ANALYZING CONVOLUTIONAL NEURAL NETWORKS FOR SPEECH ACTIVITY DETECTION IN MISMATCHED ACOUSTIC CONDITIONS

Samuel Thomas, Sriram Ganapathy, George Saon and Hagen Soltau

IBM T.J. Watson Research Center, Yorktown Heights, USA.

{sthomas,ganapath,gsaon,hsoltau}@us.ibm.com

ABSTRACT

Convolutional neural networks (CNN) are extensions to deep neural networks (DNN) which are used as alternate acoustic models with state-of-the-art performances for speech recognition. In this paper, CNNs are used as acoustic models for speech activity detection (SAD) on data collected over noisy radio communication channels. When these SAD models are tested on audio recorded from radio channels not seen during training, there is severe performance degradation. We attribute this degradation to mismatches between the two dimensional filters learnt in the initial CNN layers and the novel channel data. Using a small amount of supervised data from the novel channels, the filters can be adapted to provide significant improvements in SAD performance. In mismatched acoustic conditions, the adapted models provide significant improvements (about 10-25%) relative to conventional DNN-based SAD systems. These results illustrate that CNNs have a considerable advantage in fast adaptation for acoustic modeling in these settings.

Index Terms— Convolutional neural networks, Speech activity detection, Neural network adaptation

1. INTRODUCTION

Speech activity detection (SAD) is the first step in most speech processing applications like speech recognition, speech coding and speaker verification. This module is an important component that helps subsequent processing blocks to focus their resources on the speech parts of the signal. In the past, several approaches have been used to build reliable SAD modules. These techniques are usually variants of decision rules based on features from the audio signal like signal energy [1], pitch [2], zero crossing rate [3] or higher order statistics in the LPC residual domain [4]. Acoustic features have also been used to train multi-layer perceptrons (MLPs) [5] and hidden Markov models (HMMs) [6] to differentiate between speech and non-speech classes. All these approaches focus on attributes of speech which differentiate it from other acoustic events that can appear in the signal.

An important step in SAD is to represent speech using features that capture its unique properties while also being robust to distortions under various noisy conditions. These features can be broadly categorized as -

(a) Short-term spectral features extracted from power spectral estimates in short analysis windows (10-30 ms) of the speech [7, 8],

This work was supported in part by Contract No. D11PC20192 DOI/NBC under the RATS program. The views expressed are those of the authors and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

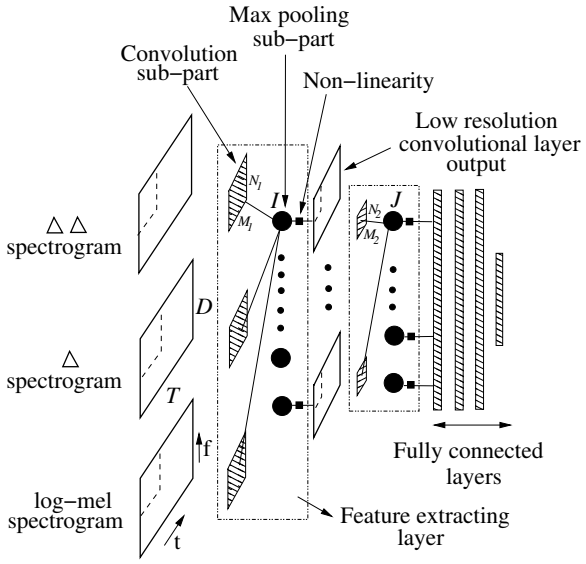

Fig. 1. Convolutional neural network with two pairs of convolution and max pooling layers used for speech recognition

(b) Long-term modulation frequency components estimated in long analysis windows spanning few hundreds of milliseconds from sub-band envelopes of speech [9, 10], and

(c) Joint spectro-temporal features derived using 2D selective filters tuned to different spectro-temporal modulations of the input spectrogram [11].

In this paper we focus on SAD using spectro-temporal features similar to (c) above. These features are however not derived using 2D filters hand-tuned to selected spectro-temporal modulations but are learnt automatically in a data-driven fashion using a convolutional neural network (CNN) [12] framework (Section 2). The filters are learnt on audio data from the DARPA RATS program collected under both controlled and uncontrolled field conditions over highly degraded, noisy communication channels [13]. Detecting speech in the presence of non-stationary and non-linear distortions introduced by these channels is a challenging task. Through a series of DARPA evaluations, significant improvements have been achieved by different sites on this task when channels involved in testing are also seen during training [14, 15, 16, 17, 18]. The focus of the program is shifting toward a more challenging problem of developing robust systems that not only perform well on noisy channels seen during training but also on unseen channels. While this paper is closely related to prior work in [16], it addresses this new program direction.

