TRAINING LOW-LATENCY TARGET SPEECH EXTRACTION FOR WEARABLE DEVICES USING PHASE- AND AMPLITUDE-ALIGNED DATA IN THE CHIME-9 ECHI TASK

Dengxiang Hu[1]*, Tomohiro Nakatani*[2]*, Naoyuki Kamo*[2]*, Marc Delcorix*[2]*, Tsubasa Ochiai*[2]*, Shoji Makino*[1]

[1]Waseda University, Japan [2]NTT, Inc., Japan

ABSTRACT

This paper presents a low-latency Target Speech Extraction (TSE) system developed for the ECHI task of the CHiME-9 challenge. The system extends the challenge baseline, which provides a TSE framework combining a causal GridNet with Feature-wise Linear Modulation (FiLM)-based target speaker conditioning. The proposed extensions include: 1) Close-to-Distant microphone (C2D) projection-based reference signal generation, which enables the use of a fine-grained objective, the Scale-Invariant Signal-to-Noise Ratio (SI-SNR), for Neural Network (NN) training, 2) the introduction of Deterministic Recursive Enhancement (DRE), which enables progressive refinement of the extracted target speech, and 3) the use of Null Beamforming (NBF) to suppress the wearer's voice, thereby facilitating more effective extraction of other speakers' voices. Experiments demonstrate that the proposed extensions significantly improve the baseline performance.

***Index Terms*—** Target speech extraction, wearable device, low latency algorithm, neural network, beamforming

1. INTRODUCTION

This paper describes a low-latency Target Speech Extraction (TSE) system that we developed for the ECHI task of CHiME-9 challenge [1] [2]. The challenge considers a recording scenario in which four speakers are seated around a table and their conversations are captured under noisy conditions using a wearable device worn by one of the speakers, referred to as the wearer. The goal is to selectively extract the voice of each speaker other than the wearer, referred to as a target speaker, by low-latency processing ($<$ 20 ms delay), thereby enabling the wearer to hear the voice more clearly and intelligibly in real time. This task of selectively extracting each target speaker's voice is referred to as TSE.

To support TSE in the challenge, an enrollment utterance for each speaker is provided a priori to capture the speaker's voice characteristics, and the device supplies multi-microphone recordings that enable spatial speaker discrimination. In such a setting, neural networks (NNs) and beamforming (BF) are two key techniques for TSE. NN-based approaches exploit speaker-specific characteristics derived from the enrollment utterances to perform TSE [3, 4]. BF approaches estimate the spatial characteristics of the target speech and interfering signals from the multi-microphone recordings and isolate the target through linear filtering [5, 6].

Despite the availability of these techniques, serious issues remain for the challenge. 1) Because the conversations are recorded in real, noisy environments, it is difficult to obtain clean reference signals that can serve as training targets for NNs. In the challenge, denoised headset recordings of each target speaker are provided as references; however, their phase and amplitude are not aligned with those of the wearable-device recordings, which limits the applicability of fine-grained NN training objectives. 2) The spatial characteristics of the target speech change dynamically and drastically due to the wearer's head movements, making the application of BF difficult. 3) Under the extremely low signal-to-noise ratio conditions of the recordings, achieving stable and accurate TSE is difficult. 4) The wearer's voice appears at a significantly higher level than the target speakers' voices in the recordings, further complicating TSE.

To address these issues, we extend the baseline TSE system [4,7–9] provided for the challenge. The baseline offers a simple TSE framework that combines a causal GridNet [4] with Feature-wise Linear Modulation (FiLM)-based target-speaker conditioning [10]. Our extension introduces several key components. First, a Close-to-Distant microphone (C2D) projection [11] generates reference signals whose phase and amplitude are aligned with the observations, enabling the use of an effective fine-grained NN training objective, Scale-Invariant Signal-to-Noise Ratio (SI-SNR) [12]. Second, Deterministic Recursive Enhancement (DRE) [11, 13] is introduced, which iteratively applies the baseline with additional recursive conditioning, enabling progressive improvement of the enhanced signals. Third, null beamforming (NBF) [14, 15] is employed to suppress the wearer's voice and used to initialize the recursive conditioning. Finally, simulated observations are generated and used for system pre-training.

