# A Study on Deep Neural Network Acoustic Model Adaptation for Robust Far-field Speech Recognition

*Seyedmahdad Mirsamadi, John H.L. Hansen*[⋆]

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX 75080-3021, U.S.A.
mirsamadi@utdallas.edu, john.hansen@utdallas.edu

10.21437/Interspeech.2015-525

## Abstract

Even though deep neural network acoustic models provide an increased degree of robustness in automatic speech recognition, there is still a large performance drop in the task of far-field speech recognition in reverberant and noisy environments. In this study, we explore DNN adaptation techniques to achieve improved robustness to environmental mismatch for far-field speech recognition. In contrast to many recent studies investigating the role of feature processing in DNN-HMM systems, we focus on adaptation of a clean-trained DNN model to speech data captured by a distant-talking microphone in a target environment with substantial reverberation and noise. We show that significant performance gains can be obtained by discriminatively estimating a set of adaptation parameters to compensate the mismatch between a clean-trained model and a small set of noisy and reverberant adaptation data. Using various adaptation strategies, relative word error rate improvements of up to 16% could be obtained on the single-channel task of the recent Aspire challenge.

**Index Terms**: far-field speech recognition, deep neural network acoustic modelling, DNN adaptation

## 1. Introduction

In the past few years, deep neural networks (DNNs) have replaced Gaussian mixture models (GMMs) for acoustic modelling in automatic speech recognition. Modern hardware has now made it possible to train networks with multiple hidden layers and millions of parameters which exhibit much stronger modelling capabilities compared to the conventional GMM-HMM systems. Although DNN-HMM systems are often trained on large databases in order to learn variabilities in the input feature space, there is still a large performance drop when mismatch exists between train and test conditions.

Far-field speech recognition is one of the tasks in which there is considerable mismatch between train and test features due to room reverberation and environmental noise. A great deal of research has been devoted in the past to noise and reverberation robustness in GMM-HMM systems [1, 2]. These efforts can be broadly categorized into three distinct groups: Feature enhancement approaches [2] aim to remove the distortions caused by reverberation and noise from the speech features and acquire enhanced feature vectors which better match the training conditions. The second group of approaches are robust feature extraction strategies [3, 4] which try to design invariant features that are less influenced by environmental distortions. Finally, model adaptation methods such as maximum likelihood linear regression (MLLR) [5] and maximum a posteriori (MAP) estimation [6] try to update the parameters of the model to best match a set of observed features. Although these methods have originally been developed for speaker adaptation, they are also able to provide significant improvements when the mismatch is due to environmental distortions.

In the context of DNN-HMM acoustic models, it has been shown that even without any explicit noise or reverberation compensation, the DNN is capable of providing results that are close to, or in some cases even better than a state of the art GMM-HMM system with various compensation strategies [7]. This is mainly a result of the fact that the higher layers in a DNN provide representations of the input data that are increasingly invariant to small changes in the input features [8]. However, in spite of this considerable improvement, the performance of a DNN-HMM system in a distant-talking scenario is still much worse compared to the close-talking counterpart. This calls for environmental robustness strategies to be used with DNN-HMM systems.

A comprehensive study on feature enhancement techniques for robust ASR with DNN-HMM systems is carried out in [9]. The results show that many of the front-end approaches previously used with GMM-HMM systems can still provide considerable improvements to a DNN-HMM system, although the relative improvements might be smaller than those with GMM-HMMs due to the inherent robustness of the DNN back-end. Furthermore, the study in [10] explores the use of robust feature extraction methods with a DNN-HMM system and reports improvements to the baseline Mel-filterbank coefficient features by using robust features such as Gammatone filterbank coefficients or normalized modulation coefficients [11].

Unlike feature enhancement and robust feature extraction methods which can be somewhat directly applied to a DNN-HMM system without any change, model adaptation techniques developed for GMM-HMM systems are inherently not suitable for DNN-based acoustic models. This is because GMMs are generative models for which adaptation consists of finding transformations which maximize the likelihood of the data given the adapted model. DNN acoustic models, in contrast, are discriminatively trained as classifiers for the acoustic characteristics of the input, and thus cannot be adapted by maximum likelihood estimations.

