# USING NEURAL NETWORK FRONT-ENDS ON FAR FIELD MULTIPLE MICROPHONES BASED SPEECH RECOGNITION

*Yulan Liu[1], Pengyuan Zhang[2] and Thomas Hain[1]*

[1] Speech and Hearing Research Group, The University of Sheffield, Sheffield, UK
[2] Key Laboratory of Speech Acoustics and Content Understanding, IACAS, Beijing, China

## ABSTRACT

This paper presents an investigation of far field speech recognition using beamforming and channel concatenation in the context of Deep Neural Network (DNN) based feature extraction. While speech enhancement with beamforming is attractive, the algorithms are typically signal-based with no information about the special properties of speech. A simple alternative to beamforming is concatenating multiple channel features. Results presented in this paper indicate that channel concatenation gives similar or better results. On average the DNN front-end yields a 25% relative reduction in Word Error Rate (WER). Further experiments aim at including relevant information in training adapted DNN features. Augmenting the standard DNN input with the bottleneck feature from a Speaker Aware Deep Neural Network (SADNN) shows a general advantage over the standard DNN based recognition system, and yields additional improvements for far field speech recognition.

***Index Terms—*** speech recognition, multiple microphone, beamforming, deep neural networks

## 1. INTRODUCTION

Reducing the ASR performance gap between far field and close-talking recordings has been an important research topic for a long time. Typically, multiple channel data is enhanced before recognition. The most representative multi-channel signal enhancement methods are beamforming techniques [1], which perform a channel based noise reduction, temporal and spatial filtering [2, 3]. Advanced beamforming also considers the correlation among different channels [4], and even extends the beamforming optimization with maximizing the speech recognition likelihood [5, 6]. [7] introduced the direct concatenation of multi-channel features and compared the performance of direct channel concatenation with beamforming using standard PLP features. The experiments showed that equal or better recognition performance can be achieved using direct channel concatenation, and that the WER tends to be higher when the speakers are standing without head movement.

Neural network based features have long been used successfully in meeting recognition [8, 9, 10]. While the early research did not involve deep layers [11], the path towards deep learning was laid in the stacking of bottleneck networks [9]. A DNN is a conventional Multi-Layer Perceptron (MLP) with many internal or hidden layers. The BottleNeck (BN) features extracted from the internal layer with a relatively small amount of neurons have been shown to effectively improve the performance of ASR systems. It is possibly due to the limited number of units which creates a constriction in the network and further forces the information pertinent to classification into a low dimensional representation [12, 13, 14]. In many ASR systems, the neural network based features and the cepstrum or spectrum based features (e.g. MFCC, PLP) are supposed to provide complementary information. Using them jointly further improved the recognition performance over any single one.

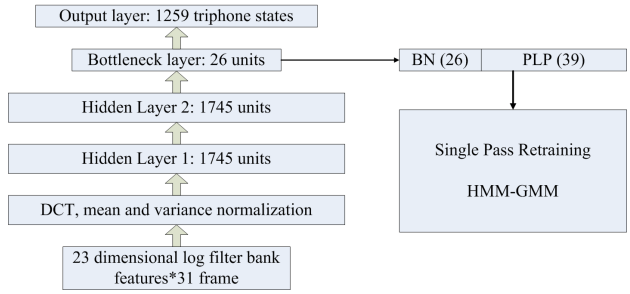


**Fig. 1**. DNN module structure

In far field ASR, the information relevant to speech recognition (meta-information) can be encoded in different forms besides the standard features. The speaker location can be encoded with the Time Difference of Arrival (TDOA) values, and the speaker identity can be represented with speaker adaptive models. Projecting standard features into a test speaker relevant space has shown to improve the ASR performance, on both the classic features [15, 16] and the DNN based features [17], both the close-talking data and the far-field data. In recent research on speaker robustness of DNN, speaker representative features like speaker code [18] and i-Vector [19] are used to perform static speaker adaptation of the DNN neuron bias.

