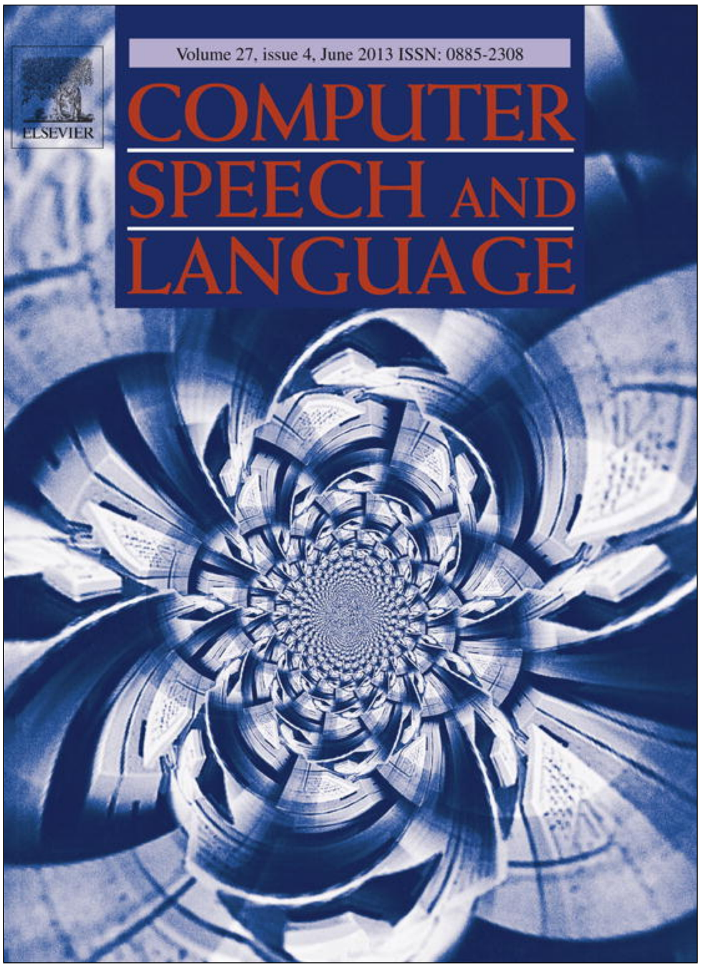
Volume 27, issue 4, June 2013 ISSN: 0885-2308
ELSEVIER
COMPUTER
SPEECH AND
LANGUAGE

Available online at www.sciencedirect.com

SciVerse ScienceDirect

Computer Speech and Language 27 (2013) 928–942

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Language model cross adaptation for LVCSR system combination[☆,☆☆]

X. Liu[*], M.J.F. Gales, P.C. Woodland

Cambridge University Engineering Department, Trumpington Street, Cambridge CB2 1PZ, England

Received 29 July 2011; received in revised form 12 July 2012; accepted 18 July 2012
Available online 27 July 2012

Abstract

State-of-the-art large vocabulary continuous speech recognition (LVCSR) systems often combine outputs from multiple sub-systems that may even be developed at different sites. Cross system adaptation, in which model adaptation is performed using the outputs from another sub-system, can be used as an alternative to hypothesis level combination schemes such as ROVER. Normally cross adaptation is only performed on the acoustic models. However, there are many other levels in LVCSR systems' modelling hierarchy where complimentary features may be exploited, for example, the sub-word and the word level, to further improve cross adaptation based system combination. It is thus interesting to also cross adapt language models (LMs) to capture these additional useful features. In this paper cross adaptation is applied to three forms of language models, a multi-level LM that models both syllable and word sequences, a word level neural network LM, and the linear combination of the two. Significant error rate reductions of 4.0–7.1% relative were obtained over ROVER and 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 (Woodland et al., 2004; Schwartz et al., 2004; Lei et al., 2009; Chu et al., 2010; Liu et al., 2010; Lamel et al., 2011). 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 (Fiscus, 1997) and confusion network combination (CNC) (Evermann and Woodland, 2000). Alternatively the second category uses acoustic model (AM) cross adaptation (Woodland et al., 1995, 2004; Peskin et al., 1999; Schwartz et al., 2004; Prasad et al., 2005). The acoustic models of one system are adapted to the recognition outputs of another. They may be viewed as an implicit form of system combination.

[☆] This paper has been recommended for acceptance by Shrikanth Narayanan.

[☆☆] This is an expanded version of a paper which was published by ISCA Interspeech 2010, and also selected as one of the top 3% papers published at the same conference.

[*] Corresponding author. Tel: +44 1223 766512; fax: +44 1223 332662.
E-mail addresses: xl207@cam.ac.uk (X. Liu), mjfg@eng.cam.ac.uk (M.J.F. Gales), pcw@eng.cam.ac.uk (P.C. Woodland).

0885-2308/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.csl.2012.07.010

The basic mechanism behind cross adaptation is to shift the decision boundary of the underlying statistical model of one system by learning the useful yet different characteristics of another. For LVCSR systems these characteristics may be exploited at multiple levels to improve system combination. The standard cross adaptation approach only makes use of the diversity among systems at acoustic model level. The output of one system is projected at the phone level when it is used to cross adapt the acoustic models of another system. However, for LVCSR systems complimentary features among diverse systems can also manifest themselves in other layers of modelling hierarchy, e.g., at the subword and word level (Hieronymus et al., 2009; Liu et al., 2010). 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 (LMs) to explicitly capture them.

In recent years a range of LM adaptation techniques have been proposed (Federico, 1999, 2003; Chen et al., 2001; Bulyko et al., 2007; Liu et al., 2008, 2009, 2010; Gildea and Hofmann, 1999; Mrva and Woodland, 2004; Chien et al., 2005; Tam and Schultz, 2005). Many of these approaches are designed for *n*-gram models, which remain the dominant form of language models used in LVCSR systems. One commonly adopted adaptation method is to construct a mixture model by combining multiple individual component *n*-gram LMs representing different styles or tasks using linear interpolation weights. These are normally estimated by minimising the perplexity of the supervision hypotheses. Standard interpolation weights are normally context independent in nature, representing a particular genre, epoch or other higher level attributes. In order to capture local variation of modelling resolution, generalisation, topics and styles, context dependent LM interpolation and adaptation can be used (Weintraub et al., 1996; Bulyko et al., 2003; Hsu, 2007; Liu et al., 2008, 2009, 2010). To incorporate richer linguistic constraints, it is also possible to train and combine LMs that model different unit sequences, for example, syllables and words (Hieronymus et al., 2009; Liu et al., 2010, 2010). The hierarchical nature of the combined LM also provides a good chance to exploit the additional, non-acoustic system diversity at the word and syllable sequence level to improve system combination (Liu et al., 2010).

