A Three-Stage System for CHiME-9 ECHI: Self-Interference Suppression, Target Speaker Extraction, and Post-Processing

Fei Zhao[1]* Changjiang Zhao[2] Zhenlong Guo[1]* Wenzheng Zhang[1]* Yongjie Yan[2] Xueliang Zhang[1]†

[1] Inner Mongolia University, Hohhot, China · [2] Elevoc, Shenzhen, China · zhaofei@mail.imu.edu.cn

01 Motivation

ECHI recovers clean target-speaker speech from 7-ch Aria glasses or 4-ch hearing aids.

Key challenges: wearer self-interference and anchor–target mismatch.

02 System Overview

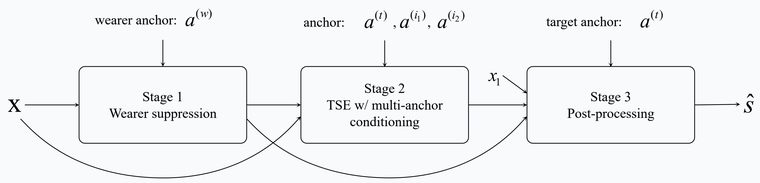

Fig. 1 · Three-stage cascade pipeline.

03 Pseudo-Label Generation

Let $s^{(k)}$ and $h^{(k)}$ $(k \in \mathcal{K} = \{t, w, i_1, i_2\})$ denote the close-talk signal and acoustic transfer function of the target speaker (t), the wearer (w), and two interfering speakers (i_1, i_2), respectively.

$$x_1[n] = \sum_{k \in \mathcal{K}} \left(s^{(k)} * h^{(k)}\right)[n].$$

Estimate per-speaker FIR filters in STFT domain via regularized least-squares. $X_k = \mathrm{STFT}(s^{(k)})$ $Y = \mathrm{STFT}(x_1)$

① Per-frequency delay vectors

$$\mathbf{x}_{k,f,t} = \left[X_k(f,t),\ \ldots,\ X_k(f,t-N+1)\right]^\top \in \mathbb{C}^N$$

② Stack all speakers & collect all frames

$$\mathbf{x}_{f,t} = [\mathbf{x}_{1,f,t}^\top, \ldots, \mathbf{x}_{4,f,t}^\top]^\top \in \mathbb{C}^{4N}$$

$$\mathbf{X}_f \in \mathbb{C}^{T \times 4N} \quad \mathbf{y}_f = [Y(f,1), \ldots, Y(f,T)]^\top \in \mathbb{C}^T$$

③ Regularized normal equations

$$\hat{\mathbf{w}}_f = \left(\mathbf{X}_f^H \mathbf{X}_f + \varepsilon\, \mathbf{I}_{4N}\right)^{-1} \mathbf{X}_f^H\, \mathbf{y}_f$$

④ Component estimate via ISTFT

$$\tilde{x}_1^{(k)}[n] = \mathrm{ISTFT}\left\{\mathbf{x}_{k,f,t}^\top\, \hat{\mathbf{w}}_{k,f}\right\}.$$

04 Network Architecture

Shared **TF-GridNet** backbone (causal, streaming). **FiLM** injects local cue + frozen **ECAPA-TDNN** global identity.

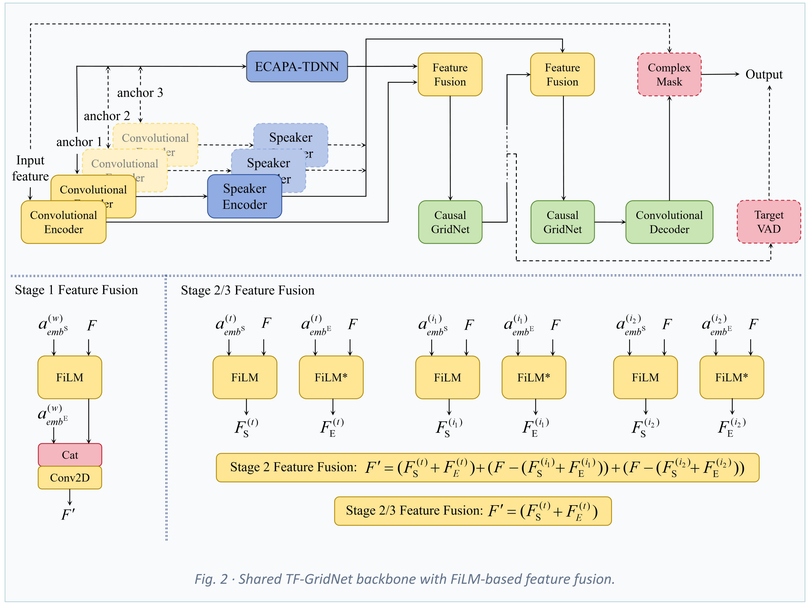

Fig. 2 · Shared TF-GridNet backbone with FiLM-based feature fusion.

