INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

The Sheffield Wargame Corpus - Day Two and Day Three

Yulan Liu[1†], *Charles Fox*[2], *Madina Hasan*[1], *Thomas Hain*[1††]

[1]MINI, SpandH, The University of Sheffield, UK
[2]The University of Leeds, UK
[†]acp12yl@sheffield.ac.uk, [††]t.hain@sheffield.ac.uk

Abstract

Improving the performance of distant speech recognition is of considerable current interest, driven by a desire to bring speech recognition into people's homes. Standard approaches to this task aim to enhance the signal prior to recognition, typically using beamforming techniques on multiple channels. Only few real-world recordings are available that allow experimentation with such techniques. This has become even more pertinent with recent works with deep neural networks aiming to learn beamforming from data. Such approaches require large multi-channel training sets, ideally with location annotation for moving speakers, which is scarce in existing corpora. This paper presents a freely available and new extended corpus of English speech recordings in a natural setting, with moving speakers. The data is recorded with diverse microphone arrays, and uniquely, with ground truth location tracking. It extends the 8.0 hour Sheffield Wargames Corpus released in Interspeech 2013, with a further 16.6 hours of fully annotated data, including 6.1 hours of female speech to improve gender bias. Additional blog-based language model data is provided alongside, as well as a Kaldi baseline system. Results are reported with a standard Kaldi configuration, and a baseline meeting recognition system.
Index Terms: distant speech recognition, multi-channel speech recognition, natural speech corpora, deep neural network.

1. Introduction

Multi-channel based speech enhancement has been shown to be effective for Distant Speech Recongition (DSR), in both classical HMM-GMM systems and state-of-art Deep Neural Networks (DNNs) based systems. Compared to using recordings from single distant microphone only, beamforming is reported to reduce word error rate (WER) by 6-10% relative in large vocabulary conversational speech recognition tasks [1–3], and up to 60% relative in specific tasks [4, 5]. Multi-channel dereverberation brings an extra 20% relative WER reduction over single channel dereverberation [6]. Recently progress in neural networks have introduced further performance improvement in a variety of tasks, particularly from three aspects: progress in novel network structures [7, 8], application-oriented neural network structure and parameter manipulation [9–12], and data manipulation for neural network training [1, 13]. While the overall WERs keep going down, the recognition performance gap remains between using recordings from close-talking microphones and from distant microphones. To reduce this gap, research effort has focused on three approaches: developing novel structures to better utilize multichannel recordings in DNN [14, 15], employing task dependent meta information [16, 17], and simulating training data for specific DSR tasks [6, 18]. However research progress is limited by lack of data that provides multichannel distant recordings accompanied with headset recordings and speaker location tracking, in a natural speech setting where speakers are allowed to move freely. To address this problem, the present study extends the first Sheffield Wargames Corpus (SWC1, [19]) with more natural speech recordings from both headsets and distant microphones in moving talker conditions, accompanied with real time speaker location tracking.

The paper is organized as follows. §2 reviews related work, §3 provides basic information about the set-up for the new recording days SWC2 and SWC3. §4 discusses dataset definitions for two different ASR tasks: adaptation and standalone training. The details about language models (LMs) are introduced in §5. §6 provides results for two tasks, using HTK, TNet and Kaldi. All WERs on eval set are above 40% for headset recordings, and above 70% for distant recordings. §7 concludes the work.

2. Multi-channel Recordings in DSR

Research on utilizing multi-channel recordings within DNN structure started with directly concatenating features from multiple channels at DNN input [1, 2]. Such method was found to perform similar or better than weighted delay and sum beamforming (wDSB) in 2 and 4 channel cases. Furthermore, joint optimization of beamforming and DNNs achieved 5.3% relative improvement over using wDSB in [15]. In [14], beamforming and standard feature pipeline are completely replaced with neural networks. Different neural networks are combined to extract information from raw signals, achieving 5.8% relative WER improvement over 8 channel delay and sum beamforming (DSB).

Meta-information can also be provided to DNNs. In [16], adding noise information provides a 5.1% relative improvement over feature enchancement. In [17], adding room information via feature augmention improves performance by 2.8% relative on the ReverbChallenge real data. In [2], geometry information was added via augmenting the concatenated multi-channel features with Time Difference of Arrival (TDOA) at DNN input. However, no improvement was observed.

