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

Sonal Joshi[1], *Ashish Panda*[1], *Meet Soni*[1], *Rupayan Chakraborty*[1], *Sunilkumar Kopparapu*[1], *Nikhil Mohanan*[2], *Premanand Nayak*[2], *Rajbabu Velmurugan*[2], *Preeti Rao*[2]

[1]TCS Innovation Labs, Mumbai
[2]Indian Institute of Technology Bombay, India

sonals.joshi@tcs.com

Abstract

In the submitted system to CHiME-5 challenge, we propose front-end enhancement of the beamformed array utterances to mitigate mismatch conditions between close-talking utterances and array utterances. Our initial experiments showed that an Acoustic Model trained by using only close-talking microphone utterances gave a superior performance than the baseline acoustic model when tested using close-talking utterances of the development set. Taking this cue, we explored the hypothesis that if array utterances are mapped to corresponding close-talking utterances, the system trained using only worn utterances will perform better. Towards this end, we trained a Time Delay Neural Network De-noising autoencoder (TDNN-DAE) using non-overlapping speech close-talking microphone utterances (targets) and their corresponding beamform utterances. However, the proposed system could not outperform the baseline.

1. Background

CHiME-5 database [1] includes conversational speech collected using close-talking and distant multi-microphone in everyday home environments. The baseline automatic speech recognition (ASR) is trained using around 149k utterances from the close-talking microphone (also called as binaural or worn microphone. Henceforth called worn microphone for the sake of brevity) and a random set of 100k utterances from the distant arrays. For our experiments, we have used the Gaussian Mixture Model (GMM) ASR baseline. This is a standard triphone based acoustic model (tri3) with linear discriminant analysis (LDA), maximum likelihood linear transformation (MLLT), and feature space maximum likelihood linear regression (fMLLR) with speaker adaptive training (SAT). [1]

The Word Error Rate (WER) for the GMM baseline for worn microphone development set was found to be 71.62%. An initial experiment by training the ASR by using only worn microphone utterances and testing using only worn microphone utterances showed that the system performance improved to 67.15%. This performance improvement can be attributed to reduction in acoustic mismatch conditions between the worn and array microphones. To this end, we propose that an acoustic model trained by using worn microphone utterances will perform better if the test data is acoustically similar to worn microphone data. Our submission to the 5^{th} CHiME Challenge's Single Device Track, Ranking A (constrained LM) uses a TDNN-DAE similar to [2] to enhance the beamformed utterances.

[1]The results for the more resource greedy LF-MMI-TDNN baseline could not be included in the paper as they were not ready at the time of submission

2. Contributions

We have trained a beamform to worn utterance TDNN-DAE [2] using Kaldi Toolkit [3]. (Please refer to Figure 1)

Figure 1: *Training the TDNN-DAE*

3. Experimental evaluation

We have trained a four hidden layer TDNN-DAE with layer-wise contexts organized as [-2,2] [-1,2] [-3,3] [-7,2] $\{0\}$ and input temporal context of [-13,9]. This configuration is similar to TDNN proposed in [2]. The TDNN-DAE is trained using 100k beamformed segments and the targets are their corresponding worn utterances. However, as the data is a truly conversational speech in a dinner party scenario, there is lot of overlapping speech. Overlapping speech means that more than one speaker speaks at a time. We do not expect the proposed front-end enhancement to do speaker separation. Hence we train the TDNN-DAE using non-overlapping speech. The next section describes the data preparation.

3.1. Data preparation

In the first step, we identify non-overlapping utterances and then in the next step find their corresponding beamformed utterances.

