SPEECH RECOGNITION IN UNSEEN AND NOISY CHANNEL CONDITIONS

Vikramjit Mitra, Horacio Franco, Chris Bartels, Julien van Hout, Martin Graciarena, Dimitra Vergyri

Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA

{vikramjit.mitra, chris.bartels, martin.graciarena, julien.vanhout, horacio.franco, dimitra.vergyri}@sri.com

ABSTRACT

Speech recognition in varying background conditions is a challenging problem. Acoustic condition mismatch between training and evaluation data can significantly reduce recognition performance. For mismatched conditions, data-adaptation techniques are typically found to be useful, as they expose the acoustic model to the new data condition(s). Supervised adaptation techniques usually provide substantial performance improvement, but such gain is contingent on having labeled or transcribed data, which is often unavailable. The alternative is unsupervised adaptation, where feature-transform methods and model-adaptation techniques are typically explored. This work investigates robust features, feature-space maximum likelihood linear regression (fMLLR) transform, and deep convolutional nets to address the problem of unseen channel and noise conditions. In addition, the work investigates bottleneck (BN) features extracted from deep autoencoder (DAE) networks trained by using acoustic features extracted from the speech signal. We demonstrate that such representations not only produce robust systems but also that they can be used to perform data selection for unsupervised model adaptation. Our results indicate that the techniques presented in this paper significantly improve performance of speech recognition systems in unseen channel and noise conditions.

Index Terms— *automatic speech recognition, unsupervised adaptation, channel- and noise-robust speech recognition, auto-encoders, bottleneck features.*

1. INTRODUCTION

Deep neural network (DNN) hidden Markov models (HMM) [1]-based automatic speech recognition (ASR) systems [2, 3] demonstrate impressive performance as long as the training and evaluation conditions are similar. Unfortunately, DNN-HMM systems are both data hungry and data sensitive [4]. DNN acoustic models can be quite sensitive to acoustic condition mismatch, where a subtle change in the background acoustic conditions due to noise, reverberation, and channel distortion can expose such models' weakness. Typically, multi-condition training supported by data augmentation is used to compensate for DNN acoustic model weakness, with the literature reporting that robust DNN acoustic models can be trained with thousands of hours of acoustic data collected from diverse sources [5]. Data augmentation [6, 7] is also found to have a significant impact. In all such conditions, the assumption is that we have *a priori* knowledge about the kind of distortion the model will see, which often may not be the case. Real-world ASR applications typically encounter diverse acoustic conditions, which often are unique and hence difficult to anticipate. One such condition is channel variation and noise, which practically is an open-set problem.

The recent MGB [8], CHiME-3 [9], and ASpIRE [10] Challenges showed how susceptible DNN-HMM acoustic models are to realistic, varying, and unseen acoustic conditions. Several studies have explored novel ways of performing unsupervised adaptation of DNN models. Unsupervised speaker adaptation of DNNs has been explored with much success in [11–13], where adaptation based on maximum likelihood linear regression (MLLR) transforms, i-vectors, etc. has shown impressive performance gains over un-adapted models. In [4], stacked bottleneck (SBN) neural network architecture was proposed to cope with limited data from a target domain, where the SBN net was used as a feature extractor. The SBN system was used to cope with unseen languages in [4] and, in [7], was extended to cope with unseen reverberation conditions. In [14], Kullback-Leibler divergence (KLD) regularization was proposed for DNN model parameter adaptation, which differs from the typically used L2 regularization [15] in the sense that it constrains the model parameters themselves rather than the output probabilities.

In this work, we focus on learning a feature-space representation by using deep autoencoder bottlenecks (DAE-BN), and employing that representation to predict the reliability of our acoustic model's decision. Unlike the SBN systems explored in the literature [4, 7], DAE-BN training requires no labeled/transcribed data and can be instead done with high volumes of unlabeled speech data. DAE-BNs take spliced (contextualized) acoustic features as input and then map that input to a differently spliced version of the same acoustic feature. In addition to DAE-BNs, we explored traditional fMLLR transforms and observed impressive performance gains for unseen channel and acoustic conditions. We investigated using the entropy measures from the DAE-BN features to generate a confidence measure, which in turn was employed to select test data and their initial ASR hypothesis for unsupervised model adaptation. Using fMLLR features in addition to model adaptation resulted in significant performance improvement.

