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

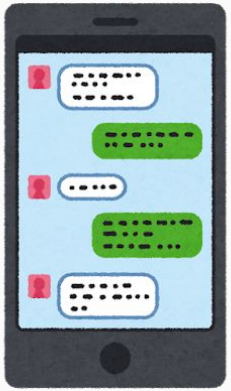
Wataru Nakata[1,2], Yuki Saito[1,2], Kazuki Yamauchi[1]
Emiru Tsunoo[1], Hiroshi Saruwatari[1]

1. The University of Tokyo
2. AIST

This work was supported by JST Moonshot JPMJMS2011, JST BOOST JPMJBY24C9, JSPS KAKENHI, Grant Number 25KJ0806, the AIST policy-based budget project ``R\&D on Generative AI Foundation Models for the Physical Domain,'' and based on results obtained from the BRIDGE Program (R7-H05), implemented by the Cabinet Office, Government of Japan.

Spoken Dialogue Langugage Models (SDLMs)

Text dialogue models



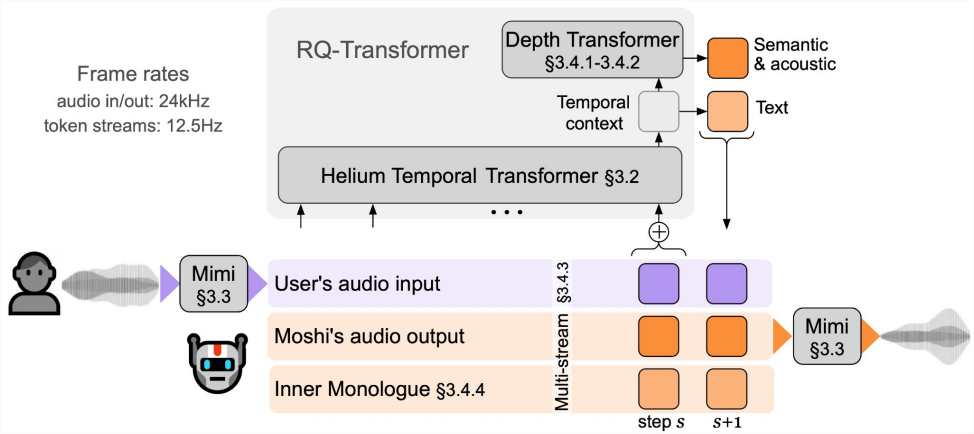
Listen then talk
Speaker turn are explicit

SDLMs: speech-based

Listen & Talk at the same time
Speaker turns are implicit

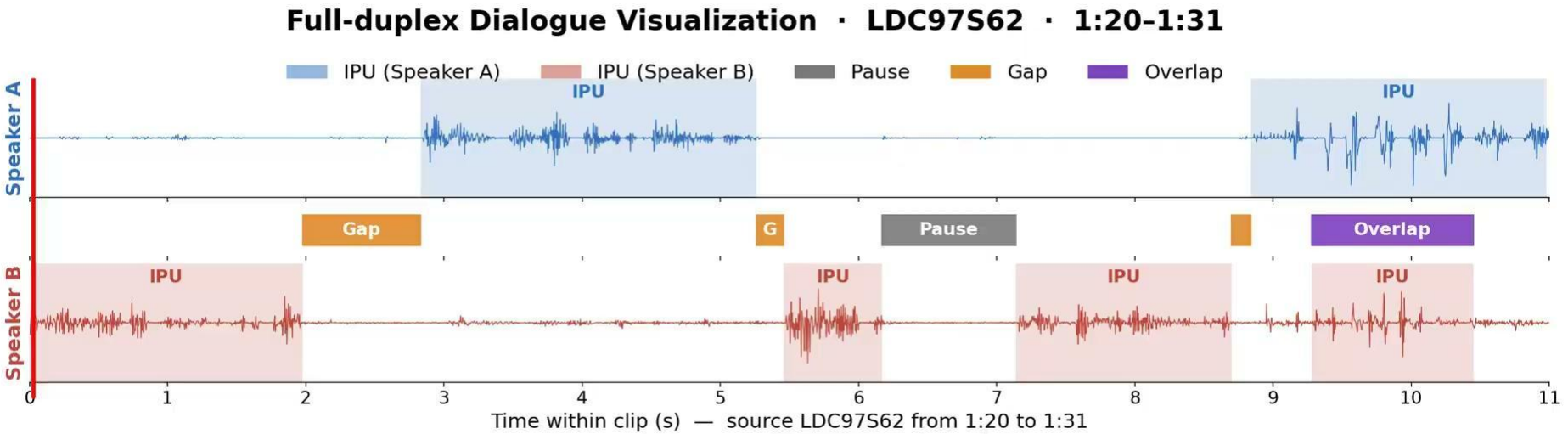
Example of SDLMs

GPT-live by OpenAI



Moshi[Defossez+24]