Work in this paper aims to extend and enhance the preliminary work in [7] with redefined training and test sets (as outlined in §2.1) and with advanced DNN based feature extraction. We followed the standard strategies where the DNN includes one input layer, two or three hidden layers, another bottleneck layer and one output layer. The BN features are extracted from the bottleneck layer of DNN trained to predict the context-dependent clustered triphone states. Figure 1 shows the architecture. It is similar to those in [20, 21, 22]. We extended the concept of direct channel concatenation to DNN. Similar work but with a different focus and system setup is referred in [17]. The DNN based front-end allows a very flexible integration of meta-information. Thus we investigated integrating different meta-information dynamically and statically into the DNN based front-end to improve far-field ASR performance. We have shown that a dynamic speaker adaptation of DNN based on 13 dimensional speaker awareness bottleneck features improves far-field recognition performance, especially when the direct channel concatenation is used at the same time.

## 2. RECOGNITION OF MEETING DATA

Recognition of speech recorded in meetings is interesting for several reasons: recordings can be made in a fairly controlled setting with multi-channel far field and close-talking recording devices; speak-

[1]This work was supported by the EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

[2] This work was supported by Chinese Academy of Sciences Fellowship for Visiting Scholars.

**Table 1**. %WER using PLP features, models are trained on *acntrain*, tested on *acntest* ("*bmit*": beamforming with BeamformIt).

| Conf | IHM | SDM | 2bmit | 4bmit | 8bmit |
|---|---|---|---|---|---|
| - | 35.6 | 66.3 | 61.8 | 60.5 | 58.2 |
| HLDA | 34.7 | 65.3 | 61.3 | 59.9 | 57.8 |

**Table 2**. %WER using PLP features, models are trained on *acftrain*, tested on *acftest*. "#o" refers to the maximal number of overlapping speakers in scoring, using NIST scoring toolkit *sctk 2.4.8*.

| #o | IHM | SDM | 2bmit | 4bmit | 8bmit |
|---|---|---|---|---|---|
| 0 | 32.3 | 61.3 | 57.1 | 56.0 | 53.8 |
| 4 | 35.4 | 65.1 | 60.4 | 59.8 | 58.2 |

ing style ranges from very formal to highly conversational; multiple speakers can speak at the same time, thus requiring the ASR to be robust to environmental effects. The following sections describe the data used and the baseline performance.

### 2.1. Data and Basic Setup

The AMI corpus [23] is used in our experiments due to its size, high quality recordings and multiple levels of annotation with meta-information like speakers' head and body movement. It includes recordings from Individual Headset Microphones (IHM) and Multiple Distant Microphones (MDM) in microphone arrays of which the first channel is referred as the Single Distant Microphone (SDM).

Segments with annotation of the speakers' movement status were included in our experiments. In the following, *H+*/*H-* refers to presence or absence of head movement, while *M+*/*M-* refers to presence or absence of body movement. Taking the balance of these movement categorisations into account, we defined a 87.7 hours training set *acftrain* and a 6.1 hour test set *acftest* which excludes the meetings seen in the training set but includes some seen speakers. A 15.8 hour sub-training set *acntrain* is defined with all the non-overlapping speech in *acftrain*. Similarly a 1.9 hour sub-test set *acntest* is defined with all the non-overlapping speech in *acftest*.

All experiments have equivalently configured 16 mixture component HMM-GMMs trained with maximum likelihood criterion. All model sets for the same training set have approximately the same amount of clustered triphone states. Viterbi decoding is performed with the AMI RT'09 trigram language model and dictionary [10].

### 2.2. Baseline Experiments

The experiments presented here are based on 39 dimensional feature vectors composed of 12 PLP coefficients plus $c_0$, and their delta and delta-delta features. Segmental Mean Normalisation (SMN) is applied in all experiments. Some experiments also include Heteroscedastic Linear Discriminant Analysis (HLDA, [24]). Beamforming is performed with the toolkit BeamformIt [25].

Tables 1 and 2 show the WER results of PLP features on non-overlapping subsets and full datasets respectively. Results on non-overlapping speech include those obtained with feature dimension reduction from 39 to 26 using HLDA, to compare with the 26 dimensional bottleneck features to show later. With more training data, WER in far field conditions reduced by around 4.4% absolute. And the WERs on the speech with a maximum of 4 overlapping speakers are around 3.7% absolute higher than the non-overlapping speech.

