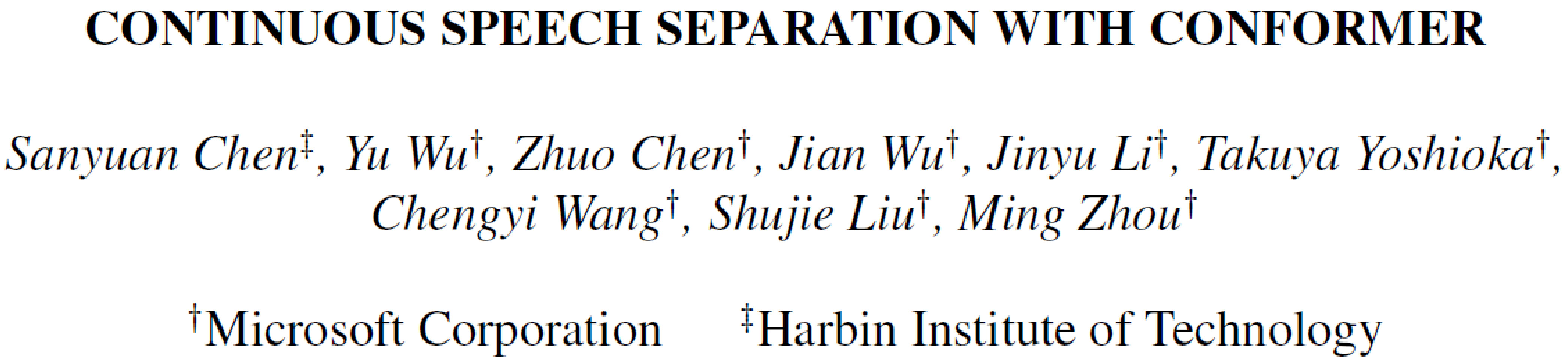


# CONTINUOUS SPEECH SEPARATION WITH CONFORMER

*Sanyuan Chen*[‡], *Yu Wu*[†], *Zhuo Chen*[†], *Jian Wu*[†], *Jinyu Li*[†], *Takuya Yoshioka*[†], *Chengyi Wang*[†], *Shujie Liu*[†], *Ming Zhou*[†]

[†]Microsoft Corporation [‡]Harbin Institute of Technology

Video demo: https://www.youtube.com/watch?v=WRfPBnWc2qQ

Accepted by IEEE ICASSP 2021

# Continuous Speech Separation

- To estimate individual speaker signals from a continuous speech input, where the source signals are fully or partially overlapped.
- Mixed signal: $y(t) = \sum_{s=1}^{S} x_s(t)$ → s-th source signal: $x_s(t)$
- (STFTs) short-time Fourier transforms: $\mathbf{Y}^1(t, f) \longrightarrow \mathbf{X}_s(t, f)$
- Speech Separation Process:

1. $\mathbf{Y}(t, f) = \mathbf{Y}^1(t, f) \oplus \mathrm{IPD}(2) \ldots \oplus \mathrm{IPD}(C) \xrightarrow{\text{Separation model}} \mathbf{M}_s(t, f)$
2. $\mathbf{X}_s(t, f) = \mathbf{M}_s(t, f) \odot \mathbf{Y}^1(t, f)$

# Conformer speech separation model

- Conformer model:
  - a state-of-the-art ASR encoder architecture
  - efficiently capture both local and global context information
- Conformer block:

$$\hat{z} = \text{layernorm}(z + \frac{1}{2}\text{FFN}(z))$$

$$z' = \text{layernorm}(\text{selfattention}(\hat{z}) + \hat{z})$$

$$z'' = \text{layernorm}(\text{conv}(z') + z')$$

$$output = \text{layernorm}(z'' + \frac{1}{2}\text{FFN}(z'')),$$

- multi-head self-attention

$$\text{Multihead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = [\mathbf{H_1} \ldots \mathbf{H}_{d_{head}}]\mathbf{W}^{head}$$

$$\mathbf{H_i} = \text{softmax}(\frac{\mathbf{Q_i}(\mathbf{K_i} + \boxed{\mathbf{pos}})^{\mathsf{T}}}{\sqrt{d_k}})\mathbf{V_i}$$