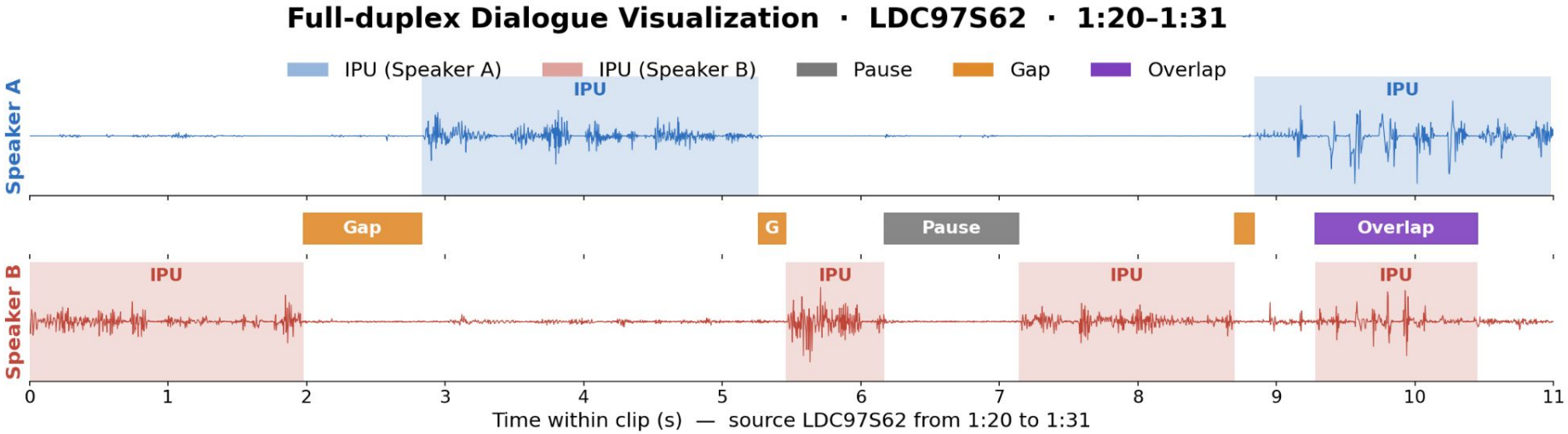
What data does SDLM require



Training requires full-duplex (FD) dialogue data

- Recording with Speaker-separated tracks

What data does SDLM require



Training requires full-duplex (FD) dialogue data

- Recording with Speaker-separated tracks

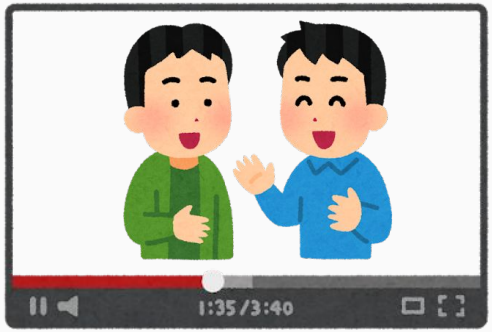
Scarcity of FD speech

Name	Recording environment	Duration [h]	Quality	Scalability
DailyTalk[Lee+23]	Studio	20	✅	👎
Fisher[Cieri+02]	Telephone	2k	👎	👎

