Towards Solving Cocktail Party Problem with Artificial Intelligence

Dr. Chenglin Xu
Kuaishou Technology, Shenzhen, China
xuchenglin03@kuaishou.com/xuchenglin28@gmail.com
Outline

- **Introduction**
- **Source Separation**
  - Frequency domain
  - Time domain
- **Speaker Extraction**
  - Audio assisted speaker extraction
  - Visual assisted speaker extraction
  - Heterogeneous speaker extraction
- **Down-stream Tasks**
  - Multi-talker speaker verification
  - Multi-talker speech recognition
- **Challenges**
Introduction

- **Cocktail Party Problem**
  - “One of our most important faculties is our ability to listen to, and follow, one speaker in the presence of others. This is such a common experience that we may take it for granted; we may call it ‘the cocktail party problem’. No machine has been constructed to do just that.” [1,2]


Picture from https://psychologenie.com/overview-of-cocktail-party-effect-in-psychology
Introduction

- Cocktail Party Problem
Introduction

- **Speech enhancement**
  - Improve the intelligibility or overall perceptual quality of degraded speech signal by removing noise and keeping the voices of human.

- **Speech deverberation**
  - Remove the reverberation of speech reflected by surfaces.

- **Echo cancellation**
  - Improve voice quality by preventing echo from being created or removing it after it is already present.

……
Introduction

- Speech separation
  - Aim to let machines have the same auditory ability as a human in a cocktail party environment.
  - Recover all original component signals from the combined signal, which is a mixture of several source signals and recorded by a single- or multi-channel microphone.

<table>
<thead>
<tr>
<th>Mixture Type</th>
<th>Speaker</th>
<th>Mixed</th>
<th>Separated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male-Female</td>
<td>Trump</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Hillary</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Video from [http://www.youtube.com/watch?v=SLWgV9tnxKQ&t=6m0s](http://www.youtube.com/watch?v=SLWgV9tnxKQ&t=6m0s)
Introduction

- What is speech separation?
  - Multi-channel speech separation
  - Single-channel speech separation
Introduction

Why do we need speech separation?

- Real world speech applications
  - Human machine interaction.
  - Hearing aids.
  - Teleconferencing.
  - Smart devices.
  - Robotics.
  - ……
Introduction

How does speech separation work?

- Auditory Model (CASA)
  - pitch continuity [3]
  - vocal tract continuity [3]
  - temporal continuity [4]
  - harmonic [5]
  - common onset/offset [6]

- Decomposition Model
  - ICA [7]
  - NMF [8]
  - sparse NMF [9]
  - convolutive NMF [10]
  - discriminative NMF [11]

- Generative Model
  - GMM-HMM [12]

- Deep Learning Model
  - conventional regression [13]
  - conventional time-frequency masking [14]
  - deep clustering [15, 16]
  - deep attractor network [17]
  - permutation invariant training [18, 19]
  - Griffin-Lim algorithm [20]
  - multiple input spectrogram inverse [21]
  - Conv-TasNet [22]

Introduction

How does speech separation work?
- State-of-the-art performance on WSJ0-2mix\(^1\) database

---

**SDR improvement (dB) on WSJ0-2mix database**

<table>
<thead>
<tr>
<th>Method</th>
<th>SDR Improvements (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DC (2016) [15]</td>
<td>5.8</td>
</tr>
<tr>
<td>DC+ (2016) [16]</td>
<td>10.3</td>
</tr>
<tr>
<td>DANN (2017) [17]</td>
<td>10.8</td>
</tr>
<tr>
<td>uPT-CNN (2017) [18]</td>
<td>10.5</td>
</tr>
<tr>
<td>uPT BLSTM (2017) [19]</td>
<td>5.2</td>
</tr>
<tr>
<td>uPT BLSTM-ST (2017) [19]</td>
<td>7.8</td>
</tr>
<tr>
<td>Contribution 1</td>
<td>9.4</td>
</tr>
<tr>
<td>Contribution 2.1</td>
<td>10.0</td>
</tr>
<tr>
<td>Contribution 2.2</td>
<td>10.5</td>
</tr>
<tr>
<td>WA-MFIS (2018) [21]</td>
<td>11.0</td>
</tr>
<tr>
<td>Conv-TabNet (2019) [22]</td>
<td>12.6</td>
</tr>
<tr>
<td>Contribution 2.2</td>
<td>15.3</td>
</tr>
<tr>
<td>Contribution 2.2</td>
<td>17.0</td>
</tr>
</tbody>
</table>

\(^1\) Available at: [http://www.merl.com/demos/deep-clustering](http://www.merl.com/demos/deep-clustering)
Outline

- Introduction

- Source Separation
  - Frequency domain
  - Time domain

- Speaker Extraction
  - Audio assisted speaker extraction
  - Visual assisted speaker extraction
  - Heterogeneous speaker extraction

- Down-stream Tasks
  - Multi-talker speaker verification
  - Multi-talker speech recognition

- Challenges
Speech Separation

- Auditory Masking
  - With the understanding of auditory mask, we can retain parts of a target sound that are stronger than the acoustic background, and discard the rest.

Simultaneous Masking

Temporal Masking
Speech Separation

- Auditory Masking
  - How to find the mask?
  - Spectro-temporal receptive fields that reflect the temporal and spectral modulations belonging to the target sound events [3].

Speech Separation

- Auditory Masking
  - Example (not the best): ideal binary mask (IBM)
Speech Separation

Auditory Masking

- Example (not the best): ideal ratio mask (IRM)
Speech Separation

- Frequency-domain Speech Separation
  - Direct mask estimation with deep neural network
Speech Separation

- Frequency-domain Speech Separation
  - DC [9], DC++ [10]

  **time-frequency bins:** $n = (t, f)$
  **input spectrogram:** $X = (x_n)$

  **one-hot speaker labels:** $Y = (y_{n,c})$

  **ideal affinity matrix:** $A = YY^T$

  **network embeddings:** $V = (v_{n,d})$
  $\mathcal{F}_\theta(X)$: a network
  $d$: embedding dimension

  **Unit length constraint:**
  $v_i = \frac{v_i}{|v_i|}$

  **estimated affinity:** $\hat{A} = VV^T$

  **Objective function:** $C(\theta, Y) = ||\hat{A} - A||_F^2 = ||VV^T - YY^T||_F^2 = ||VV^T||_F^2 + ||YY^T||_F^2 - 2||V^TY||_F^2$

Minimize the error during training, then k-means clustering on the learned embeddings during inference.

**DC++**: extend DC by incorporating an enhancement network after mask estimation with soft clustering, and performing end-to-end training of embedding and clustering.