Experiments using the challenge Aria device show that the proposed system significantly outperforms the baseline on multiple objective and subjective metrics across both the development (dev) and evaluation (eval) sets.

2. SYSTEM DESCRIPTION

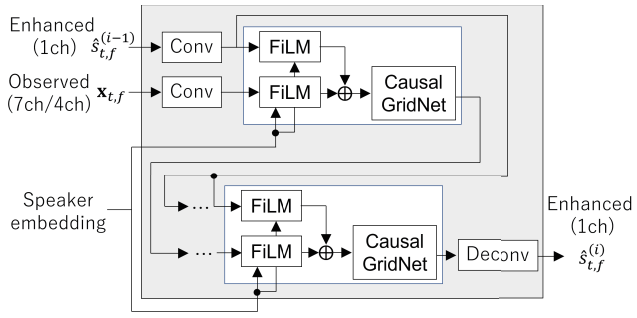
This section provides a detailed description of our proposed system. Section 2.1 defines the system architecture. The training and inference schemes are presented in Section 2.2 and 2.3, respectively.

In the following, $\mathbf{x}_{t,f} \in \mathbb{C}^M$ and $\hat{s}_{t,f} \in \mathbb{C}$ denote the multi-channel observed signal captured by M microphones and a single-channel (1-ch) enhanced signal, respectively, in the Short-Time Fourier Transform (STFT) domain, where t and f denote indices of time frames and frequencies.

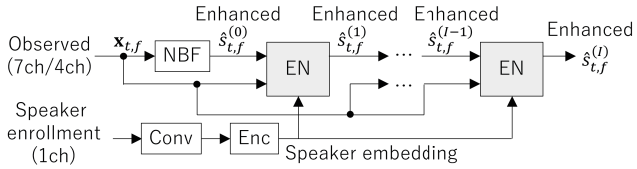
2.1. Deterministic Recursive Enhancement (DRE)

Figure 1(a) illustrates the fundamental processing unit of the proposed system, referred to as Enhancement Network (EN), which can itself perform TSE. EN takes a multi-channel observed signal $\mathbf{x}_{t,f}$ and an initial 1-ch enhanced signal $\hat{s}_{t,f}^{(i-1)}$ as inputs, and outputs a refined 1-ch enhanced signal $\hat{s}_{t,f}^{(i)}$, where i denotes the recursion index, as explained later. To condition the TSE on characteristics of a target speaker's voice, EN also receives a speaker embedding extracted from the target speaker's enrollment utterance.

We adopt an EN architecture that is nearly identical to the challenge baseline. As in the baseline, a 2D convolution (Conv) unit first extracts an observed feature vector from the observed signal. A FiLM unit [10] fuses this vector with the speaker embedding, which



(a) Enhancement Network (EN)



(b) Deterministic Recursive Enhancement (DRE)

Fig. 1. Processing flow of the proposed system.

is extracted using the same speaker encoder as the baseline, and then a causal GridNet [4] processes the fused representation to produce an enhanced feature vector. By iteratively applying this same processing, the enhanced feature vector is refined and is finally decoded into the enhanced signal by a 2D deconvolution (Deconv) unit.

A key extension beyond the baseline is the use of the 1-ch enhanced signal as an additional conditioning input. For this extension, a Conv unit extracts an enhanced feature vector from this input, which is then fused with the speaker embedding by a FiLM unit. The resulting fused vector is summed with the observed feature vector using learnable mixing weights and fed into the causal GridNet. This processing is also repeated in subsequent iterations.

To achieve improved TSE performance over the baseline, the proposed system incorporates EN into DRE [11, 13], as illustrated in Fig. 1(b). With DRE, NBF is first applied to the observed signal to obtain an initial 1-ch enhanced signal $\hat{s}_{t,f}^{(0)}$, in which the wearer's voice is suppressed (see also Section 2.1.1). DRE then recursively applies EN to the observed signal for recursion indices $i = 1, \ldots, I$, using the 1-ch enhanced signal $\hat{s}_{t,f}^{(i-1)}$, obtained either from NBF or from the previous EN application, as an additional input. This recursive procedure enables progressive refinement of the enhanced signal. Since all EN instances in the recursion share the same model parameters, DRE achieves this improvement without increasing the overall model size.