Relative position embedding

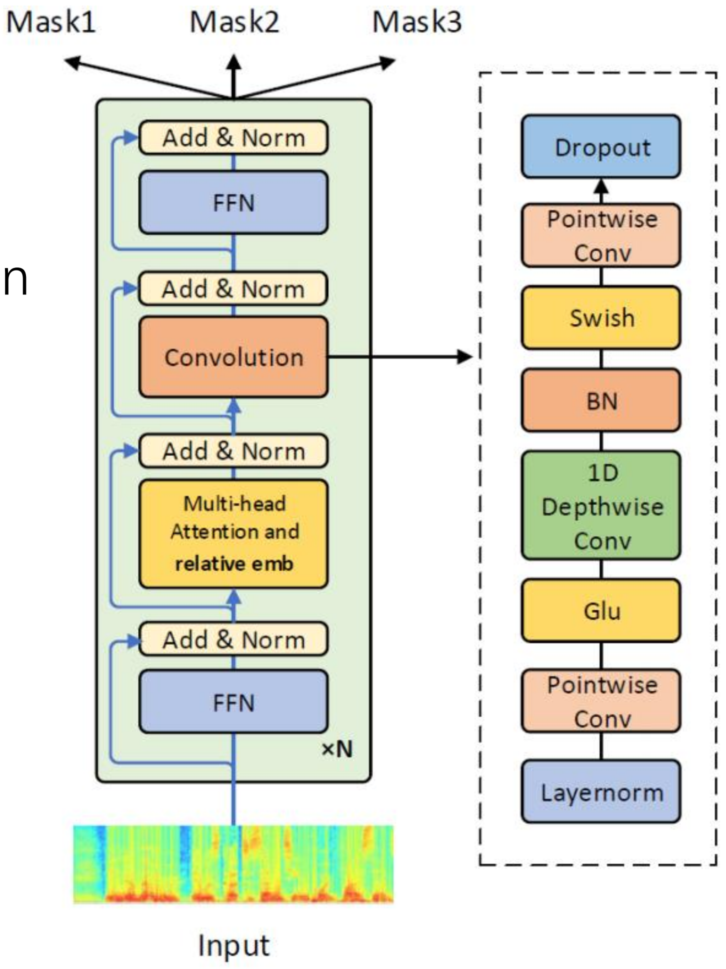
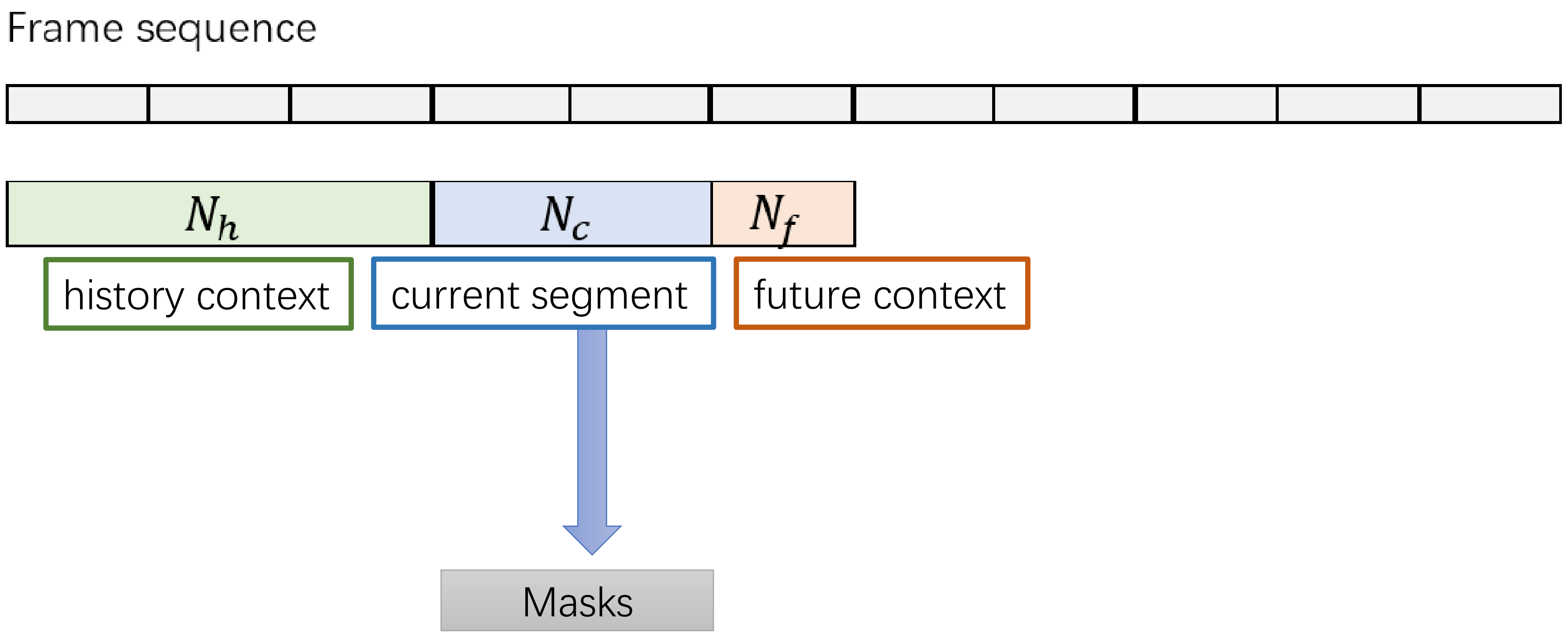


