as the CU system presented in table 1. The baseline cross adapted system involves cross adapting the 4 CU acoustic models to the outputs from BBN and LIMSI separately using confidence scored based MLLR, before a final 8-way CNC combination [10]. Optionally using an n-gram LM alone, or combined with an NNLM, and further with or without LM cross adaptation for either or both, gives a total of six combined systems shown in table 2. LM cross adaptation was performed at at audio document level.

NGLM	NNLM	**d07**	**d08**	**d09s**	**p2ns**	**p3ns**
Base	-	7.5	6.7	7.8	7.1	10.1
Base	Base	7.3	6.4	7.6	6.9	9.9
Base	XA Proj	7.2	6.4	7.6	6.8	9.8
XA CD	-	7.0	6.5	7.4	6.7	9.6
XA CD	Base	7.0	6.4	7.3	6.6	9.3
XA CD	XA Proj	6.9	6.4	7.3	6.5	9.2

Table 2: Performance of 2009 AGILE systems. "Base" stands for no LM adaptation, "XA" for cross adaptation, "CD" for context dependent. "Proj" for projection layer NNLM adaptation.

Context dependent cross adaptation of multi-level n-gram LM gave significant CER gains. For example, absolute CER reductions of 0.4%-0.5% (5.0%-6.7% rel.) were obtained on all test sets except **d08** using the "XA CD/-" system over the comparable baseline "Base/-". This "XA CD/-" AGILE system was used in the 2009 GALE evaluation. These gains were maintained when further combined with an NNLM. For example, the "XA CD/Base" system outperformed the "Base/Base" configuration by 0.3%-0.6% absolute on all test sets except **d08**. The use of NNLMs gave consistent but smaller gains, for example, 0.3% absolute on **p3ns** for the "XA CD/Base" system over the "XA CD/-" baseline, compared with the results shown in table 1 for the self-adapted "SA CD/Base" system. This indicates that NNLMs and acoustic model across adaptation may have a similar, but non-additive, effect on improving generalization performance. Consistent improvements were further obtained using

NNLM cross adaptation, for example, CER gains of 0.1% absolute using the "XA CD/XA Proj" system against the comparable "XA CD/Base" baseline on **d07**, **p2ns** and **p3ns**. This fully cross adapted "XA CD/XA Proj" system gave the best CER performance among all combined systems shown in table 2. The overall CER gains over the second acoustic only cross adaptation baseline "Base/Base" in table 2, which used a fixed n-gram and NNLM, were 0.3%-0.7% (4.0%-7.1% rel.) absolute on **d07**, **d09s**, **p2ns** and **p3ns**.

The 2010 AGILE Chinese LVCSR system were then used to conduct a similar set of cross adaptation experiments based on the results in table 2. The overall cross adaptation based system architecture remains the same. BBN and LIMSI provided outputs generated using their respective updated systems [14, 8]. Four MPE trained word position dependent (PD) phone or syllable acoustic models [12] with 12k tied states were used in the 2010 AGILE and CU systems before a 4-way CU internal or cross site adapted 8-way CNC combination:

- P3a: GD PLP+MLP PD quinphone
- P3b: SAT Gaussianized PLP tri-syllable
- P3c: GD Gaussianized PLP+MLP PD triphone
- P3d: GD PLP full covariance PD quinphone

Three new GALE Chinese speech test sets: 6 hour **d10c**, 12 hour **d10r**, 12 hour **d10d** and 3 hour **p4ns** were also used. Site specific and cross adapted performance of the 2010 AGILE systems are shown in tables 3 and 4 respectively. Based on the results in table 1, the 2010 CU system used a "SA CD/Base" LM configuration. CER reductions of 0.2%-0.4% over the best single branch were obtained after the 4-way CU internal CNC.

System	**d09s**	**d10c**	**d10r**	**d10d**	**p4ns**
BBN	7.8	12.2	7.9	24.6	7.2
LIMSI	8.5	13.4	8.3	25.5	7.9
CU-P3a	8.3	12.3	7.9	24.4	7.5
CU-P3b	8.2	12.3	7.9	24.2	7.4
CU-P3c	8.2	12.3	7.9	24.4	7.3
CU-P3d	8.2	12.2	7.8	24.1	7.3
CU	7.8	11.9	7.6	23.7	7.1

Table 3: Performance of 2010 AGILE sub-systems.

