The Cambridge University 2014 BOLT Conversational Telephone Mandarin Chinese LVCSR System for Speech Translation

Xunying Liu, Federico Flego, Linlin Wang, Chao Zhang, Mark Gales & Philip Woodland

University of Cambridge Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ, U.K.

{xl207,ff257,lw519,cz277,mjfg,pcw}@eng.cam.ac.uk

Abstract

This paper presents the development of the 2014 Cambridge University conversational telephone Mandarin Chinese LVCSR system for the DARPA BOLT speech translation evaluation. A range of advanced modelling techniques were employed to both improve the recognition performance and provide a suitable integration with the translation system. These include an improved system combination technique using frame level acoustic model combination via joint decoding. Sequence trained deep neural network (DNN) based hybrid and tandem systems were combined on-the-fly to produce a consistent decoding output during search. A multi-level paraphrastic recurrent neural network LM (RNNLM) modelling both alternative paraphrase expressions and character sequences while preserving a consistent character to word segmentation was also used. This system gave an overall character error rate (CER) of 29.1% on the BOLT **dev14** development set.

Index Terms: conversational speech transcription, speech translation, system combination, RNNLM, character LM

1. Introduction

Conversational telephone speech (CTS) translation still remains a challenging task to date. Highly variable speech data is collected under limited bandwidth and mixed with complex background acoustic events. The informal speech style and rich choice of expressions introduce further variation. When only limited in-domain training material is available, this task becomes even more challenging. Advanced modelling techniques and system combination approaches are required to handle the above rich variabilities and obtain state-of-the-art speech recognition performance. The quality of the speech recognition system outputs significantly impacts the performance of downstream machine translation (MT) systems [1]. In addition to a desired minimum error rate, there are also important design issues that need to be considered when developing speech recognition systems for speech translation.

First, conventional system combination techniques normally perform hypothesis combination at word level using a combination of voting counts and confidence scores [2, 3]. As a consistent sequence level decoding output from any component speech recognizers is no longer retained, the resulting combined recognition outputs are likely to be sub-optimal for the downstream translation system. Hence, system combination methods that can produce a single consistent recognition output are preferred. Second, a consistent form of word tokenization is required for both the recognition and translation stages for the Mandarin speech translation task considered in this paper [1, 4]. The Chinese language is syllable based and has no natural word boundaries available. It is important to use a common consistent word segmentation during both speech recognition and translation [4]. As ambiguity can occur in the character to word segmentation process [5], language modelling techniques that not only use a single consistent character to word segmentation, but also can implicitly learn alternative word segmentations, for example, incorporating character sequence LMs [6], are preferred.

This paper presents the development of the 2014 Cambridge University conversational telephone Mandarin Chinese large vocabulary continuous speech recognition (LVCSR) system for the DARPA Broad Operational Language Translation (BOLT) speech translation evaluation. A range of techniques were employed in the system to take the above issues into account. These include the use of conversation side based vocal tract length normalization (VTLN) [7] and cepstral parameter normalization in front-end processing; improved pitch extraction and smoothing for Mandarin; efficient HTK based sequence level discriminative training [8] of deep neural network [9] based hybrid [10] and tandem [11, 12] acoustic models; a multi-level paraphrastic recurrent neural network LM (RNNLM) [13, 14] that models both alternative paraphrase expressions and character sequences; improved system combination using frame level acoustic model combination. The evaluation system gave an overall character error rate (CER) of 29.1% on the BOLT **dev14** development set.

The rest of this paper is organized as follows. The description of the acoustic training and test data used for system development is presented in section 2. The acoustic modelling, language modelling techniques and system combination approaches used to construct the system are described in sections 3, 4 and 5. The evaluation system's architecture and performance are presented in section 7. Section 8 concludes and discusses possible future work.

2. Task Description

This section describes various data resources that were used for the development of the CU Mandarin CTS LVCSR system.

