REAL-TSE Challenge Track 2 Submission

Han Han[1], Séverin Baroudi[3], Joonas Kalda[1], Mattias Di Bernardo[1], Antoine Laurent[1], Ricard Marxer[2,3], Hervé Bredin[1]

[1]pyannoteAI
[2]ILLS, CNRS, France
[3]Université de Toulon, Aix Marseille Univ, LIS, CNRS, France

***Abstract*—This report presents *pyannhance-tse*, a target speaker extraction system that isolates a speaker's voice from a multi-speaker monophonic mixture given a short enrollment recording. The system operates on 8-second mixture segments with 4-second enrollment audio, and is built around a deep attractors-based separation network guided by a pretrained end-to-end diarization model. We evaluate our submission on the four official challenge metrics: word error rate (WER) for speech intelligibility, speaker similarity between the enrollment and extracted audio, DNSMOS for perceptual quality, and F1 timing for target speaker activity accuracy.**

I. System Architecture

Our system architecture consists of four components: (1) a short-time Fourier transform (STFT) encoder-decoder that converts between waveforms and complex STFT coefficients, (2) a self-supervised learning (SSL) model pretrained with masked speech prediction for feature extraction, (3) a pretrained end-to-end diarization (EEND) model that estimates frame-level diarization, and (4) a deep attractors-based separation model that predicts the target speaker mask in the complex STFT domain.

A. Self-Supervised Speech Embedding

Self-supervised speech representations pretrained in a BERT-style fashion have gained traction for their effectiveness across various downstream audio tasks, including speech recognition, diarization and separation [1]–[4]. In our submitted systems, we adopt either WavLM or an extension of BestRQ as our SSL speech representation model. Building upon the original BestRQ framework and the denoising pretext task introduced in WavLM, we propose two advancements to learn speech representations that are more robust to additive noise and better suited for downstream multi-speaker speech tasks. We name this multi-talker optimized model SepRQ. First, we introduce an on-the-fly utterance mixing strategy that leverages a large corpus of single-speaker utterances. During pretraining, utterances are mixed with varying gain levels to simulate cocktail-party scenarios. The model is then trained to predict the pseudo-labels of the corresponding clean utterances. This denoising objective encourages the model to learn speaker-disentangled representations in the discrete latent space, thereby improving the robustness of the learned speech representations. Second, we jointly train the model with the same masked language modeling (MLM) objective used in HuBERT, applied directly to the synthetic mixtures. By learning both the contextual representations of multi-speaker mixtures and how to recover speaker-specific information through the denoising objective, the resulting representations are expected to be more effective for downstream target speech extraction and speech separation tasks.

B. End-to-End Diarization (EEND) Model

After the SSL model is trained, we inject it as a feature extraction backbone into an end-to-end local segmentation model as described in [5]. The model processes 20-second mixture audio chunks and derives frame-wise features from the pretrained SSL model by computing a weighted sum of the selected intermediate transformer layers. These features are then passed through 4 bidirectional LSTM layers with hidden size 128, followed by 2 linear layers with hidden size 128 and ReLU activation, and finally a powerset classifier head with sigmoid activation. We set the maximum number of local speakers to 3 per chunk and 2 per frame, and optimize the EEND model with the powerset multi-class cross-entropy loss [6]. The SSL model backbone is finetuned jointly with the EEND model during training.

For all SSL embeddings, we compute the learnable weighted sum of all intermediate transformer layers during the EEND training.

