3-D CNN MODELS FOR FAR-FIELD MULTI-CHANNEL SPEECH RECOGNITION

Sriram Ganapathy

Indian Institute of Science, Bangalore.
sriramg@iisc.ac.in

Vijayaditya Peddinti†

Google, Inc., U.S.A.
vpeddinti@google.com

ABSTRACT

Automatic speech recognition (ASR) in far-field reverberant environments, especially when involving natural conversational multi-party speech conditions, is challenging even with the state-of-the-art recognition methodologies. The two main issues are artifacts in the signal due to reverberation and the presence of multiple speakers. In this paper, we propose a three dimensional (3-D) convolutional neural network (CNN) architecture for multi-channel far-field ASR. This architecture processes time, frequency & channel dimensions of the input spectrogram to learn representations using convolutional layers. Experiments are performed on the REVERB challenge LVCSR task and the augmented multi-party (AMI) LVCSR task using the array microphone recordings. The proposed method shows improvements over the baseline system that uses beamforming of the multi-channel audio along with a 2-D conventional CNN framework (absolute improvements of 1.1 % over the beamformed baseline system on AMI dataset).

Index Terms— Far-field speech recognition, 3D CNN modeling, Multi-party conversational speech.

1. INTRODUCTION

Multi-speaker conversations in far-field environments pose a significant challenge to automatic speech recognition systems even when employing state-of-the-art speech recognition systems [1]. *e.g.* Peddinti *et al.,* [2] report a 75% rel. increase in word error rate (WER) when signals from a far-field array microphone are used in place of those from headset microphones in the ASR systems, both during training and testing. This degradation could be primarily attributed to reverberation and multi-speaker overlaps [3, 4]. The availability of multi-channel signals can be leveraged for alleviating these issues.

Beamforming is a popular approach for multi-channel signal based enhancement of speech [5, 6]. In this paper, we develop a neural network architecture consisting of 3-D convolutional models as a front-end for processing multi-channel speech. The spectrogram representation of speech is extracted from each channel independently. The 3-D CNN is fed with three dimensional input representation consisting of time, frequency and channel dimensions. The feature maps from the initial CNN layers are then fed to time delay neural network layers (TDNN) [7] followed by a final layers of sequence modeling with LSTM networks. The entire model is trained using a sequence cost function with lattice free minimum mutual information (MMI) criterion [8].

The research reported here was conducted at the 2017 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Carnegie Mellon University and sponsored by Johns Hopkins University with unrestricted gifts from Amazon, Apple, Facebook, Google, and Microsoft. † - Work performed as an open source contributor to Kaldi.

Experiments are primarily performed on the augmented multi-party speech database [9]. In these experiments, the proposed method improves the performance over a baseline system using a 2-D CNN-TDNN-LSTM architecture which uses either single channel speech or beamformed output of multi-channel speech as inputs. This model was shown to provide the best reported results on the single channel AMI LVCSR task [10]. We also contrast the performance of the proposed approach in the case of a single speaker setting (using the REVERB Challenge corpus [4]) with multi-party conversations in AMI dataset.

The rest of the paper is organized as follows. Prior work is discussed in § 2, proposed architecture is discussed in § 3, experimental setup is described in § 4 and the results for multi-channel speech recognition are reported in § 5. This is followed by a discussion § 6 and conclusion in § 7.

2. PRIOR WORK

Beamforming fundamentally relies on estimating the time delay between speech signals recorded from multiple channels, and designing a spatial filter to perform a delay and sum operation for generating an enhanced single channel signal [11]. This can be used in improving the downstream speech systems, including ASR. These methods can also be modified for maximizing the likelihood [12].

With the advent of neural network based acoustic models multi-channel acoustic models have also been explored. Recently Swietojanski *et al* [13] proposed the use of features from each channel of the multi-channel speech directly as input to a convolutional neural network based acoustic model. Here the neural network is seen as a replacement for conventional beamformer. Joint training of a more explicit beamformer with the neural network acoustic model has been proposed by Xiao *et al.,* [14]. Training of neural networks, which operate on the raw signals and are optimized for the discriminative cost function of the acoustic model, has also been recently explored. These approaches are termed *Neural Beamforming* approaches as the neural network acoustic model subsumes the functionality of the beamformer [15, 16, 17].

In this paper, we develop a neural network architecture consisting of 3-D convolutional models as a front-end for processing multi-channel speech. 3-D CNN architectures have shown promising results in video signal based human action segmentation [18] and more recently in the bio-medical imaging applications for lesion segmentation [19].

