

DialogueSidon: Recovering Full-Duplex Dialogue Tracks from In-the-Wild Dialogue Audio

Wataru Nakata^{1,2}, Yuki Saito^{1,2},
Kazuki Yamauchi¹, Emiru Tsunoo¹, Hiroshi Saruwatari¹,

¹The University of Tokyo, Tokyo, Japan,

²National Institute of Advanced Industrial Science and Technology (AIST), Tokyo, Japan

Correspondence: nakata-wataru855@g.ecc.u-tokyo.ac.jp

Abstract

Full-duplex dialogue audio, in which each speaker is recorded on a separate track, is an important resource for spoken dialogue research, but is difficult to collect at scale. Most in-the-wild two-speaker dialogue is available only as degraded monaural mixtures, making it unsuitable for systems requiring clean speaker-wise signals. We propose DialogueSidon, a model for joint restoration and separation of degraded monaural two-speaker dialogue audio. DialogueSidon combines a variational auto-encoder (VAE) operates on the speech self-supervised learning (SSL) model feature, which compresses SSL model features into a compact latent space, with a diffusion-based latent predictor that recovers speaker-wise latent representations from the degraded mixture. Experiments on English, multilingual, and in-the-wild dialogue datasets show that DialogueSidon substantially improves intelligibility and separation quality over a baseline, while also achieving much faster inference.

1 Introduction

Building natural spoken dialogue systems remains a key challenge in dialogue research. Recent systems have begun to exhibit human-like conversational behaviors such as backchannels, overlap, and flexible turn-taking, enabling fluent human-AI conversation (Nguyen et al., 2023; Défossez et al., 2024; Roy et al., 2026). However, modeling such phenomena robustly still requires large amounts of suitable conversational speech data.

In particular, full-duplex dialogue recordings are important resources for spoken dialogue research. In these recordings, two speakers are recorded on separate tracks, making it possible to analyze and model conversational behaviors while preserving clean speaker-wise signals. Such data are especially valuable for studying natural interaction and for developing human-like dialogue systems.

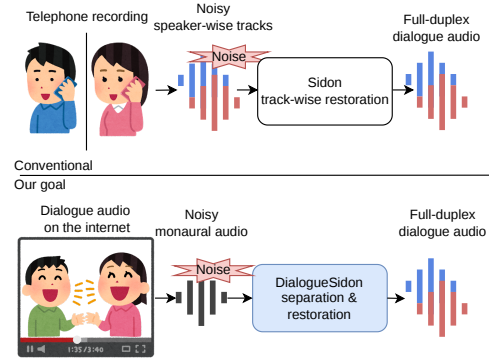


Figure 1: Comparison of current full-duplex dialogue audio acquisition and our goal

Despite their importance, full-duplex dialogue data are difficult to collect at scale. A common approach is to record telephone conversations (Cieri et al., 2004; LDC, 2008b) or controlled two-party interactions (Agrawal et al., 2025; Lee et al., 2023), but such pipelines are costly to build, often suffer from latency and channel artifacts, and are not easily scalable. Another possibility is to synthesize conversations using text-to-speech systems (Défossez et al., 2024). However, because these systems are usually driven by turn-based text dialogues, the resulting audio often lacks spontaneous interactional phenomena such as natural overlap, backchannels, and timing variability.

Meanwhile, large-scale in-the-wild audio from the Internet provides a potentially rich source of conversational speech (Nakata et al., 2026b). Internet recordings contain diverse speakers, speaking styles, and interaction patterns that are difficult to reproduce in controlled data collection. In fact, such data are already used for tasks such as large-scale (SSL) (Zhang et al., 2023), automatic speech recognition (Radford et al., 2022; Peng et al., 2025), and speech synthesis (Koizumi et al., 2023a).

However, such recordings are difficult to use directly for training full-duplex dialogue systems. First, Internet audio is heavily degraded. Recordings may contain background music, environmen-

tal noise, reverberation, clipping, and compression. Second, Internet dialogue audio is usually available only as mixed recordings rather than isolated speaker-wise tracks. Jointly addressing both problems—restoring signal quality and separating speakers—is therefore necessary to recover speaker-wise signals from in-the-wild dialogue audio.

A natural approach is to apply speech separation (Luo and Mesgarani, 2019). However, existing separation methods are typically developed for mixtures of monologue clean speech and often do not generalize well to spontaneous conversational speech, where overlap patterns, prosody, and interaction timing are substantially different.

Speech restoration methods (Liu et al., 2022; Karita et al., 2025; Nakata et al., 2026a), on the other hand, have recently shown strong performance in removing noise and recording artifacts. In fact, some of open corpora are based on the restored samples by speech restoration models (Koizumi et al., 2023a; Ma et al., 2024; He et al., 2024). Inspired by these approaches, the joint modeling of speech restoration and separation has been proposed (Asai et al., 2026; Zhang et al., 2024). Nevertheless, existing work has mainly focused on monologue and monaural speech rather than mixed dialogue recordings. This limits their applicability to full-duplex dialogue recovery from in-the-wild two-speaker dialogue audio.

In this paper, we propose *DialogueSidon*, a model for joint restoration and separation of in-the-wild dialogue as shown in Figure 1. The key idea is to extend the previous open-source speech restoration model Sidon (Nakata et al., 2026a) to the dialogue setting by incorporating speaker separation, allowing degraded mixed conversational recordings to be transformed into cleaner speaker-wise dialogue signals by leveraging SSL model pretrained on large-scale speech dataset. Experimental results show that DialogueSidon improves both restoration and separation quality compared with conventional unified restoration and separation baselines, while also achieving faster inference.

Our contributions are as follows: (i) We propose DialogueSidon, a model that jointly restores and separates degraded monaural two-speaker dialogue audio into clean speaker-wise tracks. (ii) DialogueSidon combines an SSL-VAE latent space with a diffusion-based latent predictor, enabling speaker-wise recovery from degraded conversational mixtures. (iii) We show through experiments on English, multilingual, and in-the-wild dialogue

data that DialogueSidon improves content preservation and separation quality over a baseline while providing substantially faster inference. Audio samples¹, code², and a live demo³ are publicly available.