As shown in table 4, cross adapting both the n-gram and NNLM gave a total CER gains of 0.2%-0.9% (3.0%-6.7% rel.) on all test sets for the "XA CD/XA Proj" system against the "Base/Base" baseline, which used unadapted n-gram and NNLMs. The gains over the "XA CD/-" system (the 2009 AGILE LM cross adaptation configuration), which used no NNLMs, are 0.2%-0.4% on **d07**, **d10c** and **d10d**. The overall gains over the 2009 AGILE combined system (4th line of table 2 and 1st line of table 4) are 0.3%-0.9% (4.0%-5.4% rel.) absolute.

5. Conclusion

Neural network language model cross adaptation was investigated in this paper to improve LVCSR system combination. Experimental results on a state-of-the-art speech recognition task suggest the proposed NNLM cross adaptation method may be useful to capture additional diversity among systems. Future research will focus on improving robustness in NNLM cross adaptation and its combination with n-gram LM adaptation.

NGLM	NNLM	**d09s**	**d10c**	**d10r**	**d10d**	**p4ns**
AGILE 2009		7.4	11.5	7.3	22.6	6.8
Base	Base	7.5	11.5	7.3	22.6	6.7
XA CD	-	7.2	11.1	7.0	22.1	6.5
XA CD	XA Proj	7.0	10.9	7.0	21.7	6.5

Table 4: Performance of 2010 AGILE systems. "Base" stands for no LM adaptation, "XA" for cross adaptation, "CD" for context dependent. "Proj" for projection layer NNLM adaptation.

6. References

[1] Y. Bengio et al. (2001). "A neural probabilistic language model", *J. of Machine Learning Research*, Vol. 3, No. 2, 2001.

[2] L. Chen et al. (2001). Language Model Adaptation For Broadcast News Transcription, in *Proc. ISCA ITRW'01*.

[3] S. M. Chu et al. (2010). The 2009 IBM GALE Mandarin Broadcast Transcription System, in *Proc. IEEE ICASSP2010*.

[4] A. Emami & L. Mangu, "Empirical study of neural network language models of Arabic speech recognition", in *Proc. IEEE ASRU2007*.

[5] G. Evermann & P.C. Woodland (2000). Posterior Probability Decoding, Confidence Estimation and System Combination, in *Proc. Speech Transcription Workshop 2000*.

[6] M. Federico (1999). Efficient Language Model Adaptation Through MDI Estimation, in *Proc. EuroSpeech'99*.

[7] J. G. Fiscus (1997). A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER). In *Proc. IEEE ASRU'97*.

[8] L. Lamel et al. (2011). Improved Models for Mandarin Speech-to-text Transcription, to appear in *Proc. IEEE ICASSP2011*.

[9] X. Liu, M. J. F. Gales & P. C. Woodland (2009). Use of Contexts in Language Model Interpolation and Adaptation, in *Proc. Interspeech'09*.

[10] X. Liu, M. J. F. Gales & P. C. Woodland (2010). Language Model Cross Adaptation For LVCSR System Combination, in *Proc. Interspeech'10*.

[11] X. Liu, M. J. F. Gales, J. L. Hieronymus & P. C. Woodland (2010). Language Model Combination and Adaptation Using Weighted Finite State Transducers, in *Proc. IEEE ICASSP2010*.

[12] X. Liu, M. J. F. Gales, J. L. Hieronymus & P. C. Woodland (2011). Investigation of Acoustic Units for LVCSR Systems, to appear in *Proc. IEEE ICASSP2011*.

[13] M. Mohri & M. Riley. Network optimizations for large vocabulary speech recognition. *Speech Communication*, 25:3, 1998.

[14] T. Ng et al. (2010). Jointly Optimized Discriminative Features for Speech Recognition, in *Proc. Interspeech'10*.

[15] J. Park, X. Liu, M. J. F. Gales & P. C. Woodland (2010). Improved Neural Network Based Language Modelling and Adaptation, in *Proc. Interspeech'10*.

[16] B. Peskin et al. (1999). Improvemnets in Recognition of Conversational Telephone Speech, in *Proc. IEEE ICASSP1999*.

[17] R. Schwartz et al. (2004). Speech Recognition in Multiple Languages and Domains: The 2003 BBN/LIMSI EARS System, in *Proc. IEEE ICASSP2004*.

[18] H. Schwenk, "Continuous space language models", *Computer Speech and Language*, Vol. 21, 2007, pp. 492–518.

[19] P. C. Woodland et al. (1995). The 1994 HTK Large Vocabulary Speech Recognition System, in *Proc. IEEE ICASSP1995*.

[20] P. C. Woodland et al. (2004). SuperEARS: Multi-site Broadcast News System, in *Proc. Rich Transcription Workshop 2004*.