MULTI-MICROPHONE NEURAL SPEECH SEPARATION FOR FAR-FIELD MULTI-TALKER SPEECH RECOGNITION

T. Yoshioka, H. Erdogan, Z. Chen, F. Alleva (Microsoft, Redmond, WA, USA)

Notice

This poster includes updated results relative to the paper in the proceedings. The details of these new results are described in a paper we have submitted to Interspeech 2018.

Highlights

The permutation invariant training (PIT) approach to single-microphone speech separation is extended to multi-microphone scenarios by using

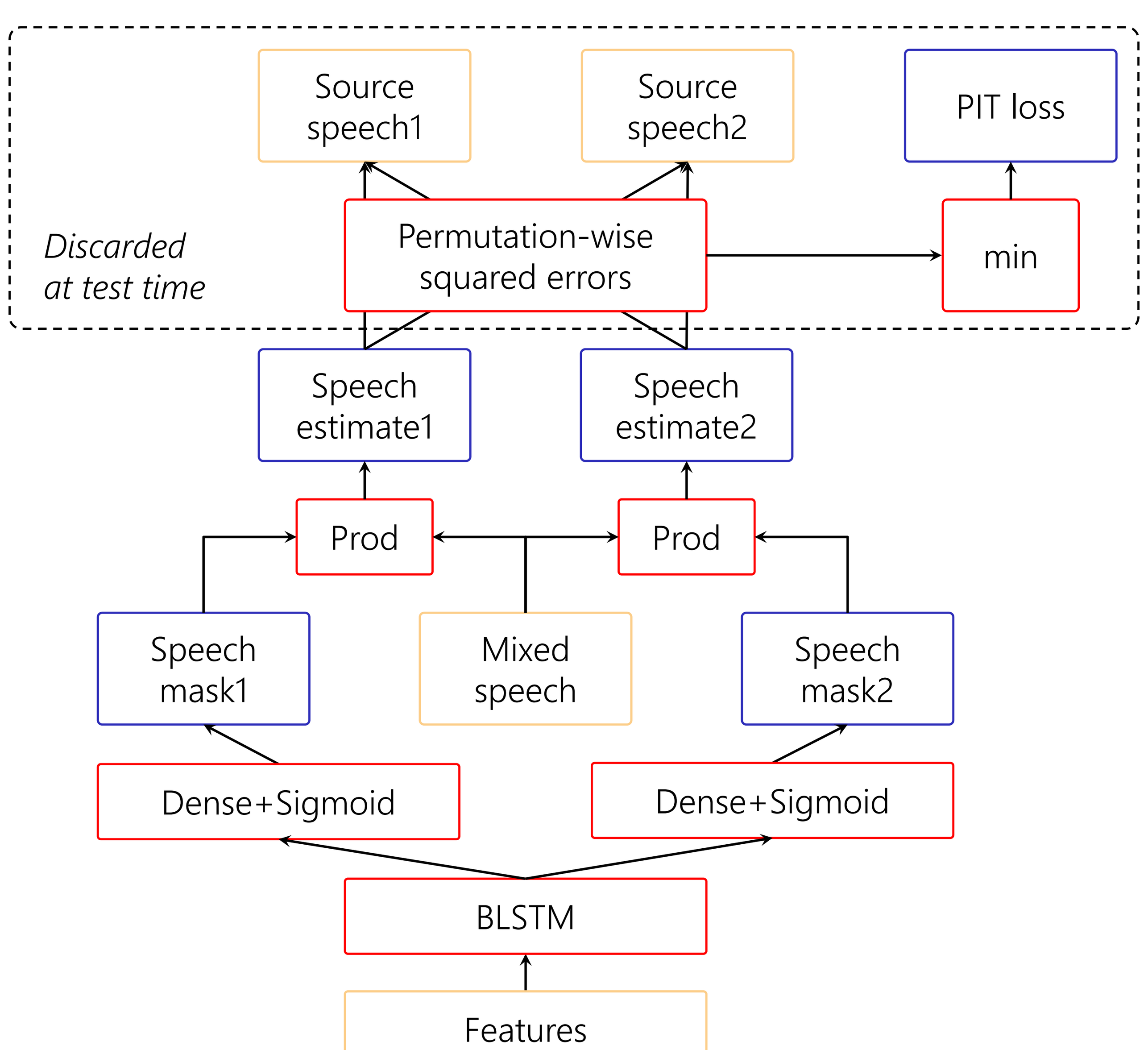
- features extracted from multiple microphones;
- beamforming instead of time-frequency masking for separation; and
- a gain adjustment mechanism to suppress duplicate outputs.