**Fig. 1.** Conformer architecture. There are three mask outputs, two for speakers and one for noise.

# Chunk-wise processing for continuous separation



Frame sequence

$N_h$

$N_c$

$N_f$

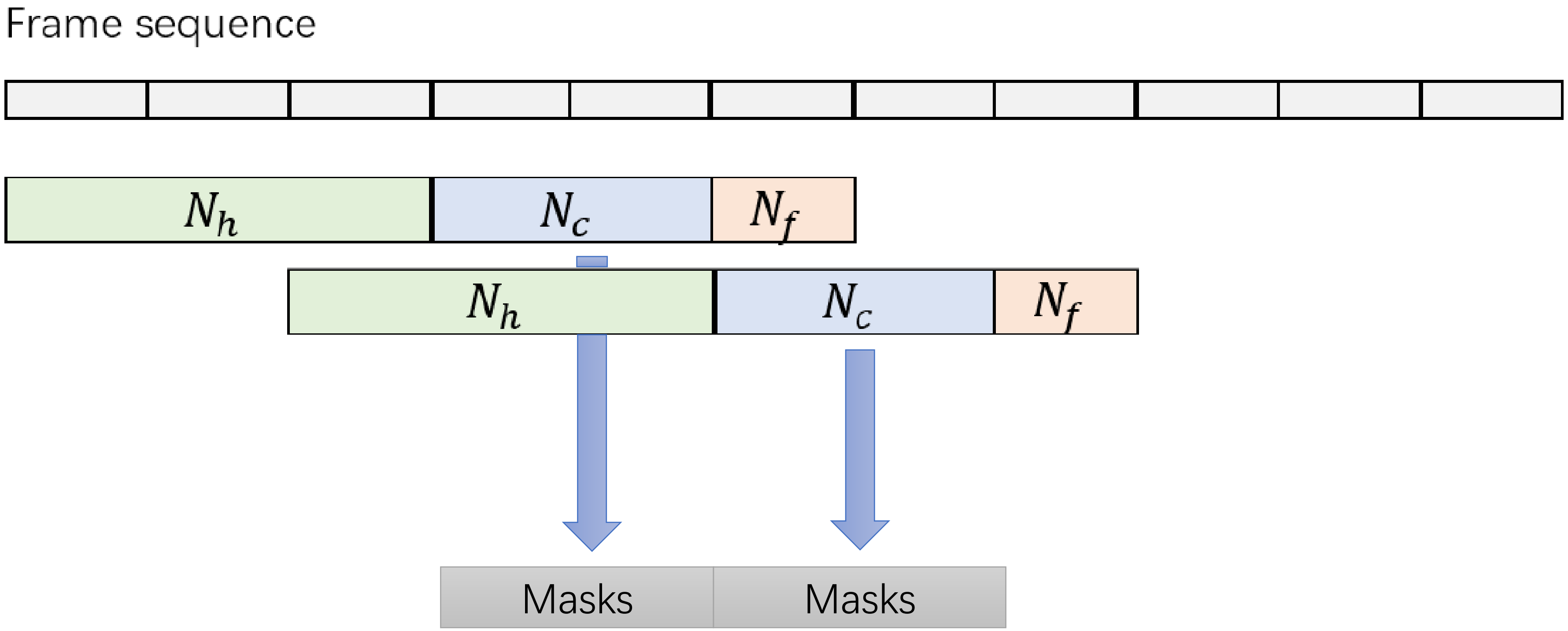
history context

current segment

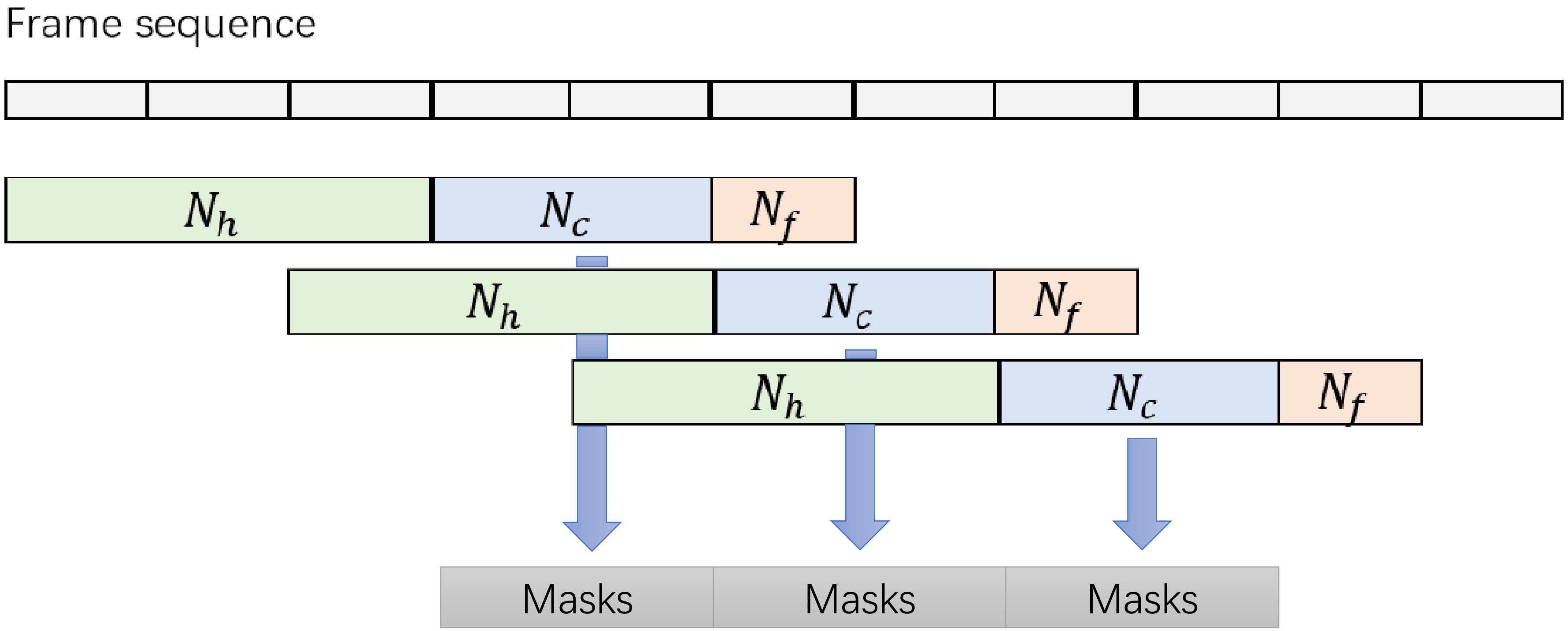
future context

Masks

# Chunk-wise processing for continuous separation



# Chunk-wise processing for continuous separation



# Chunk-wise processing for continuous separation

- **Self-attention** considering the history information beyond the current chunk:
  - **Q** is obtained by the current chunk
  - **K** and **V** are the concatenations of the key and value in the previous and current chunks.

$$\mathbf{H_i} = \mathrm{softmax}\left(\frac{\mathbf{Q_i}(\mathbf{K_i} + \mathbf{pos})^{\intercal}}{\sqrt{d_k}}\right)\mathbf{V_i}$$

$$\Downarrow$$

$$\mathrm{softmax}\left(\frac{\mathbf{Q_i}(\mathbf{K_i}\boxed{\oplus\,\mathbf{K_{cache,i}}} + \mathbf{pos})^{\intercal}}{\sqrt{d_k}}\right)(\mathbf{V_i}\boxed{\oplus\,\mathbf{V_{cache,i}}})$$

# Experiments

- Training data: 219 hours of artificially reverberated and mixed utterances that sampled randomly from WSJ1
- Evaluate data: LibriCSS (real recordings created by concatenating and mixing LibriSpeech utterances with various overlap ratios)
- Separation models:

| Models | # Params | Layers | Hidden dimension | Attention heads | Attention dimension |
|---|---|---|---|---|---|
| BLSTM | 21.80M | 3 | 512 | - | - |
| Transformer-base | 21.90M | 16 | 2048 | 4 | 256 |
| Transformer-large | 58.33M | 18 | 2048 | 8 | 512 |
| Conformer-base | 22.07M | 16 | 1024 | 4 | 256 |
| Conformer-large | 58.72M | 18 | 1024 | 8 | 512 |