However, when only a limited amount of text data is available for adaptation, the generalisation ability of *n*-gram LM based approaches remains limited. For any unseen context that only occurs in the test data, component *n*-gram models must back-off to lower order distributions, or make use of class information. 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 a continuous space (Bengio and Ducharme, 2003; Schwenk, 2007; Emami and Mangu, 2007; Gildea and Hofmann, 1999; Federico, 2003; Mrva and Woodland, 2004; Chien et al., 2005; Tam and Schultz, 2005; Blei et al., 2003) are preferred for LM adaptation. Along this line, a cascaded network based neural network LM (NNLM) adaptation method was investigated in Park et al. (2010).

This paper investigates cross adaptation of language models to improve LVCSR system combination. The rest of the paper is organised as follows. ROVER based hypothesis level combination is reviewed in Section 2.1. Standard acoustic model cross adaptation is presented in Section 2.2. A generic WFST based LM combination scheme is reviewed in Section 3. Context dependent cross adaptation of multi-level *n*-gram modelling both syllable and word sequences is proposed in Section 4.1. A cascaded network based NNLM cross adaptation scheme is presented in Section 4.2. In Section 6 LM cross adaptation schemes are evaluated on a range of Chinese LVCSR sub-systems used in the DARPA GALE phase 4 and 5 evaluations.

2. System combination

As discussed in Section 1, most LVCSR system combination schemes can be categorised into hypothesis level combination and acoustic model cross adaptation based approaches.

2.1. Hypothesis level system combination

One commonly used form of hypothesis level combination is ROVER (Fiscus, 1997). Hypotheses from a total of *S* sub-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,

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

where $N_{1:S}(w_s)$ is number of systems that output word w_s, and $c_w^{(s)}$, the confidence score[1] assigned by the sth system, and α is a tunable parameter to balance the contribution between voting counts and confidence scores. When combining 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 (Gales et al., 2007; Ng et al., 2008). 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.[2] 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 (Gales et al., 2007). 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 large difference in the error rate performance of the individual systems, acoustic model cross adaptation provides a useful alternative to hypothesis level combination. It was initially used as an implicit form of within site system combination (Woodland et al., 1995; Peskin et al., 1999). As greater diversity can be exploited among speech recognition systems developed at different sites, acoustic model cross adaptation was also adopted in later research for cross site combination, often together with hypothesis level combination techniques (Woodland et al., 2004; Schwartz et al., 2004; Prasad et al., 2005; Lei et al., 2009; Chu et al., 2010; Liu et al., 2010). Word level outputs from one system are mapped to phone model sequences first using a lexicon and a forced alignment process. Then MLLR (Leggetter and Woodland, 1995) or CMLLR (Gales, 1998) based linear transforms are estimated using the resulting phone level supervision. The number of transform parameters balances the trade-off between learning sufficient information and a bias to the supervision. To improve robustness to the adaptation supervision quality, it is possible to use statistics weighted by confidence scores during cross adaptation (Anastasakos and Balakrishnan, 1998; Wallhoff et al., 2000). The associated auxiliary function is

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

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

Standard acoustic model 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 differences in error rate, cross adaptation is in general less sensitive than ROVER to the performance difference between component systems. As the same language model and vocabulary are used, consistent word tokenisations are also preserved. Due to this advantage over ROVER, cross adaptation is considered a "safe" choice to combine LVCSR systems for speech translation tasks (Gales et al., 2007; Lei et al., 2009; Chu et al., 2010).

[1] ROVER allows either the average or maximum confidence scores to be used (Fiscus, 1997). In the ROVER experiments of this paper, maximum character level confidence scores are used throughout.

[2] It is preferable to use character level sub-system outputs for character level ROVER combination. However, for the experiments conducted in this paper only word level outputs were provided from other research sites. Hence, a simple approximation is used here.

3. Language model combination

In state-of-the-art LVCSR systems language models (LMs) are often constructed by combining multiple separately trained LMs together. The precise nature of these component LMs determines the appropriate form of combination to use. For example, when building word level LMs, in order to improve context coverage and generalisation, a linear interpolation between component word n-gram, class n-gram, or neural network LMs (Schwenk, 2007; Emami and Mangu, 2007; Park et al., 2010) could be used. When using 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 (Liu et al., 2010). In order to capture the local variation of modelling resolution, generalisation, topics and styles among component LMs, context dependent LM interpolation and adaptation can be used (Liu et al., 2008, 2009). 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 weighted finite state transducers (WFSTs) (Mohri and Riley, 1998; Liu et al., 2010). They provide a generic and well-defined framework to represent different types of LMs and their combined forms.

Linear model combination: is based on a *union* of all the individual probabilistic experts. It is often referred to as a mixture of experts (MoE) model in machine learning literature (Hinton, 1999, 2002). It tends to give a broader distribution than individual components alone and thus improves generalisation. Let w_i denote the ith word of a L word long sequence $\mathcal{W} = \langle w_1, w_2, \ldots, w_i, \ldots, w_L \rangle$. The LM log-probability combined over M component models for the complete word sequence is given by

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

where h_i^{n-1} represents the ith word's history of $n - 1$ words, $\langle w_{i-n+1}, \ldots, w_{i-1} \rangle$, and λ_m is the global, context independent weight for the mth component under a positive and sum-to-one constraint. 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*.[3]

Log-linear model combination: provides an *intersection* of individual experts and yields a high likelihood only when all component models agree. It is often referred to as a product of experts (PoE) model in the machine learning literature (Hinton, 1999, 2002). For the example above, the log-linearly interpolated LM probability, is

$$\ln\ P(\mathcal{W}) = \sum_{i=1}^{L} \sum_{m=1}^{M} \lambda_m\ \ln\ P_m(w_i | h_i^{n-1}) - \ln(Z) \tag{4}$$

where Z is a normalisation term to ensure that the interpolated probability is a valid distribution. Its exact computation for general forms of log-linear models is non-trivial. Analytical solutions are available only when certain forms of density functions, for example, Gaussian distributions (Gales and Airey, 2006), are used for component models. The normalisation term may be ignored when the interpolation is performed at complete word sequence level under a discriminative framework, for example, *maximum entropy* models and *logistic regression* (Darroch and Ratcliff, 1972; Rosenfeld et al., 2001; Och and Ney, 2002), as considered in this paper. λ_m is the context independent log-linear weight for the mth component. The log-linear interpolation weights are no longer subject to a positive and sum-to-one constraint. For a simple two way log-linear combination between word and character based LMs is considered in this paper, these weights are fixed as equal in all experiments. 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 supports not only the use of global, context independent weights in LM combination, but also a more general case when context dependent weights are employed. In previous research this method was used for multi-level n-gram LM