Given the increasing popularity of DNN-HMMs as acoustic models, there have recently been a number of studies addressing DNN adaptation for ASR. These include linear transformation and its variants [12, 13, 14], conservative training in

⋆ This project was funded by AFRL under contract FA8750-12-1-0188 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

This paper makes use of IARPA ASpIRE data for evaluation.

the form of either regularized objective functions [15] or conventional L1/L2 regularization [16], factorized adaptation [17], and also the use of auxiliary features in the form of noise-aware or speaker-aware training and decoding [7, 18]. These methods, which have mostly been developed for speaker adaptation, aim to adjust a small number of adaptation parameters without harming or erasing the information learned by the network during training. Although these approaches have been shown to provide improvements in speaker adaptation, fewer studies have investigated their performance in environment adaptation.

In this study, we investigate the use of DNN adaptation methods for improved far-field speech recognition in reverberant and noisy environments. Given a DNN-HMM acoustic model trained on clean data as well as some adaptation data from the target environment, our goal is to investigate how effective DNN adaptation approaches are in reducing the mismatch between the model and the test data. We employ various adaptation strategies and evaluate the resulting ASR performance with different amounts of adaptation data. Specifically, we are interested to find out which adaptation strategy is more suitable for handling the types of distortion encountered in distance-based speech capture (i.e. room reverberation and environmental noise). We also evaluate the performance of various adaptation methods based on the amount of data they require for a reliable estimation of the parameters.

The rest of this paper is organized as follows. In Sec. 2 we briefly review the hybrid context-dependent DNN-HMM acoustic modelling technique. In Sec. 3, we present different adaptation strategies that can be used with DNNs. We present the results of ASR experiments in Sec. 4, and provide a summary of the work and concluding remarks in Sec. 5

# 2. A brief review of hybrid DNN-HMM systems

A deep neural network can be trained to predict posterior probabilities for the tied HMM states (senones) in a speech recognition system. Given a concatenated feature vector $\mathbf{x}_t$ from a context window of frames, the DNN applies multiple layers of nonlinear transformations of the form:

$$\mathbf{v}_t^{l+1} = \sigma(\mathbf{W}^l \mathbf{v}_t^l + \mathbf{b}^l), \tag{1}$$

where $\sigma()$ is often a sigmoid function and $\mathbf{v}_t^l$ represents the hidden unit values at layer $l$, with $\mathbf{v}_t^0 = \mathbf{x}_t$. The activations from the last sigmoid layer ($\mathbf{v}_t^L$) are processed by a logistic regression model (softmax layer) to produce senone posteriors:

$$p(s|\mathbf{x}_t) = \frac{\exp(\mathbf{w}_s^L \mathbf{v}_t^L + b_s^L)}{\sum_j \exp(\mathbf{w}_j^L \mathbf{v}_t^L + b_j^L)}, \tag{2}$$

where $\mathbf{w}_s^L$ represents the weights belonging to the output node representing senone $s$. By performing forced-alignment using a conventional GMM-HMM system, frame-level senone labels are obtained for the training data. The cross-entropy between these ground-truth targets (denoted by the distribution $p_{GT}(s)$) and the posteriors in Eq. (2) is often used as an optimization criterion for DNN training:

$$C = -\frac{1}{T} \sum_{t=1}^{T} \sum_{s=1}^{N} p_{GT}\left(s|\mathbf{x}_t\right) \log\Big(p\big(s|\mathbf{x}_t\big)\Big). \tag{3}$$

The DNN parameters are tuned by stochastic gradient descent using error back propagation to minimize (3). For decoding, posteriors generated by the DNN for the test data are first converted to scaled likelihoods using

$$p\big(\mathbf{x}_t|s\big) \sim \frac{p\big(s|\mathbf{x}_t\big)}{p(s)}, \tag{4}$$

where the statistics $p(s)$ are collected from the force-aligned training data. A Viterbi decoding on these likelihoods is finally used to produce the recognized hypothesis.

# 3. DNN adaptation

## 3.1. Linear Transformations

The simplest and most common method for DNN adaptation is to apply an affine transformation to either the input features, activations of a hidden layer, or the inputs to the softmax layer of the network. In the case of input features, the approach is referred to as Linear Input Network (LIN), and is similar in form to the feature-space MLLR (fMLLR), which is a common adaptation technique for GMMs. Note that in spite of this similarity in form, LIN is fundamentally different from fMLLR in that the parameters of the affine transformation are tuned discriminatively using frame-level senone labels of the adaptation data.