# Experiments on LibriCSS

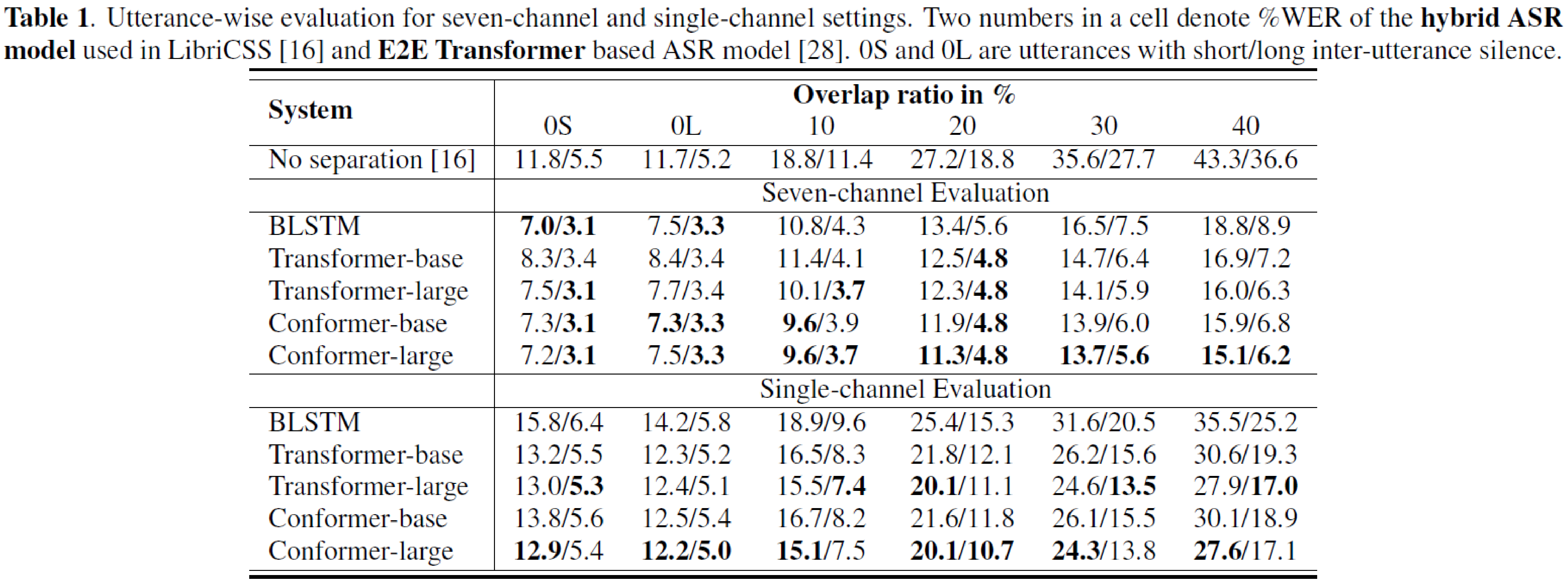


**Table 1**. Utterance-wise evaluation for seven-channel and single-channel settings. Two numbers in a cell denote %WER of the **hybrid ASR model** used in LibriCSS [16] and **E2E Transformer** based ASR model [28]. 0S and 0L are utterances with short/long inter-utterance silence.