2. DATA

The speech dataset used in our experiments was collected by the Linguistic Data Consortium (LDC) under DARPA's RATS program, which focused on speech in noisy or heavily distorted channels in two languages: Levantine Arabic (LAR) and Farsi. The data was collected by retransmitting telephone speech (denoted as source channel) through eight communication channels [16] (denoted as A, B, C, D, E, F, G, and H), each of which had a range of associated distortions. The DARPA RATS dataset is unique in that the noise and channel degradations were not artificially introduced by performing mathematical operations on the clean speech signal; instead, the signals were rebroadcast through channel- and noise-degraded ambience and then rerecorded. Consequently, the data contained several unusual artifacts, such as nonlinearity, frequency shifts, modulated noise, and intermittent bursts—conditions under which the traditional noise-robust

978-1-5090-4117-6/17/$31.00 ©2017 IEEE

approaches developed in the context of additive noise may not have performed well.

For this paper, we focused only on the LAR dataset for our reported experiments. For LAR acoustic model (AM) training, we used approximately 250 hours of retransmitted conversational speech (LDC2011E111 and LDC2011E93). For language model (LM) training, we used various sources: 1.3M words from the LDC's EARS (Effective, Affordable, Reusable Speech-to-Text) data collection (LDC2006S29, LDC2006T07); 437K words from Levantine Fisher (LDC2011E111 and LDC2011E93); 53K words from the RATS data collection (LDC2011E111); 342K words from the GALE (Global Autonomous Language Exploitation) Levantine broadcast shows (LDC2012E79); and 942K words from web data in dialectal Arabic (LDC2010E17). We used a held-out set for LM tuning, which was selected from the Fisher data collection and contained approximately 46K words. To evaluate ASR and keyword-spotting (KWS) performance for LAR, we used two test sets—referred to here as dev-1 and dev-2. Each test set consisted of 10 hrs of held-out conversational speech. Dev-2 did not come with reference transcriptions and was meant solely for KWS evaluation and as we focus only on ASR, we will be reporting our results on dev-1 only. Note that approximately 2K segments from each channel condition were used as a held-out validation set for model training and optimization.

The LAR data had eight channels denoted by A through H. In our experiments, we removed channels A and B from the training set (referred to here as "no A-B train") and evaluated the models across all eight channels as well as the source data (non-retransmitted data) that were distributed as the dev-1 set in the DARPA RATS distributions. In addition to the LAR data, 2500 hrs of communication-channel-degraded Mandarin data was also used to train the DAE-BN system. We observed that the performance of the DAE-BN system improved with addition of the Mandarin training data.

3. ACOUSTIC FEATURES

We used gammatone filterbank energies (GFBs) as one of the acoustic features for our experiments. Gammatone filters are a linear approximation of the auditory filterbank found in the human ear. For the GFB processing, the speech was analyzed by using a bank of 40 gammatone filters equally spaced on the equivalent rectangular bandwidth (ERB) scale. Within an analysis window of approximately 26 ms, the power of the bandlimited time signals was computed at a frame rate of 10 ms. The subband powers were then root-compressed by using the 15th root, and the resulting 40-dimensional feature vector was used as the GFBs.

We also used normalized modulation coefficients (NMCs) [20] as a candidate feature in our experiments. NMCs capture the amplitude modulation (AM) information from bandlimited speech signals. NMCs track the AM trajectories of subband speech signals in a time domain by using a Hamming window of 26 ms with a frame rate of 10 ms. The powers of the AM signals were root compressed by using the 15th root. The resulting 40-dimensional feature vector was used as the NMC feature in our experiments

In addition to the above feature sets, we also used standard mel-filterbank energies (MFBs) and mel-frequency cepstral coefficients (MFCCs) as candidate feature sets.

4. DEEP AUTOENCODER BOTTLENECK (DAE-BN) SYSTEM

The DAE-BN system is a five-hidden-layer, fully connected DNN system, with the third hidden layer containing a bottleneck of eighty neurons. The remaining hidden layers had 1024 neurons. The hidden layers had sigmoid activations, whereas the output layer had linear activation. The DAE-BN was trained by using mean squared error (MSE) backpropagation. The input to the DAE-BN system was 40 GFBs with a splicing of 11, resulting in 440 dimensional features, whereas the output was the same 40 GFBs but with a splicing of five.