Assuming $\mathbf{x}_t$ to be a context-dependent feature vector (concatenation of multiple consecutive frames), LIN applies an affine transformation of the form,

$$\mathbf{y}_t = \mathbf{W}\mathbf{x}_t + \mathbf{b}. \tag{5}$$

The senone posteriors associated with $\mathbf{x}_t$ are computed based on this transformed representation and compared to the ground truth label obtained from the force-aligned adaptation data. Using the cross-entropy objective function in Eq. (3), and by back-propagating the error to the LIN layer, $\mathbf{W}$ and $\mathbf{b}$ are updated using stochastic gradient descent.

Given the fact that $\mathbf{x}_t$ is a concatenation of features from multiple frames, the transformation matrix is sometimes constrained to be block-diagonal, with the parameters of the diagonal blocks tied together [14]. The resulting adaptation strategy is referred to as feature-discriminative linear regression (fDLR). Note that this is reasonable for speaker adaptation where we are interested in transforming a single feature vector to replace speaker characteristics, and can lead to improvements specially with small adaptation data due to the use of fewer parameters. However, in adaptation for environmental distortions, specially in reverberant environments, context information and the correlations between adjacent frames are important information to be used for adaptation. It is therefore desirable in such cases to use the unconstrained transformation matrix $\mathbf{W}$, although this will require more adaptation data for estimating a larger number of adaptation parameters.

Similar transformations can be applied to the activations of a hidden layer in the network, or the input to the softmax layer. The resulting adaptation strategies are called Linear Hidden Network (LHN) and Linear Output Network (LON), respectively. LHN and LON are equivalent to adding an extra layer with linear activation to the network. LIN, on the other hand, can be considered either as an extra layer appended to the DNN (model adaptation), or simply a transformation on the features (feature-space adaptation).

Although LIN, LHN, and LON are similar in nature, their adaptation capability in different tasks may be quite different depending on the amount of available adaptation data and the nature of the distortion. While the number of adaptation parameters in LHN is determined by the number of nodes in the

hidden layers, LIN and LON parameter sizes are dependent on feature dimension and the total number of senones in the system, respectively. This can influence the amount of adaptation data required for each method.

### 3.2. Factorized Adaptation

When dealing with noise and channel distortions, the following relationship is often considered between the clean speech feature vector $\mathbf{x}$ and the distorted observation $\mathbf{y}$ (we drop the time index for simplicity)

$$\mathbf{x} = \mathbf{y} + \mathbf{g}(\mathbf{y}, \mathbf{n}, \mathbf{h}), \tag{6}$$

where $\mathbf{g}(\cdot)$ is a nonlinear function of the feature vector $\mathbf{y}$, noise factor $\mathbf{n}$, and the channel factor $\mathbf{h}$. In factorized adaptation (FA), we assume a similar relationship for the activations of the last sigmoid layer in the DNN [17]:

$$\tilde{\mathbf{v}}^L \simeq \mathbf{v}^L + \mathbf{g}'(\mathbf{y}, \mathbf{n}, \mathbf{h}), \tag{7}$$

$\tilde{\mathbf{v}}^L$ and $\mathbf{v}^L$ represent the activations of the last sigmoid layer resulting from the clean and noisy feature vectors, respectively. Borrowing from the idea of Vector Taylor Series (VTS) expansion [19], the nonlinear relationship in Eq. (7) can be approximated by

$$\tilde{\mathbf{v}}^L \simeq \mathbf{v}^L + \mathbf{A}\mathbf{y} + \mathbf{B}\mathbf{n} + \mathbf{C}\mathbf{h} + \mathbf{d}, \tag{8}$$

where $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{C}$ are the Jacobian matrices of $\mathbf{g}'(.)$ w.r.t. $\mathbf{y}$, $\mathbf{n}$ and $\mathbf{h}$, respectively, and $\mathbf{d}$ is the sum of all constant terms in the VTS expansion.

Assuming we have estimates of the noise and channel components, the adaptation problem consists of estimating the transformation matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$, as well as the offset vector $\mathbf{d}$. Effectively, we are adding new nodes to the last sigmoid layer whose values are fixed to our prior estimates of $\mathbf{y}$, $\mathbf{n}$ and $\mathbf{h}$. The weights connecting these nodes to the softmax layer (i.e. the values of $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ and $\mathbf{d}$) are tuned by back-propagation to minimize the objective function in Eq. (3).