| **System** | **Overlap ratio in %** | | | | | |
|---|---|---|---|---|---|---|
| | 0S | 0L | 10 | 20 | 30 | 40 |
| No separation [16] | 11.8/5.5 | 11.7/5.2 | 18.8/11.4 | 27.2/18.8 | 35.6/27.7 | 43.3/36.6 |
| | Seven-channel Evaluation | | | | | |
| BLSTM | **7.0/3.1** | 7.5/**3.3** | 10.8/4.3 | 13.4/5.6 | 16.5/7.5 | 18.8/8.9 |
| Transformer-base | 8.3/3.4 | 8.4/3.4 | 11.4/4.1 | 12.5/**4.8** | 14.7/6.4 | 16.9/7.2 |
| Transformer-large | 7.5/**3.1** | 7.7/3.4 | 10.1/**3.7** | 12.3/**4.8** | 14.1/5.9 | 16.0/6.3 |
| Conformer-base | 7.3/**3.1** | **7.3/3.3** | **9.6**/3.9 | 11.9/**4.8** | 13.9/6.0 | 15.9/6.8 |
| Conformer-large | 7.2/**3.1** | 7.5/**3.3** | **9.6/3.7** | **11.3/4.8** | **13.7/5.6** | **15.1/6.2** |
| | Single-channel Evaluation | | | | | |
| BLSTM | 15.8/6.4 | 14.2/5.8 | 18.9/9.6 | 25.4/15.3 | 31.6/20.5 | 35.5/25.2 |
| Transformer-base | 13.2/5.5 | 12.3/5.2 | 16.5/8.3 | 21.8/12.1 | 26.2/15.6 | 30.6/19.3 |
| Transformer-large | 13.0/**5.3** | 12.4/5.1 | 15.5/**7.4** | **20.1**/11.1 | 24.6/**13.5** | 27.9/**17.0** |
| Conformer-base | 13.8/5.6 | 12.5/5.4 | 16.7/8.2 | 21.6/11.8 | 26.1/15.5 | 30.1/18.9 |
| Conformer-large | **12.9**/5.4 | **12.2/5.0** | **15.1**/7.5 | **20.1/10.7** | **24.3**/13.8 | **27.6**/17.1 |