**Table 3**. %WER using linear BN features (*acntrain*, *acntest*).

| Conf | IHM | SDM | 2bmit | 4bmit | 8bmit |
|---|---|---|---|---|---|
| *2ML0* | 29.9 | 53.7 | 51.6 | 51.2 | 50.5 |
| *2TL* | 26.6 | 49.5 | 46.8 | 46.3 | 45.6 |
| *3TL* | 26.8 | 49.3 | 47.8 | 46.9 | 45.8 |

## 3. NEURAL NETWORK BASED FEATURES

### 3.1. Configurations

All DNNs are trained feed-forward with the TNET toolkit[1] on GTX690 based GPUs. In a default TNET setup, 31 adjacent frame log filter bank features are decorrelated and compressed with DCT into a dimension of 368 ($31\times23\rightarrow16\times23$). Global mean and variance normalization are performed on each dimension before feeding as the DNN input. The 5 layered DNN structure is shown in Figure 1. For 6 layered DNN, an extra hidden layer composed of 1745 units will be inserted before the 26 dimensional bottleneck layer. On average 10% data in (*acntrain* and *acftrain*) is reserved for cross validation in DNN training. The training stops automatically when the improvement of frame-based target classification accuracy on the cross validation set drops to below 0.1%.

The bottleneck layer is placed just before the output layer, as in our initial experiments this topology gives the best performance. DNNs are trained on classification targets of monophone (*M*) or triphone states (*T*). In the bottleneck layer, linear (*L*) or sigmoidal (*S*) BN features are extracted respectively before or after the sigmoid activation. Hence a configuration abbreviation of "*2TS*" denotes BN features extracted after the sigmoid function from a 5 layered DNN (2 extra hidden layers plus an input layer, an bottleneck layer and an output layer) trained on triphone targets classification. If a DNN is initialized randomly, a "*0*" is attached at the end of configuration abbreviation (e.g. "*2ML0*" in table 3). Otherwise it is trained layer by layer. Standard HMM-GMM models are trained on the BN features with Single Pass Retraining (SPR) from the corresponding PLP models with HLDA (Table 1), with 8 iteration Baum-Welch re-estimation followed. SMN is performed in all experiments.

### 3.2. Bottleneck Features Only

Table 3 shows the results for BN features of different configurations on IHM, SDM and MDM with beamformed audio by BeamformIt. All the DNNs are trained with the triphone state targets force-aligned using IHM data on the *acntrain* training set. The decoding is performed on the *acntest* test set. On average there is 20% relative WER improvement over the standard PLP features (Table 1). Performance for more than 2 extra hidden layers seems to decrease. Note that linear BN features are used because the variance value of sigmoidal BN features is very small, which leads to complications in Baum-Welch training due to overly dominant posteriors.

### 3.3. Concatenation of DNN and PLP Features

Table 4 shows the results for BN+PLP feature concatenation. Overall there are 8.1% absolute WER improvement for IHM data and 12.5% for 8 channel beamformed data over the PLP baseline.

Experiments here addressed three issues in the DNN front-end: network depth, monophone or triphone target, and linear or sigmoidal feature output. While these seem rather technical, we have observed mixed results depending on the exact nature of the input throughout experiments. Hence these results are reported here.

[1] http://speech.fit.vutbr.cz/software/neural-network-trainer-tnet

**Table 4**. %WER using BN+PLP feature concatenation under different DNN configuration (*acntrain*, *acntest*).

| | Conf | IHM | SDM | 2bmit | 4bmit | 8bmit |
|---|---|---|---|---|---|---|
| - | *2TL* | 26.8 | 49.9 | 48.0 | 47.4 | 45.7 |
| - | *2TS* | 26.7 | 49.9 | **46.9** | 46.8 | 45.3 |
| - | *3TL* | **26.3** | **49.4** | 47.6 | 47.0 | 45.6 |
| - | *3TS* | 26.5 | 49.5 | 47.2 | **46.5** | **44.5** |
| STC | *2TS* | 26.5 | 49.6 | 47.1 | 47.0 | 45.4 |
| STC | *3TS* | 26.0 | 48.9 | 46.9 | 46.2 | 44.9 |

**Table 5**. %WER using BN+PLP concatenation (*acftrain*, *acftest*).

| Conf | #o | IHM | SDM | 2bmit | 4bmit | 8bmit |
|---|---|---|---|---|---|---|
| *2TS* | 0 | 22.1 | 43.5 | 41.8 | 41.2 | 39.5 |
| *2TS* | 4 | 23.9 | 48.5 | 46.8 | 46.9 | 45.1 |

As shown, the recognition performance improves with more microphones involved in beamforming. Among different configurations of DNN and BN features, the best performance is observed with *3TS* for 4 and 8 channel MDM, with *2TS* for 2 channel MDM, and with *3TL* for IHM and SDM data. Interestingly the difference between using linear BN features (Table 3) and using BN+PLP concatenated features is small and it varies among different recording channels. The concatenation with PLP gave a slight improvement in the majority with *3TL* configuration, but not with *2TL*.