Speech Separation

- Frequency-domain Speech Separation
  - DANet [11]

Speech Separation

- Frequency-domain Speech Separation
  - PIT-DNN, PIT-CNN [15]

---

Speech Separation

- Frequency-domain Speech Separation
  - uPIT-BLSTM [12]

Speech Separation

- Frequency-domain Speech Separation
  - WA-MISI-5 \[^{[13]}\]

---

**Algorithm 2: MISI algorithm**

**Input:** mixture signal $y$, estimated magnitudes $\hat{S}_c$ for $c = 1, ..., C$, and the number of iterations $K$

**Output:** Reconstructed signal $\hat{s}_c^{(K)}$ for $c = 1, ..., C$

**Initialize:**

$\phi_c^{(0)} = \angle \text{STFT}(y)$,

$\hat{s}_c^{(0)} = i\text{STFT}(\hat{S}_c, \phi_c^{(0)})$, for $c = 1, ..., C$

**for** $i = 1, ..., K$ **do**

$e^{(i-1)} = y - \sum_{c=1}^{C} \hat{s}_c^{(i-1)}$

$\phi_c^{(i)} = \angle \text{STFT}(\hat{s}_c^{(i-1)} + \frac{e^{(i-1)}}{C})$

$\hat{s}_c^{(i)} = i\text{STFT}(\hat{S}_c, \phi_c^{(i)})$, for $c = 1, ..., C$

**end**

---

Speech Separation

- The SDC-G-MTL framework (as an example)
Speech Separation

- Temporal objective function

\[ J_p = \frac{1}{T} \sum_{c=1}^{C} \| \mathcal{F}_{sdc}(M_c \otimes |Y|) - \mathcal{F}_{sdc} \left( |S_{\phi_p(c)}| \otimes \cos(\theta_y - \theta_{\phi_p(c)}) \right) \|_{\hat{p}}^2 \]

\[ \mathcal{F}_{sdc}(t) = [\mathcal{F}_d(t), \mathcal{F}_d(t + l), ..., \mathcal{F}_d(t + (K-1)\times l)] \]

\[ \mathcal{F}_d(t) = \frac{\sum_{i=1}^{L} l \times (v(t + l) - v(t - l))}{\sum_{i=1}^{L} 2 \times l^2} \]

Minimizing:

\[ \hat{p} = \min_{p \in P} J_p \]

\[ J_{sdc} = J_{\hat{p}} \]
Speech Separation

- Spectro-temporal features by a grid LSTM

\[
i^{(j)}_{t,k} = \sigma(W^{(j)}_{iz} y_{t,k} + W^{(t)}_{ih} h^{(t)}_{t-1,k} + W^{(k)}_{ih} h^{(k)}_{t,k-1} + W^{(t)}_{ic} c^{(t)}_{t-1,k} + W^{(k)}_{ic} c^{(k)}_{t,k-1} + b^{(j)}_i)
\]
\[
f^{(j)}_{t,k} = \sigma(W^{(j)}_{fx} y_{t,k} + W^{(t)}_{fh} h^{(t)}_{t-1,k} + W^{(k)}_{fh} h^{(k)}_{t,k-1} + W^{(t)}_{fc} c^{(t)}_{t-1,k} + W^{(k)}_{fc} c^{(k)}_{t,k-1} + b^{(j)}_f)
\]
\[
c^{(t)}_{t,k} = f^{(t)}_{t,k} \odot c^{(t)}_{t-1,k} + i^{(t)}_{t,k} \odot g(W^{(t)}_{cx} y_{t,k} + W^{(t)}_{ch} h^{(t)}_{t-1,k} + W^{(k)}_{ch} h^{(k)}_{t,k-1} + b^{(t)}_c)
\]
\[
c^{(k)}_{t,k} = f^{(k)}_{t,k} \odot c^{(k)}_{t-1,k} + i^{(k)}_{t,k} \odot g(W^{(k)}_{cx} y_{t,k} + W^{(k)}_{ch} h^{(t)}_{t-1,k} + W^{(t)}_{ch} h^{(k)}_{t,k-1} + b^{(k)}_c)
\]
\[
o^{(j)}_{t,k} = \sigma(W^{(j)}_{ox} y_{t,k} + W^{(t)}_{oh} h^{(t)}_{t-1,k} + W^{(k)}_{oh} h^{(k)}_{t,k-1} + W^{(t)}_{oc} c^{(t)}_{t,k} + W^{(k)}_{oc} c^{(k)}_{t,k} + b^{(j)}_o)
\]
\[
h^{(j)}_{t,k} = o^{(j)}_{t,k} \odot \sigma(c^{(j)}_{t,k})
\]
\[
h_t = [h^{(k)}_{t,1}, ..., h^{(k)}_{t,B}, h^{(t)}_{t,1}, ..., h^{(t)}_{t,B}]
\]

Pictures from the paper:
Speech Separation

- Multi-task learning
  - Loss for the subtask

\[ J_{ce} = -\frac{1}{T} \sum_{t=1}^{T} g_t^H \times \log \hat{g}_t \]

- Total loss

\[ J_{mtl} = (1 - \lambda) \times J_{sdc} + \lambda \times J_{ce} \]
Speech Separation

Dataset

- The WSJ0-2mix \(^1\) database with the sampling rate at 8 kHz.
  - Training set: 20,000 utterances \(\approx 30\)h.
  - Development set: 5,000 utterances \(\approx 8\)h.
  - Test set: 3,000 utterances \(\approx 5\)h.

- Features
  - 129-dimensional spectral magnitude features computed by a STFT with a normalized square root of 32ms length hamming window and 16ms widow shift.

\(^1\) Available at: http://www.merl.com/demos/deep-clustering
Speech Separation

Evaluation Metrics

- The global normalized signal-to-distortion ratio (GNSDR, same as SDR improvement) with the toolbox [15].

\[
\hat{s} = s_t + e_i + e_n + e_a \\
SDR = 10 \log_{10} \frac{\|s_t\|^2}{\|e_i + e_n + e_a\|^2} \\
GNSDR = \frac{1}{N} \sum (SDR_{sep} - SDR_{mix})
\]

- Signal-to-interferences ratio (SIR).

\[
SIR = 10 \log_{10} \frac{\|s_t\|^2}{\|e_i\|^2}
\]

- Signal-to-artifact ratio (SAR).

\[
SAR = 10 \log_{10} \frac{\|s_t + e_i + e_n\|^2}{\|e_a\|^2}
\]

- A/B preference test.

Speech Separation

Evaluation:

- The performance of speech separation with temporal objective function is better than that of static feature based objective function ($p<0.05$).
- Spectro-temporal features by a grid LSTM could improve the performance ($p<0.05$).
- With the subtask to predict an attribute for each time-frequency bin, the performances of all multi-task learning systems are better than those without multi-task learning ($p<0.05$).