3.1.1. Step 1 :Identify non-overlapping utterances

In the first stage, we obtain worn utterances that are non-overlapping i.e no two speakers speak at the same time. This is done by splitting all the worn microphones training utterances (*train_worn*) into different sessions. Each session is then the sorted in ascending order of time. Later, we use a simple algorithm, to decide whether a segment is non-overlapping. Suppose we have N utterances indexed as $i = 1, 2, 3, ...N$. An utterance $x(i)$ is said be non-overlapping if the start time of the next utterance $x(i+1)$ does not lie in the range of start and end times of utterance $x(i)$. This method can be easily explained using the following flowchart:

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

↓

Step 2 :
Sort each session in ascending order of time

↓

Step 3 :
Find if the segment is non-overlapping.
If N utterances are indexed as $i = 1, 2, 3, ...N$, 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* $(x(i))$

Flowchart 1: *Obtaining non-overlapping worn utterances*

3.1.2. Step 2: Obtain worn to beamform mappings

The second stage of data preparation is to find beamformed utterances corresponding to obtained non-overlapping worn utterances. These can be easily obtained using timings and utterance transcriptions. Thus, we get mappings between beamform and worn segments.

3.2. Training TDNN-DAE

A random set of 100k [2] 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).

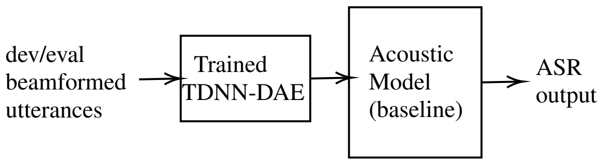

Figure 2: *Block diagram of System 1 (Using CHiME5 baseline Acoustic Model) with TDNN-DAE*

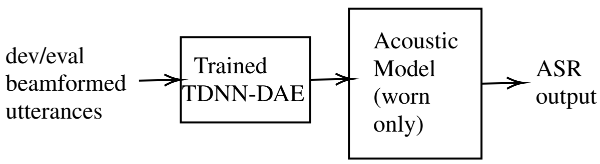

Figure 3: *Block diagram of System 2 (using only worn utterances for Acoustic Model) with TDNN-DAE*

[2]We did not observe any significant improvement by using more data

4. Results

The overall WER(%) for both the systems without using TDNN-DAE is shown in Table 1. We observe that the enhanced features do not perform well when using worn only AM for training (System 2) as compared to baseline AM (System 1). This maybe because System 2's AM is trained using very less data, only 149k utterances, which is 100k less utterances than System 1. We tried to re-train the AM by passing 100k utterances of array train data, but we did not observe any improvement.

The overall WER(%) for both the systems using TDNN-DAE is shown in Table 2. We expect that the enhanced utterances are more matched to the worn only AM and the results are in sync. System 2 performs better than System 1 by 1.54% absolute WER. Table 3 gives the results for the proposed System 2 with TDNN-DAE per session and location.
Tables 1-3 are the results obtained after scoring the ASR hypothesis locally. Table 4 shows the official results given by the organisers. The mismatch in entries of Table 3 and 4 are due to a file mixup at the time of result submission.

Table 1: *Overall WER (%) for the systems tested on the development test set without using TDNN-DAE*

Track	System	WER
Single	System 1	90.82
	System 2	92.31

Table 2: *Overall WER (%) for the systems tested on the development test set using TDNN-DAE*

Track	System	WER
Single	System 1	95.52
	System 2	93.98

Table 3: *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.*

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

Table 4: *Official 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.*

Track	Session		Kitchen	Dining	Living	Overall
Single	Dev	S02	99.10	96.23	94.35	95.52
		S09	95.02	94.77	92.38	
	Eval	S01	100.90	91.76	139.25	104.67
		S21	100.38	97.55	105.22	

5. Conclusion and ongoing work

A performance improvement is observed when using TDNN-DAE enhanced features with worn only AM. However, the results do not look very promising. One of the key things we did not take into consideration is the inherent reverberation in the array microphone utterances. Our ongoing experiments aim at evaluating the effectiveness of the proposed method after performing dereverberation as a pre-processing step.

6. References

[1] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth CHiME Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, Hyderabad, India, Sep. 2018.

[2] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[3] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*, no. EPFL-CONF-192584. IEEE Signal Processing Society, 2011.