3. PROPOSED 3-D CNN ARCHITECTURE

The block schematic of the proposed architecture is shown in Fig. 1. The multi-channel recordings are converted into spectrogram representation containing 40 bands of log-mel filtered filter bank energies

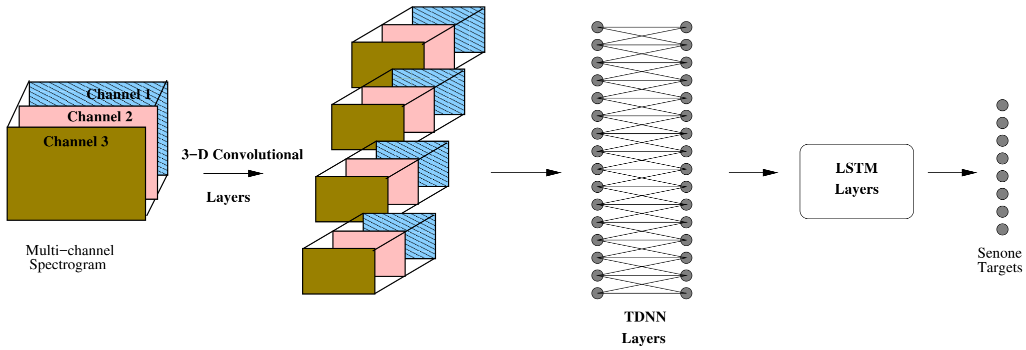

Fig. 1. Block schematic of the proposed 3-D CNN based network architecture for AMI speech recognition.

sampled at every 25 ms of the audio file with a shift of 10 ms. The multi-channel audio segments are stacked in a 3-D fashion and fed as input to the neural network model. The neural network architecture consists of convolutional layers which have 3-D kernel. The CNN layers perform the following convolutional operation,

$$Y(i,j,k) = \sum_{x=1}^{N_x}\sum_{y=1}^{N_y}\sum_{z=1}^{N_z} X(i+x, j+y, k+z)K(x,y,z) \quad (1)$$

where K is the 3-D kernel, X is the input multi-channel spectrogram, Y is the output of the feature map and (N_x, N_y, N_z) represents the kernel size. The operation is performed without any padding on the input spectrogram so that output dimensions are reduced in each convolutional operation. In our case, we perform two layers of 3-D convolutions.

The feature maps generated by the CNN layers are flattened along the frequency and channel dimensions, every time step, and are fed to successive layers of time delay neural networks [7]. While standard feed forward networks process the entire input contexts, TDNN architecture captures narrow temporal context in the initial layers and then approximates the wider temporal context in the final layers. This has shown to improve the ASR performance for far-field reverberant speech [20].

The higher TDNN layers are interleaved with the LSTM layers as a combination of these layers has been shown to be optimal for low latency temporal context modeling [2]. The final output of the TDNN-LSTM stack is mapped to the senone (context dependent HMM state) targets using a fully connected layer.

4. EXPERIMENTAL SETUP

4.1. AMI LVCSR task

The AMI meeting corpus [21] contains conversational speech with the training and test configuration chosen similar to [22]. The training data corpus consist of about 100 hours of meetings recorded in instrumented meeting rooms at three sites in the UK, the Netherlands, and Switzerland. Each meeting usually has four participants and the language is English, albeit with a large proportion of non-native speakers. The recording involves multiple parallel microphones, including individual headset microphones (IHM), lapel microphones, and one or more microphone arrays. Every recording used a primary 8-microphone uniform circular array (10 cm radius), as well as a secondary array whose geometry varied between sites. In this work we use the first three microphone recordings of the primary array for our single distant microphone (SDM) experiments.

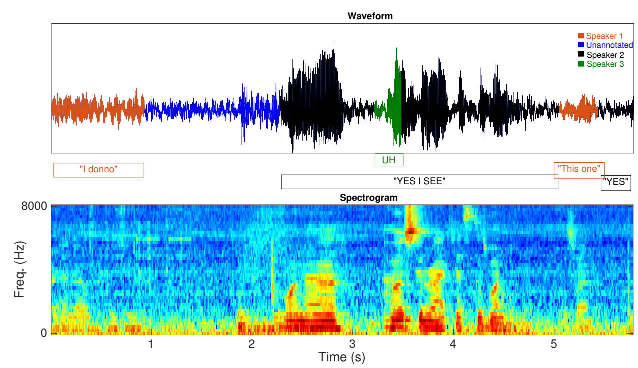

Fig. 2. Portion of meeting speech and corresponding spectrogram. The speech regions with their labels and speaker index are highlighted.

