Language Model Cross Adaptation For LVCSR System Combination

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 subsystems developed at different sites. Cross system adaptation can be used as an alternative to direct hypothesis level combination schemes such as ROVER. In normal cross adaptation it is assumed that useful diversity among systems exists only at acoustic level. However, complimentary features among complex LVCSR systems also manifest themselves in other layers of modelling hierarchy, e.g., subword and word level. It is thus interesting to also cross adapt language models (LM) to capture them. In this paper cross adaptation of multi-level LMs modelling both syllable and word sequences was investigated to improve LVCSR system combination. Significant error rate gains of 6.7% relative were obtained over ROVER and acoustic model only cross adaptation when combining 13 Chinese LVCSR subsystems used in the 2010 DARPA GALE evaluation.

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 [5, 6, 2]. Increasing the diversity among sub-systems often leads to larger combination gains. Two major categories of techniques are often used: 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 [4] and confusion network combination (CNC) [3]. Alternatively the second category uses acoustic model (AM) cross adaptation [14, 11, 15, 13, 12]. They may be viewed as an implicit form of system combination. Standard cross adaptation assumes that useful diversity among component systems exists exclusively at acoustic model level. The output of one system is projected at the phone level when it's used to cross adapt another system. However, for LVCSR systems diversity may also be exploited at other levels. Complimentary features among diverse systems can also manifest themselves in other layers of modelling hierarchy, e.g., at the subword and word level [8]. These are not addressed under the conventional acoustic-only cross adaptation framework. For example, homophone and parameter tying related acoustic confusions will unduly discard part of the word level system diversity. It is thus useful to also cross adapt language models (LM) to explicitly capture them.

To address this issue, this paper investigates cross adaptation of language models to improve LVCSR system combination. The rest of the paper is organized as follows. ROVER based hypothesis level combination is reviewed in section 2.1. Standard acoustic model cross adaptation is presented in section 2.2. Confidence measure based context dependent cross adaptation of a multi-level LM modelling both syllable and word sequences is proposed in section 3. In section 4 various cross adaptation and ROVER combination schemes are evaluated on a total of 13 Chinese LVCSR systems used in the DARPA GALE phase 4 evaluation.

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.

2. System Combination

2.1. Hypothesis Level System Combination

One commonly used form of hypothesis level combination is ROVER [4]. Hypotheses from a total of S component systems are iteratively aligned to create word transition networks. An interpolation between voting counts and confidence scores is then used to find the optimal word sequence within the network. For any set of confusions in the network this is given by,

$$\hat{w} = \arg\max_{w_s} \left\{ \alpha \frac{N_{1:S}(w_s)}{S} + (1-\alpha) c_w^{(s)} \right\} \quad (1)$$

where $N_{1:S}(w_s)$ is number of systems that output word w_s, and $c_w^{(s)}$ the confidence score assigned by the sth system, and α is a tunable parameter to balance the contribution between voting counts and confidence scores. When component systems using different word segmentation schemes, a direct combination between their outputs is problematic, for example, in Chinese where different character to word segmentations are used. Hence, for the Mandarin speech recognition tasks considered here, the most successful approach is to perform a character level combination [5, 10]. This requires the mapping of word level outputs to subword, character level. The confidence score of each word is assigned to each character it contains. One major issue with character level ROVER is it does not preserve a consistent character to word segmentation in the final outputs, and thus affects machine translation performance for speech translation tasks [5]. In general, hypothesis level combination methods such as ROVER also require the error rate performance of components systems to be close in order to be effective in combination.

2.2. Acoustic Model Cross Adaptation

When there is a large difference in the error rate performance of component systems, acoustic model cross adaptation provides an alternative to hypothesis level combination. It was initially used as an implicit form of within site system combination [14, 11]. In later research it was also adopted for cross site combination, often together with hypothesis level combination techniques [15, 13, 12, 6, 2]. Word level outputs from

one system are mapped to phone model sequences first using a lexicon and a forced alignment process. Then MLLR or CMLLR based linear transforms are estimated using the resulting phone level supervision and sufficient statistics. The number of transform parameters balances the trade-off between learning sufficient information and a bias to the supervision. To improve robustness to the supervision quality, it is possible to use statistics weighted by confidence scores during cross adaptation [1]. The associated auxiliary function is

$$\mathcal{Q}_{\mathsf{conf}}(\lambda, \tilde{\lambda}) = \sum_{j,t} c_t \gamma_j(t) \log p\left(\boldsymbol{o}_t; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j, \boldsymbol{W}_{r_j}\right) \quad (2)$$

where $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$ are the j^{th} Gaussian component's mean and covariance, $\boldsymbol{W}_{r_j}$ the linear transform it is assigned to, and $\gamma_j(t)$ the posterior probability of frame $\boldsymbol{o}_t$ at component j. c_t is the frame confidence score and normally set equal to that of word level, as considered in this paper.