[3] This operation is different from the standard WFST union operation performed at sequence level.

adaptation by intersecting LMs with context dependent adaptation at word and syllable level (Liu et al., 2010, 2010). This approach can be extended to further include NNLMs in a context free, linear combination,

$$P(w_i|h_i^{n-1}) = \lambda P_{\mathsf{NG}}(w_i|h_i^{n-1}) + (1-\lambda)\tilde{P}_{\mathsf{NN}}(w_i|h_i^{n-1}) \tag{5}$$

where $\tilde{P}_{\mathsf{NN}}(w_i|h_i^{n-1})$ is the normalised NNLM probabilities to ensure a sum-to-one constraint for the combined probabilities. Here, λ is the context independent interpolation weight assigned to the multi-level n-gram distribution $P_{\mathsf{NG}}(\cdot)$. It is fixed as $\lambda = 0.5$ in all experiments of this paper.

4. Language model adaptation

In this section, context dependent adaptation of multi-level n-gram LMs modelling both syllable and word sequences is reviewed. A cascaded network based NNLM adaptation scheme is also presented.

4.1. Multi-level n-gram LM adaptation

In current LVCSR systems LMs are often constructed by training and combining multiple component n-gram LMs in a mixture model (Ng et al., 2008; Lei et al., 2009; Chu et al., 2010; Liu et al., 2010; Lamel et al., 2011). In order to improve robustness to varying styles or tasks, unsupervised LM adaptation to a particular broadcast show, for example, may be used (Chen et al., 2001; Federico, 1999). As directly adapting n-gram probabilities is impractical on limited amounts of data, the standard adaptation schemes only involve updating the context independent, linear interpolation weights in Eq. (3). 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, generalisation, topic and style, are poorly modelled. To handle this issue, context dependent LM interpolation and adaptation can be used (Liu et al., 2009). A set of discrete context dependent back-off weights are used to dynamically adjust the contribution from component LMs. Thus Eq. (3) is extended to

$$\ln\, P(\mathcal{W}) = \sum_{i=1}^{L} \ln\left(\sum_{m=1}^{M} \boldsymbol{\phi}_m(h_i^{n-1}) P_m(w_i|h_i^{n-1})\right) \tag{6}$$

where $\boldsymbol{\phi}_m(h_i^{n-1})$ is the mth component weight for context h_i^{n-1}. Both MAP based maximum likelihood and discriminative schemes are available to robustly estimate these weight parameters (Liu et al., 2009). Taking the ML based adaptation as an example, this is given by

$$\hat{\boldsymbol{\phi}}_m(h_i^{n-1}) = \frac{\mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau\hat{\boldsymbol{\phi}}_m(h_i^{n-2})}{\sum_m \mathcal{C}_m^{\mathsf{ML}}(h_i^{n-1}) + \tau} \tag{7}$$

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, $\hat{\boldsymbol{\phi}}_m(h_i^{n-2})$, before intersecting with a high resolution training data prior (Liu et al., 2009, 2011).

To improve robustness to the supervision quality, it is possible to use confidence score weighted sufficient statistics when estimating context independent, and context dependent interpolation weights. The log-likelihood in Eq. (6) is thus modified as

$$\ln\, \check{P}(\mathcal{W}) = \sum_{i=1}^{L} c_i \,\ln\left(\sum_{m=1}^{M} \boldsymbol{\phi}_m(h_i^{n-1}) P_m(w_i|h_i^{n-1})\right) \tag{8}$$

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 independent weights in Eq. (3). To further improve robustness during context dependent LM 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 a threshold will be pruned during weight estimation. This approach is used 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 (Hieronymus et al., 2009; Liu et al., 2010; Alumäe and Kurimo, 2010). Context dependent interpolation of LMs was performed separately at word and syllable level before the two resulting LMs are intersected using Eq. (4) to yield a final combined multi-level LM. This LM leverages both linear and log-linear forms of model combination and aims to achieve a good balance between generalisation and discrimination. Its hierarchical nature also provides a good chance to exploit the additional, non-acoustic system diversity at the word and syllable sequence level to improve system combination. Hence, it is considered in the LM cross adaptation experiments of the following section.

4.2. Neural network LM adaptation

As discussed in Section 1, the use of a continuous space representation of words in NNLMs provides a stronger generalisation 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 adaptation scheme is considered in this paper (Park et al., 2010). 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 generic NNLM, before it being specialised to, for example, a particular broadcast show. The precise location of such adaptation layer is determined by two factors. First, in conventional NNLMs fewer nodes are often used for projection and hidden layers than input and output layers (Schwenk, 2007; Emami and Mangu, 2007). Given very limited amounts of data, the number of parameters to adapt needs to be relatively small. 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 genre or broadcast show.

To reduce computational cost, conventional NNLMs only model the probabilities of a small and more frequently occurring subset of the complete vocabulary, commonly referred to as the *shortlist*. The output layer only contains nodes for in-shortlist words. A similar approach may also be used at the input layer when a large vocabulary is used, for example, for languages such as Arabic (Emami and Mangu, 2007; Park et al., 2010). 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 NNLM architecture with an additional output node explicitly modelling the probability mass of OOS words is used (Park et al., 2010). 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 Fig. 1. During adaptation, only the part of the network connecting the adaptation layer and projection layer are updated (shown in dashed lines in Fig. 1), while other parameters are fixed.