2 Related Work

2.1 Spoken Dialogue Language Modeling

Recent spoken dialogue language models (Nguyen et al., 2023; Défossez et al., 2024; Roy et al., 2026; Veluri et al., 2024) can capture acoustic and temporal phenomena such as overlap, backchannels, and turn-taking timing by modeling spoken interactions directly. However, these models still lag behind human conversation (Nguyen et al., 2023; Défossez et al., 2024; Roy et al., 2026), partly due to the scarcity of large-scale full-duplex dialogue data—recordings where each speaker is on a separate track. Existing datasets are typically limited to telephone corpora such as Fisher (Cieri et al., 2004) (2k hours) and proprietary recordings (Défossez et al., 2024), far smaller than the scale used in recent speech generation research which frontier models use millions of hours of speech (Zhang et al., 2023; Peng et al., 2025). This data scarcity motivates our work on converting Internet audio into usable full-duplex dialogue data.

2.2 Speech Separation and Restoration

Speech separation (Luo and Mesgarani, 2019; Wang et al., 2023; Scheibler et al., 2025) aims to recover individual speaker signals from multi-speaker mixtures. However, most methods are developed on clean monologue mixtures and do not generalize well to in-the-wild conversational audio with background music, noise, reverberation, and compression artifacts.

Speech restoration removes degradations from corrupted recordings (Liu et al., 2022; Koizumi et al., 2023b; Karita et al., 2025; Nakata et al., 2026a), making it useful for large-scale dataset cleansing (Koizumi et al., 2023a; Ma et al., 2024). Recent methods such as Sidon (Nakata et al., 2026a) perform restoration in the latent space of SSL model, improving robustness to diverse in-the-wild degradations. However, they mainly target single-speaker speech and do not address speaker separation.

¹<https://hf.co/spaces/Wataru/dsidonsamples>

²<https://github.com/sarulab-speech/Sidon>

³<https://hf.co/spaces/sarulab-speech/DialogueSidon-demo>

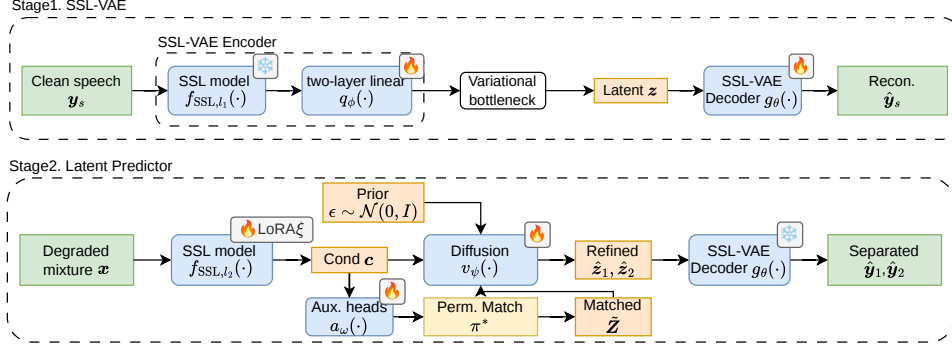


Figure 2: Overview of DialogueSidon training. ❄ indicates frozen modules and 🔥 indicates trained modules.

A straightforward cascade of restoration and separation is generally inadequate for this setting. Applying restoration first can suppress or distort overlapping speech, because existing restoration models are typically designed for monologue and tend to treat overlap as corruption to be removed. Applying separation first is also unreliable, since most separation models are not designed for heavily degraded in-the-wild mixtures. This motivates unified modeling of separation and restoration rather than a simple cascade.

More recently, several studies have explored unified modeling of speech separation and restoration (Zhang et al., 2024; Asai et al., 2026). These approaches are promising because both problems must be addressed to recover usable speaker-wise signals from degraded mixtures. However, existing methods are primarily developed for mixtures of monologue speech and may not generalize well to in-the-wild dialogue audio, where overlap patterns, backchannels, and interaction timing differ substantially from monologue mixtures.

3 DialogueSidon

DialogueSidon extends the speech restoration model Sidon (Nakata et al., 2026a) to the dialogue setting by introducing speaker-wise recovery from degraded conversational mixtures. Given a degraded monaural two-speaker dialogue mixture, the goal is to reconstruct a clean waveform for each speaker track. As shown in Figure 2, the model consists of two components: (i) an SSL-VAE that defines a compact latent space for clean speech, and (ii) a diffusion-based latent predictor that estimates the speaker-wise latent representations from the degraded mixture. The predicted latents are then decoded into separated clean waveforms.

Formally, given a degraded monaural two-speaker dialogue mixture with T samples, $\mathbf{x} \in \mathbb{R}^T$, our goal is to recover the corresponding clean full-

duplex waveforms $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^T$, where each waveform corresponds to one speaker. In this work, we focus on the two-speaker case.

DialogueSidon first maps each target signal $\mathbf{y}_s$ to a compact latent representation $\mathbf{z}_s \in \mathbb{R}^{L \times D}$ using the SSL model followed by two layer linear layers with ReLU activation, where $s \in \{1, 2\}$, L is the latent sequence length, and D is the latent dimensionality. The latent predictor is then trained to estimate $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2$ from the degraded mixture $\mathbf{x}$. Finally, the SSL-VAE decoder reconstructs the corresponding waveforms $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2$ from the predicted latents.

Training is performed in two stages. In the first stage, the SSL-VAE is trained to construct a compact and speech-relevant latent space from clean dialogue tracks. In the second stage, the latent predictor is trained in this latent space to recover the speaker-wise clean representations from degraded mixed dialogue audio. At inference time, the latent predictor first estimates the latent representation for each speaker track from the degraded input, and the SSL-VAE decoder then reconstructs the final waveforms.

3.1 SSL-VAE

The SSL-VAE is designed to compress SSL model representations into a compact latent space suitable for diffusion modeling. Previous studies on speech restoration (Koizumi et al., 2023b; Karita et al., 2025; Guimarães et al., 2025) have shown that features extracted from large SSL models are effective for representing speech content and acoustic characteristics. However, directly applying diffusion models to such high-dimensional (e.g., 1,536-dimensional) feature spaces is computationally expensive. To address this issue, we adopt a latent diffusion framework (Dhyani et al., 2025) in which SSL features are first compressed with a pretrained VAE encoder.