Above approaches all require large data sets for training. One main challenge in DSR is the variety in environment conditions of real recordings. Even within the same room, speakers may move around a room, resulting in continually changing room impulse responses (RIRs). One method to address this issue is multi-condition training [6], by simulating data of different environment conditions with different RIRs and by adding background noise to clean speech. The RIRs can be either generated by simulation or measured in real environments [20–22]. Examples of corpora with simulated environment effects are Aurora [23–25], DIRHA-GRID [26] and DIRHA-ENGLISH [27]. Another method is to select targeted RIRs that match best to the test scenario [18]. However there is a lack of corpora

http://dx.doi.org/10.21437/Interspeech.2016-98

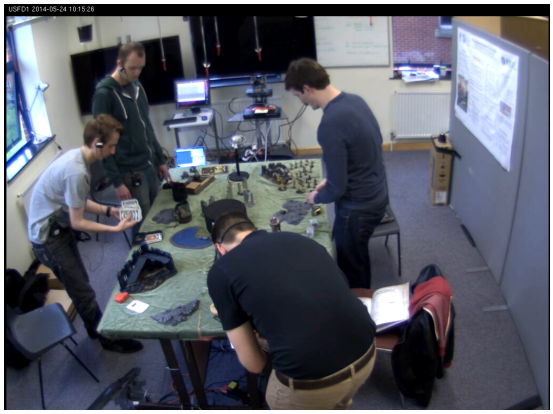

Figure 1: *SWC2 recording (from Camera C1 in Fig. 2).*

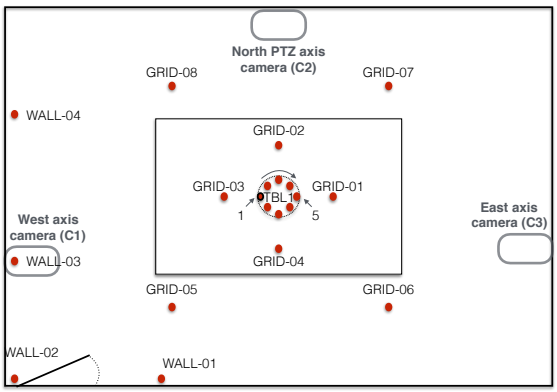

Figure 2: *Video location in SWC2.*

covering different environment conditions that also have natural speech. Existing research corpora of real multi-channel distant recordings often use artificial scenarios, read speech and re-recorded speech. Examples are the real recording part of MC-WSJ-AV corpus [28] used in ReverbChallenge 2014 [6], or the CHiME corpora [29]. Other corpora are recorded with controlled environment and speaker movement, such as the meeting corpora AMI [30] and ICSI [31].

The first Sheffield Wargame Corpus (SWC1, [19]) released in 2013 is a natural, spontaneous speech corpus of native English speakers who are constantly speaking and moving while playing tabletop games. It includes 3-channel video recordings and 96-channel audio recordings from headsets and distant microphones at static known locations in the room. Besides, it includes ground truth head location, providing a reference for research on localization and beamforming algorithms. The task is challenging as it represents everyday colloquial conversation among friends, with emotional speech, laugher, overlapping speech fragments as well as body movement while speaking.

The size of SWC1, 8 hour speech, limits its usefulness for training and adaptation. In addition, SWC1 contains male speech only. This paper releases, for free use in the research community, the extended Sheffield Wargame Corpus recording Day 2 (SWC2) and Day 3 (SWC3). In addition, it releases blog and wikipedia based text data to build in-domain LMs, along with a well built set of in-domain LM and dictionary. SWC3 provides 6.1h of female speech to provide a gender balance. Combined with SWC1, the corpora form a total of 24.6h speech database. Standard datasets are defined to enable comparative research on combined corpora for two scenarios: adapting existing acoustic models (AMs) to SWC data, and standalone training of AMs with SWC data only. An open-source Kaldi recipe is provided for standalone training. Baseline experiment results are reported for both standalone and adaptation systems.