C. Separation Model

Inspired by the deep attractors network [7], our separation model derives the target speech mask from the scaled dot-product similarity between a learned speaker-specific attractor and the mixture embedding. First, the mixture STFT is split into sub-bands of size 16 and encoded by a strided Conv2D layer into a 256-dimensional embedding space, yielding a tensor of shape $16 \times \text{frames} \times 256$. Frame-level SSL features from the EEND backbone are projected via FiLM layer as a global conditioning signal for the mixture embedding. Simultaneously, we initialize a target speaker attractor vector of the same dimension 256, and refine it with 2 residual self-attention blocks. Then, the mixture embedding and attractor are jointly passed through 4 triple-path blocks, each comprising (i) a sub-band bidirectional LSTM (hidden size 128) operating along the frequency axis, (ii) a chunked (in time with chunk size of 125 frames) multi-head attention module where the

attractor is prepended as attention sink, and (iii) an inter-chunk bidirectional LSTM that propagates information across chunks. Finally after the last block, we compute the scaled dot-product similarity between the refined attractor and each time–frequency bin of the mixture embeddings, pass them through sigmoid gate and multiply with the mixture embedding. The final complex STFT speech mask is obtained via a linear projection head, and the target waveform is reconstructed by multiplying the learned speech mask onto the mixture STFT and decoding via iSTFT.

The target speaker attractor is initialized by fusing two complementary sources of speaker information. First, the enrollment audio is passed through the SSL backbone and mean-pooled over speech-active frames to yield an enrollment embedding. Second, the EEND model's frame-wise segmentation prediction is used to extract single-speaker speech segments from the mixture. Among these, we identify the mixture-derived speaker query by selecting the segment whose ResNet34 [8] embedding maximizes cosine similarity to the enrollment. The two SSL embeddings (from the mixture-derived segment and enrollment audio) are summed and L2-normalized to form the final attractor, which combines information from acoustic context of both the enrollment audio and the noisy mixture. We find empirically that this fusion method yields better separation than using enrollment embedding alone.

For all SepRQ-based experiments, we adopt a different strategy. We assume that, as suggested in [4], different tasks may benefit from representations extracted from different regions of the SSL architecture. Based on this assumption, rather than sharing a single SSL encoder between the enrollment branch and the EEND, we use two SSL encoders. The first is trained to process the enrollment and mixture signals, with its own learnable weighted average of the SSL layers, while the second is dedicated to the SepRQ module within the EEND. This design allows each task to exploit the SSL representations that are most beneficial for its objective.

II. Data Usage and Preparation

A. SSL Embedding Model

For all our results, we experiment with three SSL models :

- the already robust open source WavLM BASE+ model, which is trained on 94k hours of speech data including Libri-light, GigaSpeech and VoxPopuli.
- a BestRQ model that we specifically pretrain on the conversational corpus assembled in the following EEND section (II-B), using the same logic as the work done in [9].
- the SepRQ approach mentioned in section I-A, also pretrained on our conversational corpus.

B. EEND

The EEND model is trained and validated on a compound dataset containing the official training and development splits of [10], Alimeeting [11], AMI [12], AVA-AVD [13], MSD-WILD [14], RAMC [15], VoxConverse [16] and NOTSOFAR-1 [17].

C. Separation Model

The separation model is trained on a compound dataset of the entire otoSpeech [18], Scalable Spontaneous Speech Dataset (SSSD) [19], and a customized training split of the Apptek dataset [20]. All three datasets are full duplex conversational telephone speech with clean single-speaker speech in each channel. We process each dataset with FastMSS [21], an emulator of realistic multi-speaker conversations from a corpus of single-speaker utterances. This yields 260 hours of noisy, reverberated, 30-second synthetic conversational mixtures of 2 to 4 speakers. Specifically, the FastMSS system models multi-speaker speech as a 4-state HMM with its 4 transition types being turn hold, turn switch, interruption, and backchannel. We use the provided transition probabilities which are fitted to real conversational speech corpora. The separation model is validated on the customized development split of the Apptek dataset, comprising 9.5 hours of audio.