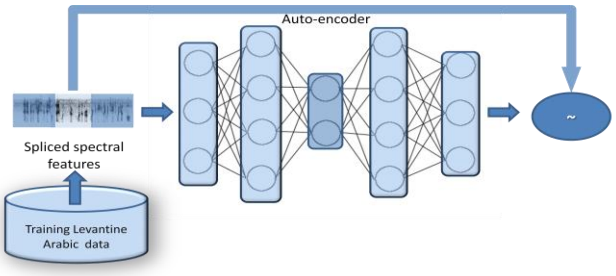

Figure 1: The DAE-BN system.

Once trained, the sigmoid activation of the BN layer was replaced by a linear activation. The BN feature from the DAE-BN system was then used to train a fully connected DNN acoustic model, as shown in Figure 2. Note that the DAE-BN system was trained with all but channels A and B, for "no A-B train" data. The BN features from the DAE-BN system described in this work are different than the previously proposed deep BN features from stacked autoencoders [24], in the sense that the autoencoder was neither trained to denoise the input features nor trained layer-wise. The DAE-BN system was trained with the same input-output features, but the feature splicing on the input side was different than that of the output side, as mentioned above.

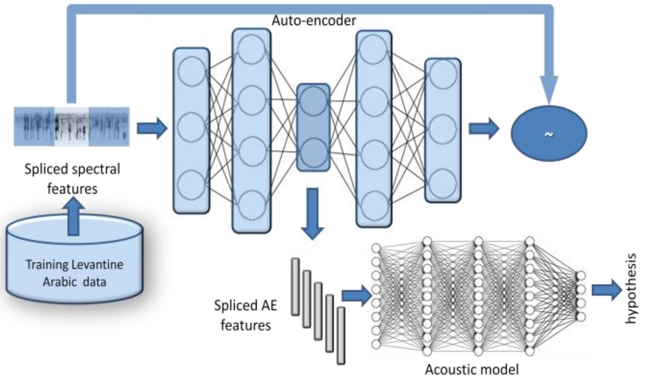

Figure 2: The DAE-BN-based DNN acoustic model training.

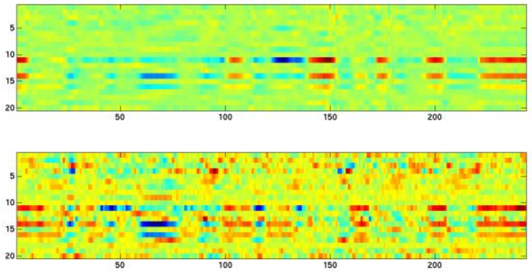

Figure 3: The DAE-BN features (the first 20 dimensions) for a source [top] and unseen retransmitted [bottom] LAR data. The DAE-BN dimensions are found on the Y-axis, and number of frames, on the X-axis.

Figure 3 shows the plot of first 20 dimensions of the DAE-BN features for source (relatively clean) and unseen (channel A) retransmitted LAR data. It is evident that for the unseen condition, several of the neurons in the BN layer are triggered, and consequently, the entropy of the BN activation outputs over a short-term window can be expected to be higher for the unseen case compared to the seen case. This observation motivated us to generate an entropy-based confidence measure, which is estimated from the BN features and can be used to select test data and their first-pass hypothesis for unsupervised model adaptation.

5. THE SPEECH RECOGNITION SYSTEM

We used the no A-B train data to train the multi-channel acoustic models, and we call the resulting models the "no-AB models". We also trained a baseline model using all the training data, which included data from the source and their eight retransmitted channel versions. Initially, we trained a three-state context dependent (CD) left-to-right GMM-HMM acoustic model, which was used to generate the senone alignments for the DNN acoustic model training. The training corpus was clustered into pseudo-speaker clusters by using unsupervised agglomerative clustering.