### 3.3. Conservative training

The most straight-forward way of adapting a DNN is to simply adapt all parameters using a few more passes of retraining on the adaptation data. However, given the small adaptation data size, this will result in overfitting and erases the information learned during training. One approach to prevent this is to force the adapted senone distribution to stay close to the unadapted distribution [15]. To achieve this, the Kullback-Leibler divergence between adapted and unadapted posteriors is added to the optimization criterion:

$$\bar{C} = (1-\rho)C + \rho\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{N} p_0(s|\mathbf{x}_t)\log\frac{p_0(s|\mathbf{x}_t)}{p(s|\mathbf{x}_t)}. \tag{9}$$

$p_0(s|\mathbf{x}_t)$ and $p(s|\mathbf{x}_t)$ are senone distributions given by the original and adapted models, respectively. Removing the constants in Eq. (9) (terms depending only on $p_0(s|\mathbf{x}_t)$), we get the following regularized cost function:

$$\bar{C} = -\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{N} p_R(s|\mathbf{x}_t)\log p(s|\mathbf{x}_t), \tag{10}$$

where

$$p_R(s|\mathbf{x}_t) = (1-\rho)p_{GT}(s|\mathbf{x}_t) + \rho p_0(s|\mathbf{x}_t). \tag{11}$$

We are thus effectively replacing the ground-truth hard alignments in the original objective function with a smoothed version which is an interpolation between the original alignments and the distribution given by the unadapted model. This results in parameter updates that are less aggressive and thus have a lower risk of erasing previously learned information.

### 3.4. Supervision in DNN adaptation

The adaptation transcripts in all of the discussed methods can either be known in advance (supervised adaptation) or obtained by decoding the data using the unadapted model (unsupervised adaptation). In the unsupervised mode, the adaptation performance is limited by the accuracy of the obtained labels for adaptation data. In spite of this limitation, in tasks such as speaker adaptation, the number of correct labels in the initial decode is often adequate to provide reasonable adaptation performance. However, in the case of environmental mismatch such as noise and reverberation, the initial decode with the unadapted model has a very high error rate and is thus unable to provide a reasonable number of correct senone labels. Furthermore, discriminative estimation of adaptation parameters is known to be more sensitive to the accuracy of the labels compared to the maximum likelihood estimations used for GMMs [20]. As a result, unsupervised DNN adaptation is not able to provide noticeable improvements in scenarios with considerable environmental mismatch. All of the experiments in this paper use supervised adaptation towards a set of transcribed adaptation data.

## 4. Experiments

### 4.1. ASR system setup and data

We evaluate the discussed DNN adaptation approaches on the single-channel data from Aspire challenge [21]. The data consists of 10-minute audio files of conversational speech from 30 different speakers recorded using far-field microphones in reverberant and noisy rooms. Half of the utterances from each recording were used as test data and the other half were kept for adaptation. A 100-hour subset of the Fisher English corpus [22] was used as training data. All experiments use a trigram language model trained on the full Fisher corpus transcripts. The speech features are 13-dimensional Mel-frequency Cepstral coefficients (MFCC), with utterance-based Cepstral mean and variance normalization used in all experiments. The input features to the DNN-HMM system are a concatenation of static MFCC vectors from a context window of 11 frames. For the GMM-HMM system, delta and double delta features were used as well, and the concatenated feature vector from 11 context frames was further transformed by Linear Discriminant Analysis (LDA) to a 40-dimensional feature vector. The Kaldi speech recognition toolkit [23] was used for training the GMM-HMM system, with a total number of 100k Gaussians and 7716 senones. We use a DNN with 6 hidden layers and 2048 nodes in each layer ($\sim$ 37M parameters). The gradient descent optimizations (both for DNN training and adaptation) use a mini-batch size of 256. The DNN training uses 50 epochs with a learning rate of 0.08 for the first 25 epochs and 0.04 for the rest. The following table shows the word error rates of the baseline GMM-HMM and DNN-HMM systems. For comparison, we also report the results obtained by using unsupervised fMLLR adaptation for the GMM models. The clean-trained (unadapted) DNN-HMM system outperforms the GMM-HMM system by an absolute error difference of 12.0%, and even performs slightly better than fmllr-adpated GMM models.