DRE is a simplified variant of Probabilistic-DRE (PDRE) [13]. Whereas PDRE estimates both the enhanced speech and its distribution, DRE estimates only the enhanced speech [11]. We adopt DRE in the proposed system due to its simplicity and its ability to deliver performance comparable to PDRE. We note that recursively applying a NN-based SE model trained for single-pass estimation (i.e., no recursion) has been explored previously [16]. In contrast, DRE additionally provides an efficient NN training scheme (see also Section 2.2.1) that is explicitly designed for recursive enhancement.

2.1.1. NBF for cancelling the wearer's voice

Because the geometrical relationship between the wearer's mouth and the microphones of the wearable device is almost fixed and does not significantly differ over different speakers, we assume that the spatial characteristics of the wearer's voice are stationary over all recordings. Under this assumption, we propose to use NBF with fixed filter coefficients to cancel the wearer's voice from the observed signal. As the wearer's voice is a major obstacle for the TSE in the challenge, the NBF's output can be a powerful initial estimate of the 1-ch enhanced signal for the DRE-based TSE.

The proposed system determined the fixed NBF coefficients in advance using the training dataset and its speaker activity labels provided for the challenge. Based on the speaker activity labels, we first extract time frames T^{W} and T^{O}, respectively, containing only the wearer's voice and the other speakers' voices, and estimate the steering vector $\mathbf{h}_f$ of the wearer's voice at each frequency f based on the covariance whitening technique [17]. Then, the coefficients of NBF is determined and applied to the observed signal:

$$\mathbf{W}_f = \mathbf{I} - (1 - \epsilon)\mathbf{h}_f \mathbf{h}_f^{\mathsf{H}} / \|\mathbf{h}_f\|_2^2, \tag{1}$$

$$\hat{\mathbf{x}}_{t,f} = \mathbf{W}_f^{\mathsf{H}} \mathbf{x}_{t,f}, \tag{2}$$

where $\hat{\mathbf{x}}_{t,f} \in \mathbb{C}^M$ is the NBF output, $\mathbf{I} \in \mathbb{R}^{M \times M}$ is the identity matrix, $(\cdot)^{\mathsf{H}}$ denotes the conjugate transpose, $\|\cdot\|_2$ denotes the Euclidean norm, and $\epsilon = 0.1$ is a flooring constant.

2.1.2. Low-latency processing with DRE

Since the baseline supports causal processing, the proposed system, which recursively applies the baseline, also operates causally. However, the Conv and Deconv units of the baseline use kernels that access one future frame, and thus each application of these units introduces additional latency equal to the frame shift of the STFT analysis window. The baseline employs an 8 ms analysis window with a 4 ms frame shift, yielding a total latency of 16 ms (= 8 ms + two 4 ms future-frame accesses by the Conv and Deconv units). Therefore, increasing the recursion order of DRE beyond one further increases the latency, which does not satisfy the challenge requirements.

To solve this problem, we modify the kernel design from the baseline [18]. Specifically, we configure the kernels to avoid accessing future frames and instead allow access to two past frames. Importantly, this modification enables us to increase the analysis window size up to the maximum allowable latency, i.e., 20 ms while complying with the challenge regulations. As a result, the proposed system can exploit a comparable amount of context to that of the baseline.

2.2. Training of network

2.2.1. Training scheme for DRE

As a general training scheme for DRE, we train EN to produce a refined estimate at each recursion step according to a specified training objective. Let $\mathbf{s}_j$ denote the reference signal for the j-th utterance in the training set, and let $\hat{\mathbf{s}}_j^{(i)}$ denote the corresponding enhanced signal obtained at the i-th recursion. Then, the parameter set θ of EN is optimized by minimizing the following cost function:

$$\mathcal{J}(\theta) = \sum_{i,j} \mu_i \mathcal{L}(\mathbf{s}_j, \hat{\mathbf{s}}_j^{(i)}; \theta), \tag{3}$$

where $\mathcal{L}(\cdot)$ denotes the training objective, and μ_i is a weight that balances the contribution of each recursion step. Under this formulation, gradients of θ are computed independently at each recursion step and accumulated across steps, while inter-step dependencies are ignored. This approximation avoids the enormous memory requirements associated with backpropagation through recursion steps, while still enabling EN to learn effective progressive refinement.

