An Initial Investigation of Long-Term Adaptation for Meeting Transcription

X. Chen [1], *M.J.F. Gales* [1], *K. Knill* [1,2], *C. Breslin* [1,2], *L. Chen* [2], *K. K. Chin* [2] *and V. Wan* [2]

[1]Engineering Department, University of Cambridge, UK
[2]Toshiba Research Europe Ltd, Cambridge, UK
Email: {xc257, mjfg, kate.knill, cb404}@cam.ac.uk
{langzhou.chen, kkchin, vincent.wan}@crl.toshiba.co.uk

10.21437/Interspeech.2014-253

Abstract

Meeting transcription is a very useful and challenging task. The majority of research to date has focused on individual meeting, or only a small group of meetings. In many practical deployments, multiple related meetings will take place over a long period of time. This paper describes an initial investigation of how this long-term data can be used to improve meeting transcription. A corpus of technical meetings, using a single microphone array, was collected over a two year period, yielding a total of 179 hours of meeting data. Baseline systems using deep neural network acoustic models, in both Tandem and Hybrid configurations, and neural network-based language models are described. The impact of supervised and unsupervised adaptation of the acoustic models is then evaluated, as well as the impact of improved language models.

Index Terms: Meeting Transcription, Unsupervised Adaptation, Confidence Score, MAP, MLLR

1. Introduction

A very useful, and challenging, task is the transcription of meetings. Accurate transcriptions would allow a number of possible applications such as meeting summarisation and audio indexing to be performed. However, it is highly expensive and slow to transcribe meetings manually. Automatic speech recognition of meeting data is an efficient and cheap alternative.

Many researchers have examined a range of approaches for automatically meeting transcription [1, 2, 3, 4, 5, 6, 7]. These have normally been applied to standard corpora, such as those used for NIST evaluations [8, 9]. These corpora allow the transcription of individual meetings or small groups of meetings, to be investigated. For practical deployed meeting transcription systems, it is expected that a sequence of related meetings will be recorded and transcribed over a long period of time. In previous work, this continuity and the existence of related topics and speakers in a series of meetings is largely ignored.

In order to explore the use of long-term information, the Speech Group at Toshiba Research Europe Ltd, Cambridge Research Laboratory, undertook the recording of meetings over a two year period. These meetings cover both speech recognition and synthesis, and were normally technical in nature. The data was recorded using a microphone array to minimise any impact that the data collection would have on the behaviour of an individual. This limited the nature of the data that could be collected: no close-talking microphone data is available; individuals were able to sit where they wanted, to move during the meetings to give presentations and to have "side" conversations. Additionally, it was agreed that this data would never be made publicly available or used to assess the performance of individuals.

The configuration of the room and some of the typical distances from the microphone[1]. are shown in figure 1.

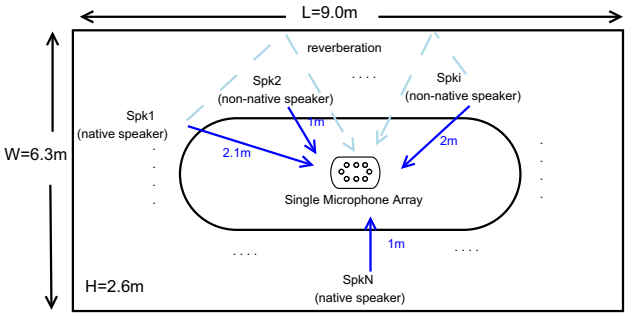

Figure 1: Toshiba Technical Meeting Recording Configuration

The rest of paper is organized as follows. The next section gives a detailed description of the meeting data involved in this paper, especially the Toshiba Technical Meetings (TTM). The scheme of unsupervised long-term adaptation is described in section 3, followed by three practical operating modes for meeting transcription systems in section 3.3. Experimental results are given in section 4 and conclusions drawn in section 5.

