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

Sonal Joshi\textsuperscript{1}, Ashish Panda\textsuperscript{1}, Meet Soni\textsuperscript{1}, Rupayan Chakraborty\textsuperscript{1}, Sunilkumar Kopparapu\textsuperscript{1}, Nikhil Mohanan\textsuperscript{2}, Premanand Nayak\textsuperscript{2}, Rajbabu Velmurugan\textsuperscript{2}, Preeti Rao\textsuperscript{2}

\textsuperscript{1}TCS Innovation Labs, Mumbai
\textsuperscript{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 beamformed utterances. However, the proposed system could not outperform the baseline.

1. Background

CHiME-5 database \cite{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 (trii) with linear discriminant analysis (LDA), maximum likelihood linear transformation (MLLT), and feature space maximum likelihood linear regression (fMLLR) with speaker adaptive training (SAT). \textsuperscript{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 microphones utterances.\textsuperscript{2}

2. Contributions

We have trained a beamform to worn utterance TDNN-DAE \cite{2} using Kaldi Toolkit \cite{3}. (Please refer to Figure 1)

![Figure 1: Training the TDNN-DAE](image)

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 \cite{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 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 \( i = 1, 2, 3, \ldots, N \). 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 \( i = 1, 2, 3, \ldots, N \), then an
utterance \( x(i) \) is non-overlapping if \( \text{start time} (x(i+1)) \)
does not lie in the range \( \text{start time} (x(i)) \) to \( \text{end time} (x(i)) \)

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 utter-
ances. 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).

---

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 utter-
ances 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 Sys-
tem 2 with TDNN-DAE per session and location.

---

We did not observe any significant improvement by using more

---

\(^2\) We did not observe any significant improvement by using more data
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