<table>
<thead>
<tr>
<th>Method</th>
<th>MTL</th>
<th>GNSDR</th>
<th>SIR</th>
<th>SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>uPIT-BLSTM</td>
<td>No</td>
<td>9.5±0.1</td>
<td>16.5±0.1</td>
<td>11.2±0.1</td>
</tr>
<tr>
<td>SDC</td>
<td>No</td>
<td>9.8±0.1</td>
<td>17.1±0.1</td>
<td>11.3±0.1</td>
</tr>
<tr>
<td>SDC-G</td>
<td>No</td>
<td>10.1±0.1</td>
<td>17.4±0.1</td>
<td>11.6±0.1</td>
</tr>
<tr>
<td>uPIT-BLSTM-MTL</td>
<td>Yes</td>
<td>10.0±0.1</td>
<td>17.3±0.1</td>
<td>11.5±0.1</td>
</tr>
<tr>
<td>SDC-MTL</td>
<td>Yes</td>
<td>10.1±0.1</td>
<td>17.6±0.1</td>
<td>11.6±0.1</td>
</tr>
<tr>
<td>SDC-G-MTL</td>
<td>Yes</td>
<td><strong>10.5±0.1</strong></td>
<td><strong>18.0±0.1</strong></td>
<td><strong>11.9±0.1</strong></td>
</tr>
</tbody>
</table>
Speech Separation

- Evaluation: same vs. different gender
  - The separation of same gender mixture is even harder than that of different gender, because speakers of same gender may have similar speech characteristics (i.e. close pitch).

<table>
<thead>
<tr>
<th>Method</th>
<th>MTL</th>
<th>GNSDR</th>
<th>SIR</th>
<th>SAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Same</td>
<td>Diff.</td>
<td>Same</td>
</tr>
<tr>
<td>uPIT-BLSTM</td>
<td>No</td>
<td>7.3</td>
<td>11.5</td>
<td>13.6</td>
</tr>
<tr>
<td>SDC</td>
<td>No</td>
<td>7.6</td>
<td>11.7</td>
<td>14.2</td>
</tr>
<tr>
<td>SDC-G</td>
<td>No</td>
<td>8.1</td>
<td>11.8</td>
<td>14.8</td>
</tr>
<tr>
<td>uPIT-BLSTM-MTL</td>
<td>Yes</td>
<td>7.7</td>
<td>11.9</td>
<td>14.3</td>
</tr>
<tr>
<td>SDC-MTL</td>
<td>Yes</td>
<td>8.0</td>
<td>12.0</td>
<td>14.3</td>
</tr>
<tr>
<td>SDC-G-MTL</td>
<td>Yes</td>
<td><strong>8.6</strong></td>
<td><strong>12.2</strong></td>
<td><strong>15.5</strong></td>
</tr>
</tbody>
</table>
Speech Separation

Evaluation: frame leakage error rate
- Insertion error, as shown in ellipse in (d) and (e), yellow in (h) and (i).
- Deletion error, as shown in rectangle in (e), blue in (h) and (i).
- Substitution error, as shown in vertical line in (d) and (e), green in (h) and (i).

<table>
<thead>
<tr>
<th>Method</th>
<th>FLER (%)</th>
<th>Pitch Continuity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VUV</td>
<td>CORR</td>
</tr>
<tr>
<td>uPIT-BLSTM</td>
<td>11.44</td>
<td>7.69</td>
</tr>
<tr>
<td>SDC</td>
<td>10.53</td>
<td>7.30</td>
</tr>
<tr>
<td>SDC-G</td>
<td>9.68</td>
<td>6.90</td>
</tr>
<tr>
<td>SDC-G-MTL</td>
<td><strong>8.47</strong></td>
<td><strong>6.44</strong></td>
</tr>
</tbody>
</table>

Example: female-female speakers’ mixture (‘050a050i_2.1935_421c020b_-2.1935’)
Speech Separation

- **Subjective evaluation: uPIT-BLSTM vs. SDC-G-MTL**
  - Randomly selected 20 pairs of listening examples.
  - A group of 10 subjects are asked to answer their preference according to the quality and intelligibility.
  - The proposed SDC-G-MTL system is preferred more than the uPIT-BLSTM baseline.

![Preference Score Chart]

- uPIT-BLSTM: 11.0%
- SDC-G-MTL: 54.0%
- Neutral: 35.0%
### Speech Separation

**Evaluation: compare with other state-of-the-art methods**

<table>
<thead>
<tr>
<th>Method</th>
<th>Opt Assign</th>
<th>Def Assign</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CC</td>
<td>OC</td>
</tr>
<tr>
<td><strong>Baselines</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DC</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DC+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DC++(Enh.)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>DANet</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PIT-DNN</td>
<td>7.3</td>
<td>7.2</td>
</tr>
<tr>
<td>PIT-CNN</td>
<td>8.4</td>
<td>8.6</td>
</tr>
<tr>
<td>uPIT-BLSTM</td>
<td>10.9</td>
<td>10.8</td>
</tr>
<tr>
<td>uPIT-BLSTM-ST</td>
<td>11.7</td>
<td>11.7</td>
</tr>
<tr>
<td><strong>Ours</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>uPIT-BLSTM*</td>
<td>10.8</td>
<td>10.7</td>
</tr>
<tr>
<td>SDC-G-MTL</td>
<td>11.4</td>
<td>11.4</td>
</tr>
<tr>
<td><strong>Upper-bounds</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IRM</td>
<td>12.4</td>
<td>12.7</td>
</tr>
<tr>
<td>IPSM</td>
<td>14.9</td>
<td>15.1</td>
</tr>
</tbody>
</table>
Speech Separation

- Evaluation: demo
  - female-female speakers’ mixture (‘050a050i_2.1935_421c020b_-2.1935’)

![Image of speech separation results]
Speech Separation

- Time-domain Speech Separation
Speech Separation

- Time-domain Speech Separation
  - Conv-TasNet [14]

Outline

- Introduction

- Source Separation
  - Frequency domain
  - Time domain

- Speaker Extraction
  - Audio assisted speaker extraction
  - Visual assisted speaker extraction
  - Heterogeneous speaker extraction

- Down-stream Tasks
  - Multi-talker speaker verification
  - Multi-talker speech recognition

- Challenges
Speaker Extraction

Motivation
• Address the unknown number of speakers in the mixture.
• Solve the global speaker permutation ambiguity problem.
• Study the influence of background noise and reverberation.

Novelty
• Mimic human’s top-down selective auditory attention.
Speaker Extraction

- Audio assisted speaker extraction
  - Frequency-domain
Audio assisted speaker extraction
- SBF-IBM [23] vs. SBF-MTSAL-Concat

Speaker Extraction

- Audio assisted speaker extraction
  - Time-domain
Speaker Extraction

- SpEx: Multi-scale time-domain speaker extraction (as an example)
Speaker Extraction

- SpEx: Multi-scale time-domain speaker extraction
  - Mimic top-down selective auditory attention
  - Multi-scale encoding and decoding
  - Multi-task learning
    - Multi-scale scale-invariant signal-to-distortion (SI-SDR) loss
      \[ J_1 = -(1 - \alpha - \beta)\rho(s_1, s) + \alpha \rho(s_2, s) + \beta \rho(s_3, s) \]
    - Cross-entropy loss
      \[ J_2 = -\sum_{i=1}^{N_s} p_i \log \hat{p}_i \]
      \[ J = (1 - \gamma)J_1 + \gamma J_2 \]
Speaker Extraction

Dataset

- WSJ0-2mix-extr ¹, and WSJ0-2mix ²(clean two-speaker mixture condition)
- WSJ0-3mix-extr ¹(clean three-speaker mixture condition)
- WHAM! ³(noisy two-speaker mixture condition)
- WHAMR! ³(reverberant two-speaker mixture condition)
- WHAMR! ³(noisy and reverberant two-speaker mixture condition)
  - Training set: 20,000 mixture utterances (each condition).
  - Development set: 5,000 mixture utterances (each condition).
  - Test set: 3,000 mixture utterances (each condition).