To achieve this goal, it is important to develop acoustic models which are not biased to any particular acoustic condition seen during

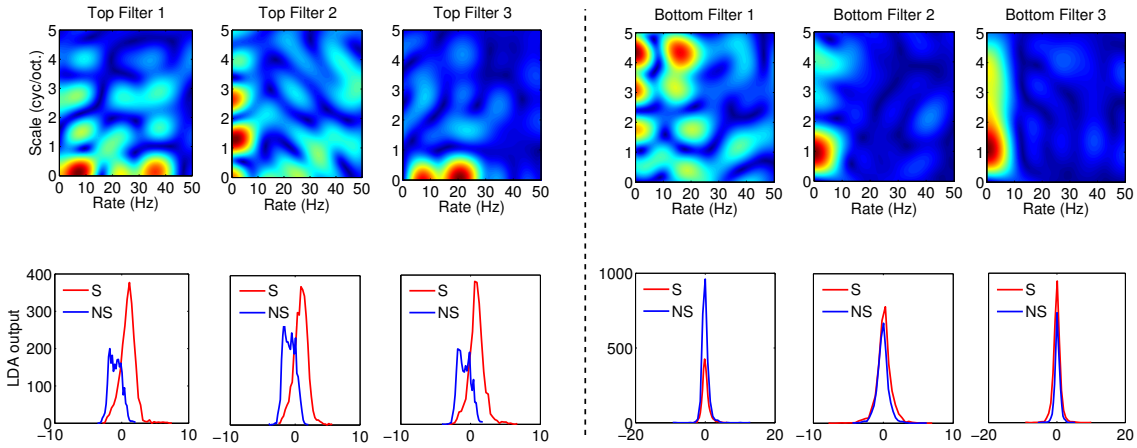

Fig. 2. Selected data-driven filters from the first convolutional layer of a CNN trained on RATS data for SAD

training. When tested on novel channels, since severe degradation in performances are anticipated, it will also be useful to have models that can be easily adapted to novel channel conditions with limited amounts of adaptation data.

Conventional DNNs are trained on fixed feature representations and have millions of parameters trained with several hundred hours of data. With very limited adaptation data, it may be infeasible to adapt this model effectively. CNNs [12] on the other hand, have data-driven *feature extracting* layers which can be reconfigured to new conditions via adaptation, followed by several hidden layers that can be trained on large amounts of data to discriminate between speech and non-speech. With this architecture, we hypothesize that these models can be better adapted and are hence well suited for this task. The proposed CNN-based SAD processing pipeline is described in Section 3. Section 4 describes the experiments we perform on several novel channels to test this hypothesis and their results. These results show that CNNs are useful acoustic models in novel acoustic conditions with their ability to adapt better with limited amounts of adaptation data. The paper concludes with a discussion in Section 5.

2. CONVOLUTIONAL NEURAL NETWORKS

Convolutional neural networks (CNN) are very similar to conventional deep neural networks - the difference between these models, being the additional CNN feature extracting layers. These layers generate features for succeeding layers instead of pre-processed features that are usually input to the DNNs. Each of the feature extracting layers consists of a pair of convolution and max pooling sub-parts [12] as shown in Figure 1. The convolution sub-part is an ensemble of filters that are locally convolved with parts of the input to produce features that are further processed by a max pooling step. The max pooling operation involves picking the maximum from adjacent filter outputs. After passing through sigmoid nonlinearities, activations from lower layers are processed by subsequent feature extracting layers with more filters and down-sampling. The extracted features are finally received by fully connected DNN layers. All the layers of the CNN are trained using the standard back-propagation algorithm to minimize the cross entropy between the targets and the activations of the output layer.