3. SWC2 and SWC3 Recordings

Following the set-up for SWC1 [19], the extended corpora are comprised of recordings where four participants play the tabletop battle game Warhammer 40K[1] (Fig. 1). The game is chosen as a close proxy for sections of business meetings and social interactions where participants are moving and talking at the same time. Such joint motion and talking is difficult to record for extended periods in those contexts but the game promotes it constantly for hours at a time, allowing much more relevant data to be captured. The game is also played by a tight community of friends, many of whom are used to wearing headset microphones from online gaming, and are generally uninhibited by recording technology. Thus they speak more naturally and colloquially during recording, and they could move while speaking. In SWC2, the final two sessions have male viewers commenting on the game, to simulate a cocktail-party scenario. In SWC3, female players (with headsets) are instructed by two male tutors (without headsets) due to less game experience.

The recordings of SWC2 and SWC3 were performed in the same meeting room as SWC1, whose geometry is detailed in [19]. The recording system has three parts: multiple microphone audio recording, multiple camera video recording and location tracking. The three corpora use the same location tracking system Ubisense, which tracks the real time 3D head location of four players during the recording process [19]. Three channels of video recordings from cameras installed at three corners of the ceiling are also provided in SWC2 (Fig. 2).

24-channel audio recordings from the integrated Sheffield IML audio recording system [32] are shared among all three corpora (Fig. 2). They contains 4 headsets for 4 game players, 8 microphones in a circular array at the center of table (diameter: 20cm), 8 microphones hanging on a grid from the ceiling and 4 microphones distributed on the walls, all synchronized at sample level [19]. In SWC2, extra audio recordings are included using a Microcone array, a circular digital MEMs microphone array and an Eigenmic array. The Microcone array has 6 microphones in a circular array (diameter: 8cm), plus the seventh microphone pointing right up to the ceiling. The MEMs digital array has 8 microphones in a circular array with a diameter of 4cm. Both Microcone array and MEMs microphone array are situated on the table. The Eigenmic array is a 32-channel sphere array (diameter: 8.4cm). Only part of Session 1 in SWC2 has Eigenmic recordings due to software failure.

Table 1 lists statistics of SWC1 [19], SWC2 and SWC3. The vocabulary of SWC3 is much smaller compared to SWC1 and SWC2. This is because the game set-up is simplified for less experienced players, leading to simpler conversation.

4. Dataset Definition

Consistent datasets have been defined for SWC1, SWC2 and SWC3. Each recording session, i.e. a continuous recording file (Table 1), is first split into three successive strips of approximately equal speech duration: A, B and C. Such "data strip" allows flexible session combination to create datasets for which results can be easily shared among researchers. Four dataset

[1] https://en.wikipedia.org/wiki/Warhammer_40,000

Table 1: *SWC statistics.*

	SWC1	SWC2	SWC3	overall
#session	10	8	6	24
#game	4	4	3	11
#unique speaker	9	11	8	22
gender	M	M	F&M	F&M
#unique mic	96	71	24	103
#shared mic	-	-	-	24
speech duration	8.0h	10.5h	6.1h	24.6h
#speech utt.	14.0k	15.4k	10.2k	39.6k
duration per utt.	2.1s	2.5s	2.2s	2.2s
#word per utt.	6.6	7.9	5.5	6.8
vocabulary	4.4k	5.7k	2.9k	8.5k
video	√	√	-	√
location	√	√	√	√

Table 2: *Dataset statistics ("spk.": speaker; "dur.": duration).*

config.	set	strips	dur.	#utt.	#spk.
AD1	dev	{1, 2, 3}.A+B	16.3h	26.2k	22
	eval	{1, 2, 3}.C	8.2h	13.3k	22
AD2	dev	1	8.0h	14.0k	9
	eval	2, 3	16.6h	25.6k	18
SA1	train	1, {2, 3}.A	13.5h	22.6k	22
	dev	{2, 3}.B	5.5h	8.5k	18
	eval	{2, 3}.C	5.6h	8.4k	18
SA2	train	1	8.0h	14.0k	9
	dev	{2, 3}.A	5.5h	8.7k	18
	eval	{2, 3}.B+C	11.1h	16.9k	18

definitions based on strips are proposed to serve for two typical tasks: adaptation and standalone training. For each task, two configurations are available with different data separation and difficulty level, as listed in Table 2.