Frequency-domain features

- 129-dimensional spectral magnitude features computed by a STFT with a normalized square root of 32ms length hamming window and 16ms window shift.

¹ Available at: https://github.com/xuchenglin28/speaker_extraction
² Available at: http://www.merl.com/demos/deep-clustering
³ Available at: http://wham.whisper.ai
Speaker Extraction

- Evaluation of speaker extraction on WSJ0-2mix-extr database
  - Benchmark against baselines
    - SpEx achieves best performance comparing with other frequency-domain speaker extraction methods.

<table>
<thead>
<tr>
<th>Method</th>
<th>Domain</th>
<th>#Paras</th>
<th>SDR</th>
<th>SI-SDR</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mixture</td>
<td>-</td>
<td>-</td>
<td>2.6</td>
<td>2.5</td>
<td>2.31</td>
</tr>
<tr>
<td>SBF-IBM</td>
<td>Frequency</td>
<td>19.3M</td>
<td>6.5</td>
<td>6.3</td>
<td>2.32</td>
</tr>
<tr>
<td>SBF-MSAL</td>
<td>Frequency</td>
<td>19.3M</td>
<td>9.6</td>
<td>9.2</td>
<td>2.64</td>
</tr>
<tr>
<td>SBF-MTSAL</td>
<td>Frequency</td>
<td>19.3M</td>
<td>9.9</td>
<td>9.5</td>
<td>2.66</td>
</tr>
<tr>
<td>SBF-MTSAL-Concat</td>
<td>Frequency</td>
<td>8.9M</td>
<td>11.0</td>
<td>10.6</td>
<td>2.73</td>
</tr>
<tr>
<td>SpEx</td>
<td>Time</td>
<td>10.8M</td>
<td><strong>15.1</strong></td>
<td><strong>14.6</strong></td>
<td><strong>3.14</strong></td>
</tr>
</tbody>
</table>
Speaker Extraction

- Evaluation of speaker extraction on WSJ0-2mix-extr database
  - Subjective evaluation
    - SpEx method is preferred more than the SBF-MTSAL-Concat approach.
Speaker Extraction

Evaluation of speaker extraction on WSJ0-2mix-extr database

- Time-domain speaker extraction: duration of reference speech
  - The longer duration of the reference speech always achieves better performance. Because the speaker embedding with better quality may be obtained with longer reference speech.

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>SDR</th>
<th>SI-SDR</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>Random</td>
<td>15.1</td>
<td>14.6</td>
<td>3.14</td>
</tr>
<tr>
<td></td>
<td>7.5s</td>
<td>15.0</td>
<td>14.6</td>
<td>3.14</td>
</tr>
<tr>
<td></td>
<td>15s</td>
<td>15.4</td>
<td>15.0</td>
<td>3.17</td>
</tr>
<tr>
<td></td>
<td>30s</td>
<td>15.5</td>
<td>15.2</td>
<td>3.19</td>
</tr>
<tr>
<td></td>
<td>60s</td>
<td><strong>15.6</strong></td>
<td><strong>15.2</strong></td>
<td><strong>3.19</strong></td>
</tr>
<tr>
<td>15s</td>
<td>15s</td>
<td>14.9</td>
<td>14.5</td>
<td>3.13</td>
</tr>
<tr>
<td></td>
<td>60s</td>
<td>15.2</td>
<td>14.8</td>
<td>3.15</td>
</tr>
</tbody>
</table>
Evaluation of speaker extraction on WSJ0-3mix-extr database
- Time-domain speaker extraction: two-speaker and three-speaker mixture

<table>
<thead>
<tr>
<th>Training</th>
<th>Test</th>
<th>#speakers</th>
<th>Duration</th>
<th>SDR</th>
<th>SI-SDR</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>speakers</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 speakers</td>
<td>2 speakers</td>
<td>15s</td>
<td>15.4</td>
<td>15.0</td>
<td>3.17</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 speakers</td>
<td>15s</td>
<td>5.2</td>
<td>5.0</td>
<td>2.35</td>
<td></td>
</tr>
<tr>
<td>3 speakers</td>
<td>2 speakers</td>
<td>15s</td>
<td>11.5</td>
<td>10.9</td>
<td>2.74</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 speakers</td>
<td>15s</td>
<td>7.9</td>
<td>7.3</td>
<td>2.40</td>
<td></td>
</tr>
<tr>
<td>2&amp;3 speakers</td>
<td>2 speakers</td>
<td>15s</td>
<td>15.0</td>
<td>14.6</td>
<td>3.14</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 speakers</td>
<td>15s</td>
<td>8.9</td>
<td>8.4</td>
<td>2.54</td>
<td></td>
</tr>
<tr>
<td>2 speakers</td>
<td>2 speakers</td>
<td>60s</td>
<td>15.6</td>
<td>15.2</td>
<td>3.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 speakers</td>
<td>60s</td>
<td>5.2</td>
<td>5.0</td>
<td>2.36</td>
<td></td>
</tr>
<tr>
<td>3 speakers</td>
<td>2 speakers</td>
<td>60s</td>
<td>12.1</td>
<td>11.6</td>
<td>2.81</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 speakers</td>
<td>60s</td>
<td>8.3</td>
<td>7.8</td>
<td>2.44</td>
<td></td>
</tr>
<tr>
<td>2&amp;3 speakers</td>
<td>2 speakers</td>
<td>60s</td>
<td>15.5</td>
<td>15.1</td>
<td>3.19</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3 speakers</td>
<td>60s</td>
<td>9.1</td>
<td>8.7</td>
<td>2.57</td>
<td></td>
</tr>
</tbody>
</table>
Speaker Extraction

- Evaluation of speaker extraction on WSJ0-2mix database
  - Time-domain speaker extraction vs. state-of-the-art speech separation (clean two-speaker mixture condition)

<table>
<thead>
<tr>
<th>Task</th>
<th>Methods</th>
<th>Domain</th>
<th>#Paras</th>
<th>SDRi</th>
<th>SI-SDR</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>DC++</td>
<td>Frequency</td>
<td>13.6M</td>
<td>-</td>
<td>10.8</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>uPIT-BLSTM-ST</td>
<td></td>
<td>92.7M</td>
<td>10.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>DANet</td>
<td></td>
<td>9.1M</td>
<td>-</td>
<td>10.5</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SDC-G-MTL</td>
<td></td>
<td>53.9M</td>
<td>10.5</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CBLDNN-GAT</td>
<td></td>
<td>39.5M</td>
<td>11.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Chimera++</td>
<td></td>
<td>32.9M</td>
<td>12.0</td>
<td>11.5</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>WA-MISI-5</td>
<td></td>
<td>32.9M</td>
<td>13.1</td>
<td>12.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BLSTM-TasNet</td>
<td>Time</td>
<td>23.6M</td>
<td>13.6</td>
<td>13.2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Conv-TasNet</td>
<td></td>
<td>8.8M</td>
<td>15.0</td>
<td>14.6</td>
<td>3.25</td>
</tr>
<tr>
<td></td>
<td>Conv-TasNet*</td>
<td></td>
<td>5.1M</td>
<td>15.6</td>
<td>15.3</td>
<td>3.24</td>
</tr>
<tr>
<td>SE</td>
<td>SpEx(Ref: Avg: 7.5s)</td>
<td>Time</td>
<td>10.8M</td>
<td>16.3</td>
<td>15.8</td>
<td>3.35</td>
</tr>
<tr>
<td></td>
<td>SpEx(Ref: 60s)</td>
<td></td>
<td>10.8M</td>
<td>17.0</td>
<td>16.6</td>
<td>3.42</td>
</tr>
</tbody>
</table>
Speaker Extraction