Further experiments in Table 4 investigated using Semi-Tied Covariance matrices transformation (STC, [26, 27]) for feature decorrelation. All models were then retrained using a full re-clustering and mixup procedure while keeping roughly the same amount of states. Similar to concatenating with PLP features, STC brought a slight improvement in most cases with *3TS*, but not with *2TS*.

### 3.4. Overlapping Speech

Speech overlap is a key feature in meeting speech recognition. While acoustic beamforming can in theory address this problem, in practice the algorithms do not allow concurrent speech recognition (e.g. due to maximum loudness target selection). Table 5 shows the performance of BN+PLP with overlapping speech in both the training set *acftrain* and test set *acftest*. Results of #o=0 scoring in Table 5 should be compared with results of the *2TS* configuration in Table 4. For non-overlapping test set, a five-fold amount increase in training data gives an average WER decrease by a 13.1% relative. Compared to using PLP features (Table 2), the WER gap between non-overlapping speech and speech with maximal 4 overlapping speakers increased on average from 3.7% to 4.6% absolute and from 6.8% to 10.6% relative.

## 4. DIRECT MULTI-CHANNEL INPUT

Compared to the direct concatenation of BN and PLP features in §3.3, a direct concatenation of multi-channel PLP features leads to very high-dimensional input for HMM-GMM training, which is problematic. For this reason, any BN+PLP feature concatenation in this section refers to using the PLP features from the first channel only[2]. Concatenation is performed on the compressed log filter bank features from multiple channels in the DNN input.

### 4.1. Concatenation Versus Beamforming in DNN Input

Table 6 shows the results of direct channel concatenation applied on PLP features and on DNN training input. The #o=0 scoring over direct channel concatenation of PLP in Table 6 should be compared with the PLP baseline results in Table 1. The 2 channel concatenation achieved 4.2% absolute WER improvement over SDM, only 0.3% less than 2 channel beamforming. However concatenation generates large feature dimensionality and HMM-GMM parameter amount, which makes higher order concatenation intractable. In a comparison between Table 6 and Table 4, the DNN front-end achieved a small but consistent improvement for the 2 and 4 channels MDM by using direct channel concatenation over standard beamforming. The performance for 8 channel concatenation degraded possibly because of large DNN input layer. Table 6 also shows that between the two different implementations of direct channel concatenation, the WER gap between the overlapping speech and the non-overlapping speech is larger when using DNN front-end and BN+PLP features than using PLP features.

[2]We tried concatenating with more channels and even beamformed channel, but there was no evident improvement over using the single first channel.

**Table 6**. %WERs using direct channel concatenation on PLP features and DNN training input, trained with non-overlapping speech (*acntrain*, *acftest*) ("*cct*": direct channel concatenation).

| Feature | Conf | #o | 2cct | 4cct | 8cct |
|---|---|---|---|---|---|
| PLP | - | 0 | 62.1 | 62.2 | - |
| PLP | - | 4 | 67.9 | 68.3 | - |
| BN+PLP | *2TS* | 0 | 46.8 | 46.5 | 47.4 |
| BN+PLP | *2TS* | 4 | 54.1 | 54.7 | 55.6 |

**Table 7**. %WER using channel concatenation on DNN training input trained with overlapping speech (*acftrain*, *acftest*).

| Conf | #o | 2cct | 4cct | 8cct |
|---|---|---|---|---|
| *2TS* | 0 | 41.1 | 40.3 | 41.7 |
| *2TS* | 4 | 46.4 | 46.2 | 47.8 |

Table 7 shows equivalent results trained with the full training set (*acftrain*). Compared to the Table 2, for 2 and 4 channels there is an an average 28% relative WER improvement on the non-overlapping test speech, and 23% relative WER improvement on the full test set. Compared to the beamforming results in Table 5, direct concatenation of 2 and 4 channel achieves better performance than beamforming on both overlapping and non-overlapping speech, while the loss in 8 channels remains. The WER gap between non-overlapping and overlapping speech also reduced compared with training on non-overlapping speech (*acntrain*) only (Table 6).