In the past, these networks have shown to produce robust representations for several image processing tasks [19]. More recently CNNs have also been applied for speech processing [20, 21]. In these approaches, CNNs show increased robustness to speaker variability and improve LVCSR performances by compensating for shifts in frequency patterns of speech exhibited across speakers. CNNs also provide significant gains when used on noisy RATS data as acoustic models for LVCSR based keyword spotting [22]. These improvements point to the ability of CNNs to learn from degraded speech as well, and hence to be potentially useful acoustic models in a task like SAD.

3. SPEECH ACTIVITY DETECTION ON RATS DATA

Convolutional neural networks described in the previous section are evaluated in terms of SAD accuracy on noisy radio communications audio provided by the Linguistic Data Consortium (LDC) for the DARPA RATS program. Most of the RATS data (about 2000 hours) released for SAD were obtained by retransmitting existing audio collections - such as the DARPA EARS Levantine/English Fisher conversational telephone speech (CTS) corpus - over eight radio channels, labeled A through H [13]. Additionally new telephone recordings in Arabic Levantine, Pashto and Urdu were collected specifically for this program, covering a wide range of radio channel transmission effects [13]. As described earlier, the program is now shifting toward developing robust systems that not only perform well on noisy channels seen during training but also on unseen channels. We address this new program direction in two steps.

3.1. Channel Independent Models

To perform well on channels seen during training, our past approach using CNNs has been to train channel specific models for each of the RATS channels [16]. This framework however is fragile and breaks down when tested against novel channels although it performs well on the seen channels. We address this issue by jointly training a single channel independent model with data from all the channels. This allows the new model to learn variabilities across all the RATS channels without being biased to any particular condition as with the channel specific framework.

As shown in Figure 1, the CNN is trained on D dimensional *log-mel* spectra augmented with Δ and $\Delta\Delta$s. The *log-mel* spectra are

extracted by first applying *mel* scale integrators on power spectral estimates in short analysis windows (25 ms) of the signal followed by the *log* transform. Each frame of speech is also appended temporally with a fixed set of T frames. All of the I nodes in the first feature extracting layer are attached with $M_1 \times N_1$ filters that are two dimensionally convolved with the input representations. The second feature extracting layer with J nodes has a similar set of $M_2 \times N_2$ filters that process the non-linear activations after max pooling from the preceding layer. The non-linear outputs from the second feature extracting layer are then passed onto the following DNN layers [22].

Figure 2 shows the response of a few filters from the first layer of a CNN trained jointly on 4 channels in the modulation domain. We learn 128 filters on the *log-mel* stream in order to process the modulation spectrum at different frequencies. To further understand the function of these filters we learn a 1 dimensional LDA transform on the output of these filters to separate speech and non-speech. The filters are then sorted based on the separability obtained using the LDA cost function. While some filters show considerable separation (top filters in Figure 2), a simple linear classifier is not capable of discriminating between outputs from all the filters (for example the bottom filters in Figure 2). The outputs of these filters in practice, pass through several non-linear transformations before they are used to produce speech/non-speech posteriors.

3.2. Adaptation to Unseen Channels

Once the channel independent networks have been trained, we hypothesize that they will be able to provide reasonable performances in unseen channels. However, the performance of the system can still improve by adapting the networks to the unseen channels. Several techniques have been proposed in the past for adapting neural networks in the context of speaker adaptation of both CNN and DNN hybrid neural models [23, 24, 25, 26]. We propose to adapt the learnt filters of the feature extracting layers of the CNN with limited amounts (up to 15 minutes) of supervised data from the novel channel.

4. EXPERIMENTS AND RESULTS

4.1. Training and Test Data

As described earlier, about 250 hours of automatically annotated audio from multiple languages is available for each of the 8 different noisy radio channels (A-H). For our experiments we keep channels A-D as unseen channels and train only on data from channels E-H. We report results on the official DEV1 test set which has about 5 hours of data for the 4 channels held out as unseen. The equal error rate (EER), defined as the operating point where the probability of Miss (P_{Miss}) equals the probability of false accept (P_{FA}), is used as the performance metric.

4.2. SAD Processing Pipeline

The proposed CNN models generate frame-level posteriors of three classes - speech (S), non-speech (NS) and non-transmission (NT). These posteriors are then used along with a HMM Viterbi decoder with a 5 state HMM topology as described in [16]. To obtain receiver operating curves, the trade-off between missed speech and falsely hypothesized speech is determined by adding a fixed offset to the S scores for every frame. The frame level scores are scaled by an acoustic weight of 0.03 for all our experiments. After decoding, the boundaries of hypothesized speech are also extended by an additional 0.1 seconds [14].