Evaluation of speaker extraction on WHAM! database

- Time-domain speaker extraction: noisy two-speaker mixture condition
  - The proposed SpEx approach achieves significant improvements comparing with other frequency-domain and time-domain speech separation methods.

<table>
<thead>
<tr>
<th>Task</th>
<th>Methods</th>
<th>Domain</th>
<th>#Paras</th>
<th>SDRi</th>
<th>SI-SDRi</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>Chimera++</td>
<td>Frequency</td>
<td>32.9M</td>
<td>-</td>
<td>9.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BLSTM-TasNet</td>
<td>Time</td>
<td>23.6M</td>
<td>-</td>
<td>9.8</td>
<td>-</td>
</tr>
<tr>
<td>SE</td>
<td>SpEx(Ref: Avg: 7.5s)</td>
<td>Time</td>
<td>10.8M</td>
<td>13.0</td>
<td>12.2</td>
<td>2.41</td>
</tr>
<tr>
<td></td>
<td>SpEx(Ref: 60s)</td>
<td>Time</td>
<td>10.8M</td>
<td>13.7</td>
<td>13.1</td>
<td>2.48</td>
</tr>
</tbody>
</table>
Speaker Extraction

- Evaluation of speaker extraction on WHAMR! database
  - Time-domain speaker extraction: reverberant two-speaker mixture condition
    - The proposed SpEx approach achieves significant improvements comparing with other time-domain speech separation methods, even the system with pre-enhancement module.

<table>
<thead>
<tr>
<th>Task</th>
<th>Methods</th>
<th>Domain</th>
<th>SDRi</th>
<th>SI-SDRi</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>Conv-TasNet</td>
<td>Time</td>
<td>-</td>
<td>7.6</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BLSTM-TasNet</td>
<td>Time</td>
<td>-</td>
<td>8.9</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Pre-Enh+BLSTM-TasNet</td>
<td>Time</td>
<td>-</td>
<td>9.9</td>
<td>-</td>
</tr>
<tr>
<td>SE</td>
<td>SpEx(Ref: Avg: 7.5s)</td>
<td>Time</td>
<td>8.7</td>
<td>9.7</td>
<td>2.68</td>
</tr>
<tr>
<td></td>
<td>SpEx(Ref: 60s)</td>
<td>Time</td>
<td>9.5</td>
<td>10.8</td>
<td>2.77</td>
</tr>
</tbody>
</table>
Evaluation of speaker extraction on WHAMR! database

- Time-domain speaker extraction: two-speaker mixture condition with both additive noise and reverberation
  - The proposed SpEx approach achieves significant improvements comparing with other time-domain speech separation methods, even the system with pre-enhancement and post-enhancement modules.

<table>
<thead>
<tr>
<th>Task</th>
<th>Methods</th>
<th>Domain</th>
<th>SDRi</th>
<th>SI-SDRi</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SS</td>
<td>Conv-TasNet</td>
<td>Time</td>
<td>-</td>
<td>8.3</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>BLSTM-TasNet</td>
<td>Time</td>
<td>-</td>
<td>9.2</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Pre-Enh+BLSTM-TasNet+Post-Enh</td>
<td>Time</td>
<td>-</td>
<td>10.1</td>
<td>-</td>
</tr>
<tr>
<td>SE</td>
<td>SpEx(Ref: Avg: 7.5s)</td>
<td>Time</td>
<td>9.5</td>
<td>10.3</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>SpEx(Ref: 60s)</td>
<td></td>
<td><strong>10.1</strong></td>
<td><strong>11.1</strong></td>
<td><strong>2.30</strong></td>
</tr>
</tbody>
</table>
Speaker Extraction

Evaluation of speaker extraction on 4 conditions

- Time-domain speaker extraction: training on 4 conditions, test on each individual condition
  - The proposed SpEx approach achieves significant improvements comparing with other time-domain speech separation methods, even the system with pre-enhancement and post-enhancement modules.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Conditions</th>
<th>SDRi</th>
<th>SI-SDRi</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpEx (Ref: Avg. 7.3s)</td>
<td>4 conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training Clean mixture</td>
<td>14.8</td>
<td>14.1</td>
<td>3.22</td>
</tr>
<tr>
<td></td>
<td>Test Noisy mixture</td>
<td>13.5</td>
<td>12.7</td>
<td>2.49</td>
</tr>
<tr>
<td></td>
<td>Test Reverberant mixture</td>
<td>8.8</td>
<td>9.8</td>
<td>2.74</td>
</tr>
<tr>
<td></td>
<td>Test Noisy and reverberant mixture</td>
<td>10.1</td>
<td>10.8</td>
<td>2.31</td>
</tr>
<tr>
<td>SpEx (Ref: 60s)</td>
<td>4 conditions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Training Clean mixture</td>
<td>15.9</td>
<td>15.4</td>
<td>3.32</td>
</tr>
<tr>
<td></td>
<td>Test Noisy mixture</td>
<td>14.3</td>
<td>13.8</td>
<td>2.57</td>
</tr>
<tr>
<td></td>
<td>Test Reverberant mixture</td>
<td>9.7</td>
<td>10.8</td>
<td>2.83</td>
</tr>
<tr>
<td></td>
<td>Test Noisy and reverberant mixture</td>
<td>10.8</td>
<td>11.7</td>
<td>2.38</td>
</tr>
</tbody>
</table>
Speaker Extraction

- Demo: female-female speaker’s mixture

![Image of speaker extraction process]
Speaker Extraction

- **SpEx+:** Fully time-domain by sharing encoder

Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang and Haizhou Li, Multi-stage Speaker Extraction with Utterance and Frame-Level Reference Signals, ICASSP 2021.
Speaker Extraction