# Experiments on LibriCSS

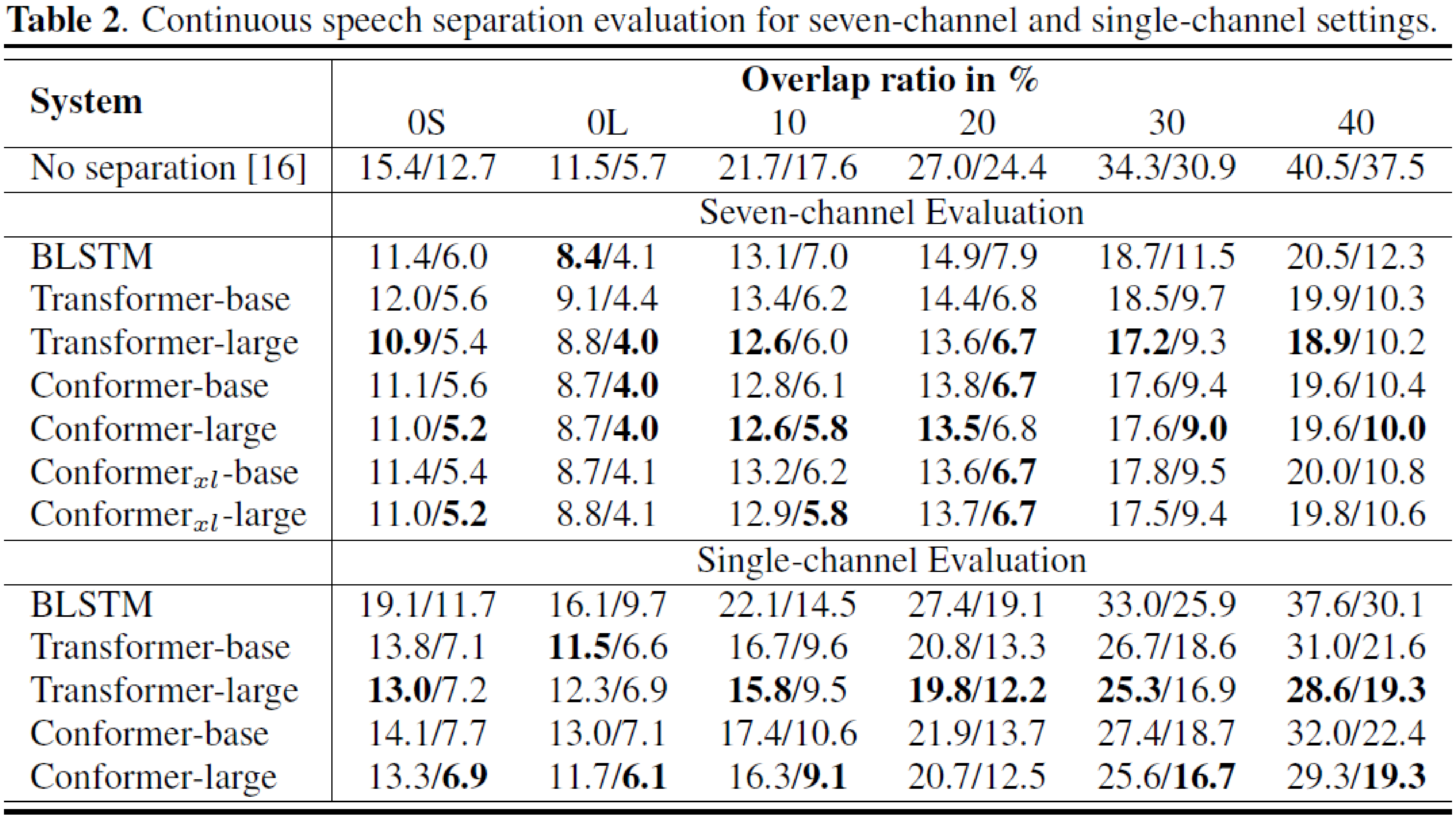


**Table 2**. Continuous speech separation evaluation for seven-channel and single-channel settings.

| System | Overlap ratio in % | | | | | |
|---|---|---|---|---|---|---|
| | 0S | 0L | 10 | 20 | 30 | 40 |
| No separation [16] | 15.4/12.7 | 11.5/5.7 | 21.7/17.6 | 27.0/24.4 | 34.3/30.9 | 40.5/37.5 |
| | Seven-channel Evaluation | | | | | |
| BLSTM | 11.4/6.0 | **8.4**/4.1 | 13.1/7.0 | 14.9/7.9 | 18.7/11.5 | 20.5/12.3 |
| Transformer-base | 12.0/5.6 | 9.1/4.4 | 13.4/6.2 | 14.4/6.8 | 18.5/9.7 | 19.9/10.3 |
| Transformer-large | **10.9**/5.4 | 8.8/**4.0** | **12.6**/6.0 | 13.6/**6.7** | **17.2**/9.3 | **18.9**/10.2 |
| Conformer-base | 11.1/5.6 | 8.7/**4.0** | 12.8/6.1 | 13.8/**6.7** | 17.6/9.4 | 19.6/10.4 |
| Conformer-large | 11.0/**5.2** | 8.7/**4.0** | **12.6/5.8** | **13.5**/6.8 | 17.6/**9.0** | 19.6/**10.0** |
| Conformer$_{xl}$-base | 11.4/5.4 | 8.7/4.1 | 13.2/6.2 | 13.6/**6.7** | 17.8/9.5 | 20.0/10.8 |
| Conformer$_{xl}$-large | 11.0/**5.2** | 8.8/4.1 | 12.9/**5.8** | 13.7/**6.7** | 17.5/9.4 | 19.8/10.6 |
| | Single-channel Evaluation | | | | | |
| BLSTM | 19.1/11.7 | 16.1/9.7 | 22.1/14.5 | 27.4/19.1 | 33.0/25.9 | 37.6/30.1 |
| Transformer-base | 13.8/7.1 | **11.5**/6.6 | 16.7/9.6 | 20.8/13.3 | 26.7/18.6 | 31.0/21.6 |
| Transformer-large | **13.0**/7.2 | 12.3/6.9 | **15.8**/9.5 | **19.8/12.2** | **25.3**/16.9 | **28.6/19.3** |
| Conformer-base | 14.1/7.7 | 13.0/7.1 | 17.4/10.6 | 21.9/13.7 | 27.4/18.7 | 32.0/22.4 |
| Conformer-large | 13.3/**6.9** | 11.7/**6.1** | 16.3/**9.1** | 20.7/12.5 | 25.6/**16.7** | 29.3/**19.3** |

# Experiments on Real Conversation dataset

- Real Conversation dataset
  - an internal real conversation corpus which consists of 15.8 hours of single channel recordings of daily group discussions
  - Significantly more complex with respect to the acoustics, linguistics, and interspeaker dynamics.
- Improvements:
  - Increase the training data amount to 1500 hours
  - Merge two channel outputs were merged when a single active speaker was judged to be present
  - we used single speaker signals corrupted by background noise as a training target