Table 1: *Baseline WERs in mismatched condition*

| Acoustic Model | WER(%) |
|---|---|
| GMM-HMM | 74.4 |
| GMM-HMM (+fMLLR) | 62.9 |
| DNN-HMM | 62.4 |

### 4.2. DNN Adaptation results

In this section, we report the recognition results obtained by adapting the clean-trained DNN model to the adaptation utterances selected from Aspire challenge data. For each 10-minute recording, a set of adaptation parameters were estimated based on half of the utterances in the recording (or a subset of them), and used to decode the rest of the utterances. All of the experiments use 10 passes of adaptation data with a minibatch size of 256 and a fixed learning rate of 0.001. In all of the reported results, LHN-$i$ refers to a linear hidden network added after $i$'th layer in the network. The transformation matrices in LIN, LHN and LON were initialized as identity matrices, and the biases as zero vectors. For factorized adaptation (FA), compensation matrices $\mathbf{A}$ and $\mathbf{B}$ were initialized as zero matrices (we do not use the channel factor). The average feature vector from the first 3 frames of each utterance was used as noise estimate in FA (this assumes a stationary noise across the utterance. Alternatively, more sophisticated strategies such as sparse decomposition [24] can be used to estimate the noise factor).

Fig. 1 compares the word error rates provided by the discussed adaptation strategies using different amounts of adaptation data. The dashed line represents the performance of the unadapted DNN model. It is observable that the discussed adaptation strategies can provide significant performance gains compared to the clean model. Using the full adaptation dataset, the relative WER improvements range from 16.6% (for LHN-1) to 3.0% (for FA). LIN and LHN have provided the largest overall improvements, particularly when adequate adaptation data is available. LON, on the other hand, has resulted in smaller improvements. This is mainly due to the large output layer size of the DNN (7716 targets), which results in a LON transformation matrix with a very large number of parameters which cannot be reliably estimated using the limited adaptation data. For comparison, we have also included the results obtained by unregularized adaptation of the DNN parameters ($\rho = 0$), which, as expected, results in a poor performance. It is interesting to observe that in this case, increasing the adaptation data size from 5 to 20 actually harms the recognition accuracy. This is because without regularization, it is possible to erase previously learned information by overfitting. While a small adaptation set is less capable of causing this with only 10 passes of adaptation data, a larger adaptation set results in a more noticeable forgetting of previously learned information. This shows the importance of adding the KLD regularization term (nonzero value for $\rho$).

Considering the superior performance of linear input and hidden transforms in Fig. 1, we did a set of experiments to identify the best position in the network for inserting the adaptation layer. Fig. 2 shows the WERs of LIN and the different types of LHN. Regardless of adaptation size, LHN-1 (i.e. a hidden transform right after the first layer) results in the lowest error rate. The better performance of LHN-1 can be attributed to the function of the first sigmoid layer. The nonlinear processing in the first layer appears to remove some of the nonlinear environmental distortions, better enabling LHN to compensate for the remaining mismatch through a linear transformation. But as

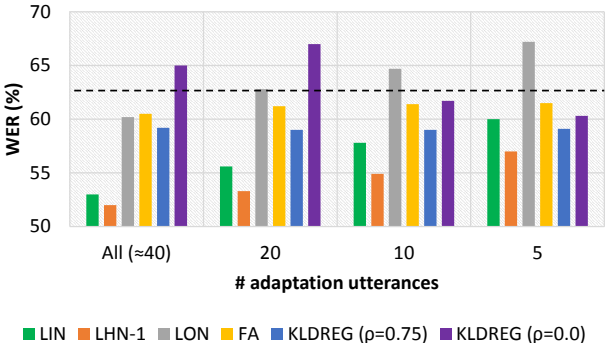


Figure 1: *Comparison of ASR performance using different adaptation strategies and different adaptation data sizes. The dashed line indicates performance of the unadapted model.*

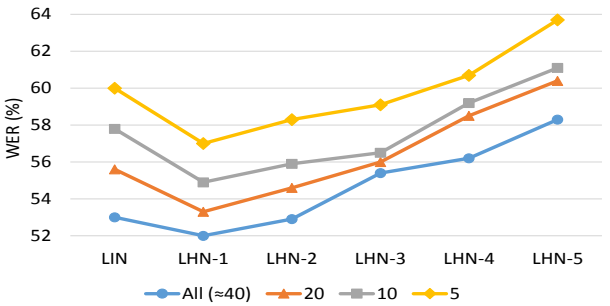


Figure 2: *Comparison of ASR performance for different positions of the linear transformation layer in the DNN*

we move the LHN layer further up in the network, the internal representations become more complex functions of the speech features and the underlying distortion, making it more difficult for a linear transformation to compensate the mismatch.