- SpEx++: Reinforcing the attention
Speaker Extraction

- **SpEx++**: Reinforcing the attention

<table>
<thead>
<tr>
<th>Given Ref.</th>
<th>Methods</th>
<th>Signal Fusion</th>
<th>#stages</th>
<th>Target Ref.</th>
<th>SDRi (dB)</th>
<th>SI-SDRi (dB)</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3s (avg)</td>
<td>SpEx++</td>
<td>×</td>
<td>1</td>
<td>Utt</td>
<td>17.2</td>
<td>16.9</td>
<td>3.43</td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>1</td>
<td>Utt</td>
<td>17.5</td>
<td>17.2</td>
<td>3.46</td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>2</td>
<td>Utt</td>
<td>17.7</td>
<td>17.3</td>
<td>3.47</td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>2</td>
<td>Fm</td>
<td>18.2</td>
<td>17.8</td>
<td>3.51</td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>2</td>
<td>Utt+Fm</td>
<td>18.3</td>
<td>17.9</td>
<td>3.52</td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>3</td>
<td>Utt+Fm</td>
<td><strong>18.4</strong></td>
<td><strong>18.0</strong></td>
<td><strong>3.53</strong></td>
</tr>
<tr>
<td>2s</td>
<td>SpEx++</td>
<td>×</td>
<td>1</td>
<td>Utt</td>
<td>16.6</td>
<td>16.2</td>
<td>3.37</td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>1</td>
<td>Utt</td>
<td>16.9</td>
<td>16.4</td>
<td>3.40</td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>2</td>
<td>Utt+Fm</td>
<td><strong>17.6</strong></td>
<td><strong>17.1</strong></td>
<td><strong>3.46</strong></td>
</tr>
<tr>
<td></td>
<td>SpEx++</td>
<td>✓</td>
<td>3</td>
<td>Utt+Fm</td>
<td><strong>17.6</strong></td>
<td><strong>17.2</strong></td>
<td><strong>3.46</strong></td>
</tr>
</tbody>
</table>
Speaker Extraction

- SpEx++: Reinforcing the attention

Noise [Image]
Reverb [Image]
Noise + Reverb [Image]
Estimated [Image]
Estimated [Image]
Estimated [Image]
Speaker Extraction

- Visual assisted speaker extraction
  - Looking attentively to hear.

Zexu Pan, Ruijie Tao, Chenglin Xu, and Haizhou Li, MuSE: multi-modal target speaker extraction with visual cues, submitted to ICASSP, 2021.
Speaker Extraction

- Visual assisted speaker extraction
  - Reentry model.

Fig. 1. The proposed audio-visual speaker extraction network, named the reentry model.
Speaker Extraction

☐ Evaluation on VoxCeleb2 dataset
  • Baselines[16][18]: Audio-only speech separation networks
  • Baselines[47][57]: Audio-visual speaker extraction networks that utilize viseme-phoneme mapping information.

<table>
<thead>
<tr>
<th>Dataset Mixtures</th>
<th>Task</th>
<th>Model</th>
<th>Att</th>
<th>SI-SDRi (dB)</th>
<th>SDRi (dB)</th>
<th>PESQI</th>
<th>STOIi</th>
<th>Params (m)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speech separation</td>
<td>Conv-TasNet [16]</td>
<td>N/A</td>
<td>10.94</td>
<td>11.27</td>
<td>0.835</td>
<td>0.230</td>
<td>8.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DPRNN [18]</td>
<td>N/A</td>
<td>11.57</td>
<td>11.88</td>
<td>0.902</td>
<td>0.235</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TDSE-O [47]</td>
<td>reentry</td>
<td>sync3</td>
<td>12.60</td>
<td>12.92</td>
<td>1.103</td>
<td>0.256</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>TDSE-I</td>
<td></td>
<td></td>
<td>11.62</td>
<td>11.96</td>
<td>0.958</td>
<td>0.240</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MuSE-O [57]</td>
<td></td>
<td></td>
<td>11.67</td>
<td>12.04</td>
<td>0.969</td>
<td>0.241</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MuSE-I</td>
<td></td>
<td></td>
<td>11.91</td>
<td>12.22</td>
<td>1.007</td>
<td>0.245</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speech separation</td>
<td>Conv-TasNet [16]</td>
<td>N/A</td>
<td>9.31</td>
<td>9.74</td>
<td>0.264</td>
<td>0.230</td>
<td>8.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>DPRNN [18]</td>
<td>N/A</td>
<td>9.83</td>
<td>10.23</td>
<td>0.283</td>
<td>0.239</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TDSE-O [47]</td>
<td>reentry</td>
<td>sync3</td>
<td>12.63</td>
<td>13.08</td>
<td>0.613</td>
<td>0.327</td>
<td>18.8</td>
<td></td>
</tr>
<tr>
<td>TDSE-I</td>
<td></td>
<td></td>
<td>11.54</td>
<td>12.04</td>
<td>0.500</td>
<td>0.303</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MuSE-O [57]</td>
<td></td>
<td></td>
<td>11.64</td>
<td>12.18</td>
<td>0.513</td>
<td>0.306</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MuSE-I</td>
<td></td>
<td></td>
<td>12.21</td>
<td>12.68</td>
<td>0.559</td>
<td>0.318</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

## Evaluation across other datasets

- Models are trained on VoxCeleb2 and tested on other datasets.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Dataset</th>
<th>Model</th>
<th>SI-SDRi</th>
<th>SDRi</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild</td>
<td>LRS2</td>
<td>TDSE-I</td>
<td>11.61</td>
<td>12.05</td>
<td>0.874</td>
<td>0.234</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MuSE-I</td>
<td>12.21</td>
<td>12.60</td>
<td>0.937</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reentry</td>
<td>12.66</td>
<td>13.08</td>
<td>1.036</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>LRS3</td>
<td>TDSE-I</td>
<td>13.39</td>
<td>13.71</td>
<td>1.102</td>
<td>0.241</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MuSE-I</td>
<td>13.69</td>
<td>13.98</td>
<td>1.142</td>
<td>0.245</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reentry</td>
<td>14.24</td>
<td>14.54</td>
<td>1.244</td>
<td>0.251</td>
</tr>
<tr>
<td>Studio</td>
<td>Grid</td>
<td>TDSE-I</td>
<td>9.33</td>
<td>11.51</td>
<td>0.822</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MuSE-I</td>
<td>10.18</td>
<td>12.06</td>
<td>0.931</td>
<td>0.090</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reentry</td>
<td>9.83</td>
<td>12.09</td>
<td>0.893</td>
<td>0.098</td>
</tr>
<tr>
<td></td>
<td>TCD-TIMIT</td>
<td>TDSE-I</td>
<td>13.70</td>
<td>14.22</td>
<td>0.989</td>
<td>0.174</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MuSE-I</td>
<td>14.36</td>
<td>14.87</td>
<td>1.067</td>
<td>0.183</td>
</tr>
<tr>
<td></td>
<td></td>
<td>reentry</td>
<td>14.96</td>
<td>15.55</td>
<td>1.118</td>
<td>0.188</td>
</tr>
</tbody>
</table>

*TABLE VI: CROSS-DATASETS EVALUATIONS FOR MODELS THAT ARE TRAINED ON THE VoxCeleb2-2mix DATASET AND TESTED ON OTHER DATASETS*
Demo video: The reentry model

https://www.youtube.com/watch?v=9HnKFUNIcfY
Speaker Extraction

- **Visual assisted speaker extraction**
  - Visual is missing, how?