Standard cross adaptation propagates complimentary information from one system to another at the phone sequence level. As there is no direct hypothesis combination between systems with potentially large difference in error rate, cross adaptation in general is less sensitive than ROVER to the performance difference between component systems. As the same language model and vocabulary are used, consistent word tokenizations are also preserved. Due to this advantage over ROVER, cross adaptation is considered a "safe" choice to combine LVCSR systems for speech translation tasks [5, 6, 2].

3. Language Model Cross Adaptation

In current LVCSR systems LMs are often constructed by training and combining multiple component n-gram LMs in a mixture model [10, 6, 2, 8]. In order to improve robustness to varying styles or tasks, unsupervised LM adaptation to a particular broadcast show, for example, may be used. 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. Let w_i denote the i^{th} word of a L word long hypothesis supervision $\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$, and h_i^{n-1} the i^{th} word's history of $n-1$ words maximum, $< w_{i-n+1}, ..., w_{i-1} >$. The aim is to optimize the LM log-probability of the supervision

$$\ln P(\mathcal{W}) = \sum_{i=1}^{L} \ln \left(\sum_{m=1}^{M} \lambda_m P_m(w_i | h_i^{n-1}) \right) \quad (3)$$

by re-estimating λ_m, the global, context free weight for the m^{th} component. By definition, when the output of an initial recognition pass is used as the supervision, LM self-adaptation is performed. When the output of another system is used as the supervision, LM cross adaptation is performed instead.

However, the above 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 [7]. A set of discrete context dependent back-off weights are used to dynamically adjust the contribution from component LMs. Thus equation (3) is extended to

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

where $\phi_m(h_i^{n-1})$ is the m^{th} component weight for context h_i^{n-1}. MAP based maximum likelihood and discriminative schemes are available to robustly estimate these weight parameters [7]. Take the ML based adaptation as an example, this is given by

$$\hat{\phi}_m(h_i^{n-1}) = \frac{\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau \hat{\phi}_m(h_i^{n-2})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau} \quad (5)$$

where $\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1})$ is ML statistics for history context h_i^{n-1}, and τ controls the contribution from a hierarchical prior, $\hat{\phi}_m(h_i^{n-2})$, before intersected with a high resolution training data prior [7].

To improve robustness to the supervision quality, it is possible to use confidence score weighted sufficient statistics when estimating context free, and dependent interpolation weights. The log-likelihood in equation (4) is thus modified as

$$\ln \check{P}(\mathcal{W}) = \sum_{i=1}^{L} c_i \ln \left(\sum_{m=1}^{M} \phi_m(h_i^{n-1}) P_m(w_i | h_i^{n-1}) \right) \quad (6)$$

where c_i is the confidence score for word w_i. By default, when using a null history the above simplifies to confidence score based adaptation of global, context free weights in equation (3). To further improve robustness during context dependent LM cross adaptation, it is also possible to impose a count cut-off for different histories, for example, the average word level confidence score computed over the supervision hypotheses. Contexts which do not have sufficient counts above such threshold will be pruned in weight estimation, as considered in this paper.

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 [8]. Context dependently interpolated LMs built at word and syllable level are intersected to yield a final combined 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 also provides a good chance to exploit the additional, non-acoustic system diversity at word and syllable sequence level to improve system combination. Hence, they are considered in the LM cross adaptation experiments of the following section.

4. Experiments and Results

In this section various ROVER combination and cross adaptation configurations are evaluated on 13 AGILE Chinese LVCSR systems used in the 2010 DARPA GALE phase 4 evaluation.

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. The system uses a multi-pass recognition and system combination framework. The overall structure of the system is shown in figure 1. 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 [8]. Hierarchical and normalized perplexity smoothing priors were used to adapt the context dependent interpolation weights [7]. This multi-level LM lattice rescoring stage used a weighted finite state transducer (WFST) based on-the-fly expansion algorithm described in [8]. The resulting lattices were then used in a "P3" acoustic re-adaptation and lattice

rescoring stage, where four different acoustic models developed on the same training data were used:

- P3a: boosted MMI GD PLP+MLP quinphone
- P3b: MPE SAT Gaussianized PLP triphone
- P3c: MPE GD Gaussianized PLP+MLP triphone
- P3d: boosted MMI GD PLP quinphone

before a final CNC combination. As discussed in section 2.1, hypothesis level combination methods require performance of components systems to be close in order to be effective. Hence, the two PLP frontend based acoustic models were cross adapted to the outputs of the two PLP+MLP models to give more balanced performance among different branches. 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 the individual branches and combined CU system are shown in table 1. LM adaptation gave consistent character error rate (CER) reductions, for example, of 0.2%-0.6% absolute for the "P3a" branch over its baseline, "P3a.base", which used an unadapted LM. Further CER reductions of 0.1%-0.4% over the best single branch were obtained in CNC combination. On d09s the final BN and BC genre specific performance are 4.1% and 12.6%.

