# System Description of Causal TF-GridNet for Real-World Target Speaker Extraction

Team Name: **insta360**
Track: **Online Track**

Yayun Liang[1,2,†], Jiupeng Gao[2,†], Zhijian Jiang[1], Zelin Wang[1], Kai Chen[2,*]

[1]Insta360 Research, Insta360, Shenzhen 518101, China

[2]Key Laboratory of Modern Acoustics, Nanjing University, Nanjing 210093, China

## 1 Model Architecture

### 1.1 System Overview

Our submitted system is based on TF-GridNet [1] and is adapted for causal target speaker extraction in the Online track. The system takes a single-channel mixture waveform and an enrollment utterance of the target speaker as inputs, and estimates the target speaker signal in the complex time-frequency domain.

The overall system consists of three parts: (1) a pre-trained Res2Net speaker encoder for extracting the target speaker embedding, (2) a lightweight speaker conditioning module for injecting the target speaker information into the separator, and (3) a causal TF-GridNet separator for estimating the enhanced complex spectrum.

Given a mixture waveform $\mathbf{x}$ and a target-speaker enrollment waveform $\mathbf{e}$, the speaker encoder first extracts a speaker embedding

$$\mathbf{v} = \mathcal{E}_\phi(\mathbf{e}), \tag{1}$$

where $\mathcal{E}_\phi$ denotes the pre-trained speaker encoder. The separator then estimates the target-speaker complex STFT from the mixture STFT and the speaker embedding:

$$\hat{\mathbf{S}} = \mathcal{F}_\theta\left(\mathrm{STFT}(\mathbf{x}), \mathbf{v}\right), \qquad \hat{\mathbf{y}} = \mathrm{iSTFT}(\hat{\mathbf{S}}), \tag{2}$$

where $\mathcal{F}_\theta$ denotes the causal TF-GridNet separator and $\hat{\mathbf{y}}$ is the final enhanced waveform.

### 1.2 Speaker Encoder and Conditioning

We use a Res2Net-based speaker encoder pre-trained for speaker verification [2]. The enrollment utterance is converted into 80-dimensional log Mel-filterbank features with a 25 ms window and a 10 ms frame shift at 16 kHz. The pre-trained speaker encoder outputs a 192-dimensional speaker embedding. The speaker encoder is kept frozen during TSE training.

To condition the separator on the target speaker, the speaker embedding is projected and injected into the TF-GridNet backbone using a FiLM-style feature modulation strategy [3]. Specifically, the speaker embedding is mapped to a channel-frequency modulation tensor:

$$\mathbf{G} = \mathrm{LayerNorm}\left(\mathbf{W}_p\mathbf{v}\right), \tag{3}$$

where $\mathbf{G}$ has the same channel-frequency dimensions as the intermediate TF-GridNet features. The modulation is applied by element-wise multiplication:

$$\mathbf{H}' = \mathbf{H} \odot \mathbf{G}, \tag{4}$$

where $\mathbf{H}$ denotes the intermediate feature map and $\odot$ denotes element-wise multiplication with broadcasting along the time dimension. In our implementation, the speaker conditioning is applied at the second TF-GridNet block.

### 1.3 Causal TF-GridNet Separator

The separator follows the general TF-GridNet architecture [1], which performs speech separation in the complex time-frequency domain. Compared with the original non-causal TF-GridNet, we modify the temporal operations to satisfy the online processing requirement.

The main causal modifications are as follows.

- **Causal convolution.** For convolutional layers involving the time dimension, we use left padding only and remove right-side temporal padding. Therefore, the output at time frame $t$ only depends on the current and past input frames.
- **Causal temporal modeling.** The inter-frame temporal modeling module is changed from bidirectional processing to unidirectional processing. In particular, the temporal LSTM is implemented as a unidirectional LSTM so that no future frames are used.
- **Causal local attention.** For the local self-attention module, a causal attention mask is used along the time dimension. Each frame is allowed to attend only to the current and previous frames within the local context window.
- **Frequency-axis modeling.** The intra-frame frequency modeling module is kept bidirectional, because it operates only along the frequency axis within the same time frame and does not introduce future temporal information.

With these modifications, the separator can process the mixture in a causal manner. The model predicts the real and imaginary components of the target speaker spectrum, which are then converted back to the waveform domain using inverse STFT.

[†]These authors contributed equally to this work.
[*]Corresponding author: chenkai@nju.edu.cn.

### 1.4 Latency Setting

For the submitted Online-track system, the STFT is computed with a 32 ms Hann analysis window and a 16 ms frame shift. A fixed lookahead of $n_{\text{future}} = 3$ STFT frames is applied on top of the STFT analysis, corresponding to an additional $3 \times 16 = 48$ ms of delay. The total algorithmic latency of the system is therefore $32\,\text{ms}$ (STFT) $+$ $48\,\text{ms}$ (lookahead) $=$ $80\,\text{ms}$. In addition, the latency check performed using the official script provided by the challenge organizers reported an average latency of 68.5 ms.

This lookahead is implemented as a fixed output delay and is applied uniformly to all evaluation utterances. No utterance-specific latency adjustment or metric-driven post-processing is used during inference. All evaluation utterances are processed using the same fixed model checkpoint and the same inference configuration.

## 2 Training Data

The model was trained using datasets that are permitted by the challenge rules. The training data were used to construct mixture and target speech pairs for supervised target speaker extraction.