Speaker Extraction

- Visual assisted speaker extraction
  - Evaluation on VoxCeleb2 dataset

### Table 1. Comparison of our proposed ImagineNET with the baseline TDSE under conditions with (√) or without (✗) partially missing visual reference during training and inference. Different visual embeddings and visual inpainting losses are used.

<table>
<thead>
<tr>
<th>System (Sys.)</th>
<th>Model</th>
<th>V missing</th>
<th>V embedding</th>
<th>V loss</th>
<th>SI-SDR</th>
<th>SDR</th>
<th>PESQ</th>
<th>STOI</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Mixture</td>
<td>✓</td>
<td>-</td>
<td>-</td>
<td>-0.09</td>
<td>0.00</td>
<td>1.886</td>
<td>0.632</td>
</tr>
<tr>
<td>1</td>
<td>TDSE</td>
<td>✓</td>
<td>✓</td>
<td>-</td>
<td>9.95</td>
<td>10.48</td>
<td>2.809</td>
<td>0.840</td>
</tr>
<tr>
<td>2</td>
<td>ImagineNET</td>
<td>✓</td>
<td>✓</td>
<td>MSE</td>
<td>10.81</td>
<td>11.33</td>
<td>2.894</td>
<td>0.854</td>
</tr>
<tr>
<td>3</td>
<td>ImagineNET</td>
<td>✓</td>
<td>✓</td>
<td>InfoNCE</td>
<td>10.89</td>
<td>11.40</td>
<td>2.903</td>
<td>0.855</td>
</tr>
<tr>
<td>4</td>
<td>TDSE</td>
<td>✓</td>
<td>✓</td>
<td>sync_v</td>
<td>-</td>
<td>9.29</td>
<td>9.87</td>
<td>2.762</td>
</tr>
<tr>
<td>5</td>
<td>ImagineNET</td>
<td>✓</td>
<td>✓</td>
<td>MSE</td>
<td>9.84</td>
<td>10.46</td>
<td>2.819</td>
<td>0.833</td>
</tr>
<tr>
<td>6</td>
<td>ImagineNET</td>
<td>✓</td>
<td>✓</td>
<td>InfoNCE</td>
<td>9.81</td>
<td>10.42</td>
<td>2.824</td>
<td>0.833</td>
</tr>
<tr>
<td>7</td>
<td>TDSE</td>
<td>✗</td>
<td>✓</td>
<td>-</td>
<td>11.88</td>
<td>12.30</td>
<td>2.991</td>
<td>0.878</td>
</tr>
<tr>
<td>8</td>
<td>ImagineNET</td>
<td>✗</td>
<td>✓</td>
<td>MSE</td>
<td>12.39</td>
<td>12.77</td>
<td>3.023</td>
<td>0.885</td>
</tr>
<tr>
<td>9</td>
<td>ImagineNET</td>
<td>✗</td>
<td>✓</td>
<td>InfoNCE</td>
<td>11.99</td>
<td>12.38</td>
<td>2.988</td>
<td>0.881</td>
</tr>
</tbody>
</table>

Speaker Extraction

- Visual assisted speaker extraction
  - Evaluation on VoxCeleb2 dataset

**Fig. 2.** The SI-SDR for different percentages of available video frames when \(v_{sr_v}\) embedding is used.

**Fig. 3.** The SI-SDR for different percentages of available video frames when \(s_{nc_v}\) embedding is used.

Speaker Extraction

- Heterogeneous speaker extraction
  - Concept?

Speaker Extraction

- Heterogeneous speaker extraction
  - Concept?

Figure 1: Illustration of the heterogeneous target speech separation task. Notice that speakers can be separated using any of the semantic concepts and speaker attributes on the left.

Speaker Extraction

Heterogeneous speaker extraction

- Gesture?

- Co-speech gesture input: A sequence of human upper-body poses consisting of the 3D coordinates of 10 spine-centered joints
- Head, neck, nose, spine, left/right (L/R) shoulders, L/R elbows, and L/R wrists.

- GSR network explicitly associates a separated speech to the target speaker

Demo video: The SEG model
Outline

- Introduction
- Source Separation
  - Frequency domain
  - Time domain
- Speaker Extraction
  - Audio assisted speaker extraction
  - Visual assisted speaker extraction
  - Heterogeneous speaker extraction
- Down-stream Tasks
  - Multi-talker speaker verification
  - Multi-talker speech recognition
- Challenges
Multi-talker Speaker Verification

Introduction

- Traditional speaker verification (SV) approaches work well when the test utterance contains speech from a single speaker.
- However, the performance of SV degrades significantly when the speech is corrupted by interference speakers.
- SV in a multi-talker scenario has gained attention with the increasing demand from real-world application.
Multi-talker Speaker Verification

**Introduction: existed strategies**

- Using speaker diarization \([24,25]\) as the front-end for multi-talker SV.
- Speaker diarization separate speakers well only if the speakers are not overlapped.

> However, speaker diarization cannot clearly separate speakers if multi-talkers speak simultaneously most of the time.

Multi-talker Speaker Verification

- **Methodology**
  - Multi-Talker speaker verification with speaker extraction (SE-SV).
  - Speaker extraction module:
    - SBF-MTSAL
    - SBF-MTSAL-Concat
    - SpEx
Multi-talker Speaker Verification

- Methodology
  - Target speaker verification (tSV).

Multi-talker Speaker Verification

- Experimental setup: speaker extraction dataset
  - Multi-Talker speaker verification with speaker extraction (SE-SV).
    - WSJ0-2mix-extr database.
    - Data distribution in different overlapped percentage
Multi-talker Speaker Verification

- **Experimental setup: speaker verification**
  - Evaluation set
    - Trials: 3,000 target and 48,000 non-target (based on test set of WSJ0-2mix-extr).
    - Mixture evaluation set & clean evaluation set (only one speaker)
  - Training set
    - Clean training set: 8,769 utterances from 101 speakers in WSJ0 corpus.
    - Extracted training set: 5,000 extracted utterances by speaker extraction from the WSJ0-2mix-extr development set.
  - Features
    - 19 MFCCs together with energy plus their 1st- and 2nd- derivatives.
  - I-vector+PLDA system
    - A gender-independent UBM with 512 mixtures.
    - Total variability matrix with 400 total factors.
    - LDA and Gaussian PLDA models with 150.
## Multi-talker Speaker Verification