05 Stage-Wise Details

STAGE 1 Wearer Suppression

Anchor: wearer $a^{(w)}$. **Output:** mask on x_1 → wearer-removed signal. **Backbone:** 2-layer.

STAGE 2 Target Extraction (Multi-Anchor)

Anchors: $a^{(t)}$, $a^{(i1)}$, $a^{(i2)}$. First two fusion blocks use contrastive

$$F' = (F_S^{(t)} + F_E^{(t)}) + (F - (F_S^{(i_1)} + F_E^{(i_1)})) + (F - (F_S^{(i_2)} + F_E^{(i_2)}))$$

Last four fusion blocks use target information:

$$F' = (F_S^{(t)} + F_E^{(t)})$$

Output: mask on x_1 → target signal, soft VAD. **Backbone:** 6-layer.

STAGE 3 Post-Processing & Inference Fallback

Maps **[x_1 ; ŝ_S1 ; ŝ_S2]** → close-talk target. **Backbone:** 2-layer.

VAD-gated fallback at inference:

$$\hat{S} = \hat{S}_{\mathrm{ct}} \odot M_{\mathrm{vad}} + \hat{R} \odot (1 - M_{\mathrm{vad}})$$

06 Training Objectives & Setup

- **Stage 1:** $\mathcal{L}_1 = \mathcal{L}$ stftm + $\lambda \cdot \mathcal{L}$ SI-MSE, $\lambda = 0.5$

$$\mathcal{L}_{\mathrm{stftm}} = \mathbb{E}\left[|A(\hat{s}) - A(s)|\right],$$
$$A(x) = |\Re\{\mathrm{STFT}(x)\}| + |\Im\{\mathrm{STFT}(x)\}|.$$

- **Stage 2:** $\mathcal{L}_2 = \mathcal{L}_1 + \alpha \cdot \mathcal{L}$ WavLM + $\beta \cdot \mathcal{L}$ vad, $\alpha = 0.1$, $\beta = 1$

$$\mathcal{L}_{\mathrm{WavLM}} = \frac{1}{L} \sum_{\ell=1}^{L} \mathbb{E}\left[\left\|\mathbf{h}^{(\ell)}(s) - \mathbf{h}^{(\ell)}(\hat{s})\right\|_2^2\right]$$

$$\mathcal{L}_{\mathrm{vad}} = \mathrm{BCEWithLogits}(\hat{v}, v).$$

- **Stage 3:** $\mathcal{L}_1 + \alpha \cdot \mathcal{L}$ WavLM
- **STFT:** win 512, hop 128, asymmetric windows → 1-hop latency
- **Optim:** Adam, lr 5e-4

07 Results (Aria, Dev · Summed)

System	FWSSNR	STOI	PESQ	Csig	Cbak	Covl
Passthrough	1.06	0.47	1.11	1.55	1.35	1.23
Baseline	2.34	0.50	1.16	1.56	1.24	1.26
2S (WS+TSE)	1.75	0.53	1.21	1.75	1.47	1.37
2S + VAD	2.22	0.54	1.21	1.74	1.49	1.37
3S (PP)	4.17	0.51	1.19	1.51	1.54	1.29
3S (PP+VAD)	3.95	0.46	1.20	1.50	1.52	1.27
3S (PP+VAD+FB) ★	**1.62**	**0.54**	**1.13**	**1.83**	**1.23**	**1.38**

★ *Proposed final configuration (highlighted).*

08 Challenge Test Set (Aria)

Intel.	SIG	BAK	OVRL
48.43%	**3.56**	**1.81**	**2.28**
vs 55.28%	*vs 2.31*	*vs 2.75*	*vs 2.16*
	+1.25		**+0.12**

Strong perceptual-quality gain (SIG +1.25, OVRL +0.12); BAK / intelligibility traded off.

09 Takeaways

- **Cascaded three-stage design** for hearable TSE.
- **Pseudo labels via linear filtering** — no clean refs needed.
- **VAD-gated fallback** balances SI reduction with continuity.
- **Future:** close anchor–target style/emotion gap.