Adaptation task ("AD") only has dev and eval datasets. The "AD1" configuration uses 16.3h speech of Strip A and Strip B from all three recordings as dev set, and the remaining 8.2h of speech from Strip C as evaluation set. This dataset definition provides the least separation of speaker and speaking style. The "AD2" configuration only uses 8.0h SWC1 as dev set, while using the whole SWC2 and SWC3 for eval set. This is representative of many real applications where significant mismatch exists between trained system and test conditions, with limited data for adaptation and a variety in speaker, speaking style, and with subtle differences in topic and vocabulary.

Standalone training task ("SA") has train, dev and eval datasets. The "SA1" configuration uses 13.5h speech for training, comprised of whole SWC1, Strip A of SWC2 and SWC3. The development set uses 5.5h speech of Strip B from SWC2 and SWC3, and evaluation set uses 5.6h speech of Strip C from SWC2 and SWC3. This dataset definition takes into account the balance in gender and speaking style across training and testing. The "SA2" configuration provides only 8h speech from SWC1 for training, 5.5h speech from Strip A of SWC2 and SWC3 for development, and the remaining 11.1h as evaluation set. This dataset definition provides the best separation of speaker, session, game and speaking style between training and testing.

5. Language Modelling and Dictionary

SWC corpora are designed for research on acoustic modelling in natural speech recognition, particularly with multi-channel distant recordings. Since the conversation topic and vocabulary differ from most existing corpora, text data is harvested from four Warhammer 40K blogs and Warhammer wikipedia pages. These data are added to the conversational web data [33] to build an in-domain LM. N-gram components are trained using SRILM toolkit [34] on a 30k vocabulary list. The vocabulary list is built from all words in the harvested text plus the most frequent words in the conversational web data. The LM components are first built on each type of data, and then interpolated using SWC1 as development set. Table 3 lists the LM components and the interpolation weights for 4-gram LM. In initial experiments it was observed that a 4-gram LM trained on 30k vocabulary performs similarly to the RT'09 50k 3-gram LM, while using 3-gram or only using a smaller vocabulary degrades recognition performance. Thus results based on 4-gram LM with a vocabulary of 30k words are reported in following experiments. The perplexity of the interpolated 4-gram LM is 173.4 on SWC1, 195.9 on SWC2, 135.0 on SWC3 and 173.3 overall. The number of words out-of-vocabulary (OOV) is 1.4k on SWC1 (1.6%), 2.8k on SWC2 (2.4%), 3.9k on SWC3 (6.9%) and 8.1k overall (3.1%). Pronunciations are obtained using the Combilex pronunciation dictionary [35]. The Phonetisaurus toolkit [36] is used to automatically generate the pronunciation for words not in Combilex.

Table 3: *LM data size and interpolation weights (4-gram).*

LM component	#words	weight
Conversational web data	165.9M	0.65
Blog 1 (addict)	21.1k	0.05
Blog 2 (atomic)	126.8k	0.05
Blog 3 (cadia)	40.4k	0.19
Blog 4 (cast)	71.2k	0.06
wikipedia (warhammer)	26.2k	0.003

6. Baseline System

6.1. Adaptation task

The acoustic models trained on AMI corpus data are used in "AD2" configuration. The experiments here are performed with HTK and TNet. TNet is used to train DNN and to generate bottleneck features. HTK is used to train HMM-GMM using bottleneck features. The configuration mostly follows the procedure presented in [2]. The AMI dataset definition however follows [1] for a better comparison with other research groups. The 368 dimensional input to DNN are compressed from 31 contextual frames of 23 dimensional log-Mel-filter bank features with DCT [2]. The DNN topology is 368:2048×3:26:1993.

When adapting AMI models to SWC data, the trained DNN is first fine-tuned with dev data using manual transcription. The alignment is obtained with AMI DNN-HMM-GMM and SWC headset recordings. Bottleneck features are then generated with the updated DNN, followed by segmental mean normalization. The AMI HMM-GMM is then maximum-a-posterior (MAP) adapted using "AD2" dev set data for 8 iterations. Neither speaker adaptation or normalization is involved. Results of the baseline systems are reported on individual headset microphone (IHM), single distant microphone (SDM) and multiple distant microphones (MDM) in Table 4. For MDM, weighted delay and sum beamforming is performed using BeamformIt [37], with the 8 channels circular array in the integrated IML recording system ("TBL1"). The scoring for IHM is performed with NIST tool *sclite*, while the scoring for SDM and MDM is performed with *asclite* with a maximum of 4 overlapping speakers.