# Experiments

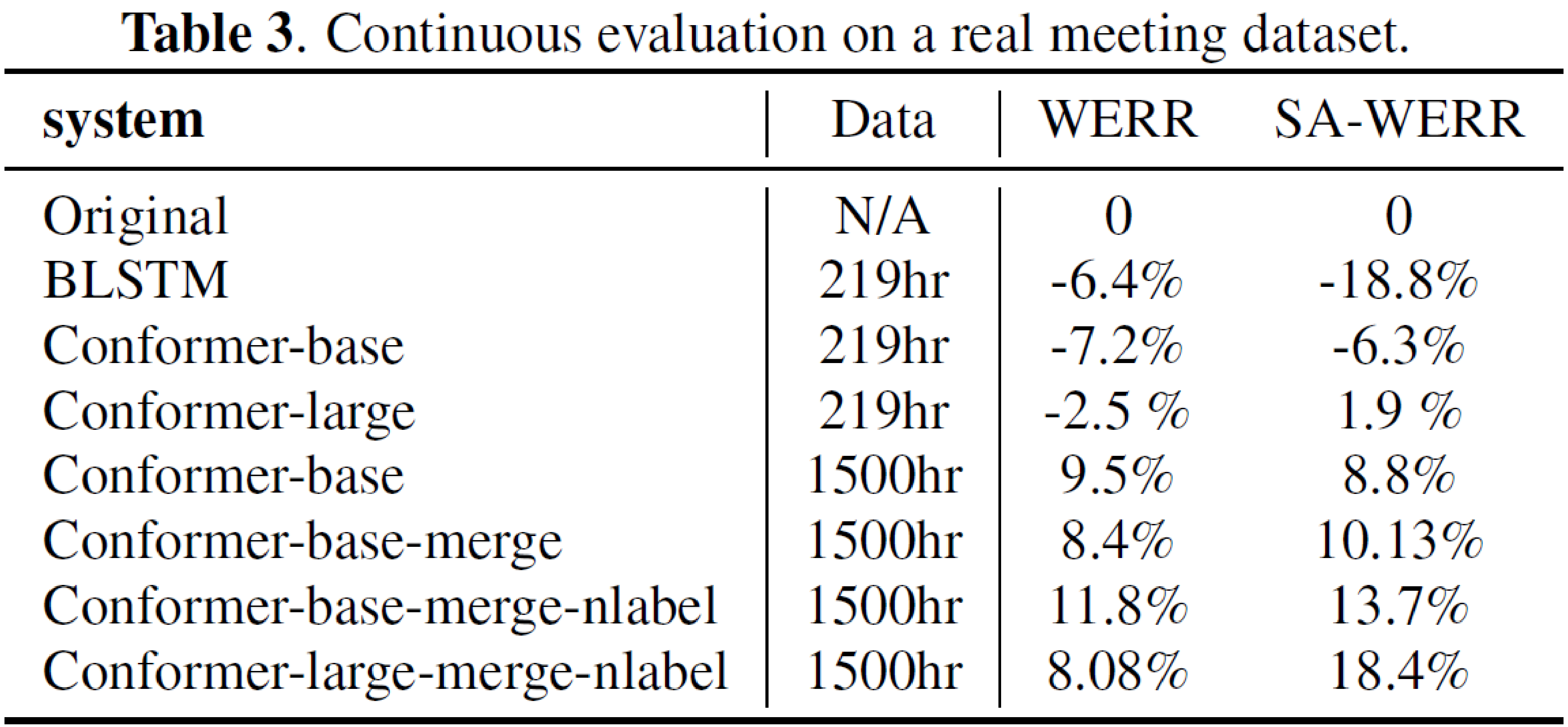


**Table 3**. Continuous evaluation on a real meeting dataset.

| **system** | Data | WERR | SA-WERR |
|---|---|---|---|
| Original | N/A | 0 | 0 |
| BLSTM | 219hr | -6.4% | -18.8% |
| Conformer-base | 219hr | -7.2% | -6.3% |
| Conformer-large | 219hr | -2.5 % | 1.9 % |
| Conformer-base | 1500hr | 9.5% | 8.8% |
| Conformer-base-merge | 1500hr | 8.4% | 10.13% |
| Conformer-base-merge-nlabel | 1500hr | 11.8% | 13.7% |
| Conformer-large-merge-nlabel | 1500hr | 8.08% | 18.4% |

# Conclusion

- We investigated the use of **Conformer** for continuous speech separation and achieve **the state of the art** on LibriCSS dataset for both the single-channel and multi-channel settings.
- We successfully achieve **significant gains** in the real meeting scenario, by introducing **several methods** (training data enlargement, mask merging scheme, and training target corruption)
- We explore **chunk-wise processing** to enable the Transformer family model to do streaming speech separation and enable the self-attention module to **consider the history information** beyond the current chunk with the previous state reused.

# Demo

What we've heard