Listening with Googlears: Low-latency neural multiframe beamforming and equalization for hearing aids

Samuel J. Yang, Scott Wisdom, Chet Gnegy, Richard F. Lyon, Sagar Savla

System E003

Google Research
E003 (preliminary listening data)

Intelligibility of speech in noise, systems in ranked by performance

<table>
<thead>
<tr>
<th>Entrant</th>
<th>Beamforming</th>
<th>DNN Noise Removal</th>
<th>Hearing Loss Compensation</th>
</tr>
</thead>
<tbody>
<tr>
<td>E003</td>
<td>RLS</td>
<td>Conv-TasNet</td>
<td>Linear, fitting formula</td>
</tr>
<tr>
<td>E021</td>
<td>Weighted LCMP</td>
<td>DNN (Deep MFMBVDR)</td>
<td>MBDRC</td>
</tr>
<tr>
<td>E019</td>
<td>Weighted LCMP</td>
<td></td>
<td></td>
</tr>
<tr>
<td>E009</td>
<td>MVDR</td>
<td>MC Conv-TasNet</td>
<td>Linear, NN optimised</td>
</tr>
<tr>
<td>E013</td>
<td>MVDR</td>
<td></td>
<td>Linear, fitting formula but AGC</td>
</tr>
<tr>
<td>E007</td>
<td>MVDR</td>
<td>Conv-TasNet</td>
<td>Linear, NN optimised</td>
</tr>
<tr>
<td>E010</td>
<td>U-Net CNN</td>
<td></td>
<td>Linear, fitting formula</td>
</tr>
<tr>
<td>E018</td>
<td>2D CNN + LSTM, WPE</td>
<td></td>
<td>Dynamic EQ</td>
</tr>
<tr>
<td>E001</td>
<td></td>
<td></td>
<td>Baseline</td>
</tr>
<tr>
<td>E005</td>
<td></td>
<td></td>
<td>Binaural Conv-Tasnet</td>
</tr>
</tbody>
</table>

Ranked by Noise

MBSTOI

Google Research
Agenda

01 Motivation
02 System description
03 Audio demos
04 MBSTOI results
05 Listening test results
Understanding speech in noise is hard (previous study with cochlear implants)

- In a small study, our application of speech enhancement helped cochlear implant (CI, a close relative of hearing aids) users' speech understanding.

<table>
<thead>
<tr>
<th>CI hackathon</th>
<th>Clarity Challenge</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-mic input</td>
<td>6-mic input</td>
</tr>
<tr>
<td>speech babble</td>
<td>single speech or noise interferer</td>
</tr>
<tr>
<td>simulated CI audio</td>
<td>audiogram-adjusted audio</td>
</tr>
<tr>
<td>speech enhancement</td>
<td>enhance + beamform</td>
</tr>
</tbody>
</table>

*Four cochlear implant users* vs *typical hearing*
Our solution: overview

1) Separate single-microphone audio from left and right into target and interference signals.
Our solution: overview

- 1) Separate single-microphone audio from left and right into target and interference signals.
- 2) Use estimate of target signal to beamform across all 6 mics with 4 context frames.
Our solution: overview

- 1) Separate single-microphone audio from left and right into target and interference signals.
- 2) Use estimate of target signal to beamform across all 6 mics with 4 context frames.
- 3) Apply linear equalizer using listener audiogram to compensate for hearing loss.
Single-mic enhancement

- Causal Conv-TasNet masking network [1] predicts a mask for input STFT.
- Trained on synthetic mixtures of target speech and interferer using TPU (next slide).

---

Training for enhancement

- Augmentation on single-microphone audio from Clarity Challenge scenes.
- Leverages cue that target starts after two seconds.
Training for enhancement


Training for enhancement

- Trained with consistent multi-resolution compressed STFT loss on target and interferer.
Causal multi-frame RLS beamforming

Causal multi-frame RLS beamforming

- Optimization problem for filter $W$ to predict target $x$ from input $y$:
  \[
  \hat{W}_t = \min_{W_t} \quad L_t(W_t) = \sum_{\tau=0}^{t} \lambda_{t,\tau} \| x_\tau - W_t^T y_\tau \|^2
  \]

  Note that the classic unweighted RLS uses $\lambda_{t,\tau} = \lambda^{t-\tau}$, where $\lambda$ is an exponential averaging weight usually chosen with value between 0.98 and 1.0.