A portion of the speech signal from the AMI corpus and the corresponding spectrogram is illustrated in Fig. 2. As seen here, the spectrogram contains reverberation artifacts and there is substantial amount of overlap speech (about 12 % of the recording duration contains multi-talker speech). Also, the dataset consists of substantial amounts short segments with large number of speaker turns.

We use the lattice-free maximum mutual information (LF-MMI) cost function to train in a purely sequence discriminative training approach [8]. The cost function is similar to connectionist temporal classification (CTC) [23], however the lattice free MMI involves a global normalization.

We use the Kaldi toolkit [24] to conduct our experiments. This recipe follows the corpus release for the training and evaluation splits. For training purposes we consider all segments (including those with overlapped speech), and the WERs of the speech recognition outputs are scored by the tool following the NIST RT recommendations for scoring simultaneous speech[1].

A HMM-GMM system is used to generate numerator lattices for LF-MMI training and also for cross-entropy regularization. The training and decoding closely follows the one described in [25]. We use the speed-perturbation technique [26] for data augmentation with 3-way speed perturbation. The log-mel features are also appended with iVectors to perform instantaneous adaptation of the neural network [27]. The WER results are reported after 4-gram LM re-scoring of lattices generated using a trigram LM. All the neural network models in AMI corpus use 40 dimensional log-mel filter bank en-

[1] http://nist.gov/speech/tests/rt/2009

Table 1. Word error rate (%) on REVERB Challenge corpus for simulated (S) and naturally reverberant conditions (R) on development (dt) and evaluation (et) datasets.Here BF corresponds to beamforming.

Model	S-dt	S-et	R-dt	R-et
DNN-Single-Chn.	12.7	13.6	31.8	37.5
CNN2D-Single-Chn	11.3	11.4	26.8	29.6
CNN2D-Multi-BF	9.7	10.0	24.8	26.4
CNN2D-Multi-BF + Dropout	10.7	11.5	26.7	27.5
CNN3D-Multi	9.8	10.3	26.7	28.4
CNN3D-Multi + Dropout	**9.1**	**9.8**	**24.6**	**25.8**

ergy features which are processed with segment level mean variance normalization. The ground truth speech activity detection (SAD) is assumed and the i-vectors of 100 dimensions are also used concatenated with the input features.

4.2. Reverb Challenge LVCSR task

To understand the impact of the proposed architecture in the single speaker setting we use the REVERB challenge [28] LVCSR task. A major deviation in this setup compared to the AMI LVCSR task is that we report the results on cross-entropy trained systems as the LF-MMI cost function has not been shown to be very effective for low resource LVCSR tasks [8]. Further in these experiments, we use only the CNN front-end followed by feed forward layers (without the TDNN and LSTM layers described in Fig.1).

The REVERB challenge [28] corpus uses the WSJCAM0 database for training. This database consists of 7861 recordings of read speech from 92 training speakers, 1488 recordings from 20 development test (dt) speakers and 2178 recordings from two sets of 14 evaluation test (et) speakers, with each speaker providing about 90 utterances. These recordings were carried out with two sets of microphone- head mounted as well as desk microphone positioned about half meter from the speaker's head.

The database consists of three subsets: training data set (Train) - for multi condition training using simulated reverb data, a simulated test dataset (Sim) and a naturally reverberant recording of the test dataset (Real). Once again the Kaldi toolkit [24] is used for the data preprocessing and the Keras [29] tool is used for training the acoustic model.

The features are the 23 dimensional log-mel filterbank energies which are mean and variance normalized. A context of 21 frames is used for generating the time-frequency representation used at the input of the CNN. For the multi-channel experiments, we use 3 channels from the array microphone recordings.

5. RESULTS

The summary of the results for various speech recognition experiments in the REVERB Challenge corpus are reported in Table 1. The CNN based architecture provides significant improvements over the DNN model for the single channel case. Here, the CNN2D architecture consist of 4 convolutional layers (2 layers of 256 filters and 2 layer of 128 filters each with a 3×3 kernel and max pooling after the second and fourth layer) and 2 dense (feed-forward) layers. The beamformed approach using the single channel CNN-2D architecture (beamformed with the 3 channels which are also used in the CNN3D architecture) further improves the performance over the single channel case. In the REVERB challenge dataset, the beamforming is more effective as each recording contains only one speaker (one source). Thus, the assumptions made in delay-sum based enhancement [11] are valid as the evaluation was done with static location of the speakers.