### Experimental results

<table>
<thead>
<tr>
<th>System No.</th>
<th>Systems</th>
<th>Training</th>
<th>Eval</th>
<th>TSE</th>
<th>EER(%)</th>
<th>DCF08</th>
<th>DCF10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Baseline)</td>
<td>SV</td>
<td>Clean(training&amp;dev)</td>
<td>Mixture</td>
<td>No</td>
<td>21.80</td>
<td>0.873</td>
<td>0.912</td>
</tr>
<tr>
<td>2</td>
<td>SE-SV</td>
<td>Clean(training&amp;dev)</td>
<td>Mixture</td>
<td>SBF-MTSAL</td>
<td>10.87</td>
<td>0.766</td>
<td>0.867</td>
</tr>
<tr>
<td>3</td>
<td>SE-SV</td>
<td>Clean(training)+Ext</td>
<td>Mixture</td>
<td>SBF-MTSAL</td>
<td>8.30</td>
<td>0.643</td>
<td>0.777</td>
</tr>
<tr>
<td>4</td>
<td>SE-SV</td>
<td>Clean(training&amp;dev)</td>
<td>Mixture</td>
<td>SBF-MTSAL-Concat</td>
<td>10.37</td>
<td>0.736</td>
<td>0.861</td>
</tr>
<tr>
<td>5</td>
<td>SE-SV</td>
<td>Clean(training)+Ext⁺</td>
<td>Mixture</td>
<td>SBF-MTSAL-Concat</td>
<td>7.77</td>
<td>0.631</td>
<td>0.747</td>
</tr>
<tr>
<td>6</td>
<td>SE-SV</td>
<td>Clean(training&amp;dev)</td>
<td>Mixture</td>
<td>SpEx</td>
<td>7.60</td>
<td>0.632</td>
<td>0.748</td>
</tr>
<tr>
<td>7</td>
<td>SE-SV</td>
<td>Clean(training)+Ext⁺</td>
<td>Mixture</td>
<td>SpEx</td>
<td>6.00</td>
<td>0.551</td>
<td>0.683</td>
</tr>
<tr>
<td>8 (upper-bound)</td>
<td>SV</td>
<td>Clean(training&amp;dev)</td>
<td>Clean</td>
<td>No</td>
<td>3.00</td>
<td>0.360</td>
<td>0.522</td>
</tr>
<tr>
<td>9</td>
<td>OSD-SV</td>
<td>Clean(training&amp;dev)</td>
<td>Mixture</td>
<td>No</td>
<td>14.60</td>
<td>0.851</td>
<td>0.908</td>
</tr>
</tbody>
</table>
Multi-talker Speaker Verification

- **Experimental results**
  - SE-SV significantly improve the performance of multi-talker SV and achieve 72.5% relative EER reduction over the zero-effort baseline.
  - SE-SV significantly outperforms oracle speaker diarization (OSD) in the overlapped multi-talker scenarios.

Upper bound: clean test speech for evaluation, only one speaker in the speech.
Multi-talker Speaker Verification

- Baseline: Zero effort
- OSD: Oracle speaker diarization
- SE-SV: Speaker verification followed by speaker extraction
- tSV: Joint optimization of speaker verification and speaker extraction
- Upper bound: Ground truth clean speech by single speaker

**Experimental results**

<table>
<thead>
<tr>
<th>Method</th>
<th>Equal Error Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>21.8</td>
</tr>
<tr>
<td>OSD-SV</td>
<td>14.6</td>
</tr>
<tr>
<td>SE-SV</td>
<td>6.0</td>
</tr>
<tr>
<td>tSV</td>
<td>5.0</td>
</tr>
<tr>
<td>Upper Bound</td>
<td>3.0</td>
</tr>
</tbody>
</table>

Kuaishou Technology @ ISCSLP 2022
Multi-talker Speech Recognition

- **VoiceFilter-Lite**

![Diagram of VoiceFilter-Lite architecture](image)


---

Kuaishou Technology @ ISCSLP 2022
Multi-talker Speech Recognition

- Full-band speech enhancement and speaker extraction system with wideband (16kHz) ASR system compatibility
Multi-talker Speech Recognition

- Visual assisted speaker extraction for audio-visual ASR
Multi-talker Speech Recognition

- Visual assisted speaker extraction for audio-visual ASR

Table 1: We report the performance of the time-domain audio-visual speaker extraction network using SI-SDR loss function and our proposed hybrid continuity loss function on the clean and noisy LRS2-mix dataset.

<table>
<thead>
<tr>
<th>Noise condition</th>
<th>Model</th>
<th>SI-SDR</th>
<th>SDR</th>
<th>PESQ</th>
<th>STOI</th>
<th>Audio-visual ASR WER</th>
<th>Audio-visual ASR CER</th>
<th>Audio-only ASR WER</th>
<th>Audio-only ASR CER</th>
<th>MAE over</th>
<th>MAE under</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clean</td>
<td>Mixture speech</td>
<td>0.79</td>
<td>1.10</td>
<td>1.985</td>
<td>0.708</td>
<td>59.2</td>
<td>36.5</td>
<td>80.9</td>
<td>54.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SI-SDR loss [27]</td>
<td>13.41</td>
<td>14.08</td>
<td>2.997</td>
<td>0.931</td>
<td>17.3</td>
<td>7.4</td>
<td>23.5</td>
<td>11.3</td>
<td>0.105</td>
<td>0.055</td>
</tr>
<tr>
<td></td>
<td>Hybrid loss (Ours)</td>
<td>13.48</td>
<td>14.12</td>
<td>3.163</td>
<td>0.934</td>
<td>16.1</td>
<td>6.8</td>
<td>21.2</td>
<td>10.0</td>
<td>0.096</td>
<td>0.060</td>
</tr>
<tr>
<td></td>
<td>Clean speech</td>
<td>88.12</td>
<td>288.88</td>
<td>4.500</td>
<td>1.000</td>
<td>11.4</td>
<td>4.4</td>
<td>12.5</td>
<td>4.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Noisy</td>
<td>Mixture speech</td>
<td>-4.43</td>
<td>-3.81</td>
<td>1.467</td>
<td>0.559</td>
<td>56.5</td>
<td>32.7</td>
<td>89.5</td>
<td>60.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>SI-SDR loss [27]</td>
<td>4.54</td>
<td>5.91</td>
<td>2.179</td>
<td>0.776</td>
<td>38.8</td>
<td>20.7</td>
<td>57.9</td>
<td>34.6</td>
<td>0.241</td>
<td>0.118</td>
</tr>
<tr>
<td></td>
<td>Hybrid loss (Ours)</td>
<td>5.12</td>
<td>6.52</td>
<td>2.276</td>
<td>0.789</td>
<td>34.5</td>
<td>17.7</td>
<td>50.8</td>
<td>29.5</td>
<td>0.223</td>
<td>0.119</td>
</tr>
<tr>
<td></td>
<td>Clean speech</td>
<td>88.12</td>
<td>288.88</td>
<td>4.500</td>
<td>1.000</td>
<td>11.4</td>
<td>4.4</td>
<td>12.5</td>
<td>4.9</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Kuaishou Technology @ ISCSLP 2022
Outline

- Introduction
- Source Separation
  - Frequency domain
  - Time domain
- Speaker Extraction
  - Audio assisted speaker extraction
  - Visual assisted speaker extraction
  - Heterogeneous speaker extraction
- Down-stream Tasks
  - Multi-talker speaker verification
  - Multi-talker speech recognition
- Challenges
Challenges

- Scene-agnostic speaker extraction

Speaker extraction task: Attend to target speaker A

- 2T-PT
- 1T-PT
- 2T-AT
- 1T-AT

Speaker A

Speaker B

Speaker C
Challenges

- Universal sound extraction
Robust speech processing
Challenges

- Real-time implementation
References