2.2.2. Generation of aligned training references

The C2D projection is a technique for estimating clean source images of target speech at distant microphones from noisy speech mixtures recorded by both close and distant microphones. The resulting source images can be effectively used as reference signals that are phase- and amplitude-aligned with the distant microphone recordings, making them suitable for training NN-based TSE models.

We slightly modify this technique to generate the reference signals for the challenge. Originally, the C2D projection was designed to denoise all close-microphone signals and use them to construct the aligned reference. However, for this challenge, we found it advantageous to selectively denoise only the 1-ch signal corresponding to the target speaker's headset and to use it for reference generation, since this channel is substantially cleaner than the others.

Let j denote the index of an utterance in the training set, and let $\mathbf{c}_{j,t,f} \in \mathbb{C}^P$ represent the recording captured by the $P = 4$ speakers' headsets. The C2D projection is then formulated as

$$s_{j,t,f} = a^*_{j,f}\hat{c}_{j,t,f} \quad \text{where} \quad \hat{c}^*_{j,t,f} = \mathbf{g}^{\mathsf{H}}_{j,f}\mathbf{c}_{j,t,f}, \tag{4}$$

where $\mathbf{g}_{j,f} \in \mathbb{C}^P$ is a Minimum-Variance Distortionless Response (MVDR) BF [19] estimated using the Guided Source Separation (GSS) approach [20, 21]; It produces a 1-ch denoised signal $\hat{c}_{j,t,f} \in \mathbb{C}$ for the target speaker's headset. $a_{j,f} \in \mathbb{C}$ is a coefficient that aligns the phase and amplitude of the target speaker's headset signal $c_{j,t,f} \in \mathbb{C}$ with those of the wearable device signal $x_{j,t,f} \in \mathbb{C}$ at its reference microphone. It is computed as

$$a_{j,f} = \left(\sum_t \lambda_{j,t,f} c_{j,t,f} x^*_{j,t,f}\right) / \left(\sum_t \lambda_{j,t,f} c_{j,t,f} c^*_{j,t,f}\right), \tag{5}$$

where $\lambda_{j,t,f}$ is a time–frequency mask for the target speech estimated by GSS, and $(\cdot)^*$ denotes the complex conjugate.

It should be noted that two preprocessing steps are applied before reference generation in the proposed system. The first is compensation for the sampling-frequency mismatch [22] between the headset and wearable-device recordings. Since the sampling frequency of the official reference signal provided for the challenge is aligned with the wearable-device recordings, we identified the mismatches between the headset and reference signals and compensated for them by resampling the headset recordings. As the second preprocessing step, we apply dereverberation to the headset and wearable-device recordings to improve the quality of the generated references. For this purpose, we employ Weighted Prediction Error (WPE) dereverberation [23, 24].

2.2.3. Frequency compensation of generated references

Because the amplitudes of the generated references are fitted to those of the distant microphones, their higher-frequency components tend to be significantly attenuated compared with those of the headset recordings. This attenuation can degrade the intelligibility of the enhanced speech. To compensate for this mismatch, we newly introduce an optional post-filter β_f. It is designed to ensure that the generated reference has, on average, the same frequency characteristics as the official reference provided for the challenge, $\tilde{c}_{j,t,f}$. It is computed and applied to the generated reference $s_{j,t,f}$ as follows:

$$\beta_f = \frac{1}{\sum_j |T_j|} \sum_{j,t\in T_j} \log_{10}\left(\frac{|\tilde{c}_{j,t,f}|^2}{|\rho_j s_{j,t,f}|^2}\right), \tag{6}$$