III. Training

A. Loss Function

We optimize our system against the weighted sum of 6 loss functions:

- **Multi-resolution STFT and time domain loss:**

$$L_{\mathrm{MSS}}(x,\tilde{x}) = \frac{0.5}{3}\sum_{k=8}^{10}\left\|\Phi^{k}_{\mathrm{STFT}}(x) - \alpha\,\Phi^{k}_{\mathrm{STFT}}(\tilde{x})\right\|_1 + \frac{0.5}{T}\sum_{t=1}^{T}\left|x_t - \alpha\,\tilde{x}_t\right| \quad (1)$$

  where Φ^{k}_{STFT} denotes the STFT magnitude computed with window size 2^k, and $\alpha = \sum_t (x_t\tilde{x}_t)/(\sum_t \tilde{x}_t^2 + \epsilon)$.
- **Scale-dependent signal-to-distortion ratio loss (SD-SDR):** defined between the target and estimated audio.
- **Speaker similarity loss:** uses a ResNet34 speaker embedding model [8] and is defined as $L_{\mathrm{spksim}}(x_e,\tilde{x}) = 1 - (\Phi_{\mathrm{R34}}(x_e)\cdot\Phi_{\mathrm{R34}}(\tilde{x}))/(\|\Phi_{\mathrm{R34}}(x_e)\|\cdot\|\Phi_{\mathrm{R34}}(\tilde{x})\|)$, where x_e is the enrollment audio.
- **ASR loss:**

$$L_{\mathrm{asr}}(x,\tilde{x}) = \sum_{n=1}^{N}\left(\frac{\Phi_{\mathrm{asr}}(x)_n}{\|\Phi_{\mathrm{asr}}(x)_n\|} - \frac{\Phi_{\mathrm{asr}}(\tilde{x})_n}{\|\Phi_{\mathrm{asr}}(\tilde{x})_n\|}\right) \quad (2)$$

  where Φ_{asr} is the Whisper-tiny [22] encoder. As Whisper-tiny operates on fixed 30-second inputs, we zero-pad the 8-second signals and retain only the corresponding frames of the output embeddings.
- **Speaker activity loss:** penalizes energy in speaker-inactive regions of the estimate,

$$L_{\mathrm{sa}}(x,\tilde{x}) = \frac{\sum_{t,f}|\tilde{X}_{t,f}|\,(1-y_t)}{\sum_{t,f}|M_{t,f}| + \epsilon} \quad (3)$$

TABLE I
COMPARISON OF SUBMITTED MODELS' METRICS ON THE PROVIDED DEV SET

Loss	SSL	TER↓	SIM↑	SIG↑	BAK↑	DNSMOS-OVRL↑	Precision↑	Recall↑	F1↑
Baseline	–	0.70	**0.52**	1.90	1.66	1.50	0.78	**0.95**	0.84
SISDR	WavLM	0.57	0.51	2.35	3.15	1.88	0.84	0.92	0.86
SISDR	SepRQ	**0.55**	0.49	2.45	**3.34**	1.94	0.85	0.92	**0.87**
SDSDR	SepRQ	**0.55**	0.49	**2.47**	3.20	**1.96**	**0.86**	0.91	**0.87**

where $y \in \{0,1\}^T$ is the ground truth speaker activity, and $\tilde{X}$, M are the STFTs of the estimate and mixture.

- **Deep feature loss:**

$$L_{\text{DFL}}(x,\tilde{x}) = \frac{1}{N}\sum_{n=1}^{N} |\Phi_{\text{SSL}}(x)_n - \Phi_{\text{SSL}}(\tilde{x})_n| \quad (4)$$

where Φ_{SSL} is the weighted sum of all intermediate layers of WavLM or BestRQ embeddings, then averaged in time, following the same implementation used in the EEND model.

The above 6 losses are weighted to constrain their magnitudes to roughly between 0.1 and 1. The final loss is defined as:

$$\begin{aligned} L_{\text{total}} = 10\, L_{\text{MSS}} + 0.1\, L_{\text{SDSDR}} + L_{\text{spksim}} \\ + 1000\, L_{\text{asr}} + L_{\text{sa}} + 0.5\, L_{\text{DFL}} \end{aligned} \quad (5)$$

B. Optimization

For the SSL models, we use the same training procedure as described in [23], using 4 H100s for a total of 180k/300k training steps for BestRQ/SepRQ respectively. All learning rates scheduler consist of a linear warmup, for 8% of the total training time, followed by a cosine decay. As our conversational dataset contains long recordings (such as the AMI dataset), we chunk each audio files into 10 seconds long extracts.