2. Data Description

Two corpora are used in this paper. The first is from the AMI project [10]. This data was used for training, and to obtain recognition results that can be compared to existing systems. The second is the TTM data, which was used for long-term adaptation and evaluation. For both corpora only the multiple distant microphone (MDM) data was used. Beamforming was performed using the *BeamformIt* tool [11] to yield a single audio channel[2].

Xie Chen would like to thank Toshiba Research Europe Ltd, Cambridge Research Lab, for funding his work. The authors would like to thank the Toshiba Cambridge Speech Group for allowing the data to be collected, also would like to thank Chao Zhang and Eric Wang for providing DNN and CMLLR transform tools.

[1]Multiple microphone arrays could address some of the issues observed with distance from the microphone.

[2]Currently there is no Wiener filtering in the front-end processing, as used for example in [4], which should yield performance gains.

2.1. Augmented Multi-party Interaction (AMI) Corpus

The AMI corpus [10] was collected for research and development of technology that may help groups interact better. The corpus consists of meetings with four participants, where each person acted a role in a product development team. While close-talking, lapel and distant microphone data was recorded, this work only makes use of the far-field multiple distant microphone (MDM) channel. Additionally overlapping speech data was removed. This yielded about 59 hours of data. In addition to the AMI corpus, 52 hours from the ICSI corpus [12] and 10 hours from the NIST corpus were also used [13]. Four meetings were held back from the AMI data to give an AMI dev (meetings IS09, ES09) and eval set (meetings IS08,ES08), each with two sets of meetings and 4 speakers per meeting. As overlapping speech was not evaluated, this yielded a total test set of 5.29 hours. The total available data for training, after removing the 4 meetings was about 121 hours. This is the same configuration, and held-out test sets, as used in [5].

2.2. Toshiba Technical Meeting Data

The second corpus was collected at Toshiba Research Europe Ltd's Cambridge Lab. The corpus was collected in a meeting room (shown in figure 1) with between 6 and 9 participating in each meeting. The Toshiba ASR and TTS technical meetings were recorded, where on-going research projects and future plans were discussed. Compared to the AMI corpus, the TTM data has a greater distance from the microphone to the speaker and a higher level of noise. A subjective estimate of SNR of the meeting is around 0 to 5dB but it varies by speaker. These differences will be seen to result in a higher baseline WER than for a typical meeting from the AMI corpus.

The TTM meeting data were recorded over a two year period. A total of 158 meetings were collected, yielding a total of 179 hours of acoustic data. The approximate breakdown of the meeting by general topic is: ASR projects 15 meetings (20 hours) TTS projects 61 meetings (57 hours); acoustic modelling 33 meetings (44 hours); group and general 47 meetings (58 hours). For the initial release of the data no information about the participants of the meetings was available.

In order to evaluate performance, seven meetings (8.88 hours) were selected for the TTM evaluation testset. Note overlapping speech was again removed. Professional transcriptions were provided as reference, and this data was transcribed early in the collection. To assess the quality of the professional transcriptions, the first recorded meeting (denote as A0001) was also transcribed by people who were present at the meeting and familiar with the attendees and their accents. This will be referred to as the "gold-standard".

Table 1: Microphone distance and WER for meeting A0001

Speaker	A	B	C	D	E	Avg
Dist (m)	2.1	2.1	1.2	1.6	1.2	—
Manual	16.4	15.7	8.2	7.6	2.9	10.6
ASR	68.6	66.4	74.9	70.2	55.4	64.8

Table 1 shows both the microphone distance and word error rate of the professional (manual) transcriptions, compared to the "gold-standard", and performance of an initial ASR system. Speakers C, D, E are native UK English speakers, while speakers A and B are non-native. The professional transcribers sometimes chose an incorrect word sequence, though the transcriptions were phonetically similar to the correct word sequence. The overall WER is 10.6%, which indicates TTM data is a highly challenging task. The transcribers had difficulty with non-native speakers (A and B), Japanese and Chinese names, technical jargon and abbreviations. These are not issues for people familiar with the participants and topics. However, these gold-standard transcriptions are very difficult to obtain as they require transcribers with expert in-domain knowledge of both the meeting topic and participants. For this work the "gold-standard" was used as the reference for meeting A0001, the manual transcriptions were used for the other six meetings. The lowest WER speaker for the ASR system was the same as the manual transcribers, a speaker close to the microphone. However, there was no consistent pattern over the other speakers.