## 5. Conclusions

We have studied DNN acoustic model adaptation for improved robustness in far-field ASR. It was shown that by discriminatively estimating a set of adaptation parameters based on a small adaptation data collected from the target environment, significant gains can be achieved in recognition accuracy. Different forms of linear transformations as well as factorized and KLD-regularized adaptation were considered. In experiments on Aspire challenge data, relative WER improvements of 16.6%, 15.0%, 3.0% and 5.1% were achieved for LHN-1, LIN, FA, and KLD-regularized adaptation. Linear transformations closer to the input layer provided consistently better results compared to the other methods. Note that the first few layers of a DNN are often considered to be feature processors generating robust invariant representations for the higher layers (which convert these representations to more discriminative features for the logistic regression model in the output). Therefore, the better performance of LIN and LHN-1 suggests that for DNN-HMM systems, improved features is still an important aspect of robust recognition. But instead of designing independent feature processing algorithms analytically, discriminative information from the DNN back-end should be incorporated into the feature processing mechanism. LIN and LHN-1 can be considered as simple examples implementing this kind of approach.

# 6. References


[1] T. Yoshioka, A. Sehr, M. Delcroix, K. Kinoshita, R. Maas, T. Nakatani, and W. Kellermann, "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 114–126, Nov 2012.

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, April 2014.

[3] S. Sadjadi and J. Hansen, "Hilbert envelope based features for robust speaker identification under reverberant mismatched conditions," in *ICASSP*, May 2011, pp. 5448–5451.

[4] C. Kim and R. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," in *ICASSP*, March 2012, pp. 4101–4104.

[5] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," 1995.

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

[7] M. Seltzer, D. Yu, and Y. Wang, "An investigation of deep neural networks for noise robust speech recognition," in *ICASSP*, May 2013, pp. 7398–7402.

[8] D. Yu, M. L. Seltzer, J. Li, J. Huang, and F. Seide, "Feature learning in deep neural networks studies on speech recognition tasks," 2013.

[9] T. Yoshioka and M. Gales, "Environmentally robust ASR front-end for deep neural network acoustic models," *Computer Speech and Language, Elsevier*, vol. 31, no. 1, pp. 65–86, Dec 2014.

[10] V. Mitra, W. Wang, H. Franco, Y. Lei, C. Bartels, and M. Graciarena, "Evaluating robust features on deep neural networks for speech recognition in noisy and channel mismatched conditions," in *Interspeech*, Sep 2014, pp. 895–899.

[11] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, "Normalized amplitude modulation features for large vocabulary noise-robust speech recognition," in *ICASSP*, March 2012, pp. 4117–4120.

[12] J. Neto, L. Almeida, M. Hochberg, C. Martins, L. Nunes, S. Renals, and T. Robinson, "Speaker adaptation for hybrid HMM-ANN continuous speech recognition system," in *4th European Conference on Speech Communication and Technology (Eurospeech)*, Sep 1995, pp. 2171–2174.

[13] D. Albesano, R. Gemello, P. Laface, F. Mana, and S. Scanzio, "Adaptation of artificial neural networks avoiding catastrophic forgetting," in *Neural Networks, 2006. IJCNN '06. International Joint Conference on*, 2006, pp. 1554–1561.

[14] F. Seide, G. Li, X. Chen, and D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, Dec 2011, pp. 24–29.

[15] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *ICASSP*, May 2013, pp. 7893–7897.

[16] X. Li and J. Bilmes, "Regularized adaptation of discriminative classifiers," in *ICASSP*, May 2006.

[17] J. Li, J.-T. Huang, and Y. Gong, "Factorized adaptation for deep neural network," in *ICASSP*, May 2014, pp. 5537–5541.

[18] 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.

[19] P. Moreno, B. Raj, and R. Stern, "A vector Taylor series approach for environment-independent speech recognition," in *ICASSP*, vol. 2, May 1996, pp. 733–736 vol. 2.

[20] K. Yu, M. Gales, and P. Woodland, "Unsupervised adaptation with discriminative mapping transforms," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 714–723, May 2009.

[21] "IARPA automatic speech recognition in reverberant environments (ASpIRE) challenge, 2015," https://www.innocentive.com/ar/challenge/9933624.

[22] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *4th International Conference on Language Resources Evaluation*, 2004, pp. 69–71.

[23] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, dec 2011.

[24] J. Gemmeke, T. Virtanen, and A. Hurmalainen, "Exemplar-based sparse representations for noise robust automatic speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2067–2080, Sept 2011.