Table 1. Performance (EER%) of DNN/CNN systems with channel D as a seen/unseen channel.

System	DNN	CNN
Channel Specific Model (Seen condition - trained on D, tested on D)	1.6	1.4
Channel Independent Model (Unseen condition - trained on E-H, tested on D)	4.8	6.6

Table 2. Performance (EER%) of CI DNN/CNN systems (trained on channels E-H) after adaptation with channel D.

System	EER (%)
DNN (L4-L5)	2.7
CNN (L4-L5)	3.6
CNN (C1-C2)	2.2

To train the CNN, 40-dimensional *log-mel* spectra with Δ and $\Delta\Delta$s are used. A file-based mean-variance normalization is additionally applied to this spectra which covers the entire 0-8Khz frequency range. Every node of the first CNN layer uses a separate 9×9 filter on each of the 3 input streams. With a temporal context of 11 frames, each node produces a combined 32 $(40-9+1) \times 3$ $(11-9+1)$ activation from the input. After a max pool operation along the frequency axis, this is reduced to an 11×3 representation. We use 128 hidden nodes in the first CNN layer. The second CNN layer with 256 nodes uses filters of size 4×3 over the outputs of the first layer. The outputs from the second CNN layer are passed on to a DNN with architecture $1024 \times 1024 \times 1024 \times 40 \times 3$. A single CNN is trained on data from 4 channels (E-H) as a channel independent (CI) network. Additionally, separate channel specific networks are also trained as baselines for channels A-D.

To compare the proposed CNN systems, we also employ a DNN-based SAD system. The DNN-based networks are trained on PLP features. A Gaussian-level LDA is applied on 17 consecutive PLP frames to project the features to 40 dimensions [16] after file-based mean-variance normalization. Channel specific and channel independent DNNs with architecture $160 \times 1024 \times 1024 \times 1024 \times 40 \times 3$ are trained on these features which are also appended with their Δ, $\Delta\Delta$s and $\Delta\Delta\Delta$s. Both the CNNs and DNNs are discriminatively pre-trained before being fully trained to convergence [22, 16].

4.3. Evaluating on Seen and Unseen Channels

The trained CNN and DNN-based SAD systems are tested on each of the held-out unseen channels. We first present results on an individual channel - channel D. Table 1 shows the performance when channel D is tested against a channel specific model trained on channel D and against a channel independent system to which channel D is unseen. We observe that at the channel specific level, CNNs trained on *log-mel* features are able to outperform DNN systems trained on PLP features. However, in the unseen case the CNN performs worse than the DNN system. We hypothesize this to be from mismatches in the feature transforms, especially those in the feature extracting layers of the CNN, since there are significant differences between channel D and channels E-H. Channel D has frequency shift distortions not present in the channels seen during training [13].

4.4. Adaptation to Unseen Channels

To further understand the behavior of the CI networks, we adapt both the CNN and DNN with 15 minutes of supervised data. With this