3. Unsupervised Task Adaptation

As an initial study for long-term adaptation, unsupervised task adaptation was investigated. For standard speaker adaptation, the canonical model obtained from Speaker Adaptive Training (SAT) [14] on the AMI corpus will be adopted for speaker adaptation on TTM evaluation testsets. Here the adaptation will not only capture the variability of speakers, but also some of the mismatch between the AMI and TTM corpora background noise, reverberation and channel distortion. However since the mismatch between the two corpora should be consistent over the whole meeting, and some attributes over all the data collected, task adaptation can also be applied. The task adaptation is followed by speaker adaptation to capture personal characters such as speaker, seating position.

3.1. Task and Speaker Adaptation

There are a number of existing studies into task adaptation, or task porting, Maximum a posterior (MAP) [15] was applied on ML and MMI trained speaker independent models for cross task adaptation in [16]. In [17], ML-MAP and MMI-MAP were investigated by porting acoustic models from Switchboard to voicemail task.

In this work, task adaptation in combination with speaker adaptation is used. MLLR and MAP (MLLR+MAP) are adopted for task adaptation, and CMLLR and MLLR (CMLLR+MLLR) for speaker adaptation. The acoustic model score for a specific speaker s can be expressed by

$$p(\boldsymbol{o}_\tau|s,t,m) = |\mathbf{A}^{(s)}|\mathcal{N}\left(\mathbf{A}^{(s)}\boldsymbol{o}_\tau + \mathbf{b}^{(s)}; \boldsymbol{\mu}^{(stm)}, \boldsymbol{\Sigma}^{(stm)}\right)$$

where $\boldsymbol{\mu}^{(stm)}$ and $\boldsymbol{\Sigma}^{(stm)}$ are the mean and covariance matrix for Gaussian component m after MLLR+MAP adaptation to the meeting(s) t and MLLR adaptation to speaker s, and $\mathbf{A}^{(s)}$ and $\mathbf{b}^{(s)}$ are the global CMLLR transforms for speaker s.

3.2. Confidence score based data selection

For distant microphone meeting transcription, the error rate may be quite high. For task adaptation, where there is more adaptation data available than in the speaker adaptation case, it is not necessary to use all available data for adaptation, as adapting to high error rate transcriptions may degrade performance. Confidence score based adaptation [18, 19] is one approach to improve unsupervised adaptation. In this work, segment-level confidence scores are used for data selection. Segments with high confidence score are kept, while the other segments are discarded. The confidence of each segment takes the arithmetic mean of word confidence score, which is obtained from word posterior probabilities in confusion network [20].

3.3. Task Adaptation Mode

In deployed meeting transcription systems, it is possible to run task adaptation in a number of different modes, depending on the requirements of the system. For example, if the room in which the meetings will take place changes, then the task adaptation should be on individual meetings. This is referred to as *independent* mode in this paper. Alternatively in *incremental* mode it is assumed that the meeting environment is consistent, but the transcriptions need to be generated in a causal fashion at the meeting level. The final model *batch* mode, again assumes a consistent meeting environment, but it is possible to transcribe all meetings in a single batch, for example if archived meetings are being transcribed. These modes are illustrated in Figure 2.