We need scalable way to collect high quality FD speech

Conventional approach & Our goal

Noisy telephone recording



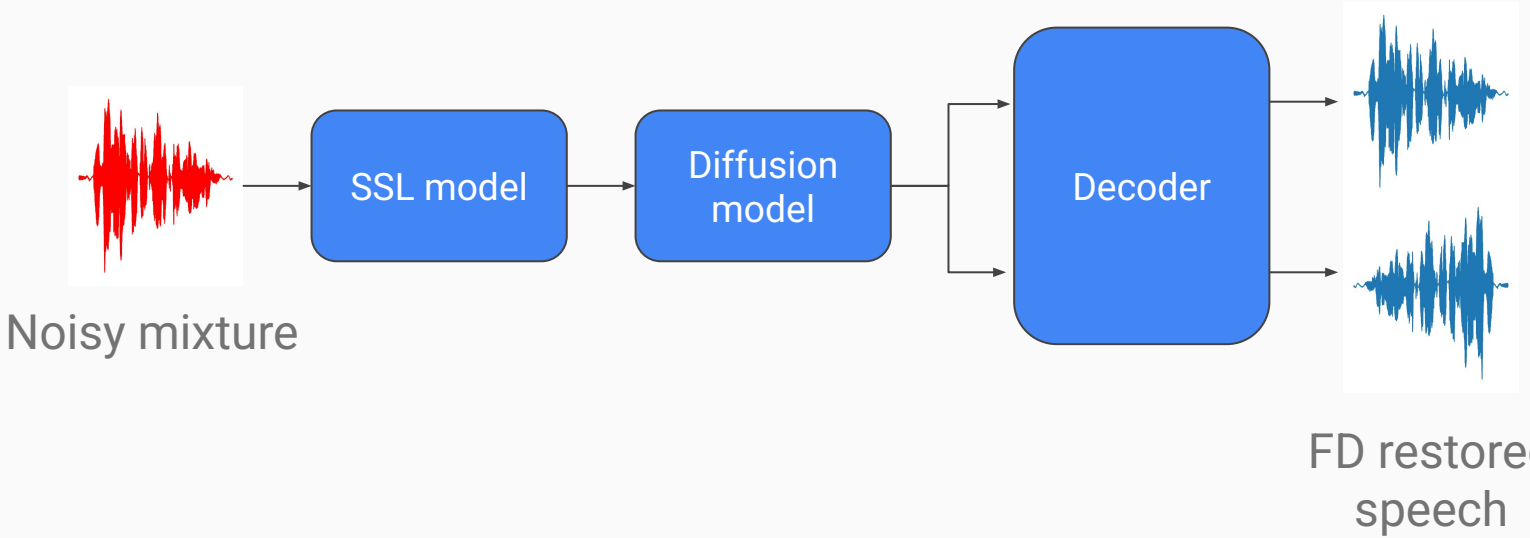
DialogueSidon (Proposed)
Restoration & Seperation

