Multi-Band PIT and Model Integration for Improved Multi-Channel Speech Separation

Lianwu Chen[1], Meng Yu[2], Dan Su[1], Dong Yu[2]

[1]Tencent AI Lab, Shenzhen, China [2]Tencent AI Lab, Bellevue, WA 98004, USA

Highlights

- Issues in multi-channnel PIT based speech separation approaches:
  - **Phase wrapping** in high frequency IPD features.
  - **Spatial ambiguity** when speakers are closely located.
- Our contributions in this paper:
  - **For phase wrapping:** a multi-band architecture for effective feature encoding in different sub-bands.
  - **For spatial ambiguity:** a model that integrates the single-channel and multi-channel PIT models in utterance level.

MULTI-CHANNEL PIT

- Input Features:
  - Spectral: Log Power Spectrum (LPS).
  - Spatial: Interchannel Phase Difference (IPD).
- Model Structure:
  - Similar as single-channel PIT

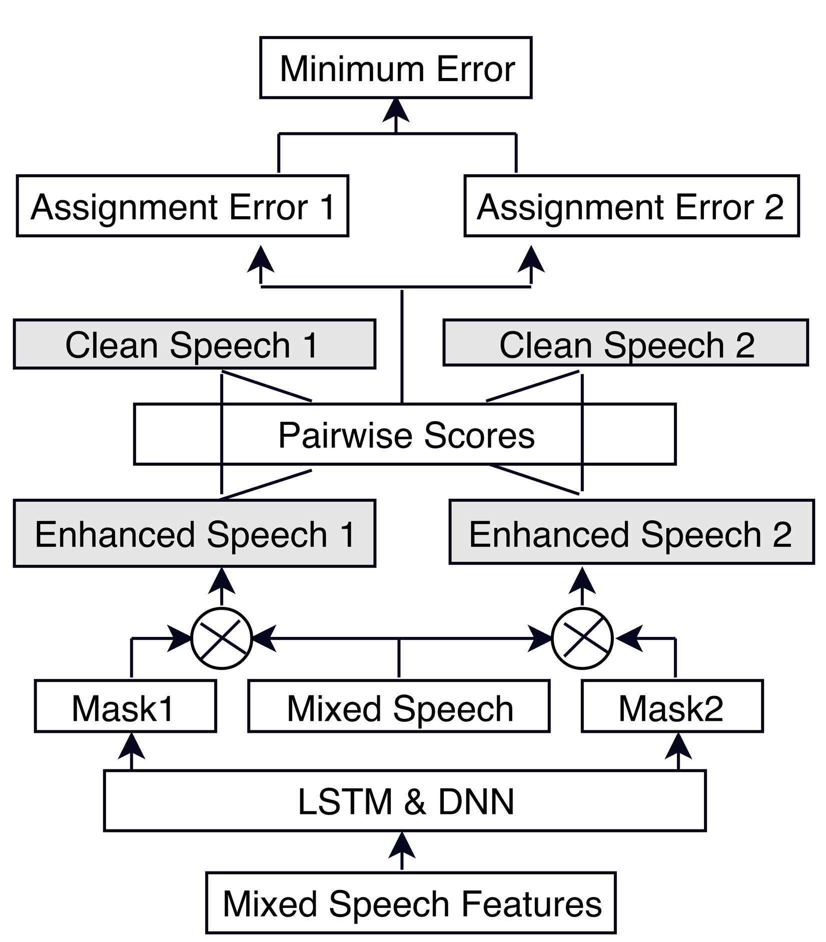

Figure 1: Conventional multi-channel PIT model

MULTI-BAND EMBEDDINGS

- Phase wrapping issue

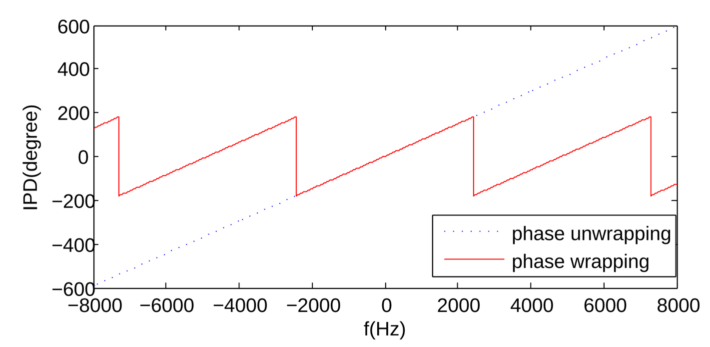

Figure 2: IPD pattern for microphone spacing 7cm.