2.1. Acoustic training, test data and audio segmentation

Acoustic models were trained on 301 hours of Mandarin Chinese conversational telephone speech data released by the LDC for the DARPA BOLT program, **bolt14train**. These include 32 hours of Call Home Mandarin data (CHM), 56 hours of Call Friend Mandarin data (CFM) and additional 213 hours of con-

The research leading to these results was supported by DARPA under the Broad Operational Language Translation (BOLT) program. The paper does not necessarily reflect the position or the policy of US Government and no official endorsement should be inferred.

versation telephone Mandarin speech collected by Hong Kong University of Science and Technology (HKUST). The training data set consists of a total of 1479 conversations. Among these, 253 CHM and CFM conversations contain multiple speakers per conversation side. A 4.5 hour BOLT development set of Mandarin Chinese conversational telephone speech data **dev14**, consisting of 57 speakers from a total of 19 conversations, was used for performance evaluation. Manual audio segmentation was also used to allow translation outputs to be accurately scored. A 72 hour training subset of the 301 hour full set used in the NIST RT04 Mandarin system, **rt04train**, and an associated 2 hour development set **dev04** containing 24 conversations were also used in the initial system development.

2.2. Lexicon, character to word segmentation and phone set

A 63k recognition word list was used in decoding. It consists of a total of approximately 52k multiple character Chinese words, 5k single character Chinese words and additional 5k frequent English words. A 44k subset of the 52k multiple character Chinese words were obtained using an LDC released Mandarin Chinese lexicon. A left to right maximum word length based character to word segmentation method [15] based on the above 52k multiple character words was applied to text data. The resulting character to word segmented acoustic transcripts contain on average 1.426 characters per word. A base phone set containing 46 toneless (124 tonal) phones was used [15].

2.3. Language model training data

The baseline 4-gram back-off LM was trained using a total of 1 billion words of text data from the following two types of text sources: 2.6M words of data from the acoustic transcripts; 1 billion words of additional web data collected by various research sites including Cambridge University, IBM Research, SRI and University of Washington under the DARPA EARS and GALE programs [1]. Numeric terms were first converted into spoken forms before the left to right maximum word length based character to word segmentation scheme described in section 2.2 was applied. A 120 million word subset of the 1 billion word full set used in the earlier RT04 CU Mandarin system [16] was also used in the initial system development.

3. Acoustic Modelling

3.1. Front-end processing

An important part of the speaker level diversity in conversational speech can be attributed to the variation of vocal tract length [7]. The first-order effect of a difference in vocal tract length can be approximated via a scaling of the formant positions. A female speaker, for example, can exhibit formants roughly 20% higher than those of a male speaker. In order to handle this problem, vocal tract length normalization (VTLN) [7] was used. VTLN is performed in a supervised mode on the training data and unsupervised manner on the test data. A maximum likelihood (ML) frequency scaling factor of the speech spectrum is estimated at a speaker level before being applied to the spectrum to produce normalized PLP [17] features. Cepstral mean and variance normalization was also used to further remove speaker level variability. The advantage of VTLN lies in its low complexity and effectiveness. It can also be efficiently implemented and applied to a range of back-end acoustic models considered in this paper, conventional GMM-HMMs, DNN hybrid and tandem systems.

In tonal languages like Mandarin Chinese, prosodic pitch variation occurs at a sentence level in the form of long and smooth contours where short and sharp lexical tones are superimposed. It is therefore important to incorporate pitch features into the acoustic front-end. Pitch features were extracted and smoothed using the Kaldi toolkit [18]. The pitch parameter along with the first and second-order differentials were mean and variance normalized at a speaker level before being augmented to the HLDA [19, 20] projected and speaker level normalized PLP. This gives a feature vector of 42 dimensions.

3.2. Baseline HMM systems

The baseline tonal triphone context dependent GMM-HMM systems was constructed using the above 42 dimensional front-end described in section 3.1. Phonetic decision tree state clustering [21] was used. In order to model complex phonological variation patterns such as tone sandhi and glottalisation, word position information was also used during decision tree tying [22]. After incorporating word level position information, the number of tonal phones is increased from 124 to 293. As expected, the use of tonal and word position dependent questions dramatically increases the number of context dependent phone units to consider during both training and decoding. As not all of them are allowed by the lexicon, only the valid subset under the lexical constraint is retained, after applying the context filtering approach proposed in [22]. The system contains a total of 12k tied HMM states with 28 Gaussians per state on average. MPE [23] based HMM parameter estimation and speaker adaptive training were performed. CMLLR [24] based speaker adaptive training (SAT) was also used. Unsupervised MLLR [25] based speaker adaptation was used.