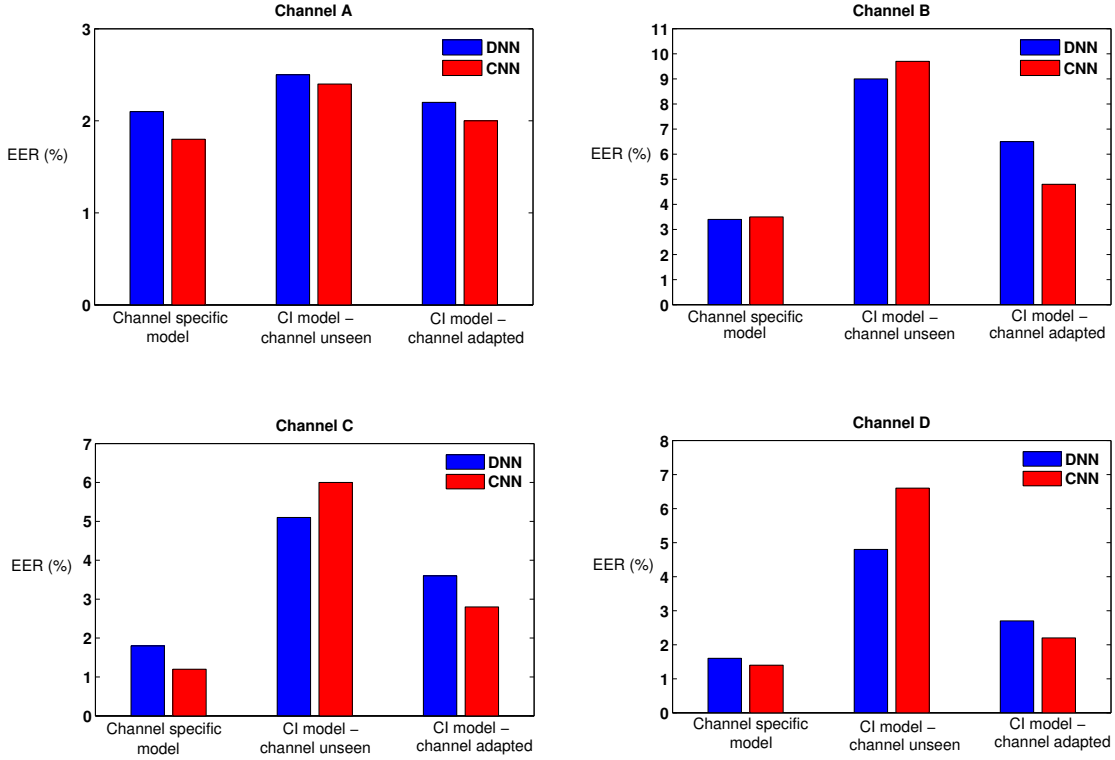

Fig. 3. Performance of DNN and CNN-based SAD systems on various RATS channels

limited amount of data, we adapt the weights and biases of only the last two layers (L4-L5) - 1024×40 and 40×3. Table 2 shows the results of this adaptation. There is a very significant improvement as the EER improves by nearly half. In a second experiment we adapt only the first two layers of the CNN (C1-C2) keeping the remaining parameters fixed. This adaptation provides further improvements confirming our hypothesis of more mismatches in the feature extracting layers of the CNN. No additional gains were observed by a similar adaptation of the front layers of the DNNs. The gains also remained same when both the front layers (C1-C2) and the back layers (L4-L5) of the CNN were both adapted. The performance improves further by an absolute 0.2% when the amount of adaptation data was increased to about 1 hour for both the models. In our remaining experiments on the other held out channels we adapt only the initial layers for the CNN and the last layers for the DNN with 15 minutes of data.

Figure 3 shows the results of these experiments on all the remaining held-out channels. On all the 4 channels we tested, we observe consistent patterns in performance -

(a) The CNNs perform better or similar to the DNNs on matched channel specific train and test conditions.

(b) In mismatched conditions both the CNNs and DNNs have a degradation in performance. However, the performance drops further for the CNNs depending on the degree of mismatch. In channels A and B for example, the performance differences between the CNNs and DNNs are not as significant as in channels C and D. We attribute these losses primarily to the CNN's data-driven feature extraction layers.

(c) Both the CNNs and DNNs respond well to minimal amounts of adaptation data. For the CNN architecture however, adaptation of the feature extracting layers has more impact than adapting the layers closer to the final output layer. There is a very significant relative improvement of up to 60% with just 15 minutes of adaptation data. This observation also points to the sensitivity of the CNN's data-driven feature extraction layers.

(d) The adapted CNN models perform better (about 10-25% relative) to DNN-based SAD systems.

5. CONCLUSIONS

We have explored the behavior of convolutional neural networks for speech activity detection on noisy data from the RATS program in the context of novel unseen channels. Our experiments show that channel independent networks can be used in this context but need to be improved via adaptation before they perform as well as channel specific networks. We have also demonstrated that CNNs are useful acoustic models in novel channel scenarios and can adapt well with limited amounts of adaptation data. We have currently restricted our adaptation experiments to limited amounts of supervised data. The data for adaptation can however be obtained via unsupervised or semi-supervised techniques. It will be useful to explore the performance of the CNN systems in self-training and co-training frameworks as well.

6. ACKNOWLEDGMENTS

The authors thank Tara Sainath and Brian Kingsbury for useful discussions.