- Effective feature encoding for different subbands with IPD and LPS.

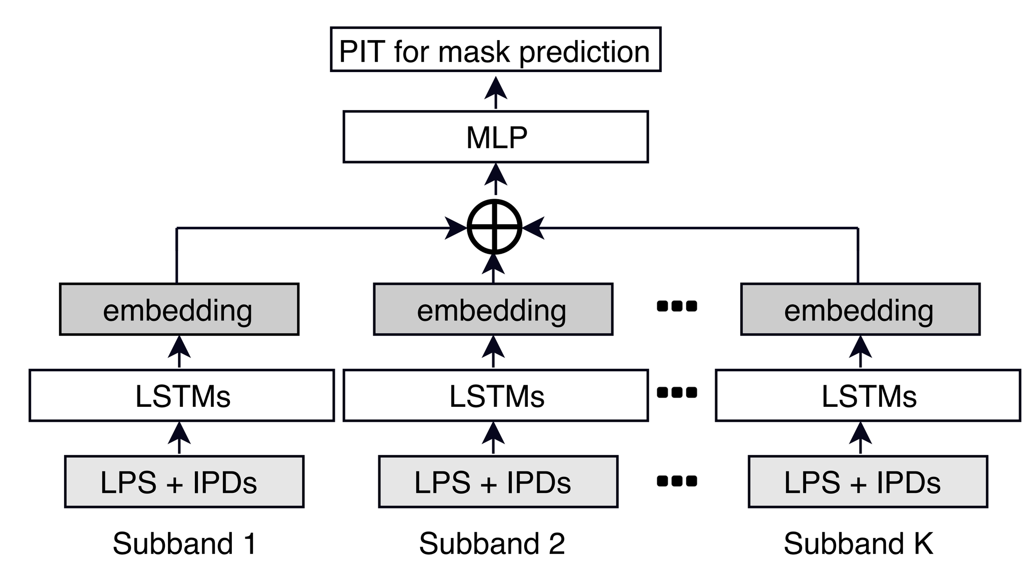

Figure 3: Multi-band feature encoding for multi-channel PIT.

MODEL INTEGRATION

- Spatial ambiguity issue:
  - IPD features fails when speakers are closely located.
  - System cannot pursue balance between IPD and LPS.
- Train a classifier to detect spatial ambiguity case for hard switching.