Let $f_{\text{SSL},l_1}(\cdot)$ denote a frozen SSL l_1 -th layer hidden feature. Given a clean monaural waveform $\mathbf{y}$, we first compute a D_h -dimensional intermediate hidden representation $\mathbf{h} = f_{\text{SSL},l_1}(\mathbf{y})$, where $\mathbf{h} \in \mathbb{R}^{L \times D_h}$. The trainable encoder q_ϕ then applies a two-layer linear layers with ReLU activation followed by a variational bottleneck to obtain a compact latent representation $\mathbf{z} \sim q_\phi(\mathbf{z} \mid \mathbf{h})$ with dimension $D < D_h$. The latent variable $\mathbf{z} \in \mathbb{R}^{L \times D}$ serves as the target representation for the second-stage latent predictor.

The trainable decoder $g_\theta(\cdot)$ reconstructs the original waveform from the latent representation: $\hat{\mathbf{y}} = g_\theta(\mathbf{z})$. The training objective consists of a reconstruction term, adversarial loss, together with the standard Kullback–Leibler regularization similar to previous work (Shi et al., 2025): $\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{rec}}(\mathbf{y}, \hat{\mathbf{y}}) + \mathcal{L}_{\text{adv}}(\mathbf{y}, \hat{\mathbf{y}}) + \beta D_{\text{KL}}(q_\phi(\mathbf{z} \mid \mathbf{h}) \parallel p(\mathbf{z}))$, where $p(\mathbf{z})$ is a prior distribution and β controls the strength of regularization.

Following prior work (Nakata et al., 2026a), we use $l_1 = 8$ -th layer hidden feature of w2v-BERT 2.0 (Communication et al., 2023) as the SSL feature extractor. This model is trained on 4.5 million hours of speech covering 143 languages and has been shown to be effective for speech-related tasks including speech translation (Communication et al., 2023) and speech restoration (Nakata et al., 2026a). For waveform reconstruction and discriminator used for adversarial training, we use the Descript Audio Codec (Kumar et al., 2023) decoder and discriminator, which is based on a HiFi-GAN-style vocoder (Kong et al., 2020) with Snake activation (Ziyin et al., 2020).

3.2 Latent Predictor

The role of the latent predictor is to estimate clean speaker-wise latent representations from a degraded monaural dialogue mixture. The latent predictor takes $\mathbf{x}$ as input and predicts $\hat{\mathbf{z}}_1, \hat{\mathbf{z}}_2$, which are then decoded by the SSL-VAE decoder to obtain the final reconstructed waveforms.

A straightforward approach would be to directly regress the target latents from the degraded input. However, in our preliminary experiments, such deterministic objectives tended to oversmooth the latent trajectories and often failed to preserve spoken content and overlapping speech. To better model the uncertainty inherent in degraded conversational mixtures, we instead adopt a diffusion-based latent predictor. This approach is reported to be effective in other speech processing tasks (Shi et al., 2025;

Guimarães et al., 2025).

Conditioning representation We first extract a conditioning representation $\mathbf{c} = f_{\text{SSL},l_2,\xi}(\mathbf{x}) \in \mathbb{R}^{L \times D_h}$ from the degraded input mixture, where $f_{\text{SSL},l_2,\xi}(\cdot)$ is an l_2 -th layer SSL model feature fine-tuned with low-rank adapter (LoRA) (Hu et al., 2022) parameterized by ξ . This representation captures the linguistic and acoustic information in the degraded mixture and is used to condition both the auxiliary latent prediction and the diffusion model.

For conditioning, we use the $l_2 = 13$ -th layer hidden feature of w2v-BERT 2.0. The LoRA adapter is applied to the last linear layer of the feed-forward network in each Conformer block, adapting the model to degraded mixtures while preserving pre-trained knowledge.

Auxiliary latent prediction A key challenge in speaker-wise latent prediction is permutation ambiguity: the two output tracks are unordered, therefore the model must not only predict the correct latent content, but also resolve speaker assignment. To reduce the burden on the diffusion model, we introduce auxiliary heads parametrized by ω and denoted a_ω that produce coarse speaker-wise latent estimates $\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2 = a_\omega(\mathbf{c})$ directly from the conditioning representation. These auxiliary heads are trained with permutation-invariant training (Yu et al., 2017) to encourage them to produce speaker-wise estimates that are close to the target latents, while allowing for flexible speaker assignment. By providing these coarse estimates, we guide the diffusion model towards better predictions and help it resolve permutation ambiguity. These auxiliary predictions are matched against the ground-truth latents in a permutation-invariant manner (Yu et al., 2017): $\pi^* = \arg \min_{\pi \in \mathcal{S}_2} \sum_{s=1}^2 \|\tilde{\mathbf{z}}_s - \mathbf{z}_{\pi(s)}\|_1$, where $\mathcal{S}_2$ is the set of all permutations over two speakers. The resulting permutation π^* defines the alignment between predicted and target speaker tracks. Using this alignment, the auxiliary latent prediction loss is $\mathcal{L}_{\text{aux}} = \sum_{s=1}^2 \|\tilde{\mathbf{z}}_s - \mathbf{z}_{\pi^*(s)}\|_1$. The same permutation π^* is then used to define the speaker ordering for the diffusion model, preventing it from suffering from the permutation ambiguity.

Diffusion-based latent refinement After permutation matching, we jointly model the two speaker latents by stacking them into $\mathbf{Z} = \text{stack}(\mathbf{z}_{\pi^*(1)}, \mathbf{z}_{\pi^*(2)}) \in \mathbb{R}^{L \times 2D}$ and similarly $\tilde{\mathbf{Z}} = \text{stack}(\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2) \in \mathbb{R}^{L \times 2D}$ for the auxiliary predictions. The stacking is performed along the latent

Table 1: Durations of dataset used for training.

Dataset	Duration (h)
CALLHOME German (LDC, 2008b)	58.1
CALLHOME English (LDC, 2008b)	76.9
CALLHOME Japanese (LDC, 2008c)	49.3
CALLHOME Spanish (LDC, 2008d)	43.6
CALLHOME Mandarin (LDC, 2008a)	39.5
Fisher (Cieri et al., 2004)	1,958.5
Total	2,225.9

dimension D . This joint representation allows the diffusion model to capture inter-speaker dependencies.