$$\tilde{s}_{j,t,f} = 10^{\frac{\alpha\beta_f}{2}} s_{j,t,f}, \tag{7}$$

where T_j denotes the set of all time frames in utterance j and $|T_j|$ denotes the number of frames. $\rho_j \in \mathbb{R}\ (> 0)$ is a time-frequency independent normalization factor that ensures $s_{j,t,f}$ has the same average level as $\tilde{c}_{j,t,f}$ for each utterance. $\alpha\ (\geq 0)$ is a hyperparameter that controls the degree of compensation. For example, $\alpha = 0.0$ disables the post-filter and $\alpha = 1.0$ forces the generated references to have the same frequency characteristics as those of the official references.

2.2.4. Two-step training

Because the amount of training data provided for the challenge may be insufficient to achieve reliable TSE, we also generated simulated data. The simulated data were used for pre-training the model, while the real data, comprising the wearable-device recordings and the references generated as described above, were used for fine-tuning.

For pre-training, we simulated mixture recordings, each comprising four speech signals drawn from the LibriSpeech dataset [25] and 10 noise signals from the CHiME-3 dataset [26]. A microphone array with the same number of microphones and geometric configuration as the wearable devices was defined and used in the simulation. The speech sources, noise sources, and microphone array were randomly placed in a room with dimensions ranging from $3\times3\times2$ m to $10\times10\times2$ m. The mixture recordings were generated using room impulse responses (RIRs) simulated with pyroomacoustics [27]. The speech-mixture-to-noise ratio and the reverberation time (T_{60}) were varied from 5 to 15 dB and from 0.15 to 0.55 s, respectively. The training, validation, and evaluation sets contained 10000, 500, and 500 utterances, respectively. Clean speech references were also simulated using the same RIRs truncated at 2 ms.

Because the phase and amplitude of the reference signals are aligned with those of the observed signals for both the simulated and real data, we can employ a training objective that evaluates fine-grained differences between the reference and estimated signals. For both datasets, we adopted a simple yet effective objective, SI-SNR [12]. It is defined as follows.

$$\mathcal{L}(\mathbf{s}, \hat{\mathbf{s}}; \theta) = 10\log_{10}\frac{\|\gamma\mathbf{s}\|_2^2}{\|\gamma\mathbf{s} - \hat{\mathbf{s}}\|_2^2} \quad \text{where} \quad \gamma = \frac{\langle \mathbf{s}, \hat{\mathbf{s}}\rangle}{\|\mathbf{s}\|_2^2}, \tag{8}$$

where $\mathbf{s}$ and $\hat{\mathbf{s}}$ are the waveforms of the reference and estimated signals, and $\langle \mathbf{s}, \hat{\mathbf{s}}\rangle$ denotes their inner product. In addition, during the final training steps using the simulated data, we employed the SNR objective by setting $\gamma = 1$ in the above equation, as this was found to slightly improve the perceived audio quality.

2.3. Inference of long recordings

For evaluation in the challenge, the proposed system is required to process long recordings of approximately 30 minutes. To handle such long recordings, we follow the inference scheme provided for the baseline. Specifically, a long recording is first segmented into 60-s segments with a 4-s overlap. The proposed system is then applied to each segment for each target speaker contained in the recording, and the enhanced segments are subsequently reconnected to form a continuous enhanced signal for each target speaker.

One issue in processing long recordings is level normalization of the enhanced signal. Because the amplitude of the enhanced signal is not explicitly constrained by the proposed system, its level must be normalized to an appropriate range. To achieve this in a low-latency manner, the proposed system continuously tracks the maximum amplitude of the enhanced signal from the beginning of the recording and normalizes each output sample by this maximum amplitude before emitting it.

Table 1. Description of the three submitted systems.

	α	I	Degree of frequency compensation
System-1	0.0	2	No compensation is applied.
System-2	0.5	3	A weak compensation is applied.
System-3	1.0	3	A full compensation is applied.

Table 2. Performance of baseline and proposed systems on the CHiME-9 ECHI dev (eval) sets using the Aria device.