Our method works well for both synthetic reverberant mixtures and real multi-party conversation recordings with far-field microphones.

Owing to PIT and the gain adjustment, our method does not require prior knowledge of the number of speakers.

Permutation Invariant Training (PIT)

- Neural net training method for speech separation (Kolbaek et al., 2017)
- Unlike deep clustering (Hershey et al., 2015), PIT does not require clustering to be performed at test time.
- While effective for anechoic mixtures, single-mic PIT performs poorly under reverberant conditions (see Tab. 1).



Multi-Microphone Features

Spectral features

$$p_{i,tf} = |y_{i,tf}|$$

i: mic index

Spatial features

$$q_{i,tf} = \mathrm{Arg}\left(\frac{y_{i,tf}}{y_{\mathrm{R},tf}}\right)$$

R: reference mic

These features are normalized on a per-utterance basis.

- The spectral features are mean- and variance-normalized.
- The spatial features are mean-normalized.

Simply feeding multi-microphone STFT coefficients resulted in performance degradation (see Tab. 2).

Speech Separation with Beamforming

- Mask-based beamforming (Heymann et al., 2016)
- Full-rank MVDR was used in our experiments.

$$\boldsymbol{w}_{i,f} = \frac{\boldsymbol{\varphi}_{\bar{i},f}^{-1}\boldsymbol{\varphi}_{i,f}\boldsymbol{e}}{\mathrm{tr}\left(\boldsymbol{\varphi}_{\bar{i},f}^{-1}\boldsymbol{\varphi}_{i,f}\right)}, \boldsymbol{e} = [1,0,\dots,0]^T$$

- Two schemes for calculating the spatial covariance matrix, $\boldsymbol{\varphi}_{i,f}$,were examined (see Tab. 3).
  - Use the masks as observation weights (mask-cov):

$$\boldsymbol{\varphi}_{i,f} = \frac{1}{\sum_t m_{i,tf}}\sum_t m_{i,tf}\boldsymbol{Y}_{tf}\boldsymbol{Y}_{tf}^H$$

  - Use masked signals (sig-cov):

$$\boldsymbol{\varphi}_{i,f} = \frac{1}{T}\sum_t \left(m_{i,tf}\boldsymbol{Y}_{tf}\right)\left(m_{i,tf}\boldsymbol{Y}_{tf}\right)^H$$

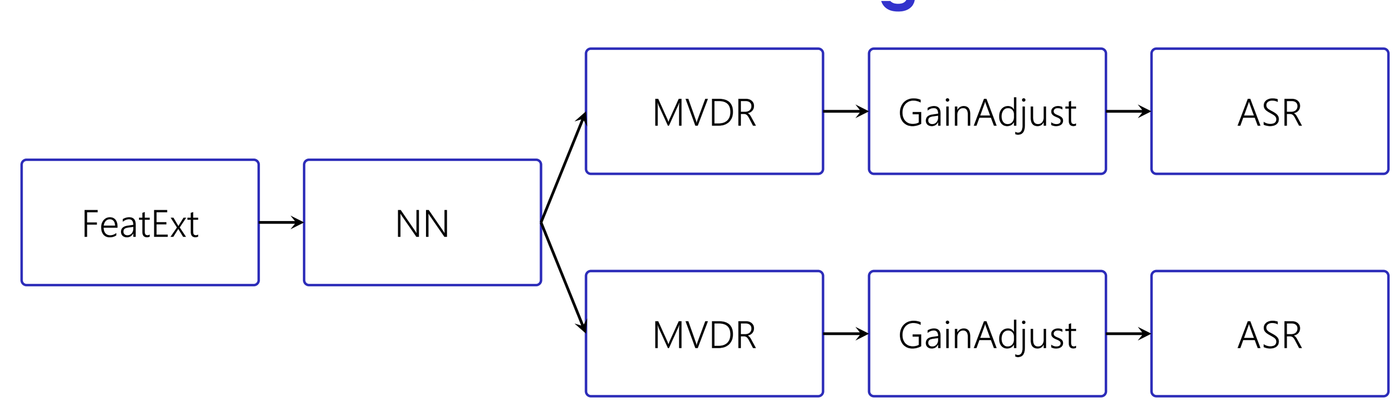
- The interference spatial covariance matrix, $\boldsymbol{\varphi}_{\bar{i},f}$, was calculated by using $1 - m_{i,tf}$ as an interference mask.

Gain Adjustment

- Changes the overall gain of the beamformed audio.
- This is needed because MVDR, which maintains a unit gain toward a certain direction, creates a degraded copy of a target signal when there is only one speaker.