As discussed in Section 3, linear interpolation between adapted multi-level *n*-gram LMs and NNLMs using Eq. (5) can be used to obtain a good context coverage and generalisation. Such combination requires that the probability mass computed from the OOS output layer node is re-distributed among all OOS words (Park et al., 2010). This can be achieved using the *n*-gram LM statistics $P_{\mathrm{NG}}(\cdot)$ as,

$$\tilde{P}_{\mathrm{NN}}(w_i|h_i^{n-1}) = \begin{cases} P_{\mathrm{NN}}(w_i|h_i^{n-1}) & w_i \in V_{\mathrm{sl}} \\ \\ \beta(w_i|h_i^{n-1})P_{\mathrm{NN}}(w_{\mathrm{oos}}|h_i^{n-1}) & \text{otherwise} \end{cases}$$
$$\beta(w_i|h_i^{n-1}) = \frac{P_{\mathrm{NG}}(w_i|h_i^{n-1})}{\sum_{\tilde{w}_i \notin V_{\mathrm{sl}}} P_{\mathrm{NG}}(\tilde{w}_i|h_i^{n-1})} \tag{9}$$

where V_{sl} is output shortlist vocabulary, and w_{oos} the OOS word. The above normalisation can be very expensive for LVCSR tasks. To improve efficiency, assuming that the OOS probability mass assigned by the NNLM and *n*-gram LM are equal, an approximate form of normalisation can be used and this is considered in this paper.

$$\tilde{P}_{\mathrm{NN}}(w_i|h_i^{n-1}) \approx \begin{cases} P_{\mathrm{NN}}(w_i|h_i^{n-1}) & w_i \in V_{\mathrm{sl}} \\ P_{\mathrm{NG}}(w_i|h_i^{n-1}) & \text{otherwise.} \end{cases} \tag{10}$$

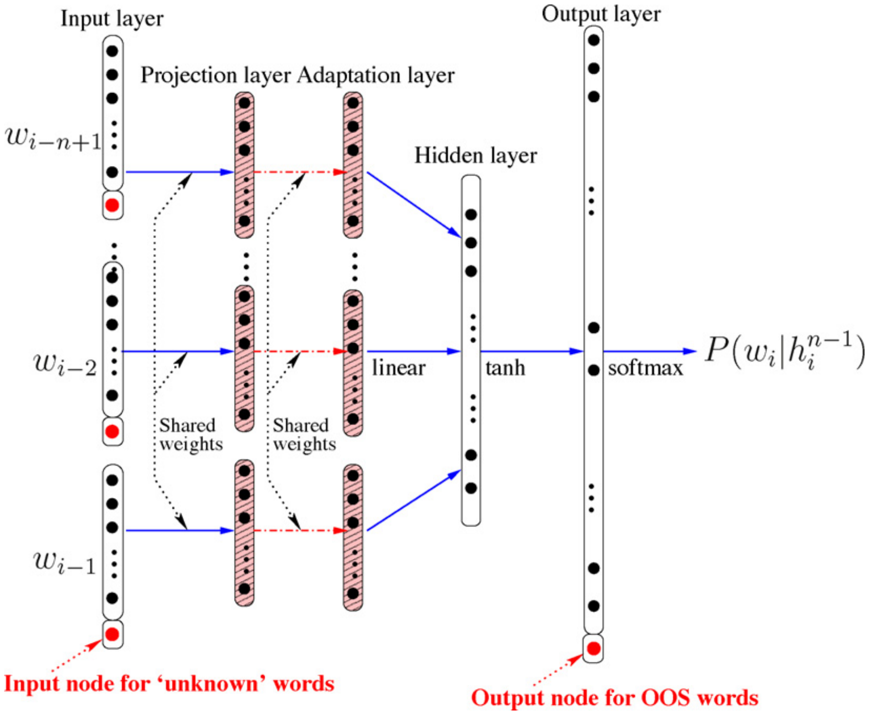

Fig. 1. Architecture of a NNLM with adaptation layer.

5. Language model cross adaptation

As discussed in Section 1, the aim of cross adaptation is to shift the decision boundary of one recognition sub-system by learning the complimentary features found in another sub-system. As these features can be exploited at multiple levels of the LVSCR modelling hierarchy, in theory many cross adaptation configurations are possible to achieve this goal. The standard cross adaptation approach only makes use of the diversity among different systems at the acoustic model level. It is also interesting to cross adapt language models to improve system combination. A suitable LM cross adaptation technique should have the following attributes: (1) sufficient power to capture complimentary features among systems; (2) inherently good generalisation; and (3) the ability to balance the trade-off between learning sufficient information and a bias to the supervision. LM cross adaptation can either be performed as a stand-alone system combination technique, or used together with acoustic model cross adaptation. One possible LM cross adaptation method considered here is based on the linear combination of adapted multi-level n-gram and neural network LMs given in Eq. (5). The combined model has all the desired features as discussed above by leveraging the strengths of both.

On one hand, the context dependent adaptation of multi-level n-gram LMs proposed in Section 4.1 has a fine modelling resolution and provides a powerful cross adaptation framework. The hierarchical nature of multi-level n-gram LMs also presents a good opportunity to exploit the additional, non-acoustic system diversity explicitly at the word and the syllable sequence level to improve system combination. In contrast, the NNLM adaptation scheme presented in Section 4.2 relies on the inserted adaptation layer whose weight parameters are globally tied over all history contexts, and thus it is potentially less powerful than the multi-level n-gram approach.

On the other hand, when using multi-level n-gram LMs for any unseen context that only occurs in the test data, component n-gram models must back-off to lower order distributions. This leads to reduced modelling resolution in the combined model, irrespective of the nature of interpolation weights being used. Such limitation may be alleviated when a combination with an NNLM component model using Eq. (5) is used, as the use of continuous space representation of words in NNLMs provides inherently stronger generalisation than n-gram LMs.

Finally, the robustness to the supervision quality, and the learning power represented by the number of distinct weight parameters to estimate, can be achieved by using confidence scores and cut-off thresholds for context dependent adaptation of multi-level *n*-gram LMs, as discussed in Section 4.1. To prevent over-fitting or a bias to the supervision during NNLM cross adaptation, a portion of the supervision data can be left as cross validation set.

6. Experiments and results

In this section various LM cross adaptation configurations are evaluated on the AGILE Chinese LVCSR systems used in the 2010 and 2011 DARPA GALE evaluations. The combined AGILE system was derived by cross adapting the Cambridge University (CU) sub-system to the outputs generated using the other sub-systems developed separately at BBN Technologies and LIMSI-CNRS. The bulk of the experiments of this section was conducted on the AGILE Chinese LVCSR systems used in the 2010 DARPA GALE evaluation. These were designed to investigate the following topics:

- performance of the multi-level *n*-gram LM adaptation and neural network LM adaptation methods presented in Sections 4.1 and 4.2 on the CU sub-systems;
- performance of the LM cross adaptation techniques proposed in Section 5 for cross site system combination;
- performance of using LM cross adaptation as a standard alone system combination technique without using acoustic model adaptation.

Finally the performance of the proposed LM cross adaptation techniques was further evaluated on the AGILE Chinese LVCSR systems used in the 2011 DARPA GALE phase 5 evaluation. For all results presented in this paper, the matched pairs sentence-segment word error (MAPSSWE) based statistical significance tests were performed at a significance level $\alpha = 0.05$.