High quality FD dialogue

Convert in-the-wild data to
high quality FD dialogue for SDLM training

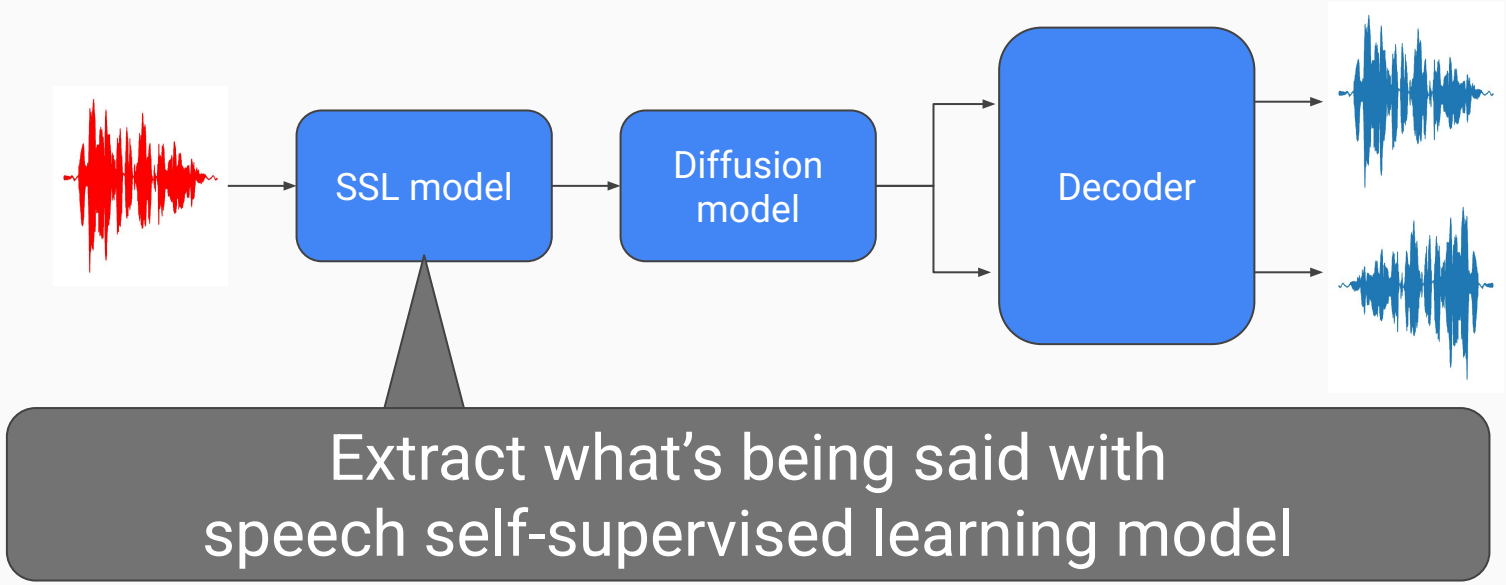
Proposed model: DialogueSidon

Extends speech restoration model Sidon[Nakata+26] with separation capability



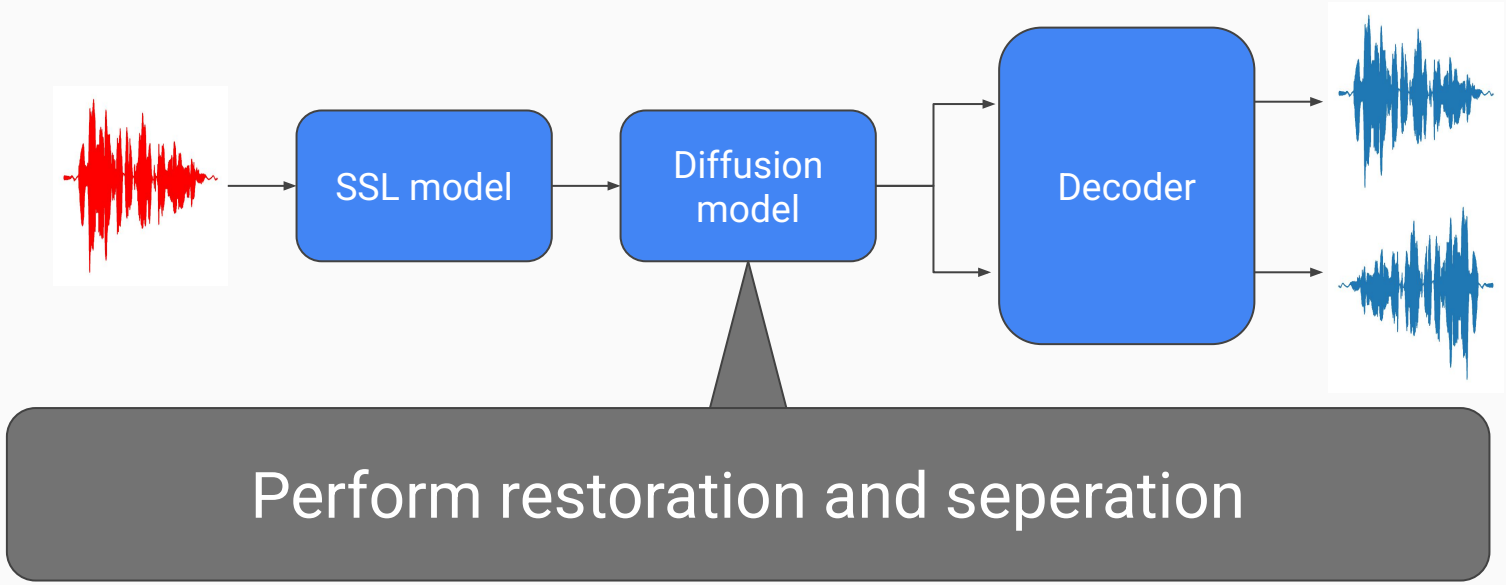
Proposed method

Extends speech restoration model Sidon[Nakata+26]