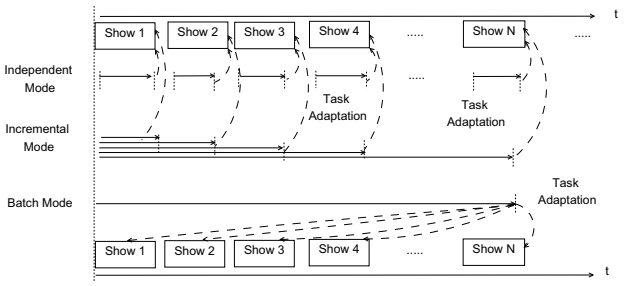

Figure 2: Operating modes for meeting transcription

4. Results

For these experiments, the AMI corpus was used for training the baseline acoustic models and to provide publicly available baseline recognition numbers. Task adaptation was only performed on the TTM data. For all test data, the automatic segmentation and speaker clustering described in [5] was used at a meeting level. A 3-gram language model (LM) trained on 2.5G words was used as the baseline language model, additional details of the training data for this LM are given in [5]. The Out of Vocabulary (OOV) rate for the AMI dev data was 2.23%, eval data 2.17% and the TTM data 1.24%.

Three forms of acoustic model were initially examined for meeting transcription. The baseline models are Gaussian mixture model (GMM) based systems using PLP features (PLP system). Additionally, two forms of acoustic model based on deep neural networks were trained. A GMM-based system using PLP and bottleneck (BN) features [21] (Tandem system) and Hybrid system [22] combining HMMs and neural network posteriors. All systems were based on state-clustered decision-tree triphone models.

The baseline PLP system was built in a similar fashion to that described in [5]; 13-dimensional PLP features with delta, delta-delta and triples appended. CMN, CVN and HLDA were then applied for feature normalisation and projection. The minimum phone error (MPE) [23] criterion was used to train the acoustic models. SAT [14] based on CMLLR [24] was also used. MLLR was then used to adapt to the target speaker for GMM-based systems. A tandem system appending bottleneck features [21] was also built. A "deep" MLP is constructed with four hidden layers, with 1000 nodes per layer. Nine-frames were spliced to form the input layer. The dimension of bottleneck feature is set to be 26. The target is 6000 context dependent states obtained from the decision tree generated for GMM-HMM system. Discriminative pretraining as in [25] was used. A semi-tied covariance (STC) [26] transform was applied to the bottleneck features prior to concatenation with the PLP features. Thus, the dimensionality of the Tandem feature in this paper is 65. The Tandem acoustic models were built using the rapid construction approach described in [27]. The final acoustic models constructed were based on hybrid systems, also known as DNN-HMM systems [22]. The training of neural network in hybrid system is similar to the training of the BN features in Tandem system. A speaker adapted hybrid system was also constructed, by applying CMLLR transforms obtained from the GMM system on PLP feature, before it is fed into DNNs. The alignment for the targets was obtained from a Tandem SAT system.

In addition to the baseline N-gram language model, feed-forward MLP (MLP) [28] and recurrent neural network (RNN) [29] language models were built. Both these neural network based language models (NNLMs) were trained on the transcriptions of acoustic training data, about 2 million words. The MLP-based NNLM used the shortlist with 6.3K words to reduce the size of LM and increase the generalization ability. An out of shortlist (OOS) symbol was also used on the output layer [30]. The MLP NNLM has two hidden layers with 600 and 400 nodes respectively. A fully connected recurrent NNLM described in [31] was adopted. The size of hidden layer was 500. Efficient lattices rescoring proposed in [32] was used to generate lattices for CN decoding. All the neural network based language model are interpolated with the standard N-Gram language model and the weight is fixed to be 0.5.

4.1. WERs for the Baseline Acoustic Systems

Initially speaker independent (SI) and speaker adapted (SA) meeting transcription was investigated. Thus only global CMLLR+MLLR adaptation at the speaker level was used. Table 2 gives WER results on the three types of acoustic models with the baseline LM. As expected, the TTM WER is significantly higher than that of the AMI testsets. Both the DNN based systems out-perform the baseline acoustic models, the Hybrid system gives comparable performance to the Tandem system in SI and SA modes. Confusion network combination (CNC) [33] of Tandem and Hybrid system provides further improvement.