3.3. HTK based hybrid DNN-HMM systems

The baseline HMM systems were then used to produce state level alignment to train hybrid DNN systems using an extended version of the HTK toolkit [26]. The artificial neural network (ANN) extension to HTK [8] can handle the training, classification, and decoding of various types of ANNs along with their mixtures. It also aims to have an integrated and efficient solution for hybrid and tandem system construction, while keeping compatibility with all previous HTK features. In this extension, a general ANN definition to cope with different types of models is adopted. Each network layer can have its input formed by a mixture of acoustic features or the outputs from any previously defined layers, while each of them allowed to have context expanded independently from adjacent frames. This allows HTK to handle a very flexible structure formed by connecting different ANN layers. The only topological constraint is that any ANN must be represented as a directed acyclic graph (DAG). The ANN extension also retains the highly modular structure of the original HTK toolkit so that the whole ANN module is treated as a front-end to the back-end HMMs. Currently both frame level DNN training using stochastic gradient descent (SGD) and sequence level discriminative training are supported. These will be included in the next public HTK release.

DNNs with five hidden layers were first trained using the cross entropy (CE) criterion on a GPU before MPE based sequence level discriminative training was performed. A layer by layer discriminative pre-training was used. The first 4 hidden layers have 2000 nodes while the 5th hidden layer use 1000 nodes. 12k output layer nodes were used. 56 dimensional input features including normalized PLP features with their differentials up to the 3rd order and pitch parameters were used. The input vector thus had 504 dimensions, which was produced by concatenating the current frame with 4 frames from both left and right contexts. A tenth of the training set was randomly selected as the held-out set for cross-validation.

3.4. Tandem systems

An alternative approach to incorporate DNNs into HMM based acoustic models is to uses a DNN as a feature extractor, trained to produce phoneme posterior probabilities. The resulting probabilistic features [11], or bottleneck features [12, 27] are used to train standard GMM-HMMs in a tandem fashion. As these features capture additional discriminative information complementary to standard front-ends, they are often combined via feature concatenation. As GMM-HMMs remain as the back-end classifier, the tandem approach requires minimum change to the downstream techniques, such as speaker adaptation and decoding, while the useful information represented by the bottleneck features can also be retained. They also provide additional useful system diversity for a combination with hybrid DNN-HMM systems [28]. DNNs with an additional bottleneck layer were trained using the same procedure described in section 3.4, before 26 dimensional bottleneck features were extracted. The resulting features were then normalized at a speaker level before de-correlated via a STC [30] transform and augmented to the standard acoustic front-ends and used in training of the back-end GMM-HMMs.

4. Language Modelling

4.1. Baseline interpolated 4-gram LM

This baseline 4-gram word level LM was trained using the 1 billion word text data and the 63k word list described in sections 2.2 and 2.3. Modified KN smoothed 4-gram LMs were estimated on the acoustic transcription data and web data sources separately before a linear interpolation was used to combine them. The interpolation weights were perplexity optimized on **dev14**, **dev04** and additional CHM and CFM data from earlier NIST evaluation sets **eval03** and **eval97**. The interpolated 4-gram LM has a total of 48M 2-grams, 133M 3-grams and 143M 4-grams. It gave a perplexity of 151 on **dev14**.

4.2. Efficient RNNLM training and lattice rescoring

An important part of the language modelling problem for speech recognition systems, and many other related applications, is to appropriately model long-distance context dependencies in natural languages. Along this line, LMs that can model longer span history contexts, for example, recurrent neural network LMs (RNNLMs) [31], have become increasingly popular for state-of-the-art LVCSR systems. In this paper, RNNLMs with non-class based, full vocabulary output layer were efficiently trained on GPU in a bunch mode [32]. An out-of-shortlist (OOS) node was also used at the output layer to model the probability mass assigned to OOS words. A detailed description of the full output RNNLM architecture can be found in [32]. A total of 512 hidden layer nodes were used. A 27k word input layer vocabulary and 20k word output layer shortlist were also used.