### 4.2. Analysis

So far the best performance on 2 and 4 channel MDM data has been achieved with direct channel concatenation as the DNN input. It is important to understand where the improvement occurs. Thus we analysed the WER in terms of head and body movement (Table 8). The recognition performance with IHM tends to be better with head or body movement than without, possibly due to cognitive load. Distant microphones are quite sensitive to speakers' body movement, and the recognition performance is evidently better when the speaker sits or only moves a little (e.g. *take_notes*). The direct concatenation tends to improve the moving speech better than beamforming, possibly because it better preserves the signals from all directions. As a result, the WER gap between with body movement (*M+*) and without body movement (*M-*) is smaller using direct channel concatenation than standard beamforming. And head movement appears to have more impact than body movement in all cases.

**Table 8**. %WER using BN+PLP (*2TS*) features; trained on *acftrain*, decoded on *acftest*, and scored on maximal 4 overlapping segments. (*M+*: 2.42 hours; *M-*: 3.67 hours; *H+*: 2.72 hours; *H-*: 2.29 hours)

| | IHM | SDM | 2bmit | 4bmit | 8bmit | 2cct | 4cct |
|---|---|---|---|---|---|---|---|
| *M+* | 22.4 | 49.4 | 48.3 | 49.0 | 47.0 | 47.0 | 46.1 |
| *M-* | 24.6 | 47.4 | 45.3 | 45.0 | 43.2 | 45.5 | 45.6 |
| *H+* | 23.3 | 45.8 | 43.9 | 43.7 | 41.5 | 43.7 | 43.5 |
| *H-* | 25.7 | 51.3 | 49.5 | 49.5 | 48.4 | 49.5 | 49.3 |

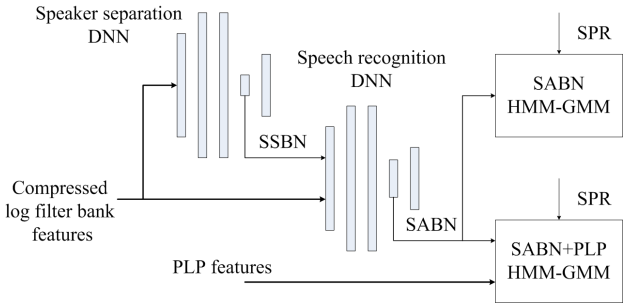


**Fig. 2**. SABN feature computation.

## 5. ADDING INFORMATION FOR DNN ADAPTATION

### 5.1. Speaker Separation Features

Similar to training DNNs for phoneme discrimination, a Speaker Separation DNN (SSDNN) to discriminate the speakers in the training set was trained. SSDNN uses the speaker identities as the targets and the same compressed log filter bank features as the input. Around 5% data from each speaker is randomly chosen and reserved for cross validation in SSDNN training. The 13 dimensional sigmoidal BN features generated from SSDNN (Figure 2) provide a projection similar to iVectors [28], but dynamically. The BN feature of this SSDNN is further referred to as the Speaker Separation BottleNeck (SSBN) feature. SSBN is appended to the compressed log filter bank features in the input of a standard feedforward DNN (Speaker Aware DNN, SADNN) for speech recognition. The BN features obtained in this form are further referred as the Speaker Aware BottleNeck (SABN) features. Figure 2 shows the process. Compared to standard bottleneck features, the SABN features requires training an extra SSDNN. As a reward it saves the effort of 2 pass decoding, as it's not necessary to estimate the speaker code or speaker dependent adaptation for the test set. This provides the possibility, flexibility and generality of fast adaptation.

As the SSBN varies from frame to frame, this operation is equivalent to a dynamic bias adaptation in the input layer. Figure 3 shows its influence on the frame accuracy improvement in each iteration of SADNN training. Table 9 compares the performance using SABN features with standard BN features. The SABN features achieved better performance than corresponding BN features in most cases, and the best performance on MDM is achieved by using SABN features together with direct channel concatenation.