Table 2: Baseline WER results on AMI and TTM testset

	System	AMI		TTM
		dev	eval	
SI	PLP	44.7	43.5	67.9
	Tandem (T1)	36.4	36.5	60.4
	Hybrid (H1)	35.9	35.6	59.8
	CNC T1⊕H1	34.1	33.9	57.9
SA	PLP	38.2	39.2	61.7
	Tandem (T2)	33.4	34.5	57.5
	Hybrid (H2)	33.2	33.0	58.4
	CNC T2⊕H2	31.8	31.6	55.7

In the following experiments, the Tandem system is used to investigate the performance for long-term adaptation as it gives slightly better performance than hybrid system in the SA mode, and Tandem system acoustic model adaptation is easier to implement.

4.2. Unsupervised Task Adaptation for TTM corpus

In this section, task adaptation in batch mode is applied with the Tandem SAT system using all TTM data. The supervision for task adaptation was generated from the baseline Tandem SAT

system. The confidence score from CN decoding [20] was used for data selection. As stated in section 3, MLLR+MAP was adopted for unsupervised task adaptation. A regression tree is used for MLLR and the maximum number of MLLR transform is set to be 128. In initial experiments, it was found that discriminative MAP [34] didn't outperform ML-MAP for unsupervised task adaptation due to the high WER of the supervision. Hence, ML-MAP is used for unsupervised task adaptation by default.

The WER results for batch-mode task adaptation are given in table 3. Task Adaptation using MLLR and MAP could reduce WER by 0.8% and 1.6% respectively with data selection. Furthermore, a combination of them gives 2.0% absolute improvement in WER when the confidence threshold is 0.8, giving a good balance between the quantity and quality of adaptation data.

Table 3: WER results for unsupervised task adaptation on Tandem SAT system in batch mode on TTM corpus

Task Adaptation	Conf Thresh	Adapt Data (hrs)	WER
-	-	-	57.5
MLLR	0.8	68.0	56.7
MAP	0.8	68.0	55.9
MLLR+MAP	0.0	179.0	56.3
	0.7	128.6	56.0
	0.8	68.0	**55.5**
	0.9	19.7	55.7

4.3. Independent, Incremental and Batch Modes

It is interesting to investigate the relative performance of each of the task adaptation models shown in Figure 2. Results are shown in table 4. For all these experiments the confidence threshold of 0.8 found in the batch-mode experiments was used. For the independent mode, WER is reduced by 0.8% using only an individual meeting, here the average amount of data selected per meeting was 0.5 hours. For the incremental mode work, only the data up to the final test meetings were used. Thus, the amount of adaptation data for these seven meetings varied from 0.5 to 8.7 hours and WER is reduced by 0.9%. Lastly, for batch mode, two configurations were run. The first experiment used the same seven meetings as the incremental system. The amount of adaptation data is 8.7 hours after data selection, giving WER of 56.3%, 0.3% absolute better than the incremental performance. The second experiment used all TTM data, 68.0 hours, and gave a WER of 55.5%. Though increasing the quantity of unsupervised data yielded improved performance, the error rate is still very high.

Table 4: WER results for unsupervised task adaptation in different operating modes

Mode	Adapt Data (hrs)	WER
—	—	57.5
Independent	0.5	56.7
Increment	0.5—8.7	56.6
Batch	8.7	56.3
	68.0	55.5

4.4. Supervised Acoustic and Language Model Adaptation

The reduction of WER with all the TTM data is only 2.0% absolute over unsupervised adaptation. For practical systems a limited amount of data may be transcribed early in the deployment to try and improve performance. To examine the impact of this supervised data, the seven meetings for which transcriptions were available were used in a cross validation fashion to adapt the acoustic and language models. This yielded an average of 7.5 hours of data, 92K words for each test meeting.

Table 5 shows the comparison of WERs for the unsupervised and supervised task adaptation. Supervised adaptation using MLLR+MAP yielded greater performance gains than using all the meetings in an unsupervised fashion. Additionally MPE-MAP gave performance gains over ML-MAP adaptation for supervised adaptation, as expected with reference transcriptions available.