6.1. Experiments on the 2009 AGILE Chinese LVCSR systems

The 2009 CU Chinese LVCSR system (Liu et al., 2010) was trained on 1960 h 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 as applied. A 63k word list was used. A word level 4-gram NNLM (Park et al., 2010) with an OOS output layer node was trained using 23 million words from acoustic transcription sources only. The size of the NNLM input and output vocabularies is 63k and 20k words respectively. The system uses a multi-pass recognition and system combination framework. The overall structure of the system is shown in Fig. 2. In the initial lattice generation stage, an interpolated 4-gram word level baseline back-off LM and MLLR adapted (Leggetter and Woodland, 1995) gender dependent cross-word triphone MPE (Povey and Woodland, 2002) acoustic models with HLDA (Kumar, 1997; Liu et al., 2003) projected PLP (Hermansky, 1990; Woodland et al., 1996) 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. LM adaptation was performed at the audio document level. Hierarchical and normalised perplexity smoothing priors were used to adapt the context dependent interpolation weights (Liu et al., 2009, 2011). The WFST based on-the-fly expansion algorithm presented in Section 3 was used (Liu et al., 2010). The resulting lattices were then used in a “P3” acoustic re-adaptation and lattice rescoring stage, as is shown in Fig. 2, where four different acoustic models were used:

- P3a: boosted MMI (Povey et al., 2008) GD PLP + MLP quinphone.
- P3b: MPE SAT Gaussianised PLP triphone.
- P3c: MPE GD Gaussianised PLP + MLP triphone.
- P3d: boosted MMI GD PLP quinphone.

before a final 4-way CNC combination (Liu et al., 2010). As discussed in Section 2.1, hypothesis level combination methods require that the performance of component 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 a more balanced performance among different branches. Five GALE Chinese speech test sets of mixed broadcast news (BN) and conversation (BC) genres: 2.6 h **d07**, 1 h **d08**, 3 h **d09s**, 2.6 h **p2ns** and 1.5 h **p3ns** were used.

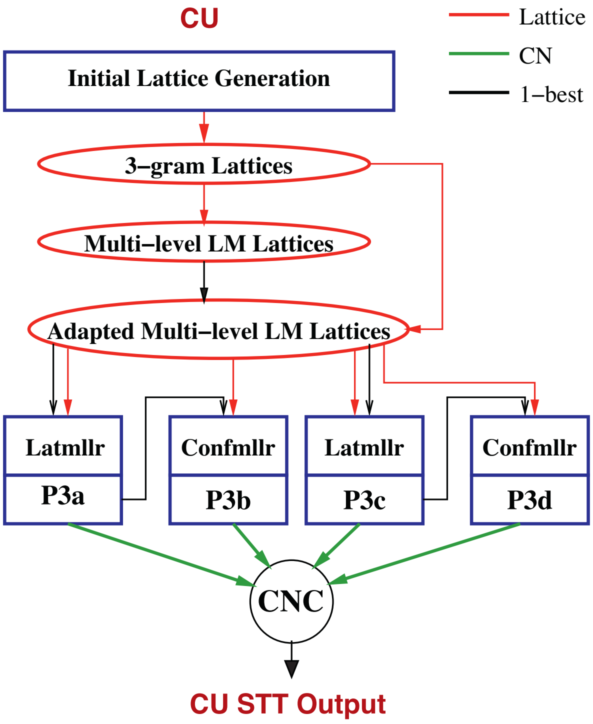

Fig. 2. The CU 2009 Chinese LVCSR system.

Table 1
CER performance of 2009 CU systems. All systems used acoustic model self-adaptation. “Base” stands for no LM adaptation, “SA” for self-adaptation. “CD” for context dependent. “Proj” for projection layer NNLM 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

6.1.1. Performance of LM adaptation

A range of language model configurations may be used to improve the performance of the CU system. As discussed in Section 3, n-gram LMs can be optionally further combined with NNLMs using Eq. (5). Adapting both or either of the two gave a total of six LM configurations to evaluate. A word level 4-gram NNLM with an out-of-shortlist output layer node was then trained using only the 23 million words of acoustic transcription sources. The size of the NNLM input and output vocabularies is 63k and 20k words respectively. A total of 400 hidden layer nodes were used. The performance of various CNC combined CU systems using these LM configuration is shown in Table 1.

The error rate of the CU 2009 system, which used only a context dependent adaptation of a multi-level n-gram LM,[4], but no NNLMs, is shown as “SA CD/-” in the 4th line of Table 1. Both context dependent multi-level n-gram LM adaptation and NNLMs gave error rate reductions and their gains were also found to be largely additive. For example, CER reductions 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 improvements over the comparable “SA CD/-” baseline. NNLM adaptation gave small CER

[4] A comparable system using a context independent interpolation weight prior during LM adaptation, similar to the method proposed in Weintraub et al. (1996) and Liu et al. (2008) rather than the context dependent weight prior given in Eq. (7) of Section 4.1, was found to lead to a slight performance degradation. Hence, context dependent weight priors are used in all experiments in this paper.

Table 2
Performance of baseline AGILE systems combined using ROVER or acoustic model only cross adaptation. “XA” for cross adaptation.

System	**d07**	**d08**	**d09s**	**p2ns**	**p3ns**
BBN	8.8	7.9	8.8	8.0	11.8
LIMSI	9.3	8.5	9.4	8.4	12.5
CU	7.9	7.3	8.4	7.6	10.6
ROVER (3 way)	7.5	7.0	8.3	7.0	10.3
ROVER (13 way)	7.3	6.8	7.8	7.1	9.8
XA (AM)	7.5	6.7	7.8	7.1	10.1

reductions 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.

6.1.2. Performance of LM cross adaptation for system combination

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 a single sub-system (Luo et al., 2009). The BBN system is more complex and used a ROVER combination between a total of 8 different sub-systems’ outputs for within site combination (Ng et al., 2008). The CER performance of the BBN and LIMSI systems are shown in the first two lines of Table 2. The CU 2009 system’s performance, previously shown in the 4th line of Table 1, is again shown in the third line of Table 2. Both the BBN and LIMSI systems used character to word segmentation schemes that are slightly different from the CU system. As discussed in Sections 2.1 and 2.2, their outputs were re-tokenised using the CU character to word segmentation scheme for cross adaptation, as well as split into character sequences for ROVER combination (Gales et al., 2007).

The *ROVER combination* performance of two AGILE systems is shown in the second section of the table. The first line 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 reductions 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, all 8 BBN component systems and the LIMSI system output. The 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 sub-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.

Table 3
Performance of 2009 AGILE systems. All systems used acoustic model cross adaptation. “Base” stands for no LM adaptation, “XA” for cross adaptation, “CI” for context independent and “CD” for context dependent. “Proj” for projection layer NNLM adaptation.