The DNNs were trained by using cross entropy employing the senone alignments. The DNNs had five hidden layers of size 2048 with sigmoid activations, except for the DNN trained on the BN features from the DAE-BN system, which had three hidden layers with 2048 neurons. The networks were trained by using an initial four iterations with a constant learning rate of 0.008, followed by learning-rate halving based on cross-validation error decrease. Training stopped when no further significant reduction in cross-validation error was noted or when cross-validation error started to increase. Backpropagation was performed using stochastic gradient descent with a mini-batch of 256 training examples.

6. RESULTS

We trained different DNN acoustic models using MFCC, MFB, and NMC features. We report system performance in terms of word error rates (WERs). To assess the performance degradation due to unseen channel conditions, we trained two DNN acoustic models: (a) including the target channels (A, B) in the training data, (b) excluding the target channels (A, B) from the training data. Table 1 shows the WERs from these two systems when channels A, B, C and E are decoded from the dev-1 evaluation set.

Table 1. WERs from DNN models trained with MFB features, for dev-1 channels A, B, C, E and the whole dev-1 data, for acoustic models trained with (a) all and (b) no A, B training data.

	dev-1 channels				dev-1
	A	B	C	E	Avg.
Train with all	70.6	68.2	75.8	70.2	62.8
no A, B train	99.5	98.0	78.2	73.4	72.6

Table 1 demonstrates the performance deterioration for the unseen channels, where the DNN models gave error rates more than 90%. The Table 1 results indicate the need for better strategies to prevent acoustic models from failing under unseen noise/channel conditions. Note that the error rates reported in Table 1 are quite high, as Arabic acoustic models typically have higher WERs, and beyond that, channels A, B, C, and E are known to be quite adverse channels that contain high levels of noise, channel degradations, and non-stationary artifacts. To obtain some insight regarding the typical WERs seen with the DARPA RATS Levantine Arabic corpus, one may refer to [22].

Next, we explored different features and investigated how the recognition rates varied for them for different channel conditions. Table 2 presents the WERs for the MFCC, MFB, and NMC features, when used with a five-hidden-layer DNN having 2048 neurons. Table 2 shows that the robust features failed to prevent the DNNs from failing under mismatched channel conditions, indicating the need for adaptation mechanisms to attain reasonable recognition accuracy.

Table 2. WERs from DNN model trained with GFB, MFB, NMC, and DAE-BN features for dev-1 channels A, B, C, and E, when trained with (a) all and (b) no A-B train data.

No A-B train	Feature	dev-1 channels				dev-1 all
		A	B	C	E	Avg.
	MFCC	100	98.5	81.6	83.6	78.8
	MFB	99.5	98.0	78.2	73.4	72.6
	NMC	92.9	93.9	76.6	73.0	70.6
	DAE-BN	79.3	82.6	80.6	78.5	71.5

Table 2 shows that the MFCC and MFB features failed for the unseen channel conditions; however, they were able to retain their performance for the seen channel conditions (comparing their performance from "All-trained" models in Table 1). The DAE-BN features were relatively robust for unseen channel conditions; however, their performance for the seen channel conditions was worse than that of the MFB and NMC features. Next, we explored using MFCC and MFB features and their fMLLR-transformed representations for training and testing the DNNs. Table 3 presents the WERs for the MFCC and MFB features, and shows that fMLLR transform resulted in a significant performance improvement.

Table 3. WERs from DNN models trained with MFCC, MFB, and NMC features with fMLLR transform, for dev-1 channels A, B, C, E, and dev-1 all, for no A-B train data.

	Dev-1 channels				dev-1 all
	A	B	C	E	Avg.
$\text{MFCC}_{\text{fmLLR}}$	75.9	80.6	76.7	73.8	67.1
$\text{MFB}_{\text{fMLLR}}$	75.7	79.1	75.3	69.8	65.4
$\text{NMC}_{\text{fMLLR}}$	76.4	79.6	75.2	70.7	65.7

Table 3 shows that the fMLLR transform significantly reduced the error rates for the unseen channels A and B, and brought them close to the error rates obtained from the seen-channel conditions reported in Table 1. It is also interesting to note that the fMLLR transformed MFB features gave lower WER than the fMLLR transformed MFCC features.

