Speech, Language and AI Lab

Yossi Keshet

May 12, 2024

SIPL 50 YEARS SIGNAL AND IMAGE PROCESSING LAB

TECHNION Israel Institute of Technology

The Andrew and Erna Viterbi Faculty of ELECTRICAL & COMPUTER ENGINEERING

Outline

- The Lab
- Speech synthesis:
  - DiffAR: Denoising Diffusion Autoregressive Model for Raw Speech Waveform Generator
  - Spectral analysis of diffusion models
  - SclaerGAN
- Speech recognition and processing
  - Self-supervised Speaker Diarization
  - Keyword spotting

DiffAR: Denoising Diffusion Autoregressive Model for Raw Speech Waveform Generator

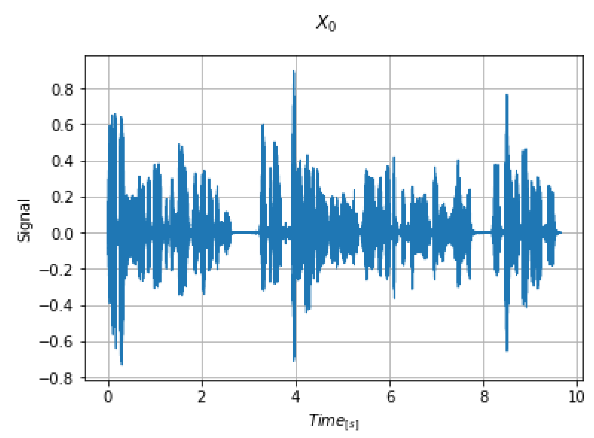
Benita, Elad, and Keshet, 2024

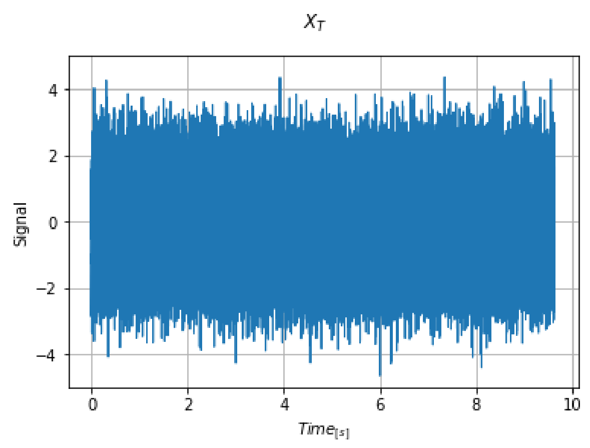
Diffusion models

Forward Markovian process (fixed)

$$x_0 \rightleftarrows x_1 \rightleftarrows x_2 \quad \cdots\cdots \quad \rightleftarrows x_T$$

Reverse Markovian process (trainable)