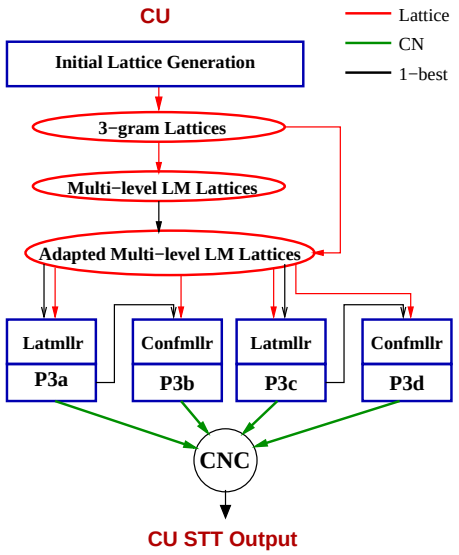

Figure 1: The CU 2009 Chinese LVCSR system.

System	d07	d08	d09s	p2ns	p3ns
P3a.base	8.6	7.7	9.2	8.3	11.7
P3a	8.3	7.5	8.8	7.9	11.1
P3b	8.1	7.4	8.6	7.9	11.0
P3c	8.2	7.5	8.6	7.9	11.0
P3d	8.4	7.9	8.9	8.1	11.2
CU	7.9	7.3	8.5	7.6	10.6

Table 1: CER performance of CU Chinese LVCSR system for different system branches. "P3a.base" used an unadapted LM.

The 2009 AGILE Chinese LVCSR system was built by combining a range of systems separately developed at Cambridge University, BBN Technologies and LIMSI-CNRS. 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 LIMSI system also employed a multi-pass architecture but only one single system [9]. The BBN system is more complex and used a ROVER combination between a total of 8 different systems' outputs for within site combination [10]. The CER performance of the BBN and LIMSI systems are shown in the first two lines of table 2. The CU system's performance is also shown in the third line of the table. Both the BBN and LIMSI systems used character to word segmentation schemes in training different from the CU system. As discussed in sections 2.1 and 2.2, their outputs were re-tokenized using the CU character to word segmentation scheme for cross adaptation, as well as split into character sequences for ROVER combination [5].

System	d07	d08	d09s	p2ns	p3ns
BBN	8.8	7.9	8.9	8.0	11.8
LIMSI	9.3	8.5	9.4	8.4	12.5
CU	7.9	7.3	8.5	7.6	10.6
ROVER(3way)	7.5	7.0	8.3	7.0	10.3
ROVER(13way)	7.3	6.8	7.8	7.1	9.8

Table 2: Performance of ROVER combined AGILE systems.

ROVER combination performance of two AGILE systems are shown in the second section of the table. The first one is a 3-way cross site combination between the final outputs of the three systems shown in the first section of the table. Absolute CER gains of 0.2%-0.6% over the best single system were obtained. The second ROVER configuration is more complicated and involved a 13-way combination between individual branch outputs of all 4 CU component systems shown in the middle section of table 1, all 8 BBN component systems and the LIMSI system's outputs. Performance of this combined system is shown in the last line of table 2. As expected, the amount of diversity and complimentary features increased as more systems were used in combination. Further CER reductions of 0.2%-0.5% were obtained on four test sets except p2ns. In particular, for test sets with higher error rates and potentially larger diversity such as d09s and p3ns, the use of more systems produced more reliable voting during ROVER, and thus a larger improvement of 0.5% over the 3-way configuration. The 13-way ROVER gave absolute CER gains of 0.5%-0.8% over the CU system.

AM	LM	d07	d08	d09s	p2ns	p3ns
XA	Base	7.5	6.7	7.9	7.1	10.1
	SA CD	7.4	6.8	7.7	7.1	9.8
	XA CI	7.3	6.8	7.6	7.0	9.8
	XA CD	***7.0***	***6.5***	***7.4***	***6.7***	***9.6***

Table 3: CER Performance of cross adapted AGILE systems. "Base" stands for no LM adaptation, "SA" for self-adaptation,"XA" for cross adaptation, "CI" for context independent and "CD" for context dependent.