We use the v-prediction (Salimans and Ho, 2022) reparametrization of diffusion model training following previous work (Guimarães et al., 2025). The forward diffusion process is $\mathbf{Z}_t = \alpha_t \mathbf{Z} + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$. Here, $t \sim \mathcal{U}(0, 1)$ represents the continuous diffusion time step, while the scalar functions α_t and $\sigma_t \in (0, 1)$ denote the signal and noise schedules, respectively. The v-prediction target is defined as $\mathbf{v}_t = \alpha_t \epsilon - \sigma_t \mathbf{Z}$. The diffusion model $v_\psi(\cdot)$ parameterized by ψ , conditioned on the SSL feature $\mathbf{c}$ and auxiliary predictions $\tilde{\mathbf{Z}}$, predicts this velocity: $\hat{\mathbf{v}}_t = v_\psi(\mathbf{Z}_t, t, \mathbf{c}, \tilde{\mathbf{Z}})$. The diffusion loss is $\mathcal{L}_{\text{diff}} = \mathbb{E}_{\mathbf{Z}, \epsilon, t} [\|\mathbf{v}_t - v_\psi(\mathbf{Z}_t, t, \mathbf{c}, \tilde{\mathbf{Z}})\|_2^2]$. This decomposition separates speaker alignment (auxiliary heads) from latent refinement (diffusion), allowing each component to focus on a distinct sub-problem.

Training objective The latent predictor is trained using the sum of the auxiliary latent loss and the diffusion loss: $\mathcal{L}_{\text{latent}} = \mathcal{L}_{\text{aux}} + \lambda_{\text{diff}} \mathcal{L}_{\text{diff}}$, where λ_{diff} controls the weight of the second term.

Inference At inference time, the auxiliary heads produce coarse speaker-wise estimates from $\mathbf{c}$, and the diffusion model refines them via the reverse process. The SSL-VAE decoder then reconstructs clean waveforms $\hat{\mathbf{y}}_s = g_\theta(\hat{\mathbf{z}}_s)$ for $s \in \{1, 2\}$.

4 Experiments

We evaluate the speech restoration and separation capabilities of DialogueSidon in English, multilingual, and in-the-wild settings. All models are trained on Sidon-restored telephone conversational corpora, which provide clean speaker-wise tracks necessary for supervised training but are not in-the-wild data. The Sidon-restored samples (48 kHz) were downsampled to 24 kHz. Evaluation is conducted on three corpora: Switchboard (SWB) (Godfrey and Holliman, 1993) for in-domain English telephone audio, CallFriend (Canavan and Zipperlen, 1996b,a,c,d; Tracey et al., 2025) for mul-

tilingual telephone audio, and OpenDialog (Zhu et al., 2025) for in-the-wild Internet audio, to assess both in-domain and out-of-domain generalization.

4.1 Experimental conditions

The statistics of the datasets used for the experiment are shown in Table 1. We used a combination of the Fisher English dialogue corpus and the English, German, Japanese, Spanish, and Mandarin variants of the CALLHOME corpus. These are conversational corpora consisting of telephone conversations on various topics. For preprocessing, the recordings were restored with Sidon (Nakata et al., 2026a), as they are telephone recordings and thus contain a wide range of noise with limited audio fidelity.

Data preparation To train a generalized speech restoration and separation model, covering a wide range of degradations is important, as we cannot make assumptions about the specific degradations present in in-the-wild dialogue audio. Therefore, similar to previous work (Saijo et al., 2025), we used diverse sets of degradations. Each degradation was applied independently to each track with the probability of 0.5 before mixing into a monaural dialogue in the final step. The details of the degradations are reported in Section A.

The noising pipeline was applied four times to each dialogue session with different random seeds. As a result, we obtained approximately 8,902 hours of paired (clean, degraded) dialogue data.

DNN architecture For the SSL feature extractor, we use w2v-BERT 2.0 as described in section 3.1. We investigate the effect of the latent channel dimensionality in SSL-VAE by varying $D \in \{8, 16, 32, 64, 128\}$ while keeping the temporal sequence length fixed. All models produce 24 kHz output audio.

For the latent predictor, the auxiliary heads consist of two separate linear projection layers applied to the conditioning feature $\mathbf{c}$, one per speaker track. In the SSL model used in the latent predictor, we performed fine-tuning with a LoRA adapter with rank $r = 64$ and scaling factor $\alpha = 16$. The architecture of diffusion model is a Diffusion Transformer (DiT) (Peebles and Xie, 2023) with hidden dimension 768, eight layers, 12 attention heads, intermediate size 3,072. Rotary positional encoding (Su et al., 2024) is used in the DiT. The conditioning feature $\mathbf{c}$ and the stacked auxiliary latent prediction $\tilde{\mathbf{Z}}$

and noised input Z_t are concatenated along the latent dimension and provided. We use a linear noise schedule with 1,000 diffusion steps during training. At inference time, we used DPM-Solver++ (Lu et al., 2025) with 30 steps for sampling.

Training details The SSL-VAE is trained for two days with a batch size of 32 using the AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 1×10^{-4} . The exponential decay learning rate schedule is used where the learning rate is multiplied by a factor $\gamma = 0.999996$ every step. The KL regularization weight is set to $\beta = 1 \times 10^{-5}$. The latent predictor is trained for two days with a batch size of 64, a learning rate of 1×10^{-4} with 2,000 warm-up steps, and diffusion loss weight $\lambda_{\text{diff}} = 1.0$. All experiments are conducted on eight NVIDIA H100 GPUs.

Evaluation data *SWB*. We use the SWB corpus, a collection of telephone conversations in English. We randomly select 100 dialogue sessions, each cropped to 20 seconds for evaluation. Since SWB and the Fisher training corpus are both telephone-band conversational corpora, SWB results reflect in-domain performance.

CallFriend. For multilingual evaluation, we use the multilingual variant of CallFriend corpus (Cavanaugh and Zipperlen, 1996b,a,c,d; Tracey et al., 2025), which includes telephone conversations in German, French (Quebec), Japanese, Spanish, and Mandarin. We randomly select 20 dialogue sessions, each cropped to 20 seconds from each language variety for evaluation.