System	STOI	PESQ	fwSNRseg	Csig	Cbak	Covl
Baseline	0.50 (0.50)	1.16 (1.11)	2.34 (-)	1.56 (1.74)	1.24 (1.08)	1.26 (1.32)
System-1	0.55 (**0.58**)	**1.27 (1.30)**	1.24 (1.47)	1.60 (1.76)	**1.42 (1.31)**	1.34 (1.43)
System-2	**0.56** (0.57)	1.23 (1.25)	2.57 (2.97)	1.83 (2.00)	1.36 (1.22)	1.43 (1.51)
System-3	0.55 (**0.58**)	1.23 (1.24)	**3.69 (4.37)**	**1.90 (2.09)**	1.41 (1.23)	**1.46 (1.55)**

Table 3. Results of subjective listening test for eval set (Aria).

System	SIG	BAK	OVRL	Intelligibility	Average
Baseline	2.31	2.75	2.16	55.28	42.14
System-2	**3.74**	**2.85**	**3.14**	**63.54**	**58.52**

3. SUBMITTED SYSTEMS

We submitted three systems to the challenge, all incorporating the components described in this paper: a combination of DRE and NBF with low-latency processing enabled by customized Conv/Deconv kernels. The window length and hop size were set to 20 ms and 10 ms, respectively, yielding an algorithmic latency of 20 ms. The systems were pre-trained on simulated data and then fine-tuned on real data using generated reference signals and the SI-SNR objective. The differences among the systems are summarized in Table 1, mainly in the post-filter parameter α for reference generation and the number of recursion steps I.

In contrast, the baseline system does not use any of these components and is trained solely using the official reference signals. Because the phase of the official references is not aligned with that of the device recordings, the baseline cannot employ phase-sensitive training objectives. Consequently, it adopts the STFTLoss provided by Auraloss [28], which evaluates only magnitude-based criteria in the STFT domain.

All models for the baseline and the proposed systems were trained for 60 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and a gradient clipping threshold of 1.0. A plateau-based scheduler dynamically adjusted the learning rate.

4. EXPERIMENTS

This section presents the evaluation results of the proposed system on the dev and eval sets of the CHiME-9 ECHI task using objective and subjective metrics. Due to time constraints, we evaluate results only for the Aria device, although the proposed method can be applied to the hearing-aid (HA) recordings provided for the challenge in the same manner.

4.1. Evaluation results based on objective measures

Table 2 presents the performance comparison between the official baseline and our three submitted systems. According to the challenge guideline, we quantified the performance using a signal distortion metric, frequency-weighted segmental SNR (fwSNRseg) [29], an intelligibility metric, Short-Time Objective Intelligibility (STOI) [30], and perceptual quality metrics, Perceptual Evaluation of Speech Quality (PESQ) [31], and the DNSMOS-based composite measures (Csig, Cbak, and Covl) [32] [33].

Overall improvement: Compared to the baseline, all proposed systems achieve substantial improvements across most metrics, with only a few exceptions. In particular, STOI improves from 0.50 (0.50) to 0.56 (0.58) on the dev (eval) sets, and composite metrics also show notable gains, e.g., Csig increases from 1.56 (1.74) to 1.90 (2.09). These results demonstrate the overall effectiveness of the processing components integrated into the proposed system.

Effect of frequency compensation post-filter: The comparison among the proposed systems reveals a trade-off controlled by the frequency compensation parameter α. 1) System-1 ($\alpha = 0.0$, no compensation) achieves the highest PESQ 1.27 (1.30) and Cbak 1.42 (1.31), indicating the most effective background noise reduction. 2) System-3 ($\alpha = 1.0$, full compensation) attains the highest fwSNRseg 3.69 (4.37), Csig 1.90 (2.09), and Covl 1.46 (1.55), while sacrificing PESQ and Cbak from System-1. This suggests that full frequency compensation best restores the spectral shape but amplifies background noise to some extent. 3) System-2 yields intermediate scores across metrics.