### 5.2. Location and Speaker Identities

We further experimented with similar DNN adaptation with information on speaker location and speaker identities. The physical location of speakers is unfortunately not available for the AMI corpus data. Instead, we used the TDOA computed by BeamformIt as a proxy since this was successfully applied to a speaker diarisation task [29]. However, adding the TDOA values directly in the input (similar to SSBN) gave no improvement over the BN+PLP baseline, with 46.8% WER on *acntest* for 4 channel concatenation. A second set of experiments constructed speaker GMMs on the test data (with 8 mixture components) on the basis of linear SSBN features. The Gaussian means were sorted by weight then added to the DNN input. However, the performance degraded substantially, to 52.2%.

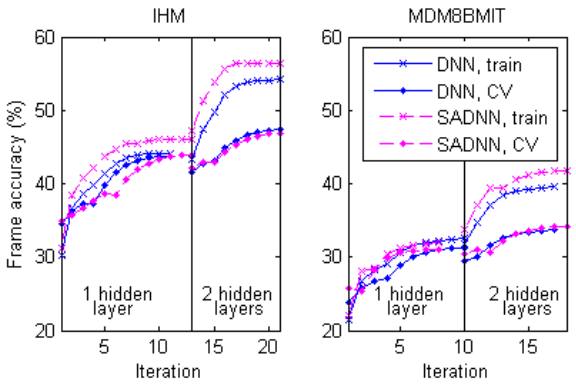


**Fig. 3**. Frame accuracy change in DNN and SADNN training (*acntrain*, *acntest*). The "*CV*" and "*train*" in the figure respectively refer to the cross validation and training subsets in *acntrain*.

**Table 9**. %WER using BN or SABN features (*acntrain*, *acntest*).

| | IHM | SDM | | 2 | 4 | 8 |
|---|---|---|---|---|---|---|
| BN+PLP (*2TS*) | 26.7 | 49.9 | bmit | 46.9 | 46.8 | 45.3 |
| | | | cct | 46.8 | 46.5 | 47.4 |
| SABN+PLP (*2TS*) | **26.1** | 49.8 | bmit | 47.4 | 46.0 | **44.7** |
| | | | cct | 46.8 | 45.5 | - |
| SABN (*2TL*) | 26.5 | **48.9** | bmit | 47.4 | 46.0 | 44.8 |
| | | | cct | **45.7** | **44.8** | - |

## 6. DISCUSSION AND CONCLUSIONS

The DNN based front-end on the AMI meeting corpus was tested in the context of far field speech recognition. Experiments showed that on both overlapping and non-overlapping speech, BN features gave an average 25% relative WER reduction over using PLP features, regardless of the number and type of microphones. As for DNN training input, using direct channel concatenation performs similar or better than using beamforming for the 2 and 4 microphone cases. Results for more microphones degrades potentially due to large feature dimensionality. An analysis of WER by speaker body and head movement shows that the WER with IHM tends to be lower with moving speakers, hinting at increased cognitive load of the speaker. Body movement challenges beamforming while the direct channel concatenation performs better on moving speech. With BN features, general performance degradation caused by overlapping speech varies from 1.8% absolute WER on IHM to 5.7% on 4 channel beamforming even with sufficient training data *acftrain* including both overlapping and non-overlapping speech. Experiments aimed at providing additional information for dynamic adaptation of DNN might be a start of further work on aiding far field recognition with dynamic meta-information. The flexibility of neural network allows using features from related tasks (e.g. speaker separation) for speech recognition tasked DNN adaptation. Shown results indicate performance improvements with speaker information, particularly in the far-field channel concatenation scenario.

## 7. REFERENCES


[1] M. Wöelfel and J. McDonough, *Distant Speech Recognition*, Wiley, 2009.

[2] D. V. Compernolle, W. Ma, F. Xie, and M. V. Diest, "Speech recognition in noisy environments with the aid of microphone arrays," *Speech Communication*, vol. 9, no. 5–6, pp. 433 – 442, 1990.

[3] K. Eneman and M. Moonen, "Multimicrophone speech dereverberation: Experimental validation," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 1, pp. 051831, 2007.