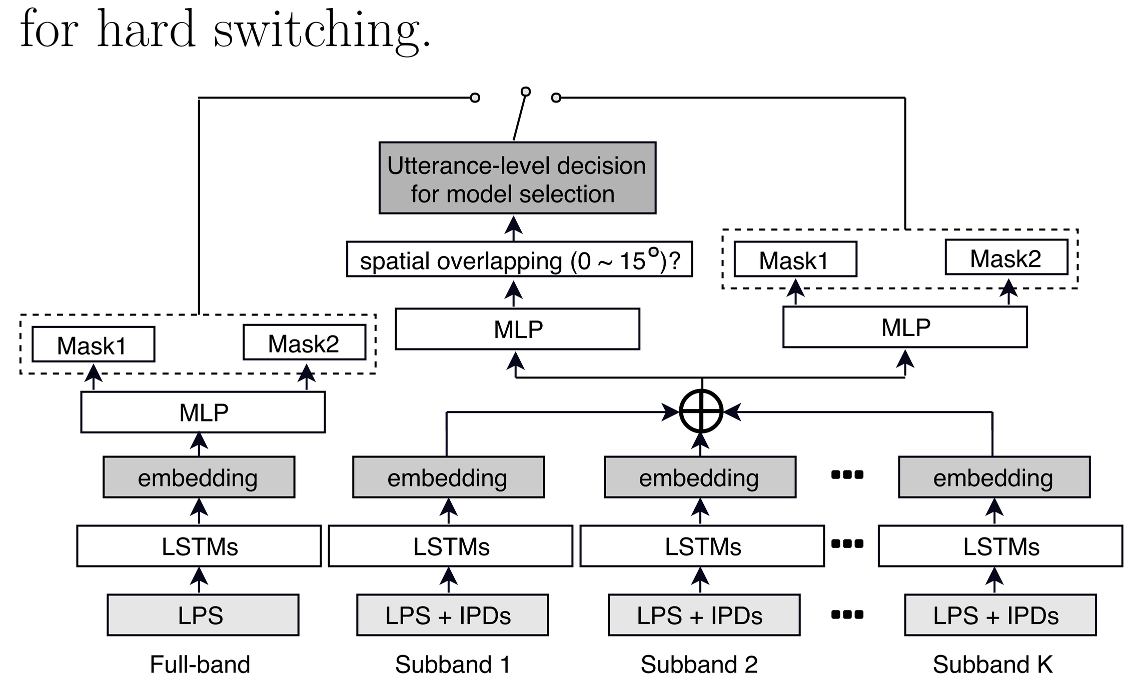

Figure 4: The architecture of model integration.

Table 1: *Evaluation of different approaches in terms of SDR (dB) on test set.*

Method	$0^\circ \sim 15^\circ$	$15^\circ \sim 45^\circ$	$45^\circ \sim 90^\circ$	$90^\circ \sim 180^\circ$	Avg.
raw	2.2	2.1	2.1	2.1	2.1
LPS[1]	8.4	8.8	8.7	8.9	8.7
LPS + 1 IPD (mic pair 1-2)[2]	8.2	8.8	9.1	9.9	9.1
LPS + 2 IPDs (mic pair 1-2, 1-4)[3]	7.1	9.8	10.9	11.5	10.2
LPS + 3 IPDs (mic pair 1-4, 2-5, 3-6)[4]	6.7	10.0	11.3	11.5	10.3
LPS + 6 IPDs (mic pair 1-4, 2-5, 3-6, 1-2, 3-4, 5-6)[5]	5.6	9.4	11.0	11.6	9.9
LPS + 6 IPDs, two-band (6k Hz)[6]	6.3	9.9	11.4	12.1	10.4
LPS + 6 IPDs, two-band (4k Hz)[7]	6.5	10.3	12.0	12.7	10.9
LPS + 6 IPDs, two-band (2k Hz)[8]	6.4	10.7	12.3	13.1	**11.2**
LPS + 6 IPDs, comparable model size[9]	6.0	9.7	11.1	11.7	10.1
LPS + 6 IPDs, four-band (2k/4k/6k Hz)[10]	6.2	10.5	12.0	12.8	11.0
LPS, two-band (2k Hz)[11]	7.9	8.3	8.2	8.3	8.2
LPS + 1 IPD, two-band (2k Hz)[12]	7.0	9.4	11.0	12.1	10.3
LPS + 2 IPDs, two-band (2k Hz)[13]	6.1	10.0	11.6	12.6	10.6
LPS + 3 IPDs, two-band (2k Hz)[14]	6.3	10.4	12.1	12.6	10.9
LPS + 6 IPDs, two-band (2k Hz), multi-task[15]	6.6	10.9	12.4	13.1	**11.3**
LPS + 6 IPDs, two-band (2k Hz), model integ.[16]	**8.3**	10.7	11.9	12.6	**11.2**

Data & Architecture

- **Corpora**:
  Mono Speech * Multi-channel Impulse Response

Table 2: Details of data set

Data	Description
Speech	WSJ-2mix Train:30h, Dev:10h, Test:5h
IR	Image method 6-mic circular array of 7cm diam 3000 rooms RT_{60} 0.05s to 0.5s Angel portion 1:2:2:2

- **Network setup**:
  The baseline PIT networks contain three LSTM layers, each with 512 units, followed by a MLP layer of 512 units and a output layer with 257 × 2 dimension mask. Phase sensitive approximation is used in loss function.

Results & Conclusion

- Spatial overlapping cases can be observed in category $0^\circ \sim 15^\circ$ (schemes 1 vs. 2-5)
- Multi-band framework improves performance (Schemes 6-8 vs. 5), and 6 IPDs achieves better results than others (Schemes 8 vs. 12-14)
- Splitting at 2kHz leads to the best result (schemes 8 vs. 6-7), which is coincident with the phase wrapping frequency 2.5kHz.
- The EER of frame-level spatial overlapping prediction is about 8%.
- With model integration (scheme 16), the spatial overlapping issue is resolved with results in category $0^\circ \sim 15^\circ$ significantly improved.

Tencent AI Lab