The multi-channel CNN architecture (CNN3D) improves significantly over the single channel set up. The best performance obtained for the multi-channel CNN3D experiments are reported here. We have used two layers of convolutional 3D filters of kernel size ($3\times3\times2$) with 256 filters followed by max-pooling and another 2 layers of 128 filters along with max-pooling. The convolutional layers are followed with 2 dense layers of 1024 dimensions. While training the models, we have observed an overfitting trend where the training and validation accuracies for frame level senone targets are drastically different. In order to circumvent this issue, we have used a regularization approach using dropout [31]. The dropout scheme provides significant improvements to the CNN3D architecture and the CNN3D outperforms the best beamforming system (average relative improvements of 2.2 % over the CNN2D-Beamform baseline).

The results for baseline experiments on the AMI corpus using the SDM recordings (single channel) are reported in Table 3 where the results are reported separately for development and evaluation meetings. The results of the baseline AMI system using HMM-GMM is reported in the first row. The HMM-GMM system is trained with linear discriminant analysis (LDA) - maximum likelihood linear transform (MLLT) approach [24] followed by a speaker adaptive training. A similar ASR system was built on the IHM recordings and alignments obtained from IHM setup (using force alignment) are used as labels for neural network models on the SDM data. The next two results reported in Table 3 are the TDNN based ASR system with four hidden layers trained with cross-entropy cost function and the sequence cost function respectively. As seen here, the performance is significantly improved using a TDNN acoustic model over the HMM-GMM system. The sequence cost function further improves the WER.All further experiments use the sequence training cost function.

A similar pipeline with CNN 2D (5 layers) followed by 2 layers of TDNN is reported next. All the CNN layers have 64 filters and the TDNN layers have 512 dimensions. The use of CNN front-end results in decrease in performance compared to the TDNN architecture. The last result in Table 3 reports the performance of the CNN2D-TDNN-LSTM setup. In this set up, two CNN layers with 256 and 128 filters respectively are used for the generating the front-end features maps. This is followed by 2 layers of TDNN architecture with 1024 dimensions and 3 layers of LSTM architecture with 1024 LSTM cells in each layer. As seen in this table, the addition of the LSTM layers provides significant improvements to the CNN2D-TDNN model as well as the baseline TDNN model (average relative improvements of about 11% over the TDNN model and about 15 % over the CNN2D-TDNN model). The single channel SDM results from two recent efforts using attention LSTM [30] as well as bidirectional LSTM with TDNN front-end [2] are also added to this table for reference.

The multi-channel experiments using the CNN3D-TDNN-LSTM model are reported in Table 2. Here, we use the first three recordings of the array microphone as input representation to the CNN3D model. Various choices of input filter sizes, weight sharing and pooling are experimented for the 3-D CNN architecture front-end. The rest of model architecture containing 2 layers of 1024 dimensions of TDNN layers and 3 LSTM layers each with 1024 cells is used as before. As seen in Table 2, increasing the number of input filters (each filter has a fixed kernel size of $(3 \times 3 \times 1)$)

Table 2. Word error rate (%) on AMI corpus for multi-channel SDM experiments using CNN3D-TDNN-LSTM architecture with various choices of model parameters.

Model	Dev.	Eval
Layer1-(64 Filt.) Layer2-(32 Filt.)	34.8	37.4
Layer1-(96 Filt.) Layer2-(32 Filt.)	34.5	37.2
Layer1-(128 Filt.) Layer2-(64 Filt.)	34.4	37.4
Layer1-(256 Filt.) Layer2-(128 Filt.)	34.9	37.9
Layer1-(256 Filt.) Layer2-(128 Filt.) + Reg.	32.7	35.7
Layer1-(256 Filt.) Layer2-(128 Filt.) + Reg. and Sharing	**32.6**	**35.4**
Layer1-(256 Filt.) Layer2-(128 Filt.) + Reg., Sharing and Avg. pool	32.6	35.7
Layer1-(384 Filt.) Layer2-(192 Filt.) + Reg. and Sharing	32.7	35.7

Table 3. Word error rate (%) on AMI corpus for single channel SDM experiments.