As RNNLMs use a complex vector space representation of full history contexts, it is non-trivial to apply them in the early stage of ASR systems, or to directly rescore the word lattices produced by them. Instead, N-best list rescoring was normally used [31, 33]. This practical constraint limits the possible improvements that can be obtained from RNNLMs for downstream applications that favor a more compact lattice representation, for example, confusion network (CN) decoding techniques [3, 34]. In order to address this issue, two efficient RNNLM lattice rescoring algorithms were proposed in [35]. The first uses an n-gram style approximation of history contexts. In this paper, this RNNLM rescoring approach was used.

4.3. Multi-level paraphrastic RNNLM

Linguistic factors influencing the realization of surface word sequences, for example, expressive richness, are only implicitly learned by RNNLMs. Observed sentences and their associated alternative paraphrases representing the same meaning are not explicitly related during training. In order to further improve the RNNLM's coverage and generalization, the 2.6M words of acoustic transcripts data were augmented with 15M words of its paraphrase variants. These paraphrases were automatically produced using the statistical paraphrase induction and generation method described in [13]. The above combined data set was then used to train a paraphrastic RNNLM [14]. In order to incorporate richer linguistic constraints, LMs that model different units, for example, syllables, words, or phrases, can be log-linearly combined in the form a multi-level LM [6, 37] to improve discrimination. In this paper, a multi-level paraphrastic RNNLM modelling both word and character sequences was constructed. It also aims to implicitly model alternative character to word segmentations, while retaining a consistent character to word segmentation. This is a useful feature for the downstream machine translation system, as discussed in section 1.

5. System Combination

State-of-the-art LVCSR systems often use system combination techniques [38, 39]. 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 [2] and confusion network combination (CNC) [3]. As discussed in section 1, hypothesis level combination is unable to retain a consistent decoding output from component systems. Alternatively the second category based on cross adaptation [41, 38, 39, 42] can be used. The acoustic and/or language models [36], of one system are adapted to the recognition outputs of another. A consistent decoding output can then be produced by decoding using the cross adapted system. Cross adaptation requires the component system to be adapted having a comparable or lower error rate than the supervision system. When this assumption is invalid, cross adaptation can lead to sub-optimal combination performance.

In order to address this issue, an improved system combination based on frame level acoustic model combination is used. The state output probabilities of an hybrid DNN-HMM system and a comparable tandem system is log-linearly combined with a weighting of 1:0.4 on-the-fly in a joint decoding [28, 29]. This combination approach has three advantages over hypothesis level combination and cross adaptation based combination schemes. First, as component acoustic model scores are dynamically combined during recognition, a consistent decoding output can then be produced for the downstream machine translation system. Second, the sensitivity towards to the error rate difference between component systems can be reduced by appropriately setting the log-linear interpolation weights. Third, as component systems are log-linear combined via an intersection of probabilities, only a reduced search space is active during recognition. This can significantly reduce the overall run time for the combined system.

6. Development Results

The performance of various HMM and tandem SAT systems trained on the 72 hour training subset **rt04train** evaluated on RT04 **dev04** are shown in table 1. Consistent improvements were obtained using VTLN for both baseline HMM and hybrid systems. The improved pitch feature extraction using Kaldi also gave further character error rate (CER) reductions over the baseline pitch feature extraction used in [15, 37].

	Front-end Processing		**dev04**
System	**VTLN**	**Kaldi Pitch**	**CER%**
HMM SAT	×	×	35.9
	√	×	34.5
	√	√	33.4
Tandem SAT	×	×	29.9
	√	×	29.2
	√	√	29.1

Table 1: Performance of HMM and tandem systems trained on 72 hour training subset **rt04train** evaluated on RT04 **dev04**.