**Training data composition.** The primary training data consists of two-speaker mixtures synthesized in-house from **LibriSpeech** [4] with background noise drawn from **WHAM!** [5]. To improve speaker diversity, domain coverage, and generalization, we additionally incorporate four existing datasets using only their officially permitted subsets for the challenge: **AliMeeting** [6], **AMI** [7], **VoxCeleb** [8, 9], and **CN-Celeb2** [10, 11]. For these external corpora, near-field recordings are used as the target speech while far-field recordings serve as enrollment, simulating realistic enrollment conditions where the reference utterance is captured under different acoustic settings.

**On-the-fly mixture simulation.** During training, each sample is constructed on the fly: a 4.0-second segment is randomly cropped from the mixture and target waveforms at the same starting point to ensure temporal alignment. The enrollment utterance is independently cropped to 10.0 seconds from a random utterance of the target speaker (selected via the `spk2utt` mapping), which is always different from the target sentence.

### 2.1 Training Objective

The model was trained with a combination of time-domain and time-frequency-domain reconstruction losses. Given the estimated waveform $\hat{\mathbf{y}}$ and the ground-truth target waveform $\mathbf{y}$, the overall training objective is defined as

$$\mathcal{L} = \mathcal{L}_{\text{MR-STFT}} + \mathcal{L}_{\text{SI-SNR}} + \lambda_{\text{mel}}\mathcal{L}_{\text{Mel}}. \tag{5}$$

The multi-resolution STFT loss $\mathcal{L}_{\text{MR-STFT}}$ is computed at four window lengths, namely 160, 320, 960, and 1920 samples, corresponding to 10 ms, 20 ms, 60 ms, and 120 ms at 16 kHz. For each resolution, we use both magnitude reconstruction loss and complex real-imaginary reconstruction loss on power-compressed spectra. This loss encourages the enhanced speech to match the target speech across multiple spectral resolutions. The SI-SNR loss $\mathcal{L}_{\text{SI-SNR}}$ is used as a waveform-level objective to directly optimize time-domain signal quality. Near-silent segments are ignored during SI-SNR computation using an energy-based silence mask. In addition, we use an auxiliary log-Mel spectrogram loss $\mathcal{L}_{\text{Mel}}$ on 80-dimensional Mel-filterbank features. This term is implemented as an $\ell_1$ loss between the log-Mel spectrograms of the enhanced and target waveforms, and is used to stabilize training and improve perceptual spectral consistency. The weight of the log-Mel loss is set to $\lambda_{\text{mel}} = 0.3$. In our final system, the loss weights were fixed during training and were not tuned on the evaluation set.

## 3 Inference Procedure

During inference, the mixture waveform and the target-speaker enrollment/reference signal were fed into the fixed model. The model generated the estimated target-speaker waveform directly. No per-utterance candidate selection, adversarial perturbation, or metric-targeted waveform refinement was applied.

## 4 Post-processing

Model inference is performed on the raw mixture waveform without pre-processing. For the REAL-TSE evaluation set, which contains far-field recordings with notably lower signal levels than the synthetic training data, the input mixture is scaled to a peak amplitude of 0.5 before the STFT to mitigate the distribution mismatch. The enhanced output is subsequently peak-normalized to 0.9, matching the convention of the official baseline system. No other post-processing such as Wiener filtering, dynamic range compression, or per-utterance adaptation is applied.

## 5 Compliance Statement

We confirm that the submitted waveforms were generated by a fixed system version. The official evaluation models, scores, gradients, embeddings, or any other feedback were not used to optimize, tune, select among candidates, or post-process individual evaluation utterances.

The evaluation set was processed only by the fixed submitted system and was not used as data for repeated metric-driven optimization. Our submission does not include metric-targeted waveform refinement, adversarial perturbation, or candidate selection with respect to DNSMOS, speaker similarity, or any other official evaluation metric.

## References

[1] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, B.-Y. Kim, and S. Watanabe, "Tf-gridnet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 3221–3236, 2023.

[2] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 2, pp. 652–662, 2019.

[3] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville, "Film: Visual reasoning with a general conditioning layer," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[5] G. Wichern, J. Antognini, M. Flynn, L. R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in *Proc. Interspeech 2019*, 2019, pp. 1368–1372.

[6] F. Yu, S. Zhang, Y. Fu, L. Xie, S. Zheng, Z. Du, W. Huang, P. Guo, Z. Yan, B. Ma, X. Xu, and H. Bu, "M2MeT: The ICASSP 2022 multi-channel multi-party meeting transcription challenge," in *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6167–6171.

[7] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, I. McCowan, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus: A pre-announcement," in *Machine Learning for Multimodal Interaction*. Springer, 2006, pp. 28–39.

[8] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[9] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[10] L. Li, R. Liu, J. Kang, Y. Fan, H. Cui, Y. Cai, R. Vipperla, T. F. Zheng, and D. Wang, "CN-Celeb: Multi-genre speaker recognition," *Speech Communication*, vol. 137, pp. 77–91, 2022.

[11] Y. Fan, J. Kang, L. Li, K. Li, H. Chen, S. Cheng, P. Zhang, Z. Zhou, Y. Cai, and D. Wang, "CN-Celeb: A challenging chinese speaker recognition dataset," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7604–7608.