For the EEND model, we train with AdamW optimizer, where the SSL component is finetuned at a lower learning rate of $1e-5$ while the remaining weights at $1e-3$. We use a batch size of 16 and train the EEND model for 25 epochs on a single H100 gpu.

For the separation model, we freeze the EEND model and train with AdamW and Muon optimizer [24], an optimizer that orthogonalizes the momentum-accumulated gradient update. Specifically, we partition the trainable weights into hidden layer weight and classifier weights according to their dimensionality. We train the hidden layer weights of two or more dimensionalities with Muon optimizer at a learning rate of $5e-3$ and the rest with AdamW at $5e-4$ learning rate and $1e-2$ weight decay. We train the separation model for 13 epochs (20 hrs) with a batch size of 32 on a single H100 gpu.

IV. RESULTS

A. EEND models

As our Target Speech Extraction (TSE) models require a strong EEND from which to extract the single speaker segments out of the mixture during training, we evaluate the performance of our different EEND backends, that are latter leveraged on each TSE models.

TABLE II
LOCAL AVERAGED DIARIZATION ERROR RATE (DER) RESULTS (%) FOR EACH SSL BACKBONE ON THE ASSEMBLED CONVERSATIONAL DATASET DEVELOPMENT SET. FA = FALSE ALARM, MISS = MISSED SPEECH, AND SC = SPEAKER CONFUSION.

Model	DER (%) ↓	FA (%)	MISS (%)	SC (%)
WavLM	13.3	4.4	6.5	2.4
BestRQ	12.5	4.7	4.2	3.5
SepRQ	**11.9**	4.1	4.7	3.1

As displayed in Table II, BestRQ shows a relative improvement of approximately 6.0% in DER over WavLM (13.3 $\rightarrow$ 12.5), demonstrating a clear benefit in using conversational datasets during SSL pretraining. This gain is noticeable through the MISS errors (6.5 $\rightarrow$ 4.2, 35% relative reduction), indicating improved speaker coverage, although it comes at the cost of slightly increased FA and SC errors.

Furthermore, SepRQ improves upon this previous result by a relative 4.8% DER reduction over BestRQ, achieving the best overall performance, showing the benefit of synthetic mixture pretraining and better pretraining losses.

Overall, BestRQ primarily improves recall-oriented errors (MISS), while SepRQ further refines segmentation quality by reducing false alarms and speaker confusions, leading to the best overall DER. For our TSE systems, we put BestRQ aside, in favor of SepRQ and the more common WavLM alternative.

B. Target Speech Extraction models

During inference time, we apply 8-second sliding window with 4-second stride size on the input mixture, extract chunk-wise target speaker audio, and reconstruct the final audio via overlap-and-add. We report the metrics on the DEV set provided by the challenge organizers, which contains samples from AISHELL-4, Alimeeting, DipCo [25] and CHIME6 [26] datasets.

ACKNOWLEDGMENT

The authors would like to acknowledge Samuele Cornell for his contribution to the early stage conceptualization and development of the system architecture, data sampling and loss objectives.

This work was granted access to the HPC resources of IDRIS under the allocations AD011016176, A0201012177, A0191014044, AD011017973, AD011016519 and AD011014274 made by GENCI.

REFERENCES

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *CoRR*, 2020.

[2] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, J. Wu, M. Zeng, and F. Wei, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *CoRR*, 2021.

[3] C. Chiu, J. Qin, Y. Zhang, J. Yu, and Y. Wu, "Self-supervised learning with random-projection quantizer for speech recognition," *CoRR*, 2022.

[4] S. Baroudi, H. Bredin, J. Razik, and R. Marxer, "On the use of self-supervised representation learning for speaker diarization and separation," in *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2025, pp. 1–7.

[5] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," 2021.