Model	Dev.	Eval
HMM-GMM (LDA-MLLT-SAT)	59.5	64.0
TDNN (CE)	41.7	46.7
TDNN (Seq.)	40.2	44.1
CNN2D-TDNN (Seq.)	41.8	46.7
CNN2D-TDNN-LSTM (Seq.)	**36.0**	**39.0**
Attention-LSTM [30]	41.3	45.8
TDNN-LSTM [2]	37.4	40.4

showed an over fitting trend compared to the single channel CNN2D case with similar number of filters. In order to circumvent this issue, we used a regularization approach [32] consisting of derivative truncation and shrinkage. This way of regularization helps to avoid overfitting and allows the models to be trained with same the number of filters as the single channel case. We also see minor improvements by sharing the weights across the channels. The use of average pooling across channels in the CNN layer did not improve the performance any further. The results reported for AMI single channel (Table 3) as well as three channel results using CNN3D model (Table 4) are the best published results for this task to the best of our knowledge.

The results reported in Table 4 compares the performance of the best single channel system (CNN2D-TDNN-LSTM) trained on channel 1 as well as the delay-sum beamformed output [11] and the CNN3D-TDNN-LSTM model[2]. The beamforming was done on the same 3 channels used in the multi-channel ASR experiments using the approach described in [11]. The proposed approach improves over the beamforming method (absolute improvements of 1.3 % on dev meetings and 0.8 % on the eval meetings over the best beamformed model).

6. DISCUSSION

The beamforming methods optimize the signal enhancement problem by delay-sum operations (spatial filtering). The acoustic modeling in speech recognition predicts the senone classes and optimizes a sequence training cost function. As seen in various speech recognition experiments in this paper, there are substantial advantages in combining the two steps by using joint CNN3D framework.

The previous efforts on CNN with multi-channel framework [13] use separate kernels on each channel which get merged in the

[2]The implementation of the proposed model can be found in $https://github.com/vijayaditya/kaldi/tree/3d\_cnn$

Table 4. Word error rate (%) in AMI corpus for multi-channel SDM experiments using proposed approach and baseline beamformed [11] approach

Model	Dev.	Eval
CNN2D-TDNN-LSTM (single)	36.0	39.0
CNN2D-TDNN-LSTM (multi beamformed)	33.9	36.2
CNN3D-TDNN-LSTM (multi)	**32.6**	**35.4**

network. The proposed approach uses 3D framework to capture the time-frequency-channel correlations. In comparison with the raw-waveform based beamforming efforts [15], the proposed approach is more robust to speaker and source location changes. While raw waveform based approaches have the advantage of using the signal phase information for multi-channel combination, the regions of speaker changes causes degradation which could potentially impede the performance. Also, the model proposed in [15] attempts to summarize the spatial dimension in the first layer of the network where as the proposed approach preserves the space dimension deep in CNN architecture (until the feed-forward/TDNN layers). Hence, we hypothesize that proposed model may be more suitable for ASR applications on natural multi-speaker conversations.

7. SUMMARY AND FUTURE WORK

In this paper, we have proposed a three dimensional neural network consisting of convolutional layers followed by LSTM layers. The 3D CNN architecture receives input from time-frequency-channel dimensions of the input multi-channel speech. Various speech recognition experiments were performed in the REVERB challenge dataset as well as the AMI speech recognition database. The main finding is that while the beamforming approach of speech enhancement is effective in simple settings involving single speaker recordings (as seen in REVERB challenge corpus), the proposed approach of 3D CNN architecture improves noticeably over the beamforming methods on REVER challenge corpus and the multi-party conversational settings. The promising results motivate us to further pursue novel modeling methods for multi-channel speech recognition involving spatial recurrence in LSTM models. The current way of combining the channels uses a time resolution of 10 ms in the spectrogram representations which can be improved by having a higher temporal granularity (frames taken below 10 ms). The higher temporal sampling rate could allow the model to have more spatial resolution and improve the 3D structure of the data. In addition, the AMI database contains 8 parallel microphones which could potentially allow greater flexibility in the 3D modeling.

8. REFERENCES

[1] Yu Zhang et al., "Highway long short-term memory rnns for distant speech recognition," in *ICASSP*. IEEE, 2016, pp. 5755–5759.

[2] Vijayaditya Peddinti, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, "Low latency acoustic modeling using temporal convolution and LSTMs," *IEEE Signal Processing Letters*, 2017.

[3] Takuya Yoshioka et al., "Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 114–126, 2012.

[4] Keisuke Kinoshita et al., "The reverb challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *2013 IEEE WASPAA*. IEEE, 2013, pp. 1–4.

[5] Matthias Wölfel and John McDonough, *Distant speech recognition*, John Wiley & Sons, 2009.

