TATA
TATA CONSULTANCY SERVICES

CHiME 2018 Workshop : Enhancing beamformed audio using Time Delay Neural Network De-noising Autoencoder

Sonal Joshi, Ashish Panda, Meet Soni, Rupayan Chakraborty, Sunilkumar Kopparapu,
Nikhil Mohanan, Premanand Nayak, Rajbabu Velmurugan, Preeti Rao
TCS Innovation Labs and IIT Bombay

Abstract

- Enhancing the beamformed utterances by using a Time Delay Neural Network De-noising autoencoder (TDNN-DAE).
- Trained the TDNN-DAE using *non-overlapping speech worn microphone utterances* (targets) and their corresponding beamform utterances.

Background

- **CHiME5 baseline** [1] :
  - Trained using around 149k worn (binaural) microphone utterances and a random set of 100k utterances from the Kinect arrays.
  - WER for *dev_worn* = 71.62% for GMM-HMM acoustic model (tri3).
- **Initial experiment** :
  - Train tusing *train_worn* microphone utterances and test using *dev_worn*
  - WER improved to 67.15%
  - This performance improvement can be attributed to acoustic mismatch conditions between the worn and array microphones.

Contribution

- We propose that an acoustic model trained by using only worn microphone utterances will perform better if the test data is acoustically similar to worn microphone data.
- A *beamform to worn* utterance TDNN-DAE [2] is trained using the Kaldi Toolkit [3].

Figure 1 :Training the TDNN-DAE

Experimental evaluation

TDNN-DAE architecture is similar to [2]. Contexts for the DAE network with four hidden layers is organized as (-2,-1,0,1,2) (-1,2) (-3,3) (-7,2) (0) and the input temporal context to [-13,9]

TDNN-DAE Training : 100k beamformed segments and the targets are their corresponding worn utterances.

Stage I: Obtain non-overlapping worn utterances

Step 1 :
Split all *train_worn* utterances into different sessions

Step 2 :
Each session is then the sorted in ascending order of time

Step 3 :
Find if the segment is non-overlapping.
If N utterances are indexed as $i = 1, 2, 3, \ldots N$, then an utterance $x(i)$ is *non-overlapping* if *start_time* $(x(i+1))$ does not lie in the range *start_time* $(x(i))$ to *end_time* $(x(i))$

Stage 2: Find mappings between beamform and worn segments using the utterance transcriptions

- Utterance transcriptions are used to find beamform utterances corresponding to the above non-overlapping worn utterances.
- A random set of 100k such mappings is used to train the TDNN-DAE.
- The worn utterances act as targets for the TDNN-DAE (Refer Figure 1).
- The development set after beamforming is enhanced using this TDNN-DAE.
- We decode the enhanced utterances using the baseline ASR (System 1, Figure 2) and another ASR trained using only worn utterances (System 2, Figure 3).

System

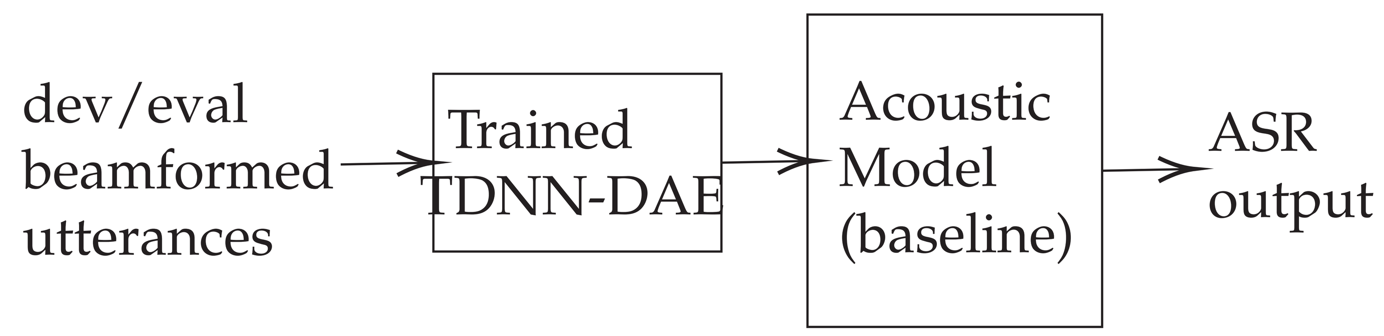

Figure 2 : Block diagram of System 1 (Using CHiME5 baseline Acoustic Model) with TDNN-DAE

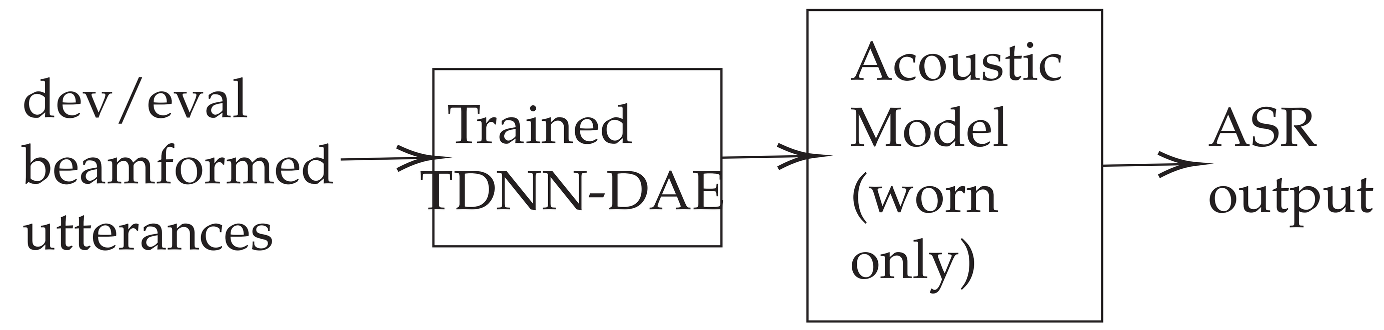

Figure 3 : Block diagram of System 2 (using only worn utterances for Acoustic Model) with TDNN-DAE

Results

The overall WER(%) for both the systems without using TDNN-DAE is shown in Table 1 and the overall WER(%) for both the systems using TDNN-DAE is shown in Table 2. Table 3 gives the results for the proposed System 2 with TDNN-DAE per session and location.

Track	System	WER
Single	System 1	90.82
	System 2	92.31

Overall WER (%) for the systems tested on the development test set without using TDNN-DAE

Track	System	WER
Single	System 1	95.52
	System 2	93.98

Overall WER (%) for the systems tested on the development test set using TDNN-DAE

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	97.10	93.53	93.21	93.98
		S09	93.43	93.23	92.06	

Results for the System 2 (using only worn utterances for Acoustic Model) with TDNN-DAE. WER (%) per session and location together with the overall WER.

References

[1] Barker et al. The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines. In *Interspeech*, 2018.

[2] Peddinti et al. A time delay neural network architecture for efficient modeling of long temporal contexts. In *Interspeech*, 2015.

[3] Povey et al. The kaldi speech recognition toolkit. In *IEEE workshop on automatic speech recognition and understanding*, 2011.