[6] A. Plaquet and H. Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *INTERSPEECH 2023*. ISCA, Aug. 2023, p. 3222–3226.

[7] Z. Chen, Y. Luo, and N. Mesgarani, "Deep attractor network for single-microphone speaker separation," *CoRR*, vol. abs/1611.08930, 2016.

[8] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023, pp. 1983–1987.

[9] S. Baroudi, T. Pellegrini, and H. Bredin, "Specializing Self-Supervised Speech Representations for Speaker Segmentation," in *Interspeech 2024*, 2024, pp. 3769–3773.

[10] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, "AISHELL-4: An Open Source Dataset for Speech Enhancement, Separation, Recognition and Speaker Diarization in Conference Scenario," in *Interspeech 2021*, 2021, pp. 3665–3669.

[11] 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 *Proc. ICASSP*. IEEE, 2022.

[12] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," in *Proceedings of Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research*, L. Noldus, F. Grieco, L. Loijens, and P. Zimmerman, Eds. Noldus Information Technology, Aug. 2005, pp. 137–140.

[13] E. Z. Xu, Z. Song, S. Tsutsui, C. Feng, M. Ye, and M. Z. Shou, "Ava-avd: Audio-visual speaker diarization in the wild," in *Proceedings of the 30th ACM International Conference on Multimedia*, ser. MM '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 3838–3847. [Online]. Available: https://doi.org/10.1145/3503161.3548027

[14] T. Liu, S. Fan, X. Xiang, H. Song, S. Lin, J. Sun, T. Han, S. Chen, B. Yao, S. Liu, Y. Wu, Y. Qian, and K. Yu, "MSDWild: Multi-modal Speaker Diarization Dataset in the Wild," in *Interspeech 2022*, 2022, pp. 1476–1480.

[15] Z. Yang, Y. Chen, L. Luo, R. Yang, L. Ye, G. Cheng, J. Xu, Y. Jin, Q. Zhang, P. Zhang, L. Xie, and Y. Yan, "Open Source MagicData-RAMC: A Rich Annotated Mandarin Conversational(RAMC) Speech Dataset," in *Interspeech 2022*, 2022, pp. 1736–1740.

[16] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the conversation: speaker diarisation in the wild," 2020.

[17] A. Vinnikov, A. Ivry, A. Hurvitz, I. Abramovski, S. Koubi, I. Gurvich, S. Peer, X. Xiao, B. M. Elizalde, N. Kanda, X. Wang, S. Shaer, S. Yagev, Y. Asher, S. Sivasankaran, Y. Gong, M. Tang, H. Wang, and E. Krupka, "Notsofar-1 challenge: New datasets, baseline, and tasks for distant meeting transcription," in *Interspeech 2024*, 2024, pp. 5003–5007.

[18] otoearth, "otospeech-full-duplex-processed-141h: Full-duplex conversational speech dataset," https://huggingface.co/datasets/otoearth/otoSpeech-full-duplex-processed-141h, 2026, license: CC BY 4.0.

[19] Z. Sheikh, S. Shimizu, S. Arora, J. Shi, S. Cornell, X. Li, and S. Watanabe, "Scalable spontaneous speech dataset (sssd): Crowdsourcing data collection to promote dialogue research," in *INTERSPEECH 2025*, 2025.

[20] E. Beck, S. Beranek, U. Moothiringote, D. Mann, W. Michel, K. Nguyen, and T. Tragemann, "Apptek call-center dialogues: A multi-accent long-form benchmark for english asr," 2026.

[21] A. Polok, I. Medennikov, S. Watanabe, J. Černocký, L. Burget, and S. Cornell, "Mind the gap: Impact of synthetic conversational data on multi-talker ASR and speaker diarization," 2026, submitted to INTERSPEECH.