7. REFERENCES

[1] K. Woo, T. Yang, K. Park, and C. Lee, "Robust Voice Activity Detection Algorithm for Estimating Noise Spectrum," *IEEE Electronics Letters*, 2000.

[2] R. Chengalvarayan, "Robust Energy Normalization using Speech/Non-speech Discriminator for German Connected Digit Recognition," in *ISCA Eurospeech*, 1999.

[3] ITU-T, "A Silence Compression Scheme for G.729 Optimized for Terminals Conforming to Recommendation V.70," in *Recommendation G.729-Annex B*, 1996.

[4] E. Nemer, R. Goubran, and S. Mahmoud, "Robust Voice Activity Detection using Higher-order Statistics in the LPC Residual Domain," *IEEE Electronics Letters*, 2000.

[5] J. Dines, J. Vepa, and T. Hain, "The Segmentation of Multichannel Meeting Recording for Automatic Speech Recognition," *ISCA ICSLP*, 2006.

[6] B. Kingsbury, G. Saon, L. Mangu, M. Padmanabhan, and R. Sarikaya, "Robust Speech Recognition in Noisy Environments: The 2001 IBM SPINE Evaluation System," *ISCA ICASSP*, 2002.

[7] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech," *The Journal of the Acoustical Society of America*, 1990.

[8] S. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1980.

[9] H. Hermansky and P. Fousek, "Multi-resolution RASTA Filtering for TANDEM-based ASR," in *ISCA Interspeech*, 2005.

[10] S. Ganapathy, P. Rajan, and H. Hermansky, "Multi-layer Perceptron based Speech Activity Detection for Speaker Verification," *IEEE WASPAA*, 2011.

[11] N. Mesgarani, M. Slaney, and S. Shamma, "Discrimination of Speech from Non-speech based on Multiscale Spectro-temporal Modulations," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[12] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient based Learning applied to Document Recognition," *Proceedings of the IEEE*, 1998.

[13] K. Walker and S. Strassel, "The RATS Radio Traffic Collection System," in *ISCA Odyssey*, 2012.

[14] T. Ng et al., "Developing a Speech Activity Detection system for the DARPA RATS Program," in *ISCA Interspeech*, 2012.

[15] S. Thomas et al., "Acoustic and Data-driven Features for Robust Speech Activity Detection," in *ISCA Interspeech*, 2012.

[16] G. Saon et al., "The IBM Speech Activity Detection System for the DARPA RATS Program," in *ISCA Interspeech*, 2013.

[17] A. Tsiartas et al., "Multi-band Long-term Signal Variability Features for Robust Voice Activity Detection," in *ISCA Interspeech*, 2013.

[18] M. Graciarena et al., "All for One: Feature Combination for Highly Channel-degraded Speech Activity Detection," in *ISCA Interspeech*, 2013.

[19] Y. Lecun, F. Huang, and L. Bottou, "Learning Methods for Generic Object Recognition with Invariance to Pose and Lighting," in *IEEE CVPR*, 2004.

[20] O. Abdel-Hamid, A. Mohamed, H. Jiang, and G. Penn, "Applying Convolutional Neural Network concepts to Hybrid NN-HMM model for Speech Recognition," in *IEEE ICASSP*, 2012.

[21] T. Sainath, A. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep Convolutional Neural Networks for LVCSR," in *IEEE ICASSP*, 2013.

[22] H. Soltau, H.K. Kuo, L. Mangu, G. Saon, and T. Beran, "Neural Network Acoustic Models for the DARPA RATS Program," in *ISCA Interspeech*, 2013.

[23] V. Abrash, H. Franco, A. Sankar, and M. Cohen, "Connectionist Speaker Normalization and Adaptation," in *ISCA Eurospeech*, 1995.

[24] R. Gemello, F. Mana, S. Scanzio, P. Laface, and R. De Mori, "Linear Hidden Transformations for Adaptation of Hybrid ANN/HMM Models," *Speech Communication*, 2007.

[25] K. Yao, D. Yu, F. Seide, H. Su, L.i Deng, and Y. Gong, "Adaptation of Context-dependent Deep Neural Networks for Automatic Speech Recognition," in *IEEE SLT*, 2012.

[26] O. Abdel-Hamid and H. Jiang, "Rapid and Effective Speaker Adaptation of Convolutional Neural Network based Models for Speech Recognition," in *ISCA Interspeech*, 2013.