4.2. Evaluation results based on subjective listening test

Table 3 presents the results of the subjective listening test on the eval set, conducted by the challenge organizer. For each of the baseline and the proposed System-2, the estimated signals of the three target speakers were summed, and native English speakers evaluated the resulting signals. Intelligibility was measured as the percentage of correctly transcribed words. Quality was evaluated by rating the naturalness of the foreground speech (SIG), the intrusiveness of background noise (BAK), and the overall quality (OVRL) following the ITU-T P.835 protocol [33]. Additional details on the evaluation setup can be found on the challenge website [1].

As shown in the table, the proposed System-2 significantly outperforms the baseline across all metrics.[1] These results clearly demonstrate the effectiveness of the proposed system in improving both intelligibility and perceptual quality.

5. CONCLUDING REMARKS

This paper presents a low-latency target speech extraction (TSE) system for the ECHI task of the CHiME-9 challenge. The proposed system integrates Close-to-Distant microphone (C2D) projection–based reference generation with Deterministic Recursive Enhancement (DRE) and Null Beamforming (NBF) to address real-world TSE with wearable devices. Experimental results show that the system significantly outperforms the official baseline on both objective and subjective metrics, improving intelligibility and perceptual quality while meeting the 20 ms latency constraint. We further analyze the effect of the frequency compensation post-filter using objective metrics, showing that it effectively restores the frequency characteristics of headset recordings, albeit with a slight increase in background noise.

Future work will focus on conducting a thorough ablation study to quantify the individual contributions of each component in the proposed system. In addition, improving TSE under extremely low-SNR conditions, particularly in the presence of many inactive speakers, remains an important challenge. The computational cost of the recursive enhancement also warrants further investigation.

[1] All subjective scores achieved by System-2, except for BAK, are the highest among all challenge submissions. However, the system is not included in the official ranking because it uses CHiME-3 noise data for pre-training, which is not in the permitted whitelist.

6. REFERENCES

[1] Jon Barker, Stefan Goetze, Robert Sutherland, Marko Lugger, Thomas Kuebert, Juan Azcarreta Ortiz, and Buye Xu, "CHiME-9 challenge - enhancing conversations to address hearing impairment," `https://www.chimechallenge.org/current/task2/index`, 2026.

[2] Robert Sutherland, Jason Clarke, Hend Elghazaly, Thomas Kuebert, Marko Lugger, Stefan Petrausch, Juan Azcarreta Ortiz, Buye Xu, Stefan Goetze, and Jon Barker, "Descriptor: Enhancing conversations for the hearing impaired in the 9th computational hearing in multisource environments challenge (CHiME9 ECHI)," *IEEE Data Descriptions*, vol. 3, pp. 73–81, 2026.

[3] Katerina Zmolikova, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu, "Neural target speech extraction: An overview," 2023.

[4] Samuele Cornell, Zhong-Qiu Wang, Yoshiki Masuyama, Shinji Watanabe, Manuel Pariente, and Nobutaka Ono, "Multi-channel target speaker extraction with refinement: The wavlab submission to the second clarity enhancement challenge," in *Proc. the 2nd Clarity Enhancement Challenge*, 2023.

[5] Barry D. Van Veen and Kevin M. Buckley, "Beamforming: a versatile approach to spatial filtering," *IEEE ASSP Magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[6] Sharon Gannot, Emmanuel Vincent, Shmulik Markovich-Golan, and Alexey Ozerov, "A consolidated perspective on multimicrophone speech enhancement and source separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 4, pp. 692–730, 2017.

[7] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "TF-GridNet: Integrating full- and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 3221–3236, 2023.

[8] Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe, "TF-GridNet: Making time-frequency domain models great again for monaural speaker separation," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023.

[9] Chengshi Zheng Fengyuan Hao, Xiaodong Li, "X-TF-GridNet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion," *Information Fusion*, vol. 112, 2024.

[10] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville, "FiLM: Visual reasoning with a general conditioning layer," in *Proc. AAAI*, 2018.

[11] Tomohiro Nakatani, Rintaro Ikeshita, Naoyuki Kamo, Marc Delcroix, and Shoko Araki, "Generating training targets for real-world speech enhancement via close-to-distant microphone projection," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

[12] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey, "SDR – half-baked or well done?," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[13] Tomohiro Nakatani, Naoyuki Kamo, Marc Delcroix, and Shoko Araki, "Hybrid probabilistic-deterministic model recursively enhancing speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025.