[6] Marc Delcroix et al., "Strategies for distant speech recognition in reverberant environments," *EURASIP Journal on Advances in Signal Processing*, vol. 2015, no. 1, pp. 60, 2015.

[7] Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *INTERSPEECH*, 2015.

[8] Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur, "Purely sequence-trained neural networks for asr based on lattice-free MMI.," in *INTERSPEECH*, 2016, pp. 2751–2755.

[9] Jean Carletta et al., "The ami meeting corpus: A pre-announcement," in *International Workshop on Machine Learning for Multimodal Interaction*. Springer, 2005, pp. 28–39.

[10] Gaofeng Cheng and Daniel Povey, "CNN-TDNN-LSTM acoustic models for AMI LVCSR," `https://github.com/kaldi-asr/kaldi/blob/master/egs/ami/s5b/local/chain/run_cnn_tdnn_lstm.sh`, 2017, [Online; accessed 3-Nov-2017].

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

[12] Michael L Seltzer, Bhiksha Raj, and Richard M Stern, "Likelihood-maximizing beamforming for robust hands-free speech recognition," *IEEE Transactions on speech and audio processing*, vol. 12, no. 5, pp. 489–498, 2004.

[13] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Convolutional neural networks for distant speech recognition," *IEEE Signal Processing Letters*, vol. 21, no. 9, pp. 1120–1124, 2014.

[14] Xiong Xiao et al., "Deep beamforming networks for multi-channel speech recognition," in *ICASSP*. IEEE, 2016, pp. 5745–5749.

[15] Tara N Sainath et al., "Multichannel signal processing with deep neural networks for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 5, pp. 965–979, 2017.

[16] Bo Li, Tara N Sainath, Ron J Weiss, Kevin W Wilson, and Michiel Bacchiani, "Neural network adaptive beamforming for robust multichannel speech recognition.," in *INTERSPEECH*, 2016, pp. 1976–1980.

[17] Tsubasa Ochiai, Shinji Watanabe, Takaaki Hori, John R Hershey, and Xiong Xiao, "A unified architecture for multichannel end-to-end speech recognition with neural beamforming," *IEEE Journal of Selected Topics in Signal Processing*, 2017.

[18] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu, "3D convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[19] Konstantinos Kamnitsas et al., "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation," *Medical image analysis*, vol. 36, pp. 61–78, 2017.

[20] Vijayaditya Peddinti, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Reverberation robust acoustic modeling using i-vectors with time delay neural networks," in *INTERSPEECH*, 2015.

[21] Jean Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus," *Language Resources and Evaluation*, vol. 41, no. 2, pp. 181–190, 2007.

[22] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, "Hybrid acoustic models for distant and multichannel large vocabulary speech recognition," in *ASRU*. IEEE, 2013, pp. 285–290.

[23] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton, "Speech recognition with deep recurrent neural networks," in *ICASSP*. IEEE, 2013, pp. 6645–6649.

[24] Daniel Povey et al., "The kaldi speech recognition toolkit," in *IEEE ASRU*. IEEE Signal Processing Society, 2011, number EPFL-CONF-192584.

[25] Vijayaditya Peddinti, Vimal Manohar, Yiming Wang, Daniel Povey, and Sanjeev Khudanpur, "Far-field asr without parallel data," in *Proceedings of Interspeech*, 2016.

[26] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition.," in *INTERSPEECH*, 2015.

[27] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, "Speaker adaptation of neural network acoustic models using i-vectors.," in *ASRU*, 2013, pp. 55–59.

[28] Keisuke Kinoshita, Marc Delcroix, Tomohiro Nakatani, and Masato Miyoshi, "Suppression of late reverberation effect on speech signal using long-term multiple-step linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 534–545, 2009.

[29] François Chollet et al., "Keras: Deep learning library for theano and tensorflow," *URL: https://keras. io/k*, 2015.

[30] Yu Zhang, Pengyuan Zhang, and Yonghong Yan, "Attention-based LSTM with multi-task learning for distant speech recognition," *INTERSPEECH*, pp. 3857–3861, 2017.

[31] George E Dahl, Tara N Sainath, and Geoffrey E Hinton, "Improving deep neural networks for lvcsr using rectified linear units and dropout," in *ICASSP*. IEEE, 2013, pp. 8609–8613.

[32] Frank E Curtis and Katya Scheinberg, "Optimization methods for supervised machine learning: From linear models to deep learning," in *Leading Developments from INFORMS Communities*, pp. 89–113. INFORMS, 2017.