OpenDialog. For in-the-wild evaluation, we use the English train set of OpenDialog (Zhu et al., 2025), a corpus of monaural dialogue audio sourced from Internet containing diverse real-world acoustic conditions. We randomly select 100 dialogue sessions for evaluation. We did not crop the samples as each sample was shorter than 30 seconds.

Evaluation metrics We evaluate the quality of restored and separated outputs using both objective and subjective metrics.

Common metrics. **DNSMOS** (Reddy et al., 2021) and **NISQA** (Mittag et al., 2021): Machine-learning-based speech quality predictors, which predicts the mean opinion score (MOS) on overall perceptual quality of denoised speech. **Speaker similarity (Spk. Sim.)**: Cosine similarity between speaker embeddings of the output and the reference, computed using the WavLM-based speaker verifi-

cation model (Chen et al., 2022)⁴. **VAD accuracy (VAD Acc.)**: The frame-level voice activity detection (VAD) accuracy on the separated output, used for English and multilingual evaluation, computed using Silero-VAD⁵. The reference VAD labels differ by corpus: for SWB, word-level timestamps from the transcription labels are used; for CallFriend, VAD is run on the ground-truth audio to obtain pseudo-labels; Note that for OpenDialog, neither speaker-wise tracks nor speaker voice activity labels were provided and we do not report the result.

Per-dataset intelligibility metrics. All intelligibility metrics are computed using Whisper-large-v3 as the ASR model. **WER**: Word error rate, used for English (SWB) and in-the-wild data (OpenDialog) evaluation where speaker-wise reference transcripts are available. **p-CER**: Pseudo-character error rate, used for the multilingual (CallFriend) evaluation as transcriptions were not available in standard orthography for some languages. Because p-CER is biased by ASR performance on the original recording, it should be interpreted as a relative metric rather than an absolute measure of intelligibility.

Subjective metric. **MOS**: We conducted a subjective listening test on SWB and OpenDialog. For SWB we compared Noisy, Sidon, GENESES and DialogueSidon and for OpenDialog, we compared GENESES (orig), GENESES and DialogueSidon. Each rater assessed 12 samples. (360 ratings per method), scoring recording and separation quality on a 5-point scale. Specifically, each rater was provided with speaker-separated recordings and asked to rate each 20-second sample based on (i) the quality of recording and (ii) speaker consistency in each track. All raters were recruited through Prolific⁶ and asked to wear headphones. The number of raters was 120 for SWB and 90 for OpenDialog.

Baselines We compare DialogueSidon against four methods. **Noisy** denotes the original telephone recordings without any processing, which serve as an upper bound for speaker identity and VAD structure but contain telephone-band degradations. **Sidon** (Nakata et al., 2026a) is a speech restoration model that enhances the monaural mixture as a whole without performing speaker separation, meaning the output is a single restored track rather than speaker-wise signals. **GENESES** and **GENESES (orig)** (Asai et al., 2026) are unified speech

⁴<https://hf.co/microsoft/wavlm-base-plus-sv>

⁵<https://github.com/snakers4/silero-vad>

⁶<https://www.prolific.com/>

Table 2: Comparison of latent size on the SWB evaluation set. We compare DialogueSidon with different latent dimensionalities D against Noisy, Sidon, and GENESES. [†]Noisy and Sidon do not perform speech separation. **Bold** and underline denote the best and worst among separation methods, respectively.

Method	WER (%) ↓ <i>Content</i>	NISQA ↑ <i>Recording Quality</i>	DNSMOS ↑ <i>Speaker</i>	Spk. Sim. ↑ <i>Speaker</i>	VAD Acc. ↑ <i>Separation</i>
Noisy [†]	60.810	2.315	3.005	—	—
Sidon [†]	57.470	3.942	3.999	0.934	0.730
GENESES (orig)	<u>79.990</u>	3.507	<u>3.582</u>	<u>0.803</u>	<u>0.812</u>
GENESES	33.540	3.720	3.680	0.853	0.923
DialogueSidon ($D = 8$)	16.550	3.394	3.586	0.887	0.936
DialogueSidon ($D = 16$)	14.950	<u>3.370</u>	3.653	0.888	0.939
DialogueSidon ($D = 32$)	14.390	<u>3.453</u>	3.641	0.887	0.938
DialogueSidon ($D = 64$)	15.040	3.443	3.642	0.887	0.939
DialogueSidon ($D = 128$)	22.720	3.393	3.630	0.876	0.906

Table 3: MOS test results on the SWB evaluation set with their standard deviation. All pairwise differences are statistically significant ($p < 0.05$, two-sided t -test).

Method	MOS ↑
Noisy	2.815 ± 0.999
Sidon	3.289 ± 1.167
GENESES	3.482 ± 1.195
DialogueSidon	3.895 ± 0.948

separation and restoration models. GENESES is a flow-matching-based model reported to be robust to complex degradations. We report results using both their official checkpoint, denoted by (orig), and a retrained model. For retraining, we trained GENESES from scratch on the same dialogue dataset as DialogueSidon using the hyperparameters reported in the original paper for fair comparison. At inference time, we used 100 inference steps, as suggested in the paper (Asai et al., 2026).

4.2 Results

Effect of latent size Table 2 presents results on the SWB evaluation set for different latent dimensionalities D . All DialogueSidon variants substantially outperform the baselines in WER. The best setting, $D = 32$, achieves 14.39% WER, compared with 33.54% for the retrained GENESES and 57.47% for Sidon. Although DialogueSidon does not achieve the best predicted perceptual quality in terms of NISQA or DNSMOS, it better preserves spoken content and speaker characteristics than GENESES, as reflected by WER and speaker similarity. Notably, GENESES (orig)—the original checkpoint trained on monologue mixtures—performs substantially worse than the retrained GENESES across all metrics (e.g., 79.99% vs. 33.54% WER), confirming that training on conversational data is critical for dialogue separation. The $D = 128$ setting degrades across all metrics, suggesting that excessive latent capacity is detrimental in this setup. Based on these results, we use $D = 32$ in the subsequent experiments.