It has been established that convolutional neural network (CNN) are typically robust against noise and channel distortions [23]; therefore, we explored CNN acoustic models for the features presented above. We explored using CNN models on fMLLR transformed MFB, NMC, and DAE-BN features. Note that convolution across feature dimension is not meaningful for DAE-BN features, as the neighboring feature dimensions may not be as correlated as the spectral features. Hence, we performed convolution across time (time-convolutional neural net (TCNN)) only and used 75 filters with a band size of 8 and max-pooling over a window size of 5. For the other spectral features, NMC and MFBs, we investigated conventional CNNs that had 200 convolutional filters with a band size of 8 and max-pooling over three frames. The convolutional layers were connected with a four-hidden-layer, fully connected neural net, where each layer had 2048 neurons. The results from the CNN models are shown in Table 4, where it can be seen that for all features, except DAE-BN, further reduction in WER was observed for both seen and unseen channel conditions compared to the DNN models.

Table 4. WERs from CNN models trained with fMLLR transformed MFB, NMC, and DAE-BN features for dev-1 channels A, B, C, E, and dev-1 all, for the no A-B train condition.

	model	dev-1 channels				dev-1
		A	B	C	E	Avg.
$\mathrm{MFB}_{\mathrm{fMLLR}}$	CNN	72.8	76.4	73.8	67.2	63.1
$\mathrm{NMC}_{\mathrm{fMLLR}}$	CNN	74.1	77.0	74.9	67.5	63.7
$\mathrm{DAE\text{-}BN}_{\mathrm{fMLLR}}$	TCNN	80.5	83.7	81.2	79.1	72.5

Table 4 shows that the CNN models gave lower WERs than the DNN models reported in Table 3. The convolution operation on the DAE-BN features did not reduce WERs as compared to the DNN model.

We also investigated bottleneck features (BN) features obtained by supervised training of a five-hidden-layer, fully connected DNN, which had a 60-dimensional BN at the third layer. The input to the BN-DNN had features spliced over 15 frames. We observed that SBNs learned from the LAR data using a vowelized dictionary gave better performance than one using the standard non-vowelized dictionary; hence, the former was used to train the BN-DNN model. The DAE-BN system also had a similar configuration as the BN-DNN system: five hidden layers with a BN at the third layer. We noticed that the BN features from the BN-DNN system performed slightly worse (0.4% relative) for the unseen channel conditions and a little better for the seen channel conditions, compared to the features from the DAE-BN system.

Next, we investigated time-frequency CNNs (TFCNNs) [25] on fMLLR transformed NMC and MFB features. TFCNNs have always shown better performance than their CNN counterparts, and here we also observed WER reduction compared to using CNN acoustic models. Table 5 shows the WERs from the TFCNN acoustic models. In addition we combined the fMLLR transformed MFB and NMC features and trained a fused CNN model (fCNN) [26], where two parallel convolutional layers are trained for each of the two individual features.

Table 5. WERs from TFCNN models trained with $\mathrm{MFB}_{\mathrm{fMLLR}}$ and $\mathrm{NMC}_{\mathrm{fMLLR}}$ features and fCNN model trained with the $\mathrm{MFB}_{\mathrm{fMLLR}}$ +$\mathrm{NMC}_{\mathrm{fMLLR}}$ features for dev-1 channels A, B, C, E, and dev-1 all, for the no A-B train condition.

	dev-1 channels				dev-1
	A	B	C	E	Avg.
$\mathrm{MFB}_{\mathrm{fMLLR}}$	72.4	76.0	73.9	66.7	62.8
$\mathrm{NMC}_{\mathrm{fMLLR}}$	73.6	77.0	75.0	67.3	63.4
$\mathrm{MFB}_{\mathrm{fMLLR}}$+$\mathrm{NMC}_{\mathrm{fMLLR}}$	72.0	75.3	73.3	65.7	61.9