with separation capability



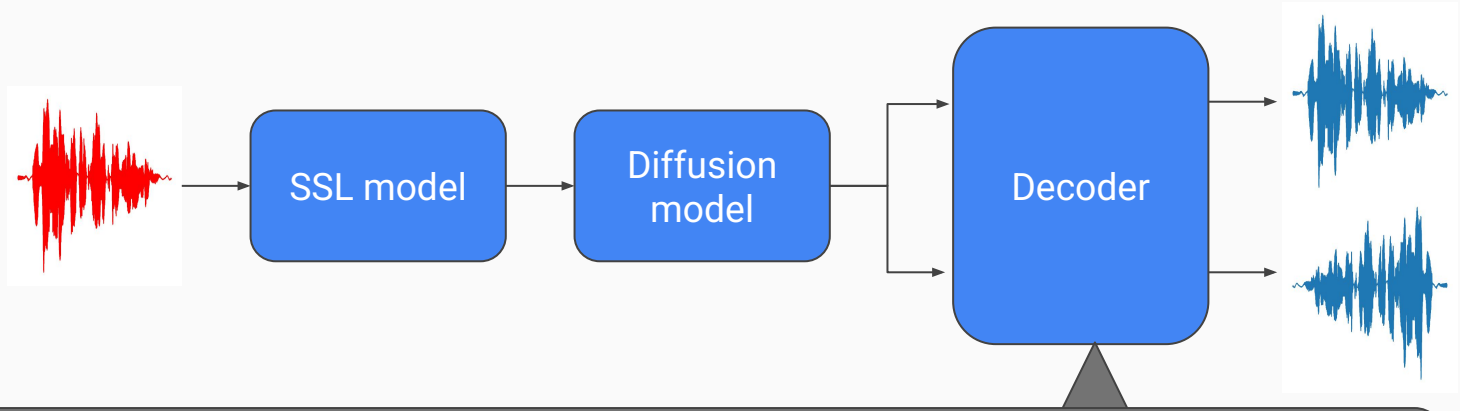
Proposed method

Extends speech restoration model Sidon[Nakata+26]
with separation capability



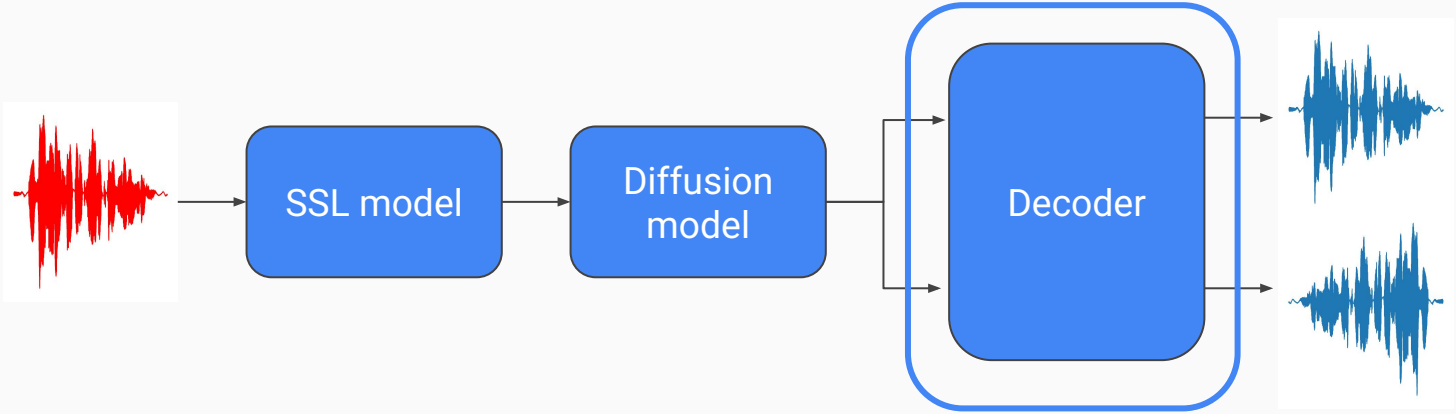
Proposed method

Extends speech restoration model Sidon[Nakata+26]
with separation capability