Subjective evaluation To further assess restoration and separation quality, we conducted a human listening test on the SWB evaluation set. As shown in Section 4.1, DialogueSidon achieves the highest MOS of 3.895, significantly outperforming GENESES (3.482), Sidon (3.289), and the unprocessed recordings (2.815) ($p < 0.05$, two-sided t -test). These results indicate that, despite not always achieving the best predicted perceptual quality according to NISQA and DNSMOS, DialogueSidon produces speaker-wise outputs that are preferred by human listeners in terms of overall recording quality and speaker consistency.

Multilingual evaluation Table 4 presents multilingual results on five CallFriend language varieties. DialogueSidon consistently outperforms GENESES in p-CER across all evaluated languages, indicating better preservation of spoken content. In addition, DialogueSidon generally achieves higher speaker similarity and VAD accuracy, suggesting more faithful speaker-wise reconstruction. Although GENESES often attains higher NISQA or DNSMOS scores, its much worse p-CER indicates a stronger tendency to distort or remove linguistic information. GENESES (orig) performs consistently worse than the retrained GENESES across all languages and metrics, further confirming that the gap between the two reflects the benefit of retraining on conversational dialogue data. These results suggest that DialogueSidon provides a better tradeoff for full-duplex dialogue data construction, where preserving content and speaker structure is critical.

In-the-wild evaluation To evaluate DialogueSidon on in-the-wild data, we apply all models to OpenDialog, a corpus of monaural dialogue audio sourced from the Internet. Table 5 presents the results. Note that Noisy and Sidon do not produce separated speaker tracks. Therefore WER cannot be measured. DialogueSidon achieves the lowest

Table 4: Multilingual evaluation on CallFriend corpus across five languages. We compare Noisy, Sidon, GENESES, and DialogueSidon with 32-dimensional VAE latents. [†]Noisy and Sidon do not perform speech separation. **Bold** denote the best and worst among separation methods, respectively.

Language	Method	p-CER (%) ↓ <i>Content</i>	NISQA ↑ <i>Recording Quality</i>	DNSMOS ↑ <i>Speaker</i>	Spk. Sim. ↑ <i>Speaker</i>	VAD Acc. ↑ <i>Separation</i>
German	Noisy [†]	—	2.131	3.001	—	—
	Sidon [†]	11.860	3.494	3.586	0.874	0.966
	GENESES (orig)	76.700	3.404	3.426	0.810	0.831
	GENESES	52.300	3.646	3.570	0.850	0.947
	DialogueSidon	14.700	3.214	3.545	0.900	0.954
French	Noisy [†]	—	2.353	3.178	—	—
	Sidon [†]	19.280	3.839	3.792	0.893	0.962
	GENESES (orig)	86.000	3.447	3.484	0.805	0.711
	GENESES	61.660	3.753	3.591	0.881	0.923
	DialogueSidon	25.440	3.369	3.617	0.905	0.945
Japanese	Noisy [†]	—	2.177	2.897	—	—
	Sidon [†]	23.110	3.366	3.455	0.904	0.901
	GENESES (orig)	120.420	3.539	3.466	0.754	0.510
	GENESES	89.110	3.728	3.460	0.824	0.785
	DialogueSidon	48.780	3.161	3.312	0.891	0.796
Spanish	Noisy [†]	—	2.235	2.919	—	—
	Sidon [†]	10.670	3.556	3.615	0.877	0.962
	GENESES (orig)	82.910	3.559	3.360	0.777	0.686
	GENESES	50.480	3.558	3.316	0.822	0.861
	DialogueSidon	18.720	3.142	3.313	0.858	0.931
Mandarin	Noisy [†]	—	2.380	3.214	—	—
	Sidon [†]	25.240	3.682	3.701	0.944	0.951
	GENESES (orig)	120.870	3.678	3.554	0.808	0.542
	GENESES	124.000	3.949	3.684	0.829	0.529
	DialogueSidon	59.960	3.436	3.526	0.868	0.700

Table 5: Evaluation on OpenDialog (in-the-wild data). [†]Noisy and Sidon do not perform speech separation; their metrics are computed on the monaural mixture. For MOS, $\pm$ indicates standard deviation and all pairwise differences are statistically significant ($p < 0.05$, two-sided t -test).

Method	WER (%) ↓ <i>Content</i>	NISQA ↑ <i>Recording Quality</i>	DNSMOS ↑ <i>Speaker</i>	MOS ↑ <i>Subjective Quality</i>
Noisy [†]	—	3.826	3.615	—
Sidon [†]	—	4.676	4.131	—
GENESES (orig)	74.510	3.427	3.479	2.611 $\pm$ 1.157
GENESES	43.790	3.809	3.620	3.131 $\pm$ 1.060
DialogueSidon	13.860	3.568	3.598	3.708 $\pm$ 1.006

WER, substantially outperforming both GENESES variants, consistent with the English and multilingual evaluations. GENESES (orig) shows markedly worse WER than the retrained GENESES (74.51% vs. 43.79%), again demonstrating the importance of conversational training data for in-the-wild dialogue. However, predicted perceptual quality (NISQA, DNSMOS) is lower for DialogueSidon than GENESES. To further validate sound quality on in-the-wild dialogue, we also report subjective evaluation result (MOS). The results were statistically significant between all compared methods. DialogueSidon achieved the highest MOS, which confirms that DialogueSidon produces higher-quality separation and restoration results than GENESES as judged by human listeners.

Inference efficiency For large-scale full-duplex dialogue data construction, inference efficiency is also important. We compared runtime on a single

NVIDIA H100 GPU for a 20-second input. DialogueSidon achieves an RTF of 0.010, compared with 0.604 for GENESES, corresponding to a 60.4 $\times$ speedup. This efficiency advantage likely stems from the relatively small diffusion model used in DialogueSidon (88M parameters), compared with GENESES (393M parameters).

5 Conclusion

We presented DialogueSidon, a model that jointly restores and separates degraded monaural two-speaker dialogue audio into clean speaker-wise tracks by combining an SSL-VAE latent space with a diffusion-based latent predictor. Experiments on English, multilingual, and in-the-wild dialogue datasets show that DialogueSidon substantially improves content preservation and separation quality over a baseline with 60 $\times$ faster inference, suggesting a practical path toward constructing full-duplex dialogue resources from large-scale in-the-wild au-

dio. Future work includes extending to more speakers and broader language coverage.