Next, we investigated the BN features from the DAE-BN network and used these to generate a confidence measure. We estimated the entropy over a running window of 21 frames (i.e., ~230 ms of temporal information) of data for each dimension of the DAE-BN features and then computed the maximum entropy for each dimension. The cumulative entropy from the top 30% percentile maximum entropies across all the dimensions was used as a measure of confidence. Note that as depicted in Figure 3, unseen data typically resulted in more spurious activations across the neurons, which resulted in higher entropy compared to seen data conditions. We used the entropy-based confidence measure to select the top 1K test segments for each channel condition that generated the lowest overall 30th percentile cumulative entropy for each channel condition and used those test segments to adapt the acoustic model. These test segments were used to retrain the previously trained TFCNN and fCNN models, using an L2 regularization of 0.02. Table 6 presents the WERs obtained from the TFCNN and fCNN model adaptation for the $\mathrm{MFB}_{\mathrm{fMLLR}}$, $\mathrm{NMC}_{\mathrm{fMLLR}}$ and $\mathrm{MFB}_{\mathrm{fMLLR}}$+$\mathrm{NMC}_{\mathrm{fMLLR}}$ features. The same retraining procedure on the DAE-BN DNN system resulted in a relative WER reduction of 4.3%.

Table 6. WERs from adapted TFCNN models trained with $\mathrm{MFB}_{\mathrm{fMLLR}}$ and $\mathrm{NMC}_{\mathrm{fMLLR}}$ features and fCNN model trained with $\mathrm{MFB}_{\mathrm{fMLLR}}$+$\mathrm{NMC}_{\mathrm{fMLLR}}$ feature for dev-1 channels A, B, C, E. and dev-1 all, for the no A-B train condition.

	dev-1 channels				dev-1
	A	B	C	E	Avg.
$\mathrm{MFB}_{\mathrm{fMLLR}}$	71.4	75.3	73.9	66.2	62.5
$\mathrm{NMC}_{\mathrm{fMLLR}}$	73.0	76.1	74.2	67.2	63.1
$\mathrm{MFB}_{\mathrm{fMLLR}}$+$\mathrm{NMC}_{\mathrm{fMLLR}}$	71.2	74.6	73.1	65.2	61.4

7. CONCLUSION

In this work, we investigated techniques to cope with unseen and noisy channel conditions for a Levantine Arabic ASR task. We observed that fMLLR transform on spectral features demonstrated significant robustness compared to the basic features (no fMLLR transform) and DNN acoustic models. We proposed a novel way to extract a confidence measure by tracking the activations from a deep autoencoder bottleneck system and demonstrated that a running-window entropy measure can provide reliable information for data selection and hence unsupervised model adaptation. Overall, 20% relative reduction in WER was obtained when fMLLR transform was used, 4% relative reduction in WER was obtained when a TFCNN model was used to replace the DNN, and 2% relative reduction was obtained when model adaptation was performed through confidence-based data selection. Combining the fMLLR transformed features was found to be useful which helped in reducing the WERs.

In the future, we plan to investigate threshold-based data selection, where such thresholds are learned by tracking the activation entropy measured from the training data. Additionally, we will investigate system combination through decision fusion. The BN features from the DAE-BN system by themselves performed better than out-of-the-box MFB and NMC features for unseen channel conditions; however, their fMLLR-transformed versions failed to perform competitively. We plan to investigate if adding more training data further improves DAE-BN feature performance. Also, we will investigate if adapting the acoustic model after DAE-BN adaptation improves the performance of the proposed system.

8. ACKNOWLEDGEMENTS

This material is based upon work partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-15-C-0037. The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

9. REFERENCES

[1] A. Mohamed, G.E. Dahl, and G. Hinton, “Acoustic modeling using deep belief networks,” IEEE Trans. on ASLP, vol. 20, no. 1, pp. 14 –22, 2012.

[2] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks,” Proc. of Interspeech, 2011.

[3] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, “Deep neural networks for acoustic modeling in speech recognition,” IEEE Signal Process. Mag., vol. 29, no. 6, pp. 82–97, 2012.

[4] F. Grézl ; E. Egorova and M. Karafiát, “Further investigation into multilingual training and adaptation of stacked bottle-neck neural network structure,” Proc. of SLT, pp. 48-53, 2014.

[5] T. Sainath, R.J. Weiss, K. Wilson, A.W. Senior and O. Vinyals, “Learning the speech front-end with raw waveform CLDNNs,” Proc. of Interspeech, 2015.