NGLM	NNLM	**d07**	**d08**	**d09s**	**p2ns**	**p3ns**
Base	–	7.5	6.7	7.8	7.1	10.1
XA CI		7.3	6.8	7.6	7.0	9.8
XA CD		7.0	6.5	7.4	6.7	9.6
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	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

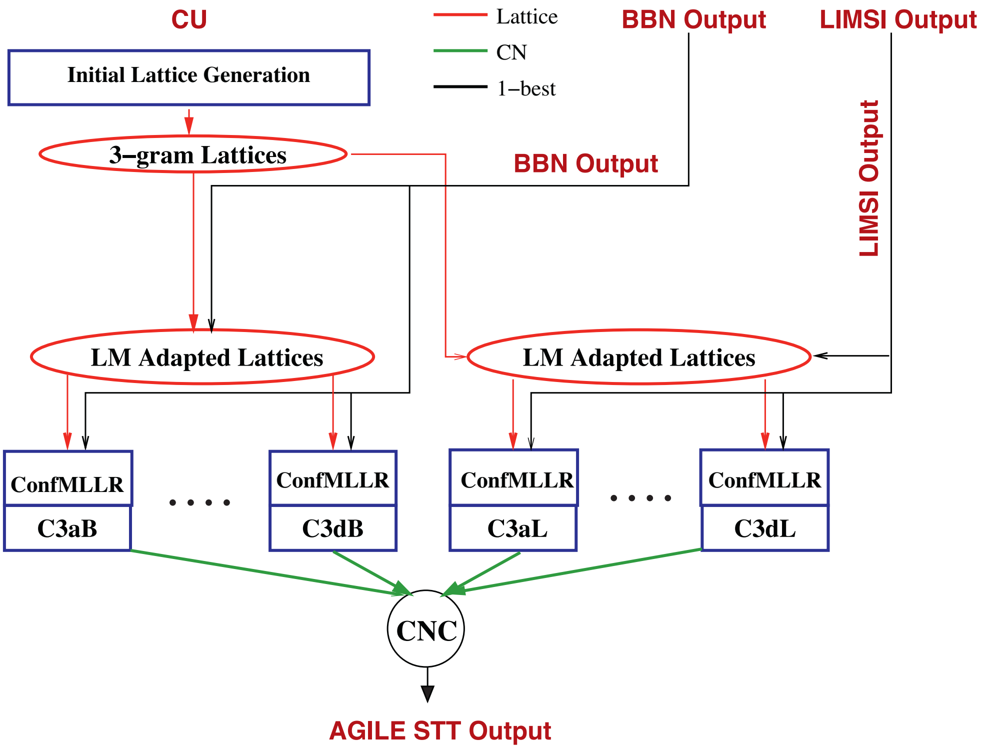

Fig. 3. Architecture of the AGILE cross adapted system.

The *cross adaptation* performance of various systems is shown in Table 3. They use either an *n*-gram LM alone, or combined with an NNLM, and further with or without LM adaptation, gives a range of combined systems shown in Table 3. The 4 CU acoustic models shown in Table 1 are 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 in total up to 4 transforms for speech states and one for silence. The resulting 8 cross adapted acoustic models were then used to rescore the lattices generated using various LM configurations before a final 8-way CNC combination. The first three systems used no NNLMs. Among these the first "Base/-" system used standard acoustic model only cross adaptation and the unadapted baseline *n*-gram LM. The performance of this system is also shown in the last line of Table 2. Compared to the 13-way ROVER combination, this baseline cross adapted system gave a comparable error rate performance.

The 2rd and 3th lines of Table 3 show the performance of two combined systems by cross adapting both the CU acoustic and multi-level 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 the audio document level using confidence score based estimation described in Section 4.1. Cross adapted word 4-gram and character 6-gram LMs were then intersected to yield the final adapted multi-level LM. The resulting two sets of cross adapted LMs were used to rebuild the CU system lattices. These were then rescored using two sets of cross adapted CU acoustic models before a final 8-way CNC combination was used as shown in Fig. 3.

The first AM + LM cross adapted "XA CI/-" system uses only context independent interpolation weights for both the 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 a 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 "XA CD/-" system used context dependent interpolation weights for both word and character layers of the CU multi-level LM. The performance of this system is shown as "XA CD/-" in the 4th line of Table 3. Further absolute CER reductions of 0.2–0.3% were obtained over the context independent LM cross adaptation based "XA CI/-" system. Significant overall CER gains of 0.4–0.5% absolute (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. Using this system the BN and BC specific performance, for example, on **d09s** are 3.5% and 11.1% respectively.

Table 4
Performance of 2009 AGILE systems. All systems used acoustic model self-adaptation. “Base” stands for no LM adaptation, “XA” for cross adaptation, “CI” for context independent and “CD” for context dependent. “Proj” for projection layer NNLM adaptation.

NGLM	NNLM	**d07**	**d08**	**d09s**	**p2ns**	**p3ns**
Base	XA Proj	7.6	6.8	8.1	7.3	10.2
XA CD	–	7.6	6.9	8.0	7.2	10.3
XA CD	Base	7.3	6.7	7.7	7.0	9.8
XA CD	XA Proj	7.3	6.7	7.7	6.9	9.8

The improvement from context dependent multi-level *n*-gram LM cross adaptation was maintained when further combined with an NNLM, as is shown in the second section of Table 3. 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, a 0.3% absolute CER reduction on **p3ns** for the “XA CD/Base” system over the “XA CD/-” baseline, compared to the improvements shown in Table 1 from the self-adapted “SA CD/Base” system over its “SA CD/-” baseline. Similar trends of reduced NNLM gains can also be found on the “Base/Base” system against the comparable “Base/-” baseline of Table 3. These trends suggest that NNLMs and acoustic model across adaptation may have a similar, but non-additive, effect on improving generalisation performance. Consistent improvements were further obtained using NNLM cross adaptation, for example, CER reductions 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 3. The overall CER reductions over the third acoustic only cross adaptation baseline “Base/Base” in Table 3, which used a fixed *n*-gram LM and NNLM, were 0.3–0.7% absolute (4.0–7.1% rel.) on **d07**, **d09s**, **p2ns** and **p3ns**.