Acknowledgments

This work was supported by JST Moonshot JPMJMS2011, JST BOOST JPMJBY24C9, JSPS KAKENHI, Grant Number 25KJ0806, and the AIST policy-based budget project “R&D on Generative AI Foundation Models for the Physical Domain.”

References

- Vasu Agrawal, Akinniyi Akinyemi, Kathryn Alvero, Morteza Behrooz, Julia Buffalini, Fabio Maria Carlucci, Joy Chen, Junming Chen, Zhang Chen, Shiyang Cheng, Praveen Chowdary, Joe Chuang, Antony D’Avirro, Jon Daly, Ning Dong, Mark Duppenthaler, Cynthia Gao, Jeff Girard, Martin Gleize, and 65 others. 2025. [Seamless interaction: Dyadic audiovisual motion modeling and large-scale dataset](#). *Preprint*, arXiv:2506.22554.
- Jont B. Allen and David A. Berkley. 1979. [Image method for efficiently simulating small-room acoustics](#). *The Journal of the Acoustical Society of America*, 65(4):943–950.
- Kohei Asai, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. 2026. [Geneses: Unified generative speech enhancement and separation](#). *Preprint*, arXiv:2601.18456.
- Alexandra Canavan and George Zipperlen. 1996a. [CALLFRIEND Canadian French](#).
- Alexandra Canavan and George Zipperlen. 1996b. [CALLFRIEND German](#).
- Alexandra Canavan and George Zipperlen. 1996c. [CALLFRIEND Japanese](#).
- Alexandra Canavan and George Zipperlen. 1996d. [CALLFRIEND Spanish Non-Caribbean Dialect](#).
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- Christopher Cieri, David Miller, and Kevin Walker. 2004. [The fisher corpus: a resource for the next generations of speech-to-text](#). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC’04)*, Lisbon, Portugal. European Language Resources Association (ELRA).
- Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. [Seamless: Multilingual expressive and streaming speech translation](#). *Preprint*, arXiv:2312.05187.
- Tushar Dhyani, Florian Lux, Michele Mancusi, Giorgio Fabbro, Fritz Hohl, and Ngoc Thang Vu. 2025. [High-resolution speech restoration with latent diffusion model](#). In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. [Moshi: a speech-text foundation model for real-time dialogue](#). *Preprint*, arXiv:2410.00037.
- Eduardo Fonseca, Xavier Favory, Jordi Pons, Frederic Font, and Xavier Serra. 2022. [FSD50K: An open dataset of human-labeled sound events](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:829–852.
- Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. [Audio set: An ontology and human-labeled dataset for audio events](#). In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.
- John J. Godfrey and Edward Holliman. 1993. [Switchboard-1 Release 2](#).
- Heitor R. Guimarães, Jiaqi Su, Rithesh Kumar, Tiago H. Falk, and Zeyu Jin. 2025. [DiTSE: High-fidelity generative speech enhancement via latent diffusion transformers](#). *Preprint*, arXiv:2504.09381.
- Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. 2024. [Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation](#). In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 885–890.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *International Conference on Learning Representations*.
- Shigeki Karita, Yuma Koizumi, Heiga Zen, Haruko Ishikawa, Robin Scheibler, and Michiel Bacchiani. 2025. [Miipher-2: A universal speech restoration model for million-hour scale data restoration](#). In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5.

- Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. 2023a. [LibriTTS-R: A Restored Multi-Speaker Text-to-Speech Corpus](#). In *Interspeech 2023*, pages 5496–5500.
- Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Yu Zhang, Wei Han, Ankur Bapna, and Michiel Bacchiani. 2023b. [Miipher: A robust speech restoration model integrating self-supervised speech and text representations](#). In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. [HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033. Curran Associates, Inc.
- Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. [High-fidelity audio compression with improved RVQGAN](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- LDC. 2008a. [CABank Chinese CallHome Corpus](#).
- LDC. 2008b. [CABank English CallHome Corpus](#).
- LDC. 2008c. [CABank Japanese CallHome Corpus](#).
- LDC. 2008d. [CABank Spanish CallHome Corpus](#).
- Keon Lee, Kyumin Park, and Daeyoung Kim. 2023. [Daillytalk: Spoken dialogue dataset for conversational text-to-speech](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Haohe Liu, Xubo Liu, Qiuqiang Kong, Qiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan Wang. 2022. [VoiceFixer: A Unified Framework for High-Fidelity Speech Restoration](#). In *Interspeech 2022*, pages 4232–4236.
- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2025. [DPM-Solver++: Fast solver for guided sampling of diffusion probabilistic models](#). *Machine Intelligence Research*, 22(4):730–751.
- Yi Luo and Nima Mesgarani. 2019. [Conv-TasNet: Surpassing ideal time–frequency magnitude masking for speech separation](#). *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266.
- Min Ma, Yuma Koizumi, Shigeki Karita, Heiga Zen, Jason Riesa, Haruko Ishikawa, and Michiel Bacchiani. 2024. [FLEURS-R: A Restored Multilingual Speech Corpus for Generation Tasks](#). In *Interspeech 2024*, pages 1835–1839.
- Daniele Mirabilli, Alexander Lodermeier, Felix Czwilong, Stefan Becker, and Emanuel A.P. Habets. 2022. [Simulating wind noise with airflow speed-dependent characteristics](#). In *2022 International Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 1–5.
- Gabriel Mittag, Babak Naderi, Assmaa Chehadi, and Sebastian Möller. 2021. [NISQA: A Deep CNN-Self-Attention Model for Multidimensional Speech Quality Prediction with Crowdsourced Datasets](#). In *Interspeech 2021*, pages 2127–2131.
- Wataru Nakata, Yuki Saito, Yota Ueda, and Hiroshi Saruwatari. 2026a. [Sidon: Fast and robust open-source multilingual speech restoration for large-scale dataset cleansing](#). *Preprint*, arXiv:2509.17052.
- Wataru Nakata, Kentaro Seki, Hitomi Yanaka, Yuki Saito, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2026b. [J-CHAT: Japanese large-scale spoken dialogue corpus for spoken dialogue language modeling](#). *Preprint*, arXiv:2407.15828.
- Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023. [Generative spoken dialogue language modeling](#). *Transactions of the Association for Computational Linguistics*, 11:250–266.
- William Peebles and Saining Xie. 2023. [Scalable diffusion models with transformers](#). In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182.
- Yifan Peng, Muhammad Shakeel, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. [OWSM v4: Improving Open Whisper-Style Speech Models via Data Scaling and Cleaning](#). In *Interspeech 2025*, pages 2225–2229.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. [Robust speech recognition via large-scale weak supervision](#). *Preprint*, arXiv:2212.04356.
- Chandan K A Reddy, Vishak Gopal, and Ross Cutler. 2021. [Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors](#). In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497.
- Rajarshi Roy, Jonathan Raiman, Sang gil Lee, Teodor-Dumitru Ene, Robert Kirby, Sungwon Kim, Jaehyeon Kim, and Bryan Catanzaro. 2026. [Personaplex: Voice and role control for full duplex conversational speech models](#). *Preprint*, arXiv:2602.06053.
- Kohei Saijo, Wangyou Zhang, Samuele Cornell, Robin Scheibler, Chenda Li, Zhaocheng Ni, Anurag Kumar, Marvin Sach, Yihui Fu, Wei Wang, Tim Fingscheidt, and Shinji Watanabe. 2025. [Interspeech 2025 URGENT Speech Enhancement Challenge](#). In *Interspeech 2025*, pages 858–862.