- Non-causal solution:
  \[
  W_t = R_{yy,t}^{-1} R_{xy,t}^T \quad \quad R_{yy,t} = \sum_{\tau=0}^{t} \lambda_{t,\tau} y_\tau y_\tau^T \quad \quad R_{xy,t} = \sum_{\tau=0}^{t} \lambda_{t,\tau} x_\tau y_\tau^T.
  \]

- Canonical causal recursive solution (no matrix inverses!):
  \[
  g_t = P_{t-1} y_t / (\lambda + y_t^T P_{t-1} y_t),
  \]
  \[
  P_t = (P_{t-1} - g_t y_t^T P_{t-1}) / \lambda,
  \]
  \[
  W_t = W_{t-1} + g_t (x_t^T - y_t^T W_{t-1}).
  \]
Linear equalizer

Beamformed STFT → Compute equalizer filter → Filter → Apply equalizer filter → Equalized STFT

Listener audiogram:
- Hearing loss (dB)
- Frequency (kHz)
Audio demos

Description: male voice target with female voice interferer (Scene S07458)

Baseline
Enhancement output before beamformer

Our submission (enhancement + beamformer + linear equalizer)

Target speech onset
Attenuated but non-zero interferer

Description: male voice target with noise interferer (Scene S08143)

2.0s silence

2.7s silence
Audio demos

Noise interferer example:  
(i.e. hairdryer, dishwasher, kettle, fan)

Speech interferer example:  
(i.e. another male or female voice**)

**interferer begins speaking immediately; the target starts speaking after 2 seconds

baseline  our submission

Description: male voice target with noise interferer

baseline  our submission

Description: male voice target with female voice interferer
MBSTOI results

Dev baseline: 0.41 mean, 0.41 median
Dev proposed: 0.632 mean, 0.642 median

Eval baseline: 0.310 mean, 0.314 median
Eval proposed: 0.644 mean, 0.6652 median

[Graphs showing MBSTOI improvement against SNR for noise and speech interferers]
Listening test results (preliminary)

- For noise interferers, +~40% boost in correctness.
  - Direction of improvement consistent with MBSTOI.
- For speech interferers, highly mixed results.
  - Next slide: investigate 2 listeners responses (p218, p219).
Lister p219 had 0% correct and had no response to highest SNR examples - possibly only heard one speaker.

Listener p218 seems to randomly alternate between correct and incorrect - possibly confusing which speaker is target.

? denotes examples where listener transcript differs significantly from actual target.
Listening test results (preliminary)

- Methodology: for each utterance, I reviewed the transcript and ground truth and made binary decision of correct or incorrect.
- 8 listeners had total scores near zero
  - 4 gave no responses for the highest SNR utterances, suggesting they were listening for the interferer and got confused when they only heard one speaker
  - 2 consistently incorrect, except for one (mid level SNR) utterance where they got it correct.
  - 2 consistently got incorrect for all examples, but appeared confident in noting many words in each utterance
- 7 listeners had non-zero total scores
  - 2 seem to alternate between incorrect and correct utterance transcripts (see p218 and p231)
  - 5 listeners appear to have completely valid responses
- Conclusion: 5 of 15 listeners appear to have completely valid responses.
Future work

- Ablations
  - Training augmentation
  - Enhancement-only

- Explore if allowing some noise in the first 2 seconds helps avoid target/interferer confusion; more generally, explore if allowing some noise allows listeners to adapt and actually enhance intelligibility.

- Real-world target identification methods (not relying on first 2 seconds of interferer)
  - Visual
  - Spatial (e.g. direction)
  - Speaker ID

- Should target/interferer speakers be from the same dataset?
Thank You
Samuel J. Yang and Scott Wisdom
Research Scientists