3-D CNN MODELS FOR FAR-FIELD MULTI-CHANNEL ASR

Sriram Ganapathy* and Vijayaditya Peddinti†

*Learning and Extraction of Acoustic Patterns (LEAP) Labs, Indian Institute of Science, Bangalore.

† Google Inc., USA.

Abstract

We propose a three dimensional (3-D) convolutional neural network (CNN) architecture for multi-channel far-field ASR which processes time, frequency & channel dimensions of the input spectrogram.

Introduction

Multi-speaker conversations in far-field environments pose a significant challenge to ASR due to reverberation and multi-speaker overlaps.

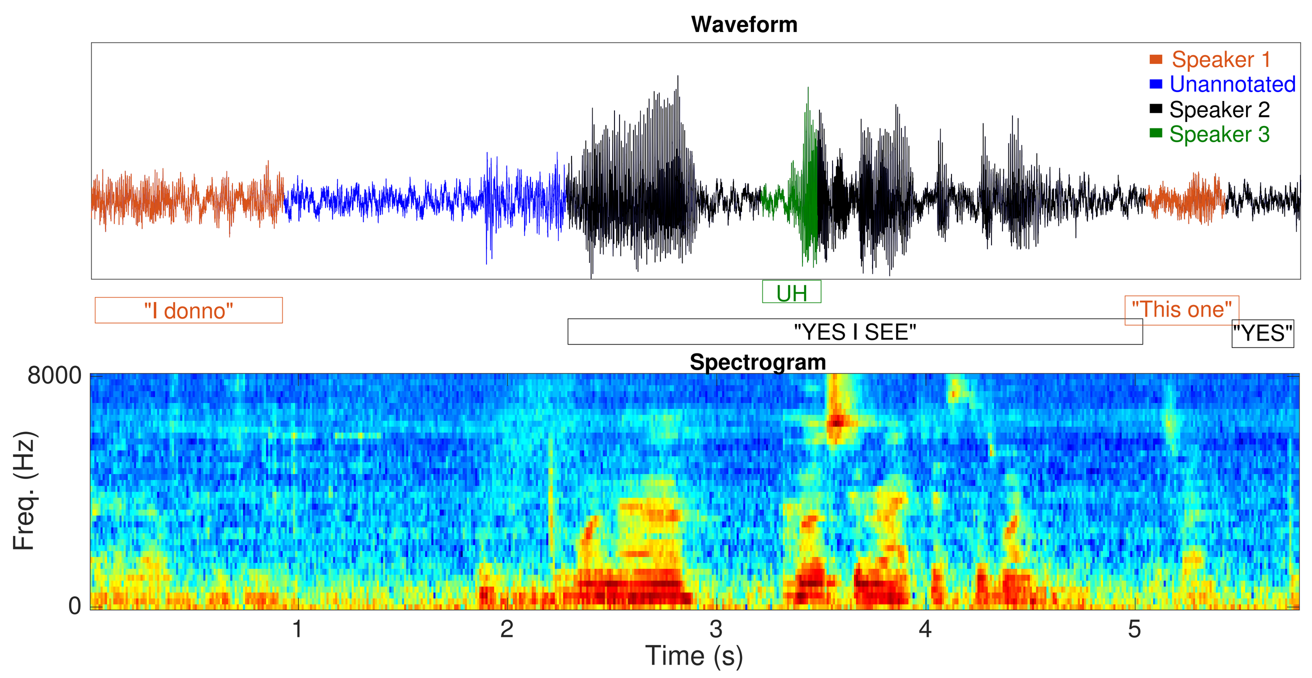


Fig. 1. Portion of meeting speech and corresponding spectrogram.

The availability of multi-channel signals can be leveraged for alleviating these issues.

