1. Problem

Problem: Independent conversations can leak into microphones
Challenge: Remove non-relevant speakers without prior knowledge
Assumption: The microphone of a person is closer than others
Idea: Consider the microphone closest to a speaker as 'clean' signal

2. Signal Model

\[ x_{\text{mix}}(n) = \sum_{c=1}^{N_s} x_c(n) + s(n) = x_c(n) * h_c(n) \]

- \( s_c(n) \) c-th speech source
- \( x_c(n) \) c - th clean audio signal
- \( h_c(n) \) the impulse response

3. Database

- LibriSpeech Dev-Clean with \( f_s = 16 \text{ kHz} \)
- Simulate office scenarios with PyRoom Acoustics package
- Width [5 m to 10 m], length [5 m to 10 m], height [2.5 m to 5 m]

4. IsoNet Architecture

- Encoder, mask-generator, enhancer, and decoder
- Input: Two audio signals
- Output: Single channel enhanced audio
- Loss function:

\[ \text{SI-SNR} := 10 \cdot \log_{10} \frac{||s_{\text{target}}||^2}{||e_{\text{noise}}||^2} \]

\( s_{\text{target}} := (\hat{s})^T s \) is the reference and \( \hat{s} \) is the estimated isolated signal.
\( e_{\text{noise}} := \hat{s} - s_{\text{target}} \) and \( || \cdot ||^2 \) is the power of the signal.

5. Results

<table>
<thead>
<tr>
<th>Method</th>
<th>PESQ</th>
<th>SI-SNR</th>
<th>SDR</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>WaveFilter</td>
<td>2.412</td>
<td>-10.45</td>
<td>5.9M</td>
<td></td>
</tr>
<tr>
<td>PercepNet</td>
<td>2.412</td>
<td>-10.45</td>
<td>8.5/26.5 M</td>
<td></td>
</tr>
<tr>
<td>VoiceFilter</td>
<td>12.6</td>
<td>-18.8 M</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conv-TasNet</td>
<td>12.2</td>
<td>-12.7 M</td>
<td>5.1 M</td>
<td></td>
</tr>
<tr>
<td>SepFormer</td>
<td>16.5</td>
<td>17.9</td>
<td>26 M</td>
<td></td>
</tr>
<tr>
<td>TDANet</td>
<td>17.4</td>
<td>17.9</td>
<td>2.3 M</td>
<td></td>
</tr>
<tr>
<td>IsoNet(ours)</td>
<td>3.7</td>
<td>18.6</td>
<td>14.1</td>
<td>3.7 M</td>
</tr>
</tbody>
</table>

* describes baseline models which are trained on the LibriMix Dataset. The reported values are the original ones from the paper, as there was no public model available for retraining.

6. Listening Tests

- Mushra Listening Test with \( N = 26 \) listeners.
- Improvement from original to IsoNet isolated signal is 47.72 Mushra points

7. Conclusions

- Speech enhancement can contribute to privacy improvements
- Voice isolation can easily be done with simple assumptions without prior information
- Iso-Net is a small, real-time capable network with 3.7M parameters

- Mutual information (MI) as indicator for residual of leaked speaker

<table>
<thead>
<tr>
<th>MI [bits]</th>
<th>Variance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before Iso-Net 3.52</td>
<td>0.177</td>
</tr>
<tr>
<td>After Iso-Net 1.39</td>
<td>0.151</td>
</tr>
</tbody>
</table>