[4] K. Kumatani, T. Gehrig, U. Mayer, E. Stoimenov, J. McDonough, and M. Wolfel, "Adaptive beamforming with a minimum mutual information criterion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 8, pp. 2527–2541, 2007.

[5] M.L. Seltzer, B. Raj, and R.M. Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *Speech and Audio Processing, IEEE Transactions on*, vol. 12, no. 5, pp. 489–498, 2004.

[6] M. L. Seltzer and R. M. Stern, "Subband likelihood maximizing beamforming for speech recognition in reverberant environments," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 14, pp. 2109–2121, 2006.

[7] D. Marino and T. Hain, "An analysis of automatic speech recognition with multiple microphones.," in *INTERSPEECH*. 2011, pp. 1281–1284, ISCA.

[8] A. Stolcke, X. Anguera, K. Boakye, O. Cetin, A. Janin, A. Mandal, B. Peskin, C. Wooters, and J. Zheng, "Further progress in meeting recognition: The ICSI-SRI spring 2005 speech-to-text evaluation system," in *Machine Learning for Multimodal Interaction*, number 3869 in LNCS, pp. 463–475. Springer Verlag, 2005.

[9] F. Grezl, M. Karafiat, and L. Burget, "Investigation into bottleneck features for meeting speech recognition," in *Proc. Interspeech 2009*, 2009, number 9, pp. 2947–2950.

[10] T. Hain, L. Burget, J. Dines, P. N Garner, F. Grezl, A. el Hannani, M. Huijbregts, M. Karafiat, M. Lincoln, and V. Wan, "Transcribing meetings with the AMIDA systems," *IEEE Transactions on Audio, Speech and Language Processing*, Aug. 2011.

[11] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring strategies for training deep neural networks," *J. Mach. Learn. Res.*, vol. 10, pp. 1–40, June 2009.

[12] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in *INTERSPEECH*, 2011, pp. 237–240.

[13] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *INTERSPEECH*, 2011, pp. 437–440.

[14] Z. Tüske, R. Schlüter, and H. Ney, "Deep hierarchical bottleneck MRASTA features for LVCSR," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Vancouver, Canada, May 2013, pp. 6970–6974.

[15] C.J. Leggetter and P.C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models," *Computer Speech and Language*, vol. 9, no. 2, pp. 171 – 185, 1995.

[16] J. P. Neto, L. B. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker-adaptation for hybrid HMM-ANN continuous speech recognition system.," in *EUROSPEECH*. 1995, ISCA.

[17] 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.

[18] O. Abdel-Hamid and Hui Jiang, "Fast speaker adaptation of hybrid NN/HMM model for speech recognition based on discriminative learning of speaker code," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, May 2013, pp. 7942–7946.

[19] 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.

[20] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and bottle-neck features for LVCSR of meetings," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 2007, vol. 4, pp. IV–757–IV–760.

[21] F. Grezl and P. Fousek, "Optimizing bottle-neck features for LVCSR," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 2008, pp. 4729–4732.

[22] K. Veselý, M. Karafiát, and F. Grézl, "Convolutive bottleneck network features for LVCSR.," in *ASRU*, David Nahamoo and Michael Picheny, Eds. 2011, pp. 42–47, IEEE.

[23] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," vol. 3869, pp. 28–39, 2006.

[24] N. Kumar and A. G. Andreou, "Heteroscedastic discriminant analysis and reduced rank HMMs for improved speech recognition," *Speech Communication*, vol. 26, no. 4, pp. 283 – 297, 1998.

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

[26] M. 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] M. Karafiát, F. Grézl, M. Hannemann, K. Veselý, and J. Černocký, "BUT BABEL system for spontaneous cantonese," in *Proceedings of Interspeech 2013*. 2013, number 8, pp. 2589–2593, International Speech Communication Association.

[28] M. Karafiát, L. Burget, P. Matejka, O. Glembek, and J. Cernocký, "iVector-based discriminative adaptation for automatic speech recognition.," in *ASRU*, David Nahamoo and Michael Picheny, Eds. 2011, pp. 152–157, IEEE.

[29] J. Ajmera, G. Lathoud, and I. Mc-Cowan, "Clustering and segmenting speakers and their locations in meetings," in *ICASSP'04*, 2004, pp. 605–608.