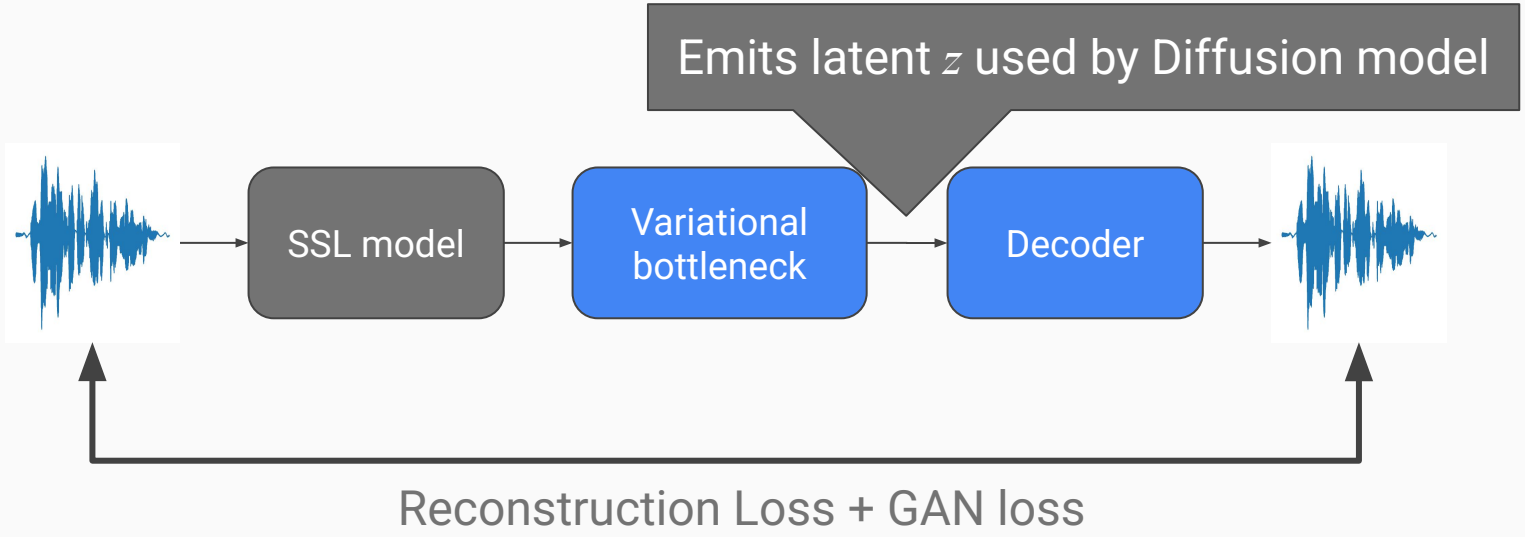
Training stage1

Extends speech restoration model Sidon[Nakata+26]
with separation capability



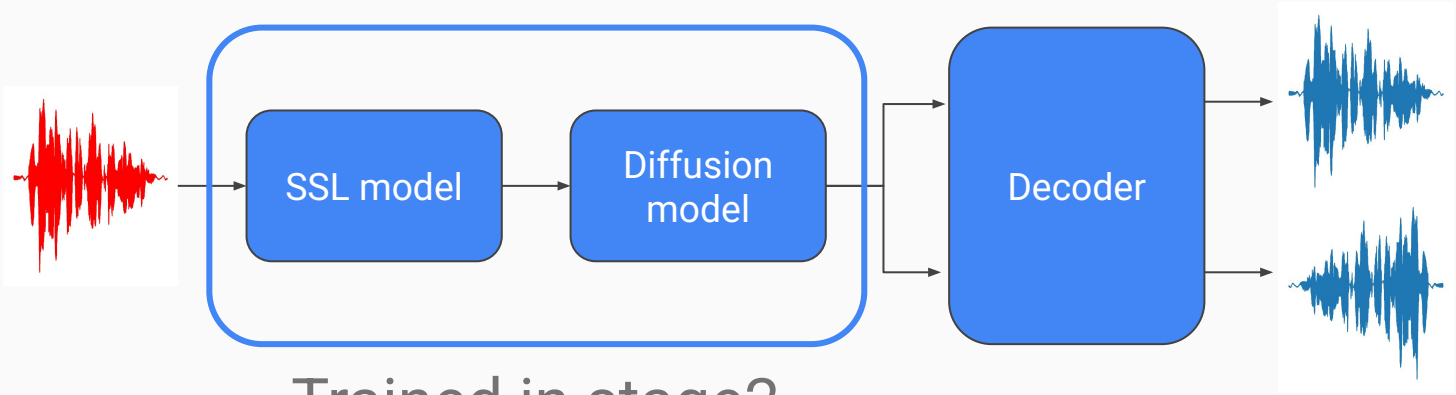
Trained in stage1

Training stage 1: SSL-VAE training



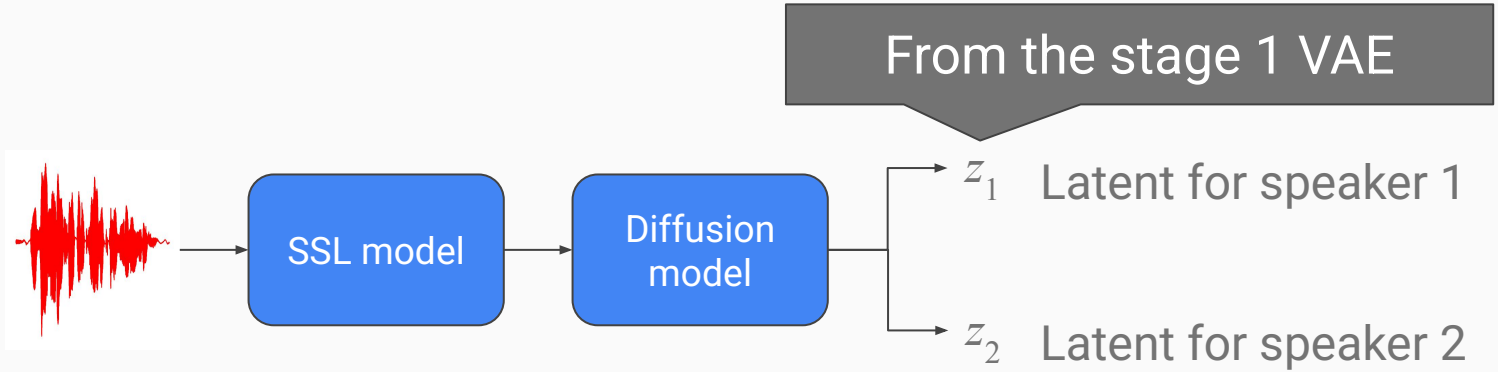
Training stage2

Extends speech restoration model Sidon[Nakata+26]