Table 5: WER results for supervised task adaptation

Adaptation Reference	MAP form (MLLR+)	Adapt Data (hrs)	WER
—	—	—	57.5
Unsupervised	ML-MAP	68.0	55.5
Supervised	ML-MAP	7.5	54.8
	MPE-MAP		54.0

The previous results have not updated the language model. For each meeting, in a cross-validation fashion, N-Gram LMs were trained from the manual references and interpolated with the baseline N-Gram LM with an interpolation weight 0.2. Table 6 shows the WER performance of using this adapted language model, as well as the interpolation of N-grams (both adapted and baseline) with feed-forward (MLP) and recurrent (RNN) neural network language models. As expected, the use of the adaption data for both the language and acoustic models, and the use of neural network language models, yields performance gains. However the overall performance is still over 50% WER.

Table 6: WER results for supervised language model adaptation

NNLM	Adapt.	PPL	WER
—	—	128.8	57.5
	LM	108.3	56.4
	LM+AM		53.5
MLP	-	124.5	56.4
	LM	111.9	56.2
	LM+AM		53.2
RNN	-	112.9	56.1
	LM	102.0	55.7
	LM+AM		53.0

5. Conclusion

This paper has described an initial investigation on long-term adaptation for meeting transcription. The meeting task selected is highly challenging, using a single microphone array in a large meeting room with participants interacting in a natural fashion. Using deep neural network based acoustic models and speaker adaptation yielded an error rate of 57.5%. Incorporating long-term unsupervised adaptation reduced this by 2.0% absolute. The limited amount of supervised data and advanced language models further reduced this by 7.8% relative, to 53.0% WER.

These initial WER improvements on long-term adaptation using standard adaptation approaches are relatively small. However in the current configurations no use has been made of additional meta-data from the meetings, such as the agenda, presentation slides, and technical documents. Furthermore there has been no explicit tracking of individuals over the meetings. Both of these approaches will be examined in future work.

6. References

[1] Hua Yu, Takashi Tomokiyo, Zhirong Wang, and Alex Waibel, "New developments in automatic meeting transcription.," in *Proc. ISCA Interspeech*, 2000, pp. 310–313.

[2] Alex Waibel, Michael Bett, Florian Metze, Klaus Ries, Thomas Schaaf, Tanja Schultz, Hagen Soltau, Hua Yu, and Klaus Zechner, "Advances in automatic meeting record creation and access," in *Proc. ICASSP*. IEEE, 2001, vol. 1, pp. 597–600.

[3] Steve Renals, Thomas Hain, and Herve Bourlard, "Recognition and understanding of meetings the AMI and AMIDA projects," in *ASRU, IEEE Workshop on*. IEEE, 2007, pp. 238–247.

[4] Thomas Hain, Vincent Wan, Lukas Burget, Martin Karafiat, John Dines, Jithendra Vepa, Giulia Garau, and Mike Lincoln, "The AMI system for the transcription of speech in meetings," in *Proc. ICASSP*. IEEE, 2007, vol. 4, pp. IV–357.

[5] Catherine Breslin, KK Chin, Mark J F Gales, and Kate Knill, "Integrated online speaker clustering and adaptation.," in *Proc. ISCA Interspeech*, 2011, pp. 1085–1088.

[6] Thomas Hain, Lukas Burget, John Dines, Philip N Garner, Frantisek Grezl, Asmaa El Hannani, Marijn Huijbregts, Martin Karafiat, Mike Lincoln, and Vincent Wan, "Transcribing meetings with the AMIDA systems," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 486–498, 2012.

[7] Stefan Kombrink, Tomas Mikolov, Martin Karafiát, and Lukas Burget, "Recurrent neural network based language modeling in meeting recognition.," in *Proc. ISCA Interspeech*, 2011, pp. 2877–2880.

[8] Jonathan G Fiscus, Jerome Ajot, Martial Michel, and John S Garofolo, *The rich transcription 2006 spring meeting recognition evaluation*, Springer, 2006.