6.1.3. Using stand-alone LM cross adaptation for system combination

So far all of the combined AGILE systems in Table 3 used acoustic model cross adaptation, irrespective of whether LM cross adaptation was performed. In order to fully evaluate the impact of LM cross adaptation on combination performance, it is necessary to de-couple these two forms of cross adaptation. In Table 4, the performance of four combined systems with acoustic model self-adaptation to CU’s outputs at the lattice generation stage shown in Fig. 2, but varying LM cross adaptation configurations, are evaluated. It is interesting to find the “XA CD/-” system, which used acoustic model self-adaptation and multi-level LM cross adaptation, gave an error rate very close to that of the “Base/-” system of Table 3, which used acoustic model cross adaptation but no LM cross adaptation. Similarly, the “XA CD/XA Proj” system, which used acoustic model self-adaptation, multi-level LM and NNLM cross adaptation, gave an error rate very close to that of the “Base/Base” system of Table 3, which used acoustic model cross adaptation, but fixed *n*-gram and neural network LMs. These results suggest that LM cross adaptation may be performed alone as a system combination technique.

6.2. Experiments on the 2010 AGILE Chinese LVCSR systems

The 2010 AGILE Chinese LVCSR system was then used to conduct a similar set of cross adaptation experiments based on the results in Table 3. The overall cross adaptation based system architecture remains the same. BBN and LIMSI provided outputs generated using their respective updated systems (Ng et al., 2010; Lamel et al., 2011). Four MPE trained position dependent (PD) phone or syllable level acoustic models (Liu et al., 2011) 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 Gaussianised PLP tri-syllable.
- P3c: GD Gaussianised PLP + MLP PD triphone.
- P3d: GD PLP full covariance PD quinphone.

Three new GALE Chinese speech test sets: 6 h **d10c**, 12 h **d10r**, 12 h **d10d** and 3 h **p4ns** were also used. The site specific and cross adapted performance of the 2010 AGILE systems are shown in Tables 5 and 6 respectively. Based

Table 5
Performance of 2010 AGILE sub-systems.

System	**d09s**	**d10c**	**d10r**	**d10d**	**p4ns**
CU 2009	8.4	12.5	7.9	24.5	7.7
BBN	7.8	12.2	7.9	24.6	7.2
LIMSI	8.5	13.4	8.3	25.5	7.9
CU	7.8	11.9	7.6	23.7	7.1

Table 6
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.

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

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. Statistically significant CER reductions of 0.6–0.8% absolute (3.3–7.8% relative) were obtained over the 2009 CU system (shown in the first line of Table 5) on **d09s**, **d10c**, **d10c** and **p4ns**.

As shown in Table 6, cross adapting both the *n*-gram LM and NNLM gave a total CER reductions 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 an unadapted *n*-gram LM and NNLM. 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 improvements since the 2009 AGILE combined system (4th line of Table 3 and 1st line of Table 6) are 0.3–0.9% (4.0–5.4% rel.) absolute. The combination gains over the 2010 CU system of Table 5 are 0.6–2.0% (8.4–10.3% rel.). This "XA CD/XA Proj" system of Table 6 was used in the 2010 GALE evaluation.

6.3. Discussion

A range of experiments were conducted to evaluate performance of the LM adaptation and LM cross adaptation techniques presented in this paper. Experimental results suggest the following:

- the multi-level *n*-gram LM adaptation and neural network LM adaptation methods presented in Sections 4.1 and 4.2 are useful to improve the performance of state-of-the-art LVCSR systems;
- the LM cross adaptation techniques proposed in Section 5 can capture additional complementary features among highly diverse sub-systems, thus are useful for improving LVCSR system combination performance;
- LM cross adaptation techniques can also be used as stand-alone system combination method. They give comparable performance to using conventional acoustic model adaptation only.

7. Conclusion

Language model cross adaptation was investigated in this paper to improve LVCSR system combination. It can either be performed as a stand-alone system combination technique, or used together with acoustic model cross adaptation. Three forms of language models, a multi-level LM that models both syllable and word sequences, a word level neural network LM, and the linear combination of the two were cross adapted. Experimental results on a state-of-the-art speech recognition task suggest complimentary features exist on multiple layers of the modelling hierarchy among highly diverse sub-systems. The significant error rate reductions obtained over ROVER and acoustic model only cross adaptation confirm that the proposed LM cross adaptation technique is highly effective in capturing additional useful diversity among sub-systems to improve system combination performance. Future research will focus on improving robustness in cross adaptation and system architecture refinement.

Acknowledgements

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.

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

References

Alumäe, T., Kurimo, M.,2010. Domain adaptation of maximum entropy language models. In: Proc. of the ACL 2010 Conference Short Papers. Association for Computational Linguistics, pp. 301–306.

Anastasakos, T., Balakrishnan, S.V., 1998. The use of confidence measures in unsupervised adaptation of speech recognizers. In: Proc. ICSLP'98, Sydney.

Bengio, Y., Ducharme, R., 2003. A neural probabilistic language model. Advances in Neural Information Processing Systems 3, 1137–1155.

Blei, D., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. Journal of Machine Learning Research 3.

Bulyko, I., Ostendorf, M., Stolcke, A., 2003. Getting more mileage from web text sources for conversational speech language modeling using class-dependent mixtures. In: Proc. HLT'03, Edmonton.

Bulyko, I., Matsoukas, S., Schwartz, R., Nguyen, L., Makhoul, J., 2007. Language model adaptation in machine translation from speech. In: Proc. IEEE ICASSP2007, Hawaii.

Chen, L., Gauvain, J.-L., Lamel, L., Adda, G., Adda, M., 2001. Language model adaptation for broadcast news transcription. In: Proc. ITRW'01, Paris.

Chien, J.T., Wu, M.S., Wu, C.S., 2005. Bayesian learning for latent semantic analysis. In: Proc. ISCA Interspeech'05, Lisbon.

Chu, S.M., et al., 2010. The 2009 IBM GALE Mandarin broadcast transcription system. In: Proc. IEEE ICASSP2010, Dallas.

Darroch, J., Ratcliff, D., 1972. Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics 43 (5), 1470–1480.

Emami, A., Mangu, L., 2007. Empirical study of neural network language models of Arabic speech recognition. In: Proc. IEEE ASRU2007, Kyoto.

Evermann, G., Woodland, P.C., 2000. Posterior probability decoding, confidence estimation and system combination. In: Proc. Speech Transcription Workshop.

Federico, M., 1999. Efficient language model adaptation through MDI estimation. In: Proc. EuroSpeech'99, Budapest.

Federico, M., 2003. Language model adaptation through topic decomposition and MDI estimation. In: Proc. IEEE ICASSP2003, Hong Kong.

Fiscus, J.G., 1997. A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER). In: Proc. IEEE ASRU1997.

Gales, M.J.F., Airey, S.S., 2006. Product of Gaussians for speech recognition. Computer Speech and Language 20 (January (1)), 22–40.