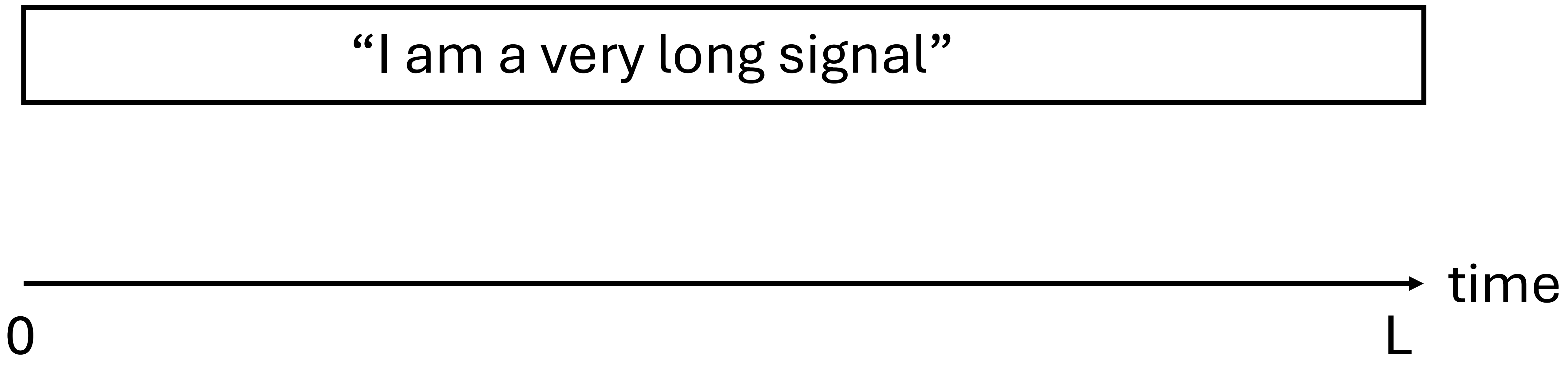


Autoregressive approach

Decompose the original problem into sub-problems

Taking advantage of the temporal behavior of the audio signal

Given that we want to produce a long signal of length L



Autoregressive approach

Decompose the original problem into sub-problems

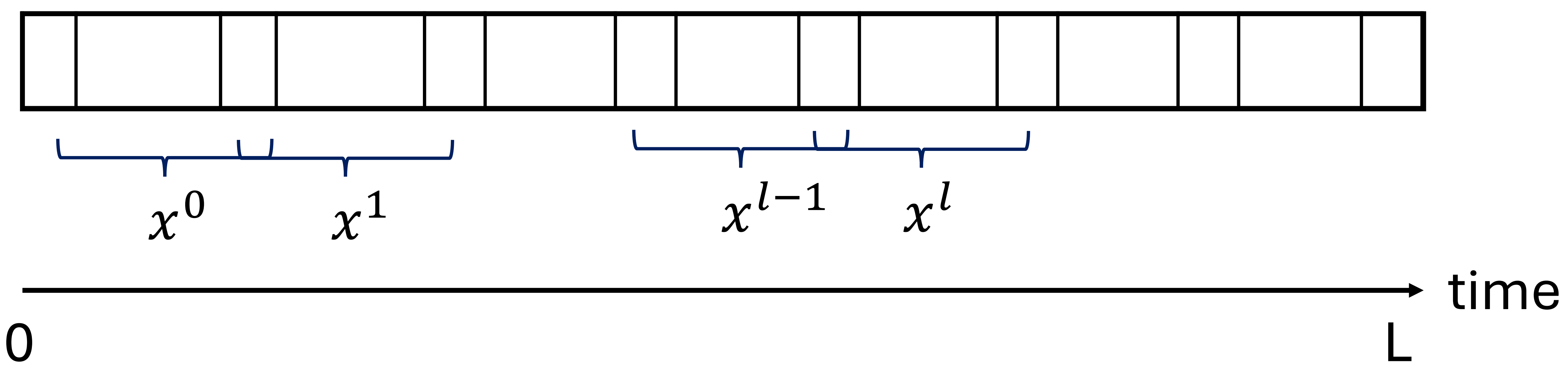
Taking advantage of the temporal behavior of the audio signal

Given that we want to produce a long signal of length L

Step 1 - Breaking down the whole signal into many small frames.

Each couple of adjacent frames are overlapping each other.

Step 2 - Generating each frame separately.



Autoregressive approach

Decompose the original problem into sub-problems

Taking advantage of the temporal behavior of the audio signal

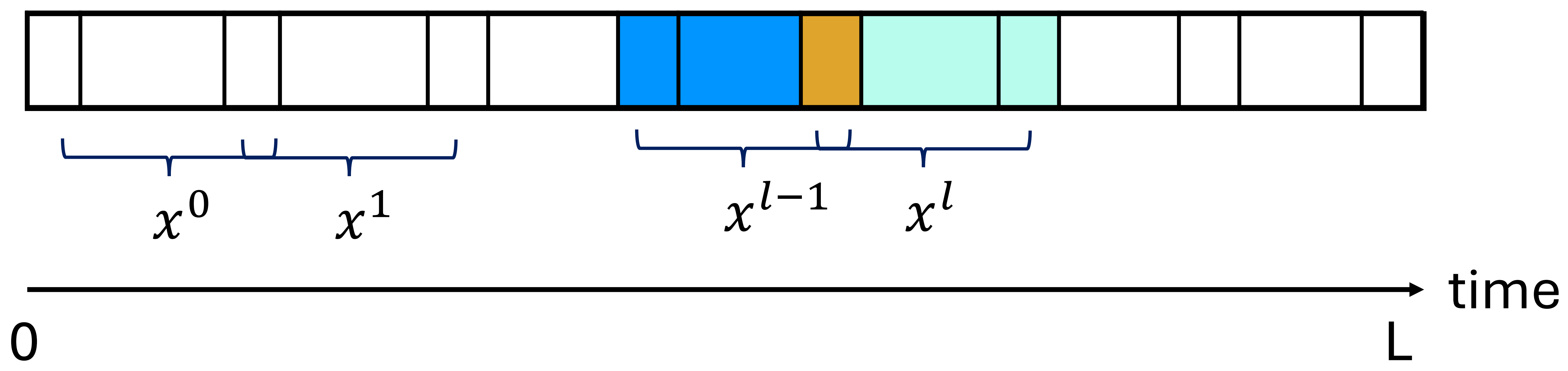
Given that we want to produce a long signal of length L

<u>Step 1 - Breaking down the whole signal into many small frames.</u>

Each couple of adjacent frames are overlapping each other.