[9] Jonathan G Fiscus, Jerome Ajot, and John S Garofolo, "The rich transcription 2007 meeting recognition evaluation," in *Multimodal Technologies for Perception of Humans*, pp. 373–389. Springer, 2008.

[10] Jean Carletta et al., "The AMI meeting corpus: A pre-announcement," in *Machine learning for multimodal interaction*, pp. 28–39. Springer, 2006.

[11] Xavier Anguera Miro, *Robust speaker diarization for meetings*, Ph.D. thesis, 2007.

[12] Adam Janin, Don Baron, Jane Edwards, Dan Ellis, David Gelbart, Nelson Morgan, Barbara Peskin, Thilo Pfau, Elizabeth Shriberg, Andreas Stolcke, et al., "The ICSI meeting corpus," in *Proc. ICASSP*. IEEE, 2003, vol. 1, pp. I–364.

[13] John S Garofolo, Christophe Laprun, Martial Michel, Vincent M Stanford, and Elham Tabassi, "The NIST meeting room pilot corpus.," in *LREC*, 2004.

[14] Tasos Anastasakos, John McDonough, Richard Schwartz, and John Makhoul, "A compact model for speaker-adaptive training," in *ICSLP*. IEEE, 1996, vol. 2, pp. 1137–1140.

[15] J-L Gauvain and Chin-Hui Lee, "Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 2, pp. 291–298, 1994.

[16] R Cordoba, Philip C Woodland, and Mark JF Gales, "Improved cross-task recognition using mmie training," in *Proc. ICASSP*. IEEE, 2002, vol. 1, pp. I–85.

[17] M. J. F. Gales, Y. Dong, D. Povey, and P. C. Woodland, "Porting: Switchboard to the voicemail task," in *Proc. ICASSP*. IEEE, 2003, vol. 1, pp. I–536.

[18] Michael Pitz, Frank Wessel, and Hermann Ney, "Improved MLLR speaker adaptation using confidence measures for conversational speech recognition.," in *Proc. ISCA Interspeech*, 2000, pp. 548–551.

[19] Kai Yu, Mark Gales, Lan Wang, and Phil C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7, pp. 652–663, 2010.

[20] Lidia Mangu, Eric Brill, and Andreas Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[21] Frantisek Grezl and Petr Fousek, "Optimizing bottle-neck features for LVCSR," in *Proc. ICASSP*. IEEE, 2008, pp. 4729–4732.

[22] George Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[23] Daniel Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, 2004.

[24] Mark Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[25] Frank Seide, Gang Li, Xie Chen, and Dong Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *ASRU, IEEE Workshop on*. IEEE, 2011, pp. 24–29.

[26] Mark J. F. Gales, "Semi-tied covariance matrices for hidden Markov models," *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 3, pp. 272–281, 1999.

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

[28] Holger Schwenk, "Continuous space language models," *Computer Speech & Language*, vol. 21, no. 3, pp. 492–518, 2007.

[29] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur, "Recurrent neural network based language model.," in *Proc. ISCA Interspeech*, 2010, pp. 1045–1048.

[30] Junho Park, Xunying Liu, Mark J. F. Gales, and P. C. Woodland, "Improved neural network based language modelling and adaptation," in *Proc. ISCA Interspeech*, 2010, pp. 1041–1044.

[31] Xie Chen, Yongqiang Wang, Xunying Liu, Mark Gales, and P. C. Woodland, "Efficient GPU-based training of recurrent neural network language models using spliced sentence bunch," in *submitted to Proc. ISCA Interspeech*. IEEE, 2014.

[32] Xunying Liu, Yongqiang Wang, Xie Chen, Mark Gales, and P. C. Woodland, "Efficent lattice rescoring using recurrent neural network language models," in *Proc. ICASSP*. IEEE, 2014.

[33] Gunnar Evermann and P. C. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*. Baltimore, 2000, vol. 27.

[34] Daniel Povey, Mark Gales, Do Yeong Kim, and P. C. Woodland, "MMI-MAP and MPE-MAP for acoustic model adaptation.," in *Proc. ISCA Interspeech*, 2003.