The performance of the baseline HMM, hybrid and tandem systems trained on the 301 hour **bolt14train** data evaluated on **dev14** using the baseline 1 billion word trained 4-gram LM and the multi-level paraphrastic RNNLM are shown in table 2. Significant performance improvements of 2.8% absolute were obtained on the hybrid SI system using the HTK based DNN MPE training of section 3.3. The best performance was obtained using a hybrid MPE SAT system shown in the last line in table 2. It was trained using the same tandem features of the tandem SAT system in table 2. This allows CMLLR transforms to be shared between the two systems during SAT training and adaptation.

		dev14 CER%	
System	**MPE**	**4-gram**	**+rnn**
HMM SAT	√	43.8	-
Tandem SAT	√	33.2	31.8
Hybrid SI	×	34.5	33.2
	√	31.8	30.4
Hybrid SAT	√	**31.4**	**29.9**

Table 2: Performance of baseline HMM, hybrid and tandem systems trained on 301 hour **bolt14train** on **dev14** set using baseline 4-gram LM and multi-level paraphrastic RNNLM.

A detailed analysis of the improvements obtained from the multi-level paraphrastic RNNLM on the tandem SAT system in table 2 are shown in table 3. Consistent improvements in both perplexity and error rate were obtained using the paraphrastic RNNLM over the baseline RNNLM trained using the original acoustic transcripts only. Using the multi-level paraphrastic RNNLM that models both alternative paraphrases and character sequences gave an overall CER reductions of 1.4% absolute over the baseline 4-gram LM trained on 1 billion words of data.

The performance of various combined systems on **dev14** are shown in table 4. The first section contains three combined systems derived from the hybrid SI and tandem SAT systems in table 2. The CNC combined and cross adapted systems are shown in the first two lines. The CER difference of the hybrid SI and tandem SAT systems before combination are quite large, as were shown in table 2. CNC combination gave the lowest CER of 29.6%. Using the joint decoding method of section 5, the same error rate was obtained. Both CNC combination and cross adapted systems require the component hybrid SI and tandem SAT systems to be used in decoding in parallel for CNC, or in sequence for cross adaptation. The joint decoding combined system only performed a single search and was 2.6x faster than both the CNC and cross adapted systems.

		dev14	
System	**LM**	**PPlex**	**CER%**
Tandem SAT	4-gram	151	33.2
	+RNNLM	134	32.6
	+para. RNNLM	127	32.2
	+para. multi-level RNNLM	-	**31.8**

Table 3: Performance of baseline 4-gram, RNNLM, paraphrastic RNNLM and multi-level paraphrastic RNNLM on **dev14**.

Combined System	**Combi.**	**dev14**
Hybrid SI ⊕ Tandem SAT	CNC	29.6
Hybrid SI → Tandem SAT	XAdapt	30.0
Hybrid SI ⊗ Tandem SAT	JointDec	**29.6**
Hybrid SAT ⊗ Tandem SAT	JointDec	**29.1**

Table 4: Performance of various combined systems evaluated on **dev14** set using component sub-systems in table 2. "⊕", "→" and "⊗" denote CNC, cross adaptation and joint decoding.

7. Evaluation System

The CU evaluation system used a multi-pass recognition framework. In the first pass the hybrid SI system of table 2 and the baseline 4-gram LM of table 3 were used to produce initial recognition outputs. These were then used to adapt both the hybrid SAT and tandem SAT systems of table 2. The joint decoding method described in section 5 was then use to combine these two systems on-the-fly at test time. After lattice rescoring using the paraphrastic multi-level RNNLM of table 3, CN decoding produced the final output and gave a CER score of 29.1% on the **dev14** set, as is shown in 4th line in table 4.

8. Conclusions

This paper described the development of the 2014 Cambridge University conversational telephone Mandarin LVCSR system used for the DARPA BOLT speech translation evaluation. A range of advanced modelling techniques including an improved system combination scheme using frame level acoustic model combination and multi-level paraphrastic RNNLMs were were used to achieve both an optimal recognition performance and a suitable integration with the translation system. Future research will focus on improving model based system combination.

9. References

[1] J. Olive, C. Caitlin and J. McCary eds. "Handbook of natural language processing and machine translation: DARPA global au-

tonomous language exploitation", Springer, 2011.

[2] J. G. Fiscus (1997). "A post-processing system to yield reduced word error rates: recogniser output voting error reduction (ROVER)," in *Proc. IEEE ASRU*, Santa Barbara, CA, pp. 347–354.