Even with supervised adaptation on dev data, it still yields a WER of 24.9% for IHM, 55.2% for SDM and 53.5% for MDM

Table 4: *AMI to SWC: "AD2" baseline (WER in %).*

	dev	eval					
	SWC1	SWC2	SWC3	overall			
				Sub.	Del.	Ins.	WER
IHM	24.9	46.4	50.5	33.4	9.3	5.0	**47.7**
SDM	55.2	75.0	85.2	53.2	19.1	6.0	**78.2**
MDM	53.5	71.6	82.4	52.4	15.4	7.3	**75.0**

with 8 channel beamforming. WER on eval data is much higher, particularly for SWC3 due to mismatch in gender and speaking style. Beamforming lowered the WER by 3.1% relative on SWC1, 4.5% relative on SWC2 and 3.3% relative on SWC3.

6.2. Standalone training task

A Kaldi recipe is released with the corpora, providing scripts to train a state-of-the-art context dependent DNN-HMM hybrid system on SWC data only. It follows the routine in other Kaldi recipes (such as AMI).

Following the default configuration, 13 dimensional MFCC features from 7 contextual frames (+/-3) are extracted and compressed with linear discriminant analysis (LDA) to 40 dimension. The output will be further referred to as "LDA features". The LDA features are used to train HMM-GMM. No external alignment is used in the recipe. Instead, the initial model training uses hypothesis timing where utterances are split into equal chunks. The alignment is updated each time the acoustic model significantly improves during the training process.

An HMM-GMM based on monophone is first trained, then an HMM-GMM based on clustered states is trained, followed by LDA and maximum likelihood linear transform (MLLT), speaker adaptive training (SAT), and maximum mutual information (MMI) training. Alignments from the system with LDA+MLLT is used for DNN training. The input of DNN is a 520 dimensional feature vector, comprised of 13 (+/-6) contextual 40 dimensional features that were used for HMM-GMM training. DNN parameters are initialized with a stack of restricted Boltzmann machines (RBMs), in a topology of 520:2048$\times$6:3804. DNN parameters are then fine-tuned to minimize cross-entropy. This is followed by 4 iterations of further fine-tuning for minimum phone error (MPE) or using the state level variant of the minimum Bayes risk (sMBR) training with updated alignment.

For IHM, results with speaker adaptation is provided. HMM-GMMs with LDA+MLLT+SAT provide the alignment and speaker feature level maximum likelihood linear regression (fMLLR) for DNN training. The DNN parameters are initialized with RBMs in a topology of 143:2048$\times$6:3710. DNN input features are comprised of 11 (+/-5) contextual 13 dimensional MFCC features with fMLLR applied.

For MDM, the weighted delay and sum beamforming is performed with BeamformIt [37] on 8 channel microphones from the circular array in the middle of the table. The automatic noise thresholding is disabled.

To reduce memory cost, the 30k 4-gram LM introduced in §5 is pruned. Table 5 shows the performance using different acoustic models and microphone channels. As shown, IHM SAT reduces the overall WER of HMM-GMM based system by 5.1% relative, while MMI did not reduce WER further. For DNN-HMM hybrid system however, speaker adaptation via fMLLR degraded the performance. The best overall WER of 42.0% on IHM is achieved with sMBR fine-tuning on DNN parameters without speaker adaptation. Therefore, fMLLR is not used in experiments with SDM or MDM hybrid system. Fine-tuning DNN with sMBR is effective for both SDM and MDM, achieving the best overall WER of 76.8% on SDM and 74.3% on MDM. Beamforming reduced the WER by 3.3% relative.