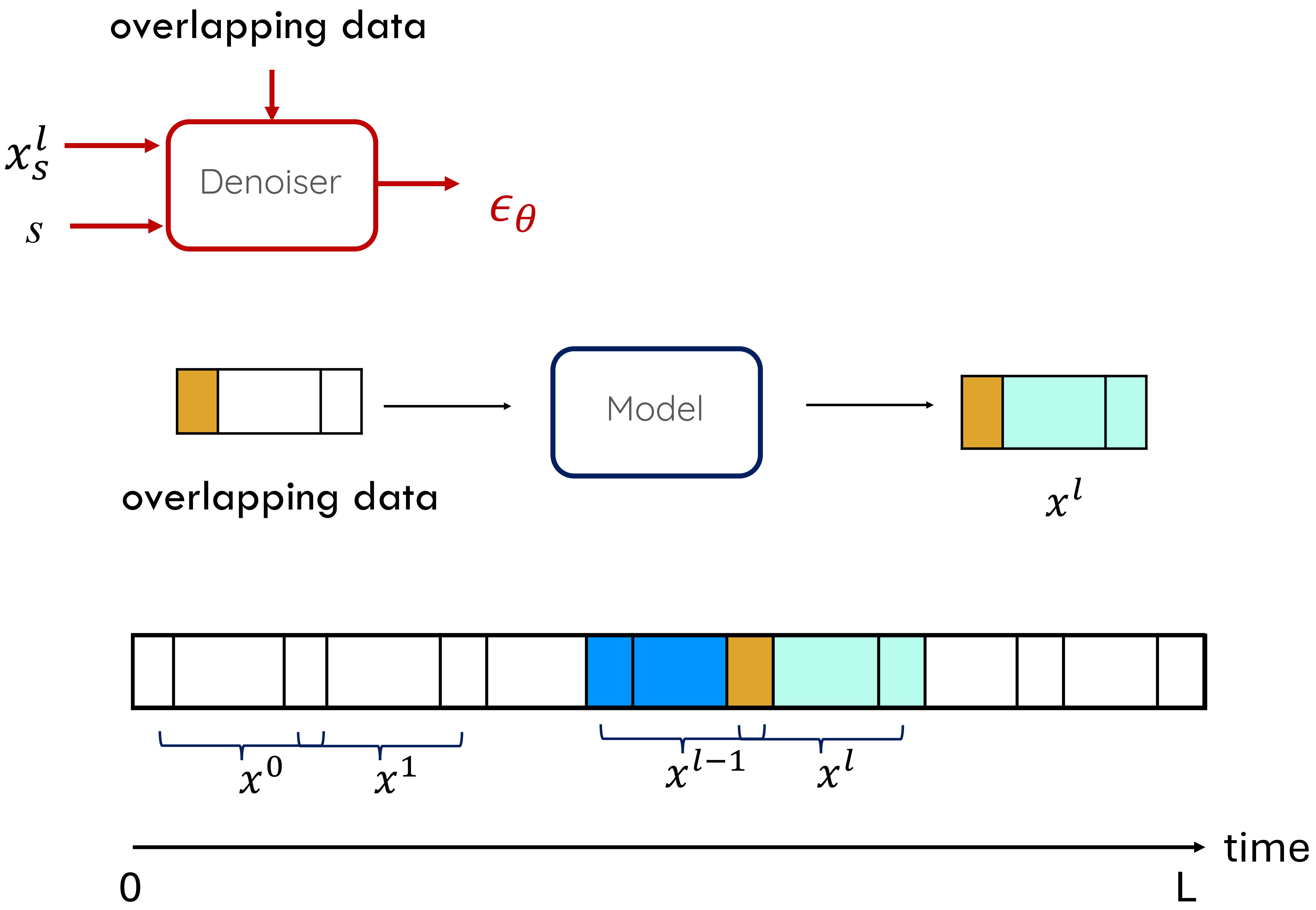
<u>Step 2 - Generating each frame separately</u>.

Generating each frame is conditioned on a portion of the previously generated one