[3] G. Evermann and P. C. Woodland (2000), "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, College Park, MD, 2000.

[4] M. J. F. Gales, X. Liu, R. Sinha, P. C. Woodland, K. Yu, S. Matsoukas, T. Ng, K. Nguyen, L. Nguyen, J.-L. Gauvain, L. Lamel, A. Messaoudi (2007). "Speech System Combination for Machine Translation, " in *Proc. IEEE ICASSP*, Honolulu, HI, 2007, vol. 4, pp. 1277–1280.

[5] R. Sproat, C. Shih, N. Chang, and W. Gale. (1996). A stocastic finite-state word-segmentation algorithm for Chinese, in *Computational Linguistics*, Vol. 22, Issue, 3, 1996, pp. 377–404.

[6] X. Liu, J. L. Hieronymus, M. J. F. Gales and P. C. Woodland (2013). "Syllable language models for Mandarin speech recognition: exploiting character sequence models", *Journal of the Acoustical Society of America*, Volume 133, Issue 1, pp. 519-528, January 2013.

[7] L. Lee, and R. C. Rose (1996) "Speaker normalization using efficient frequency warping procedures," in *Proc. IEEE ICASSP* , Atlanta, GA, 1996, vol. 1, pp. 353–356.

[8] C. Zhang, and P. C. Woodland (2015). "A general artificial neural network extension for HTK", in submission to *ISCA Interspeech*.

[9] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pretrained deep neural networks for large vocabulary speech recognition", in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, January 2012.

[10] H. A. Bourlard and N. Morgan (1993). "Connectionist speech recognition: a hybrid approach", Kluwer Academic Publishers, Norwell, MA, USA, 1993.

[11] H. Hermansky, D. Ellis and S. Sharma (2000). "Tandem connectionist feature extraction for conventional HMM systems", in *Proc. IEEE ICASSP*, Istanbul, Turkey, vol. 3, pp. 1635–1638.

[12] D. Yu and M. L. Seltzer (2011). "Improved bottleneck features using pretrained deep neural networks", in *Proc. ISCA Interspeech*, Florence, Italy, 2011, pp. 237–240.

[13] X. Liu, M. J. F. Gales, and P. C. Woodland (2014). "Paraphrastic language models", *Computer Speech and Language*, vol. 28, Issue 6, pp. 1298–1316, November 2014.

[14] X. Liu, M. J. F. Gales, and P. C. Woodland (2015), "Paraphrastic recurrent neural network language models," in *Proc. IEEE ICASSP*, Brisbane, Australia, 2015.

[15] R. Sinha, M. J. F. Gales, D. Y. Kim, X. Liu, K. C. Sim, and P. C. Woodland (2006). "The CU-HTK Mandarin broadcast news transcription system," in *Proc. IEEE ICASSP*, Toulouse, France, 2006, vol. 1, pp. 1077–1080 .

[16] M. J. F. Gales, B. Jia, X. Liu, K. C. Sim, P. C. Woodland, and K. Yu (2005). "Development of the CUHTK 2004 Mandarin conversational telephone speech transcription system," in *Proc. IEEE ICASSP*, Philadelphia, PA, 2005, vol. 1, pp. 841–844.

[17] P. C. Woodland, M. J. F. Gales, D. Pye, and S. J. Young (1996). "The development of the 1996 HTK broadcast news transcription system", in *Proc. DARPA Speech Recognition Workshop*, Arden House, NY, US, pp. 73–78.

[18] The Kaldi speech recognition toolkit. *http://kaldi.sourceforge.net*

[19] N. Kumar (1997). *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*, PhD Thesis, John Hopkins University.

[20] X. Liu, M. J. F. Gales, and P. C. Woodland (2003). "Automatic complexity control for HLDA systems", in *Proc. IEEE ICASSP*, Hong Kong, China, vol. 1, pp. 132–135.

[21] S. J. Young, J. J. Odell, and P. C. Woodland (1994). Tree-based State Tying for High Accuracy Acoustic Modeling, in *Proc. ARPA Human Language Age Technology Workshop*, Morgan Kaufman, 1994, pp. 307–312.