Table 5: *SWC "SA1" baseline (WER in %).*

		dev	eval	overall			
				Sub.	Del.	Ins.	WER
IHM	LDA+MLLT	50.9	51.8	35.9	8.9	6.4	51.3
	+SAT	48.7	**48.8**	34.4	8.1	6.3	**48.7**
	+MMI	48.8	49.1	34.4	8.8	5.7	48.9
	DNN	44.4	44.3	30.5	9.7	4.1	44.4
	+sMBR	42.0	**42.0**	29.5	7.6	5.0	**42.0**
	+fMLLR	48.1	48.1	32.9	11.4	3.8	48.1
	+sMBR	44.9	44.8	31.2	9.8	3.8	44.9
SDM	DNN	78.9	80.5	53.9	21.4	4.4	79.7
	+sMBR	76.4	**77.3**	39.1	35.5	2.2	**76.8**
MDM	DNN	76.0	77.9	53.3	18.2	5.5	76.9
	+sMBR	73.8	**74.9**	36.0	36.0	2.4	**74.3**

7. Conclusions

This paper presents the extended recordings for Sheffield Wargame Corpus, which is freely available for research use in the speech community, and which is designed for distant speech recognition work with multi-channel recordings. It includes unique ground truth annotation of speaker location. The extended corpus adds up to around 24.6h of multi-media and multi-channel data for natural native English speech. Four dataset definitions are provided for two different tasks: low resource adaptation of existing acoustic model and standalone training of acoustic model. A Kaldi recipe is provided for standalone training. Performance of baseline deep neural network systems for each task is illustrated. The WERs on the eval sets are above 40% for all systems, suggesting a high difficulty level in SWC corpora compared to other corpora. The WERs for SDM on eval set are all above 70%. Beamforming reduced the WER by 3-4% relatively. The best overall WER obtained is 42.0% for IHM, 76.8% for SDM and 74.3% for MDM.

8. Acknowledgements

This research was supported by EPSRC Programme Grant EP/I031022/1, Natural Speech Technology (NST). The results reported in this work could be accessed from `https://dx.doi.org/10.15131/shef.data.3119743`. More details and some samples of SWC recordings can be found at `http://mini-vm20.dcs.shef.ac.uk/swc/SWC-home.html`.

9. References

[1] P. Swietojanski, A. Ghoshal, and S. Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 285–290.

[2] Y. Liu, P. Zhang, and T. Hain, "Using neural network front-ends on far field multiple microphones based speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 5542–5546.

[3] T. Yoshioka, X. Chen, and M. J. F. Gales, "Impact of single-microphone dereverberation on dnn-based meeting transcription systems," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 5527–5531.

[4] K. Kumatani, J. McDonough, B. Rauch, D. Klakow, P. N. Garner, and W. Li, "Beamforming with a maximum negentropy criterion," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 5, pp. 994–1008, July 2009.

[5] K. Kumatani, J. McDonough, and B. Raj, "Maximum kurtosis beamforming with a subspace filter for distant speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, Dec 2011, pp. 179–184.

[6] M. Delcroix, T. Yoshioka, A. Ogawa, Y. Kubo, M. Fujimoto, N. Ito, K. Kinoshita, M. Espi, T. Hori, T. Nakatani *et al.*, "Linear prediction-based dereverberation with advanced speech enhancement and recognition technologies for the REVERB challenge," in *Proc. REVERB Workshop*, 2014.

[7] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 10, pp. 1533–1545, Oct. 2014.

[8] H. Sak, A. W. Senior, and F. Beaufays, "Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition," *CoRR*, vol. abs/1402.1128, 2014.

[9] G. E. Dahl, T. N. Sainath, and G. E. Hinton, "Improving deep neural networks for LVCSR using rectified linear units and dropout," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*, 2013, pp. 8609–8613.

[10] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*, Dec 2013, pp. 55–59.

[11] Y. Zhang, E. Chuangsuwanich, and J. Glass, "Extracting deep neural network bottleneck features using low-rank matrix factorization," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 185–189.

[12] M. Ravanelli and M. Omologo, "Contaminated speech training methods for robust DNN-HMM distant speech recognition," in *INTERSPEECH*. ISCA, 2015, pp. 756–760.

[13] P. Zhang, Y. Liu, and T. Hain, "Semi-supervised dnn training in meeting recognition," in *2014 IEEE Spoken Language Technology Workshop (SLT 2014)*, South Lake Tahoe, USA, December 2014.