Prior Work

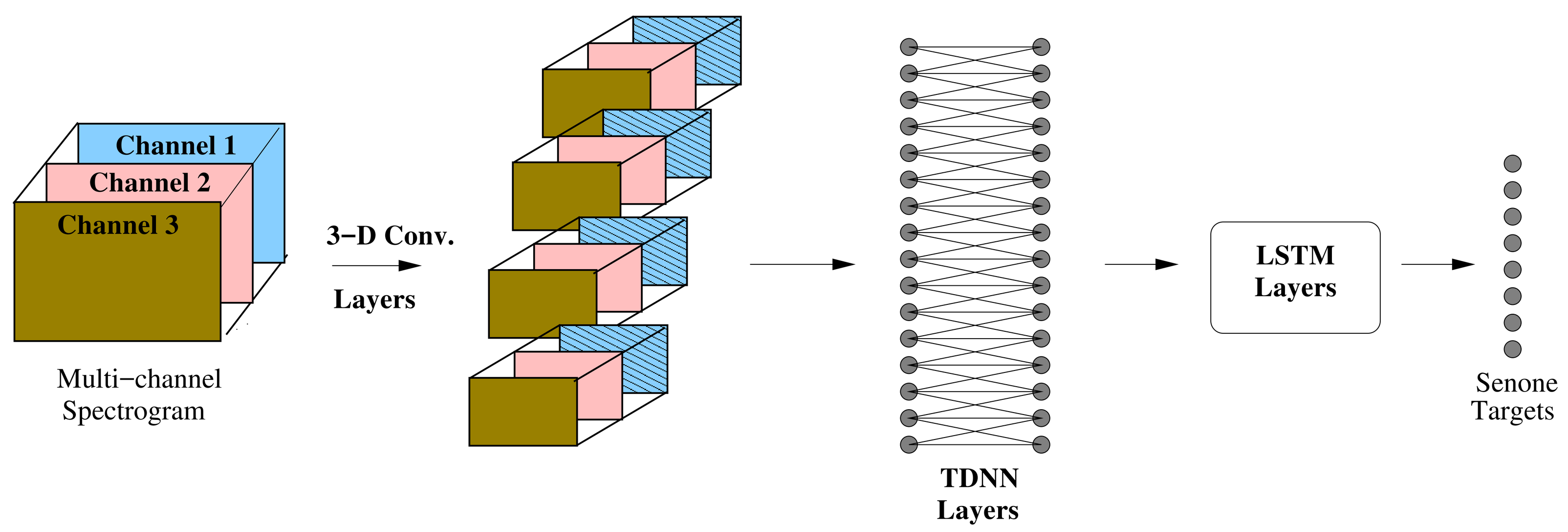
- Beamforming - designing a spatial filter to perform a delay and sum operation [1]
- Swietojanski *et al* [4] proposed the use of features from each channel speech directly as input.
- Training of neural networks on the raw signals optimized for the discriminative cost function of the ASR[3].

Proposed 3-D CNN Architecture

The multi-channel audio segments are stacked in a 3-D fashion and fed as input to the neural network model. The CNN layers perform the following 3-D 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.



Experiments and Results

Reverb Challenge LVCSR task

For the single speaker far-field experiments, we use the REVERB challenge LVCSR task with first three microphones.

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**

AMI Single Distant Microphone ASR

The performance of AMI-SDM experiments, shown below, 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.

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 [5]	41.3	45.8
TDNN-LSTM [2]	37.4	40.4

AMI Multi Microphone ASR

Here, we use the first three recordings of the array microphone as input representation to the CNN3D model.

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

Comparing with beamforming [1] 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**

Summary

In this paper, we have proposed a three dimensional neural network consisting of convolutional layers that receives input from time-frequency-channel dimensions of the input. The CNN3D model improves the beamforming methods for multi-channel ASR.

Acknowledgement

The research reported here was conducted at the 2017 Frederick Jelinek Memorial-JHU Workshop hosted at Carnegie Mellon University.

References

[1] Xavier Anguera, Chuck Wooters, and Javier Hernando. Acoustic beamforming for speaker diarization of meetings. *IEEE TASLP*, 2007.

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

[3] Tara N Sainath et al. Multichannel signal processing with deep neural networks for automatic speech recognition. *IEEE/ACM TASLP*, 2017.

[4] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals. Convolutional neural networks for distant speech recognition. *IEEE SPL*, 2014.

[5] Yu Zhang, Pengyuan Zhang, and Yonghong Yan. Attention-based LSTM with multi-task learning for distant speech recognition. *INTERSPEECH*, 2017.