- Tim Salimans and Jonathan Ho. 2022. [Progressive distillation for fast sampling of diffusion models](#). In *International Conference on Learning Representations*.
- Robin Scheibler, Eric Bezzam, and Ivan Dokmanić. 2018. [Pyroomacoustics: A python package for audio room simulation and array processing algorithms](#). In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 351–355.
- Robin Scheibler, John R. Hershey, Arnaud Doucet, and Henry Li. 2025. [Source separation by flow matching](#). In *2025 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5.
- Bowen Shi, Andros Tjandra, John Hoffman, Helin Wang, Yi-Chiao Wu, Luya Gao, Julius Richter, Matt Le, Apoorv Vyas, Sanyuan Chen, Christoph Feichtenhofer, Piotr Dollár, Wei-Ning Hsu, and Ann Lee. 2025. [SAM Audio: Segment anything in audio](#). *Preprint*, arXiv:2512.18099.
- Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. [Roformer: Enhanced transformer with rotary position embedding](#). *Neurocomputing*, 568:127063.
- Jennifer Tracey, David Graff, Song Chen, and Stephanie Strassel. 2025. [BOLT CTS CALLFRIEND CALL-HOME Mainland Mandarin Chinese Audio](#).
- Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. [Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21390–21402, Miami, Florida, USA. Association for Computational Linguistics.
- Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. 2023. [TF-GRIDNET: Making time-frequency domain models great again for monaural speaker separation](#). In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Gordon Wichern, Joe Antognini, and 1 others. 2019. WHAM!: Extending speech separation to noisy environments. In *Proc. Interspeech*.
- Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. 2017. [Permutation invariant training of deep models for speaker-independent multi-talker speech separation](#). In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245.
- Yu Zhang, Wei Han, and 1 others. 2023. Google USM: Scaling automatic speech recognition beyond 100 languages. *arXiv*, abs/2303.01037.
- Zizheng Zhang, Chen Chen, Hsin-Hung Chen, Xiang Liu, Yuchen Hu, and Eng Siong Chng. 2024. [Noise-aware speech separation with contrastive learning](#). In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1381–1385.
- Han Zhu, Wei Kang, Liyong Guo, Zengwei Yao, Fangjun Kuang, Weiji Zhuang, Zhaoqing Li, Zhifeng Han, Dong Zhang, Xin Zhang, Xingchen Song, Long Lin, and Daniel Povey. 2025. [Zipvoice-dialog: Non-autoregressive spoken dialogue generation with flow matching](#). *Preprint*, arXiv:2507.09318.
- Liu Ziyin, Tilman Hartwig, and Masahito Ueda. 2020. [Neural networks fail to learn periodic functions and how to fix it](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1583–1594. Curran Associates, Inc.

A Details of the degradation pipeline

We used the degradation pipeline in the following order.

1. Reverberation: We used pyroomacoustics (Scheibler et al., 2018) for simulating room impulse responses (RIRs). Specifically, random RT60 and rectangular cuboid room dimensions were drawn from $\mathcal{U}(0.1, 1.0)$ seconds and $\mathcal{U}(2, 20)$ m respectively. Based on the drawn RT60 and room dimensions, wall absorption and maximum order of the image-source method (Allen and Berkley, 1979) were calculated using Sabine’s equation. Then, RIRs were simulated.
2. Background noise: We formed a noise pool from AudioSet (Gemmeke et al., 2017), Free Music Archive, WHAM! (Wichern et al., 2019), FSD50K (Fonseca et al., 2022), and synthetic wind noise generated by SC-Wind-Noise-Generator (Mirabilli et al., 2022). For each clean utterance, we randomly sampled a single noise recording from this pool. The selected noise was looped to exceed the utterance duration and then truncated to match exactly, and it was added at an SNR drawn from $\mathcal{U}(-5, 20)$ dB.
3. Band limitation: The input speech was randomly resampled at {8, 16, 22.05, 24, 44.1, 48} kHz sampling rate before being converted back to the original sampling rate.
4. Clipping: The input speech was randomly clipped by setting its new minimum value

to the value corresponding to a quantile uniformly chosen between the 0th and 10th percentiles, and its new maximum value to the value corresponding to a quantile uniformly chosen between the 90th and 100th percentiles of the original signal.

5. Codec: We applied the MP3 compression with a random average bitrate ranging from 65 kbps to 245 kbps.
6. Packet loss: Random 9% segments of speech were selected for packet loss. For each segment duration sampled from $\mathcal{U}(20, 200)$ milliseconds were selected to be replaced with zeros to simulate packet loss.
7. Mixing: From the degraded tracks $\tilde{y}_1, \tilde{y}_2$, monaural dialogue audio was created as a weighted sum $x = w \cdot \tilde{y}_1 + (1 - w) \cdot \tilde{y}_2$, where the weight w is drawn from $\mathcal{U}(0.3, 0.7)$.