NN based speech separation

Speech and noise separation

- Speech
- Noise
- Neural network
- Mixture

Multiple speaker separation

- Speaker A
- Speaker B
- Neural network
- Mixture
Issues

1. dependency on number of speakers
2. label permutation
3. speaker tracing
Popular approaches

1. **permutation invariant training**
   permutes outputs of the network during training

2. **deep clustering**
   projects T-F points to embedding space

<table>
<thead>
<tr>
<th></th>
<th>PIT</th>
<th>DC</th>
</tr>
</thead>
<tbody>
<tr>
<td># of speakers</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>label permutation</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>speaker tracing</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Speaker aware neural network

- Speaker information
  - informs the network about target speaker
  - extracted from an adaptation utterance

Solves issues
1. independent of number of speakers
2. no label permutation
3. tracks the speaker
Informing the neural network

auxiliary feature

factorized layer

scaled activations
Informing the neural network

- appending speaker information as additional input
- (Saon et al. 2013; Senior et al. 2014)
Informing the neural network

- splitting one of the layers into sublayers
- sublayers combined with weights derived from speaker info
- (Delcroix et al. 2015; Wu et al. 2015)
Informing the neural network

- activations in one layer scaled by weights derived from speaker info
- (Swietojanski et al. 2014; Samarakoon et al. 2016)
Extracting the speaker information

- speaker information extracted from adaptation utterance with auxiliary network
- average pooling to create utterance-wise vector from frame-wise features
- jointly trained with the main network
Datasets

- **WSJ0-2mix** (Hershey et al. 2016)
  about 10 second long fully overlapped mixtures
  based on WSJ utterances

- **WSJ0-2mix-long**
  same mixing process as WSJ0-2mix
  three utterances from each speaker
  about 1 minute long mixtures
Experimental settings

Network configurations

- smaller configuration

```
BLSTM 300 FC 1000 FC 1000 FC
```

- larger configuration

```
BLSTM 900 BLSTM 900 BLSTM 900 FC
```

- magnitude STFT as input
- predicting T-F mask, MSE objective
### Comparing adaptation methods

- WSJ0-2mix, smaller NN configuration

<table>
<thead>
<tr>
<th>method</th>
<th>$\Delta$SDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>auxiliary feature</td>
<td>-2.2</td>
</tr>
<tr>
<td>factorized layer</td>
<td>6.2</td>
</tr>
<tr>
<td>scaled activations</td>
<td>5.7</td>
</tr>
<tr>
<td>IBM</td>
<td>12.8</td>
</tr>
</tbody>
</table>

#### Diagrams

- **Auxiliary feature**
- **Factorized layer**
- **Scaled activations**
Comparison with DC and PIT

- larger NN configuration, scaled activations method

<table>
<thead>
<tr>
<th>method</th>
<th>2mix</th>
<th>2mix-long</th>
</tr>
</thead>
<tbody>
<tr>
<td>SpeakerBeam</td>
<td>8.2</td>
<td>12.2</td>
</tr>
<tr>
<td>PIT</td>
<td>8.2</td>
<td>9.9</td>
</tr>
<tr>
<td>DC</td>
<td>8.7</td>
<td>10.0</td>
</tr>
<tr>
<td>SpeakerBeam+DC</td>
<td>9.1</td>
<td>12.6</td>
</tr>
<tr>
<td>IBM</td>
<td>12.8</td>
<td>17.1</td>
</tr>
</tbody>
</table>

2mix

- A
- B

2mix-long

- A
- A
- A
- A
- B
- B
- B
- B
• for 0.5 seconds, -0.9 dB SDR degradation
Extracted speaker representations
Conclusions

- Additional speaker information can help to avoid problems of NN based speech separation and do speaker tracing.

- Methods adapting parameters of entire layer work well.

- This can be combined with deep clustering to enhance its accuracy, especially on longer mixtures.

Thank you!

izmolikova@fit.vutbr.cz