[22] X. Liu, M. J. F. Gales, J. L. Hieronymus and P. C. Woodland (2011). "Investigation of acoustic units for LVCSR systems", in *Proc. IEEE ICASSP*, Prague, Czech Republic, pp. 4872–4875.

[23] D. Povey and P. C. Woodland (2002). "Minimum phone error and I-smoothing for improved discriminative training", in *Proc. IEEE ICASSP*, Orlando, FL, 2002, vol. 1 105–108.

[24] M. J. F. Gales (1998). "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, 12(2): 75-98, 1998.

[25] C. J. Leggetter and P. C. Woodland (1995). "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs", *Computer Speech and Language*, 9(2): 171-185, 1995.

[26] S. Young G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. C. Woodland. "The HTK Book Version 3.4.1", 2009.

[27] F. Grezl, M. Karafiat, S. Kontar and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings", in *Proc. IEEE ICASSP*, Honolulu, HI, 2007, vol. 4, pp. 757–760.

[28] P. Swietojanski, A. Ghoshal, and S. Renals (2013). "Revisiting hybrid and GMM-HMM system combination techniques," in *IEEE ICASSP*, Vancouver, Canada, 2013, pp. 6744–6748.

[29] H. Soltau, G. Saon, and T. N. Sainath (2014). "Joint training of convolutional and non-convolutional neural networks," in *IEEE ICASSP*, Florence, Italy, 2014, pp. 5572–5576.

[30] M. J. F. Gales (1999). "Semi-tied Covariance Matrices for Hidden Markov Models", *IEEE Transactions on Speech and Audio Processing*, pp. 272–281, vol. 7, 1999.

[31] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky, and S. Khudanpur (2010), "Recurrent neural network based language model," in *Proc. ISCA Interspeech*, Makuhari, Japan, 2010, pp. 1045–1048.

[32] X. Chen, Y. Wang, X. Liu, M. J. F. Gales and P. C. Woodland (2014). "Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch", in *Proc. ISCA Interspeech*, Singapore, 2014, pp. 641–645.

[33] Y. Si, Q. Zhang, T. Li, J. Pan, and Y. Yan (2013), "Prefix tree based n-best list re-scoring for recurrent neural network language model used in speech recognition system," in *Proc. ISCA Interspeech*, Lyon, France, 2013, pp. 3419–3423.

[34] L. Mangu, E. Brill, and A. Stolcke (2000). "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech and Language*, 14(4): 373-400, 2000.

[35] X. Liu, Y. Wang, X. Chen, M. J. F. Gales, and P. C. Woodland (2014), "Efficient lattice rescoring using recurrent neural network language models," in *Proc. IEEE ICASSP*, Florence, Italy, 2014, pp. 4941–4945.

[36] X. Liu, M. J. F. Gales, and P. C. Woodland (2013), "Use of contexts in language model interpolation and adaptation," *Computer Speech and Language*, vol. 27, no. 1, pp. 301–321, January 2013.

[37] X. Liu, M. J. F. Gales & P. C. Woodland (2013). "Language model cross adaptation for LVCSR system combination", *Computer Speech and Language*, vol. 27, no. 4, pp. 928-942, June 2013.

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

[39] R. Schwartz et al. (2004). Speech Recognition in Multiple Languages and Domains: The 2003 BBN/LIMSI EARS System, in *Proc. IEEE ICASSP*, Montreal, Canada, 2004, vol. 3, pp. 753–756.

[40] S. M. Chu et al. (2010). "The 2009 IBM GALE Mandarin Broadcast Transcription System," in *Proc. IEEE ICASSP*, Dallas, TX, 2010, pp. 4374–4377.

[41] P. C. Woodland, C. J. Leggetter, J. J. Odell, V. Valtchev, and S. J. Young. (1995). "The 1994 HTK Large Vocabulary Speech Recognition System, " in *Proc. IEEE ICASSP*, Detroit, MI, pp. 73–76.

[42] R. Prasad et al. (2005). "The 2004 BBN/LIMSI 20xRT English conversational telephone speech recognition system," in *Proc. ISCA Interspeech*, Lisboa, Portugal, 2005, pp. 1645–1648.