Our system focuses on implementing a better front-end for the Automatic Speech Recognition (ASR) system

Single-channel enhancement using non-negative matrix factorization (NMF) followed by multi-channel minimum variance distortionless response (MVDR) beamformer

Alternate model to enhance the MVDR output signal by a novel NMF based enhancement.

Challenge Setup And Baseline

- Distant speech recognition with natural conversational speech [1]:
  - Microsoft Kinects arrays, 4 microphones each, placed at different locations.
  - Session has 6 such arrays, 2 each at locations: living, kitchen and dining.
  - Session has 4 speakers, in the same room at a particular instant wearing a close-talking binaural mic.
- Our results are for the single-array track (Ranking A) and focuses on acoustic robustness.
- We use baseline acoustic model (AM) and language model (LM)

Baseline enhancement system

- Single channel noise filtering using Weiner Filtering
- Source localization by GCC-PHAT followed by Viterbi algorithm.
- Delay Sum Beamformer (DSB)

Proposed System

- MVDR + NMF:
  - GCC-PHAT computes TDOA's.
  - Minimum Variance Distortions Response Beamforming (MVDR)
    - For removal of directional noise
    - Covariance matrix computed using noisy frames located using VAD
  - Non-negative Matrix Factorization (NMF) [3] used to enhance MVDR output.

  - Drawback:
    - No improvement in terms of ASR.

  - Possible reason: noisy TDOA's fed as steering vector

  - Modified system: enhance each channel using NMF filtering followed by MVDR beamforming

NMF + MVDR system:

- Input array signals were using NMF and fed to MVDR.
- Supervised approach: clean speech and noise bases learnt from the degraded data

Resolved beamformit+NMF, Beamformit+RNMF

Results and Analysis

- Training using the baseline AM, a mixture of both close-talking microphones and array channels.

- Total of 100k (61349 close talking and 38651 array) utterances of this mixture.

- Magnitude spectrogram obtained using a 64ms Hamming window with a 32ms hop.

- TDOA estimates obtained from NMF filtered channel Beamformit used compute steering vector for MVDR

- Enhanced: distance used for ASR.

Single Channel NMF

- Noise bases learning
  - Clean speech bases learned using unsupervised approach

- MVDR output used for feature extraction and decoded by ASR system.

- Degraded (reverb and noisy) speech spectrogram: \( T = \mathbf{W} \mathbf{X} + \mathbf{Z} \)

- Reverb spectrogram \( V = \mathbf{W} \mathbf{x} \), Noise spectrogram \( Z = \mathbf{W} \mathbf{x} \).

- Reverb bases and activations related to corresponding clean bases and activations

<table>
<thead>
<tr>
<th>Track</th>
<th>System</th>
<th>Session</th>
<th>Kitchens</th>
<th>Using</th>
<th>Dining</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single Microphone Array</td>
<td>NMF</td>
<td>$93.91$</td>
<td>$95.06$</td>
<td>$94.77$</td>
<td>$94.43$</td>
<td>$94.56$</td>
</tr>
<tr>
<td>Single Microphone Array</td>
<td>NMF+MVDR</td>
<td>$93.41$</td>
<td>$95.06$</td>
<td>$94.77$</td>
<td>$94.43$</td>
<td>$94.56$</td>
</tr>
</tbody>
</table>

- Enhancements done on GMH-HMM acoustic models:
  - Beamformit: Baseline enhancement by DSB beamforming
  - Beamformit+NMF: Beamformit followed by NMF de-noising for noise suppression
  - Beamformit+RNMF: Noise suppression using MVDR beamformit followed by TDOA's computed via GCC-PHAT
  - NMF+MVDR+RNMF: noise suppression using MVDR beamformit followed by NMF

Acknowledgements

Part of the work supported by IIT Centre for Communication in IIT Bombay, Council of Scientific and Industrial Research (CSIR), India and Tata Consultancy Services (TCS), India

References