[14] Shoko Araki, Shoji Makino, Ryo Mukai, and Hiroshi Saruwatari, "Equivalence between frequency domain blind source separation and frequency domain adaptive null beamformers," in *Proc. EUROSPEECH*, 2001.

[15] Wen-Yuan Ting, Syu-Siang Wang, Yu Tsao, and Borching Su, "IANS: Intelligibility-aware null-steering beamformingfor dual-microphonearrays," in *Proc. international workshop on machine learning for signal processing*, 2023.

[16] Yen-Ju Lu, Samuele Cornell, Xuankai Chang, Wangyou Zhang, Chenda Li, Zhaoheng Ni, Zhong-Qiu Wang, and Shinji Watanabe, "Towards low-distortion multi-channel speech enhancement: the ESPNET-SE-submittion to the L3DAS22 challenge," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022.

[17] Shmulik Markovich-Golan, Sharron Gannot, and Israel Cohen, "Multichannel eigenspace beamforming in a reverberant noisy environment with multiple interfering speech signals," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 6, pp. 1071–1086, 2009.

[18] Wendi Sang, Kai Li, and Runxuan Yang, "A fast and lightweight model for causal audio-visual speech separation," *arXiv:2506.06689*, 2025.

[19] Mehrez Souden, Shoko Araki, Keisuke Kinoshita, Tomohiro Nakatani, and Hiroshi Sawada, "A multichannel mmse-based framework for speech source separation and noise reduction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1913–1928, 2013.

[20] Christoph Boeddeker, Jens Heitkaemper, Joerg Schmalenstroeer, Lukas Drude, Jahn Heymann, and Reinhold Haeb-Umbach, "Front-end processing for the CHiME-5 dinner party scenario," in *CHiME5 Workshop, Hyderabad, India*, 2018.

[21] Nobutaka Ito, Shoko Araki, and Tomohiro Nakatani, "Complex angular central Gaussian mixture model for directional statistics in mask-based microphone array signal processing," in *Proc. 2016 24th European Signal Processing Conference (EUSIPCO)*. IEEE, 2016, pp. 1153–1157.

[22] Shigeki Miyabe, Nobutaka Ono, and Shoji Makino, "Blind compensation of inter-channel sampling frequency mismatch with maximum likelihood estimation in stft domain," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013.

[23] Tomohiro Nakatani, Takuya Yoshioka, Keisuke Kinoshita, Masato Miyoshi, and Biing-Hwang Juang, "Speech dereverberation based on variance-normalized delayed linear prediction," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 7, pp. 1717–1731, 2010.

[24] Takuya Yoshioka and Tomohiro Nakatani, "Generalization of multichannel linear prediction methods for blind MIMO impulse response shortening," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 10, pp. 2707–2720, 2012.

[25] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015.

[26] Jon Barker, Ricard Marxer, Emmanuel Vincent, and Shinji Watanabe, "The third CHiME speech separation and recognition challenge: Dataset, task and baselines," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.

[27] Robin Scheibler, Eric Bezzam, and Ivan Dokmanić, "Pyroomacoustics: A python package for audio room simulation and array processing algorithms," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[28] Christian J. Steinmetz and Joshua D. Reiss, "auraloss: Audio-focused loss functions in Pytorch," in *149th Audio Engineering Society Convention*, Oct. 2020.

[29] Yi Hu and Philipos C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 229–238, 2008.

[30] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4214–4217.

[31] Antony W. Rix, Jan G. Beerends, Michael P. Hollier, and Andries P. Hekstra, "Perceptual evaluation of speech quality (PESQ): A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2001, pp. 749–752.

[32] Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler, "DNSMOS P.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," *arXiv preprint arXiv:2110.01763*, 2021.

[33] "ITU-T Recommendation P.835: Subjective test methodology for evaluating speech communication systems," 2019, International Telecommunication Union.