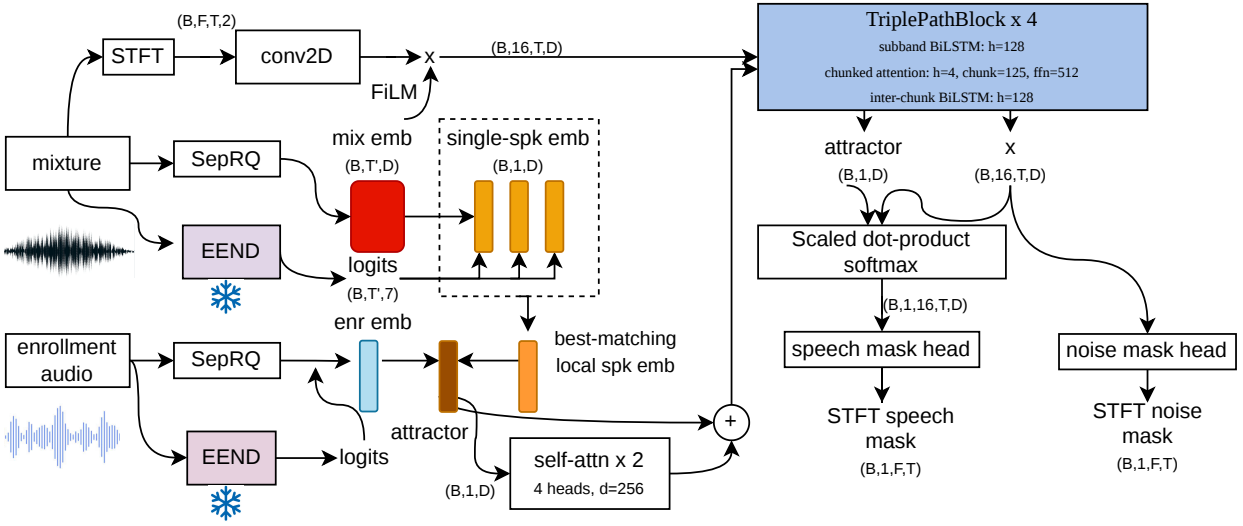
[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[23] R. Whetten, T. Parcollet, M. Dinarelli, and Y. Estève, "Implémentation ouverte et étude de BEST-RQ pour le traitement de la parole," in *Actes des 35èmes Journées d'Études sur la Parole*, M. Balaguer, N. Bendahman, L.-M. Ho-dac, J. Mauclair, J. G Moreno, and J. Pinquier, Eds. Toulouse, France: ATALA and AFPC, 7 2024, pp. 412–420. [Online]. Available: https://aclanthology.org/2024.jeptalnrecital-jep.42/

[24] K. Jordan, Y. Jin, V. Boza, Y. Jiacheng, F. Cesista, L. Newhouse, and J. Bernstein, "Muon: An optimizer for hidden layers in neural networks," 2024.

[25] M. Van Segbroeck, A. Zaid, K. Kutsenko, C. Huerta, T. Nguyen, X. Luo, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, "Dipco—dinner party corpus," *Proc. Interspeech 2020*, pp. 434–436, 2020.

[26] S. Watanabe, M. Mandel, J. Barker, E. Vincent, A. Arora, X. Chang, S. Khudanpur, V. Manohar, D. Povey, D. Raj *et al.*, "Chime-6 challenge: Tackling multispeaker speech recognition for unsegmented recordings," in *CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments*, 2020.


(B,F,T,2)
STFT
conv2D
x
(B,16,T,D)
FiLM
TriplePathBlock x 4
subband BiLSTM: h=128
chunked attention: h=4, chunk=125, ffn=512
inter-chunk BiLSTM: h=128
mixture
SepRQ
mix emb
(B,T',D)
single-spk emb
(B,1,D)
attractor
x
(B,1,D)
(B,16,T,D)
EEND
logits
(B,T',7)
Scaled dot-product softmax
(B,1,16,T,D)
enrollment audio
SepRQ
enr emb
best-matching local spk emb
speech mask head
noise mask head
STFT speech mask
(B,1,F,T)
STFT noise mask
(B,1,F,T)
EEND
logits
attractor
(B,1,D)
self-attn x 2
4 heads, d=256
+