[14] T. N. Sainath, R. J. Weiss, K. W. Wilson, A. Narayanan, M. Bacchiani, and A. Senior, "Speaker location and microphone spacing invariant acoustic modeling from raw multichannel waveforms," in *ASRU 2015*. IEEE, December 2015.

[15] X. Xiao, S. Watanabe, H. Erdogan, L. Lu, J. Hershey, M. L. Seltzer, G. Chen, Y. Zhang, M. Mandel, and D. Yu, "Deep beamforming networks for multi-channel speech recognition," in *Acoustics Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, March 2016.

[16] M. L. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7398–7402.

[17] R. Giri, M. L. Seltzer, J. Droppo, and D. Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning." IEEE Institute of Electrical and Electronics Engineers, April 2015.

[18] M. Ravanelli and M. Omologo, "On the selection of the impulse responses for distant-speech recognition based on contaminated speech training," in *INTERSPEECH*. ISCA, 2014, pp. 1028–1032.

[19] C. Fox, Y. Liu, E. Zwyssig, and T. Hain, "The sheffield wargames corpus." in *Proceedings of Interspeech 2013*, Lyon, France, August 2013.

[20] J. Wen, N. D. Gaubitch, E. Habets, T. Myatt, and P. A. Naylor, "Evaluation of speech dereverberation algorithms using the MARDY database," in *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*, Paris, France, 2006.

[21] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Digital Signal Processing, 2009 16th International Conference on*, July 2009, pp. 1–5.

[22] R. Stewart and M. Sandler, "Database of omnidirectional and B-format room impulse responses," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, March 2010, pp. 165–168.

[23] D. Pearce, H. günter Hirsch, and E. E. D. Gmbh, "The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions," in *in ISCA ITRW ASR2000*, 2000, pp. 29–32.

[24] S.-K. Au Yeung and M.-H. Siu, "Improved performance of Aurora 4 using HTK and unsupervised MLLR adaptation," in *Proceedings of Interspeech 2004—ICSLP: 8th International Conference on Spoken Language Processing, Jeju Island, Korea*, 2004, pp. 161–164.

[25] H. G. Hirsch and H. Finster, "The simulation of realistic acoustic input scenarios for speech recognition systems," in *in Proc. ICSLP*, 2005.

[26] M. Wolf and C. Nadeu, "Channel selection measures for multi-microphone speech recognition," *Speech Communication*, vol. 57, pp. 170 – 180, 2014.

[27] M. Ravanelli, L. Cristoforetti, R. Gretter, M. Pellin, A. Sosi, and M. Omologo, "The DIRHA-ENGLISH corpus and related tasks for distant-speech recognition in domestic environments," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, Dec 2015, pp. 275–282.

[28] M. Lincoln, I. McCowan, J. Vepa, and H. Maganti, "The multi-channel wall street journal audio visual corpus (MC-WSJ-AV): specification and initial experiments," in *Automatic Speech Recognition and Understanding, 2005 IEEE Workshop on*, Nov 2005, pp. 357–362.

[29] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'CHiME' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2015)*, Scottsdale, AZ, United States, Dec. 2015.

[30] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The AMI meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[31] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, E. Shriberg, A. Stolcke, and C. Wooters, "The ICSI meeting corpus," 2003, pp. 364–367.

[32] C. Fox, H. Christensen, and T. Hain, "Studio report: Linux audio for multi-speaker natural speech technology." in *Proc. Linux Audio Conference*, 2012.

[33] T. Hain, V. Wan, L. Burget, M. Karafiat, J. Dines, J. Vepa, G. Garau, and M. Lincoln, "The ami system for the transcription of speech in meetings," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–357.

[34] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in *ICSLP'02: Proc. of International Conference on Spoken Language Processing*, 2002.

[35] K. Richmond, R. Clark, and S. Fitt, "On generating combilex pronunciations via morphological analysis," in *Proceedings of ISCA Interspeech*, Makuhari, Japan, 2010, pp. 1974–1977.

[36] J. Novak, N. Minematsu, and K. Hirose, "WSFT–based grapheme–to–phoneme conversion: Open source tools for alignment, model–building and decoding," in *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, San Sebastián, Spain, 2012.

[37] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, Sept 2007.