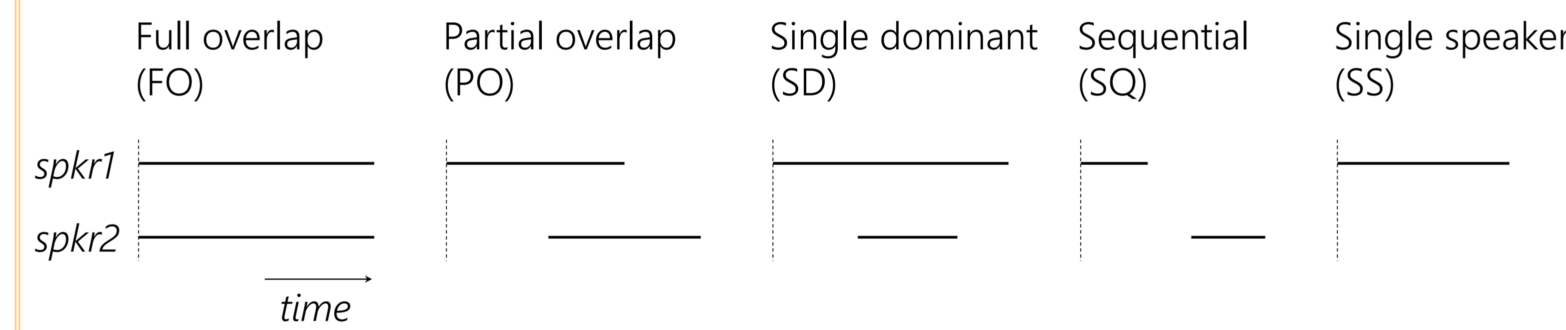
$$x_{i,tf}^{*} = x_{i,tf}\frac{E_i}{\sum_j E_j} \qquad E_i = \sqrt{\sum_{tf}\left|m_{i,tf}y_{\mathrm{R},tf}\right|^2}$$

Overall Processing Flow



Data

- 7-channel circular mic array
- 5 testing conditions



  - Signals were reverberated with randomly generated RIRs.
- Separation network training: 43.7 (x1) or 216 (x5) hours of reverberant speech mixtures created by using SI-284 utterances
- AM: Teacher-student model trained on 6.8K hours of noisy/clean speech audio

Results

- The proposed method substantially reduced the WER compared with the single-mic PIT.
- For SS, one of the output signals was successfully zeroed-out as indicated by a high inter-channel energy ratio (ICER).

Table 1. %WERs of different speech separation systems. ICERs in dB are also shown for proposed system trained on x5.

Separation system	Perf. Metrics	Mixing configurations FO	PO	SD	SQ	SS
Oracle	WER	16.6	17.7	16.4	18.8	16.8
Mixed speech	WER	83.0	83.8	56.8	107.3	16.8
1-mic PIT, x1	WER	63.0	50.6	48.5	31.0	19.3
Proposed, x1	WER	30.6	31.8	24.9	32.5	24.0
Proposed, x5	WER	26.3	31.3	24.0	31.1	19.6
	ICER	0.20	0.14	2.21	0.56	46.2

- The proposed spatial features were much more effective than simply using the raw multi-mic STFT coefficients.

Table 2. %WER comparison for different network inputs. Separation networks were trained on x1.

Network input	Mixing configurations FO	PO	SD	SQ	SS
1 mic	42.4	42.0	34.6	36.3	25.1
7 mics, raw	45.2	43.0	35.2	36.1	24.1
7 mics, magnitude+IPD	30.6	31.8	24.9	32.5	24.0

- The sig-cov scheme slightly outperformed mask-cov.

Table 3. %WER comparison for different enhancement schemes. Separation networks were trained on x5.

Enhancement	Mixing configurations FO	PO	SD	SQ	SS
TF masking	45.6	34.6	35.5	18.4	17.5
MVDR, mask-cov	30.2	33.8	24.8	31.6	17.2
MVDR, sig-cov	26.3	31.3	24.0	31.1	19.6

- Our method works for real far-field multi-party conversations with some modifications (details to be published later).

Table 1: *%WER of different front-ends.*

System	%WER Overall	Segments with overlaps
No processing (mic0)	44.6	48.0
WPE (Yoshioka et al., 2012)	42.1	45.5
+BeamformIt (Anguera et al., 2007)	43.2	45.9
+MaskBF (Heymann et al., 2016)	37.9	42.9
+Proposed separation	**33.8**	**37.0**