[6] V. Peddinti, G. Chen, V. Manohar, T. Ko, D. Povey and S. Khudanpur, “JHU ASpIRE system : Robust LVCSR with TDNNs, i-vector adaptation and RNN-LMS,” Proc. of ASRU, 2015.

[7] M. Karafiát, F. Grézl, L. Burget, I. Szöke and J. Cernocký “Three ways to adapt a CTS recognizer to unseen reverberated speech in BUT system for the ASpIRE challenge,” Proc. of Interspeech, pp. 2454–2458, 2015.

[8] P. Bell, M.J.F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester and P.C. Woodland, “The MGB challenge: Evaluating multi-genre broadcast media recognition,” Proc. of ASRU, 2015.

[9] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, “The third ‘CHiME’ speech separation and recognition challenge: Dataset, task and baselines,” Proc. of ASRU, 2015.

[10] M. Harper, “The automatic speech recogition in reverberant environments (ASpIRE) challenge,” Proc. of ASRU, 2015.

[11] T. Yoshioka, A. Ragni, M.J.F. Gales, “Investigation of unsupervised adaptation of DNN acoustic models with filterbank input,” Proc. of ICASSP, pp. 6344–6348, 2014.

[12] G. Saon, H. Soltau, D. Nahamoo and M. Picheny, “Speaker adaptation of neural network acoustic models using i-vectors,” Proc. of ASRU, pp. 55–59, 2013.

[13] S.H.K. Parthasarathi, B. Hoffmeister, S. Matsoukas, A. Mandal, N. Strom and S. Garimella, “fMLLR based feature-space speaker adaptation of DNN acoustic models,” Proc. of Interspeech, 2015.

[14] D. Yu, K. Yao, H. Su, G. Li and F. Seide, “KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition,” Proc. of ICASSP, 2013.

[15] X. Li and J. Bilmes, “Regularized adaptation of discriminative classifiers,” Proc. ICASSP’06, 2006.

[16] K. Walker and S. Strassel, “The RATS radio traffic collection system,” Proc. of Odyssey 2012-The Speaker and Language Recognition Workshop, 2012.

[17] A. Stolcke, “SRILM—An extensible language modeling toolkit,” Proc. of ICSLP, pp. 901–904, 2002.

[18] A. Venkataraman and W. Wang, “Techniques for effective vocabulary selection,” Proc. Eighth European Conference on Speech Communication and Technology, pp. 245–248, 2003.

[19] A. Mandal, J. van Hout, Y-C. Tam, V. Mitra, Y. Lei, J. Zheng, D. Vergyri, L. Ferrer, M. Graciarena, A. Kathol, and H. Franco, “Strategies for high accuracy keyword detection in noisy channels,” Proc. of Interspeech, pp. 15-19, 2013.

[20] V. Mitra, H. Franco, M. Graciarena, and A. Mandal, “Normalized amplitude modulation features for large vocabulary noise-robust speech recognition,” Proc. of ICASSP, pp. 4117–4120, 2012.

[21] V. Mitra, H. Franco, and M. Graciarena, “Damped oscillator cepstral coefficients for robust speech recognition,” Proc. of Interspeech, pp. 886–890, 2013.

[22] T. Ng, R. Hsiao, L. Zhang, D. Karakos, S.H. Mallidi, M. Karafiat, K. Vesely, I. Szoke, B. Zhang, L. Nguyen, and R. Schwartz, “Progress in the BBN Keyword Search System for the DARPA RATS Program,” Proc. of Interspeech, pp. 959-963, 2014.

[23] 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 Proc. of Interspeech, pp. 895-899, Singapore, 2014.

[24] J. Gehring, Y. Miao, F. Metze, A. Waibel, “Extracting deep bottleneck features using stacked auto-encoders,” in Proc. of ICASSP, 2013.

[25] V. Mitra, and H. Franco, “Time-frequency convolution networks for robust speech recognition,” Proc. of ASRU, 2015.

[26] V. Mitra, J. VanHout, W. Wang, C. Bartels, H. Franco, D. Vergyri, A. Alwan, A. Janin, J. Hansen, R. Stern, A. Sangwan and N. Morgan, “Fusion Strategies for Robust Speech Recognition and Keyword Spotting for Channel- and Noise-Degraded Speech,” Proc. of Interspeech 2016.