with separation capability



Trained in stage2

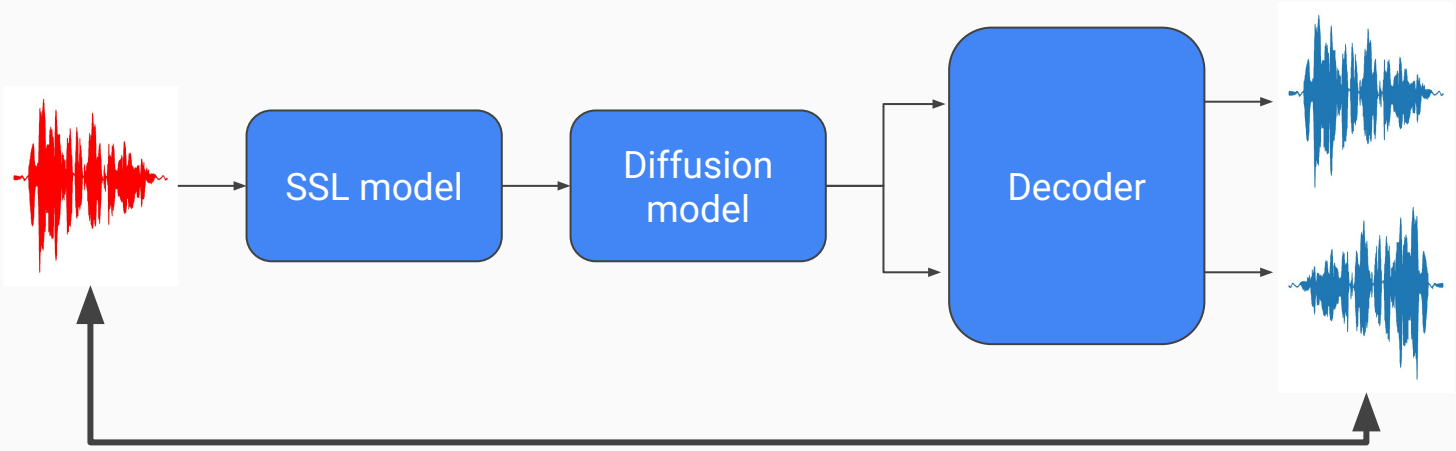
Training stage 2: Diffusion model training



Learn mapping from degraded speech
to separated & clean latent

Paired data simulation

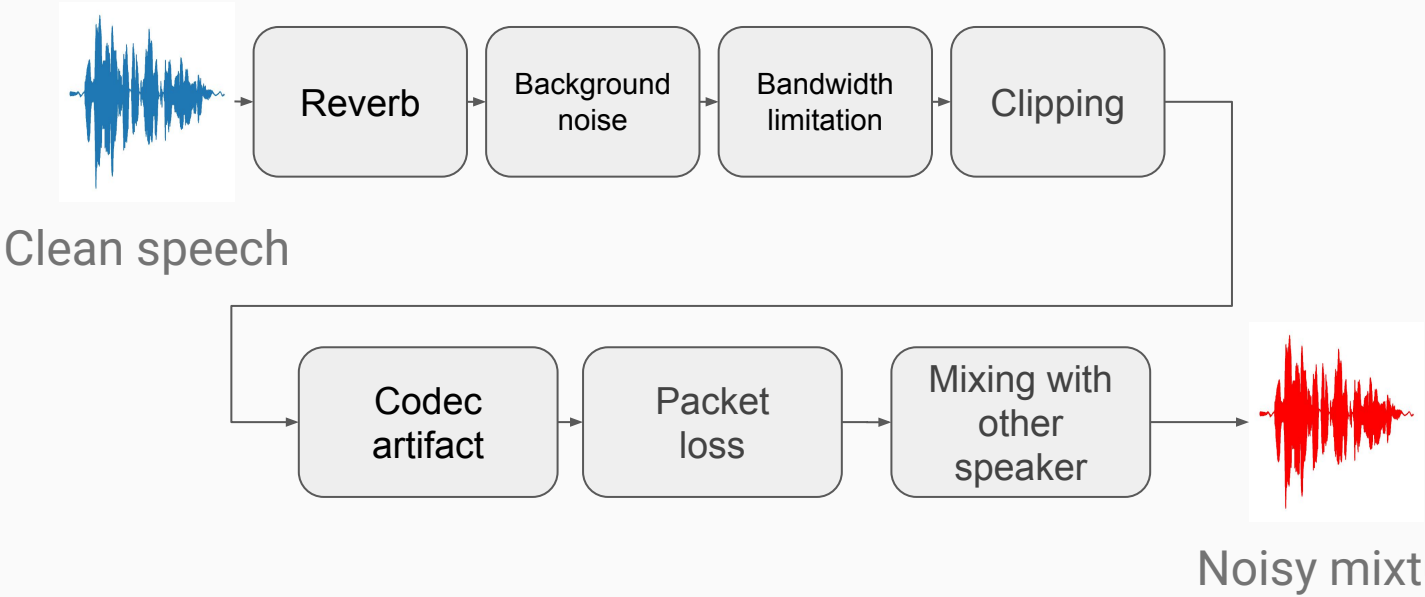
Extends speech restoration model Sidon[Nakata+26] with separation capability