Cross adaptation performance of various systems are shown in table 3. The first two systems used standard acoustic model only cross adaptation. The 4 CU acoustic models shown in table 1 were each adapted to the outputs from BBN and LIMSI separately using confidence scored based MLLR discussed in

section 2.2. A regression class tree was used that can generate a maximum total number of 4 transforms for speech states and one for silence. The resulting 8 cross adapted acoustic models were then used to rescore the CU system's lattices shown in figure 1 before a final 8-way CNC combination. Optionally using the unadapted baseline LM, or the CU self-adapted LM gave the "Base" and "SA CD" systems in table 3 respectively. Some gains from LM self-adaptation are still maintained on this acoustic cross adaptation setup. The "SA CD" system gave 0.1%-0.3% CER reductions on d07, d09s and p3ns over the "Base" system which used no LM adaptation of any form. This "SA CD" system also gave CER performance very close to the 13-way ROVER system of table 1.

The rest of table 3 shows performance of two combined systems by cross adapting both the CU acoustic and language models to the BBN and LIMSI outputs. In addition to the acoustic only cross adaptation described above, the interpolation weights of the CU multi-level LM were also adapted to the BBN and LIMSI outputs separately at audio document level using confidence score based estimation described in section 3. The resulting two sets of cross adapted LMs were used to rebuild the CU system lattices. These were then rescored using 4 BBN or LIMSI cross adapted CU acoustic models before a final 8-way CNC combination, as is shown in figure 2. The first AM+LM cross adapted system uses only context free interpolation weights for both word and character layers of the CU multi-level LM. As is shown in the second line of table 3, this "XA CI" system only gave 0.1% CER reduction on d07, d09s and p2ns against the acoustic only cross adapted "SA CD" system. In order to capture more of the LM diversity among different systems, the second AM+LM cross adapted system used context dependent interpolation weights for both word and character layers of the CU multi-level LM. Performance of this system is shown in the last line of table 3. Further absolute CER reductions of 0.2%-0.3% were obtained over context free LM cross adaptation. This fully cross adapted "XA CD" system gave the best CER performance among all combined systems shown in tables 2 and 3. The overall CER gains over the first acoustic only cross adaptation baseline system in table 3 were 0.2%-0.5% absolute across all test sets. In particular, significant CER reductions of 0.4%-0.5% were obtained on d07, d09s, p2ns and p3ns (5.0%-6.7% rel.). Using this system the BN and BC specific performance on d09s are 3.5% and 11.1%.

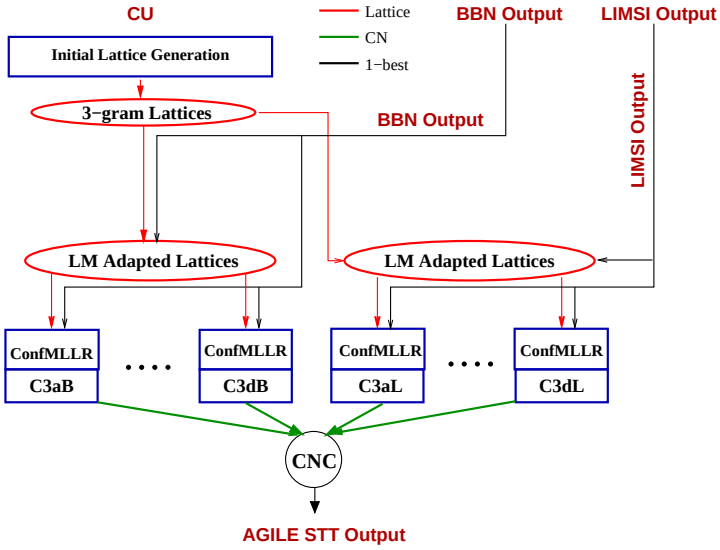

Figure 2: Architecture of the AGILE cross adapted system.

5. Conclusion

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 complimentary features exist on multiple layers of modelling hierarchy among highly diverse systems. The proposed LM cross adaptation method may be useful to capture additional diversity. Future research will focus on improving robustness in cross adaptation and system architecture refinement.

6. Acknowledgments

The authors would like to thank BBN Technologies and LIMSI-CNRS for sharing their outputs and the 13-way ROVER results.

7. References

[1] T. Anastasakos & S.V. Balakrishnan (1998). The Use of Confidence Measures in Unsupervised Adaptation of Speech Recognizers, in *Proc. ICSLP'98*.

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

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

[4] 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*.

[5] M. J. F. Gales et al. (2007). Speech System Combination for Machine Translation, in *Proc. IEEE ICASSP2007*.

[6] X. Lei et al. (2009). Development of the 2008 SRI Mandarin Speech-to-text System for Broadcast News and Conversation, in *Proc. Interspeech'09*.

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

[8] 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*.

[9] J. Luo, L. Lamel & J-L Gauvain (2009). Modeling Characters versus Words for Mandarin Speech Recognition, in *Proc. ICASSP2009*.

[10] T. Ng et al. (2008). Progress in the BBN 2007 Mandarin Speech to Text System, in *Proc. IEEE ICASSP2008*.

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

[12] R. Prasad et al. (2005). The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech Recognition System, in *Proc. ICSLP'05*.

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

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

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