Gales, M.J.F., et al., 2007. Speech system combination for machine translation. In: Proc. IEEE ICASSP2007, Hawaii.

Gales, M.J.F., 1998. Maximum likelihood linear transformations for HMM-based speech recognition. Computer Speech and Language 12 (2), 75–98.

Gildea, D., Hofmann, T., 1999. Topic-based language models using EM. In: Proc. Eurospeech'99, Budapest.

Hermansky, H., 1990. Perceptual linear prediction (PLP) of speech. Journal of the Acoustic Society of America 87 (4), 1738–1752.

Hieronymus, J.L., Liu, X., Gales, M.J.F., Woodland, P.C., 2009. Exploiting Chinese character models to improve speech recognition performance. In: Proc. ISCA Interspeech'09, Brighton.

Hinton, G., 1999. Products of experts. In: Proc. ICANN.

Hinton, G., 2002. Training products of experts by minimizing contrastive divergence. Neural Computation 14, 1771–1800.

Hsu, B., 2007. Generalized linear interpolation of language models. In: Proc. IEEE ASRU2007, Kyoto.

Kumar, N., 1997. Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition. Ph.D. Thesis. John Hopkins University, Baltimore.

Lamel, L., et al., 2011. Improved models for Mandarin speech-to-text transcription. In: Proc. IEEE ICASSP2011, Prague.

Leggetter, C.J., Woodland, P.C., 1995. Maximum likelihood linear regression for speaker adaptation of continuous density HMMs. Computer Speech and Language 9, 171–186.

Lei, X., et al., 2009. Development of the 2008 SRI Mandarin speech-to-text system for broadcast news and conversation. In: Proc. ISCA Interspeech'09, Brighton.

Liu, X., Gales, M.J.F., Woodland, P.C., 2003. Automatic complexity control for HLDA systems. In: Proc. IEEE ICASSP2003, vol. 1, Hong Kong, pp. 132–135.

Liu, X., Gales, M.J.F., Woodland, P.C., 2008. Context dependent language model adaptation. In: Proc. ISCA Interspeech'08, Brisbane.

Liu, X., Gales, M.J.F., Woodland, P.C., 2009. Use of contexts in language model interpolation and adaptation. In: Proc. ISCA Interspeech'09, Brighton.

Liu, X., Gales, M.J.F., Hieronymus, J.L., Woodland, P.C., 2010. Language model combination and adaptation using weighted finite state transducers. In: Proc. IEEE ICASSP2010, Dallas.

Liu, X., Gales, M.J.F., Woodland, P.C., 2010. Language model cross adaptation for LVCSR system combination. In: Proc. ISCA Interspeech'10, Makuhari.

Liu, X., Gales, M.J.F., Hieronymus, J.L., Woodland, P.C., 2011. Investigation of acoustic units for LVCSR systems. In: Proc. IEEE ICASSP2011, Prague.

Liu, X., Gales, M.J.F., Woodland, P.C. Use of contexts in language model interpolation and adaptation. Computer Speech and Language, in press.

Luo, J., Lamel, L., Gauvain, J.-L., 2009. Modeling characters versus words for Mandarin speech recognition. In: Proc. ICASSP2009, Taipei.

Mohri, M., Riley, M., 1998. Network optimizations for large vocabulary speech recognition. Speech Communication 25 (3).

Mrva, D., Woodland, P.C., 2004. A PLSA-based language model for conversational telephone speech. In: Proc. ISCA Interspeech'04, Jeju.

Ng, T., et al., 2008. Progress in the BBN 2007 Mandarin speech to text system. In: Proc. IEEE ICASSP2008, Las Vegas.

Ng, T., et al., 2010. Jointly optimized discriminative features for speech recognition. In: Proc. ISCA Interspeech'10, Makuhari.

Och, F.J., Ney, H., 2002. Discriminative training and maximum entropy models for statistical machine translation. In: Proc. ACL02', Philadelphia, pp. 295–302.

Park, J., Liu, X., Gales, M.J.F., Woodland, P.C., 2010. Improved neural network based language modelling and adaptation. In: Proc. ISCA Interspeech'10, Makuhari.

Peskin, B., et al., 1999. Improvements in recognition of conversational telephone speech. In: Proc. IEEE ICASSP1999, Phoenix.

Povey, D., Woodland, P.C., 2002. Minimum phone error and I-smoothing for improved discriminative training. In: Proc. IEEE ICASSP2002, Orlando.

Povey, D., Kanevsky, D., Kingsbury, B., Ramabhadran, B., Saon, G., Visweswariah, K., 2008. Boosted MMI for model and feature-space discriminative training. In: Proc. IEEE ICASSP2008, Las Vegas, pp. 4057–4060.

Prasad, R., et al., 2005. The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system. In: Proc. ISCA Interspeech'05, Lisboa.

Rosenfeld, R., Chen, S.F., Zhu, X., 2001. Whole-sentence exponential language models: a vehicle for linguistic–statistical integration. Computers Speech and Language 15 (1).

Schwartz, R., et al., 2004. Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system. In: Proc. IEEE ICASSP2004, Montreal.

Schwenk, H., 2007. Continuous space language models. Computer Speech and Language 21 (July (3)), 492–518.

Tam, Y.C., Schultz, T., 2005. Dynamic language model adaptation using variational Bayes inference. In: Proc. ISCA Interspeech'05, Lisboa.

Weintraub, M., et al.,1996. LM95 project report: fast training and portability. In: 1995 Language Modeling Summer Research Workshop Technical Reports, Research Note 1. Center for Language and Speech Processing, Johns Hopkins University, Baltimore, http://www-speech.sri.com/cgi-bin/run-distill?papers/lm95-report.ps.gz (accessed 26.03.12).

Woodland, P.C., Leggetter, C.J., Odell, J.J., Valtchev, V., Young, S.J., 1995. The 1994 HTK large vocabulary speech recognition system. In: Proc. IEEE ICASSP1995, Detroit.

Woodland, P.C., Gales, M.J.F., Pye, D., Young, S.J., 1996. The development of the 1996 HTK broadcast news transcription system. In: Proc. DARPA Speech Recognition Workshop, Arden House, NY, US, pp. 73–78.

Woodland, P.C., et al., 2004. SuperEARS: multi-site broadcast news system. In: Proc. Rich Transcription Workshop 2004, Palisades, NY.

Wallhoff, F., Willett, D., Rigoll, G., 2000. Frame discriminative and confidence-driven adaptation for LVCSR. In: Proc. IEEE ICASSP2000, Istanbul.