How can we get this paired data🤔

Dataset preparation

Simulate real world noisy data similar to previous work[Saijo+25]



Training Data

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

Collect multilingual FD corpus

Preprocess with speech restoration for better quality

Evaluation

Goal: Convert in-the-wild data to high quality FD dialogue for SDLM training

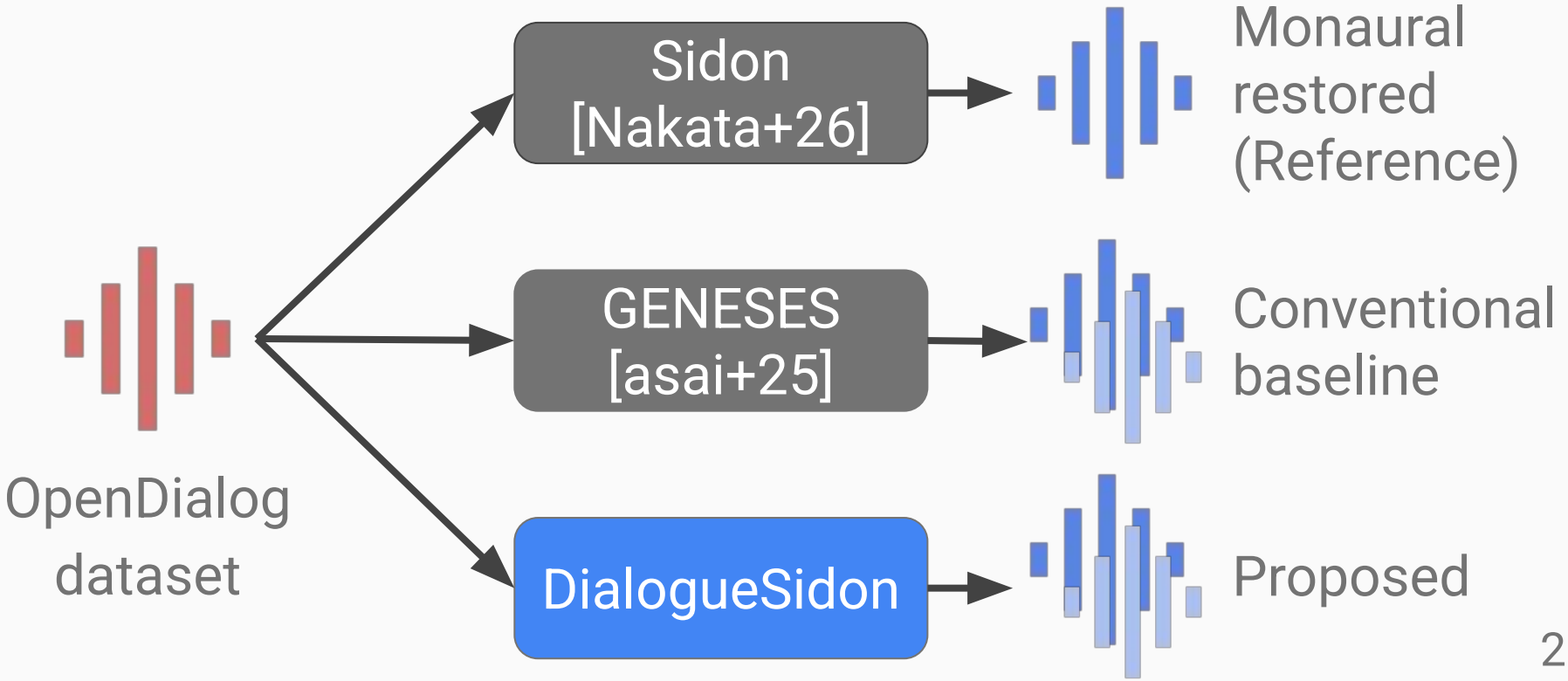
Q1: Are restoration & separation working?

- WER: Is the content preserved?
  - word error rate between the ground truth transcript vs. ASR model inferred transcript
- Human mean opinion score (MOS): Perceptually good?
  - Asked 90 raters to rate the quality of restored & separated speech in 1-5 scale

Q2: Is it fast enough to be scaled?

- RTF: ratio of processing time / processed time

Compared methods



Results

Method	WER [%]	MOS	RTF
Noisy	—	—	
Sidon	—	—	
GENESES	43.790	3.131	0.064
DialogueSidon	**13.860**	**3.708**	**0.010**

Best in MOS, WER and RTF

Summary

DialogueSidon:

- Joint separation & restoration model for scaling out SDLM training data
- Achieves better results than previous model

Future work:

Train SDLM to see its effectiveness