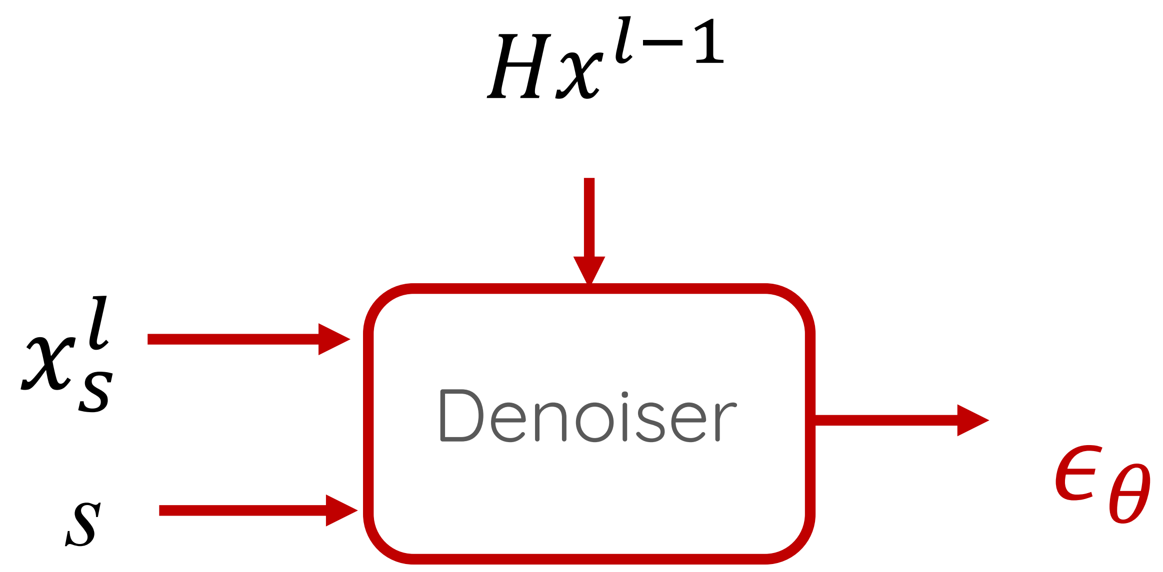
How can it be formulated?

Modeling


overlapping data
x_s^l
s
Denoiser
ϵ_θ
Model
overlapping data
x^l
x^0
x^1
x^{l-1}
x^l
time
0
L

Modeling

Inpainting problem:

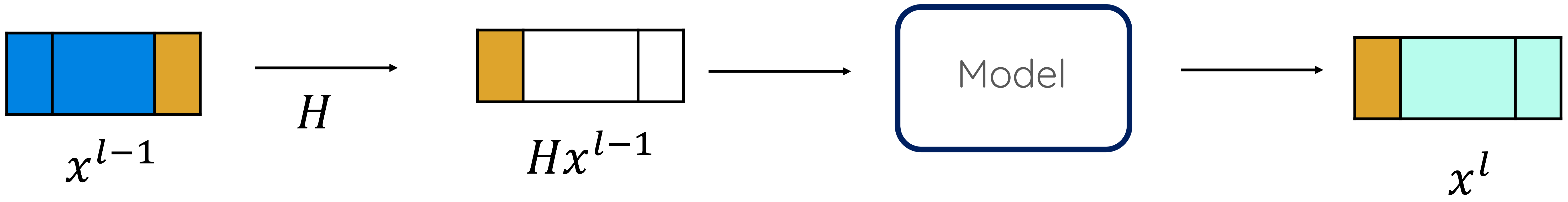


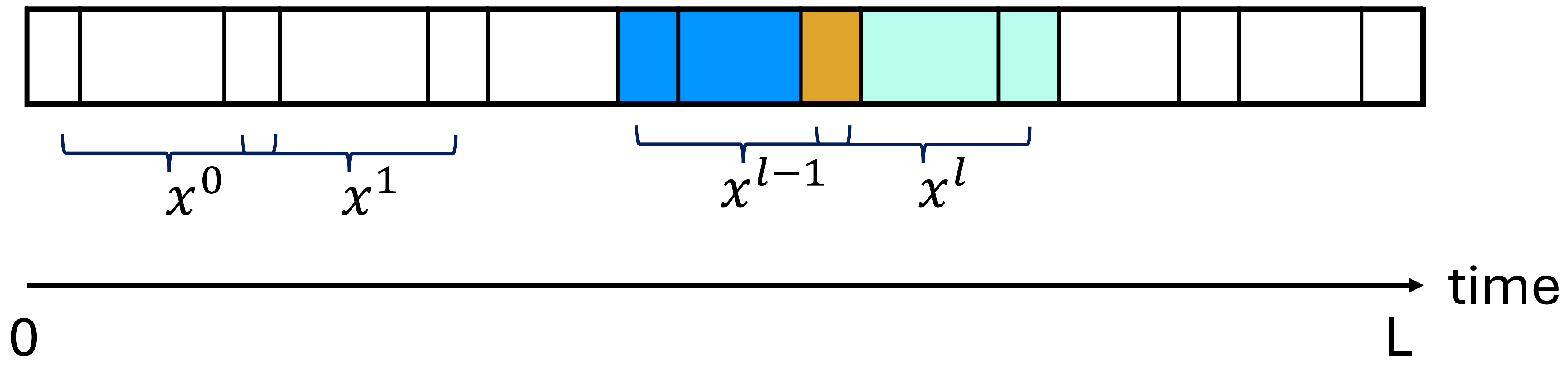
Training procedure:

$$\mathcal{L}_s = \mathbb{E}_{\mathbf{x}_0^l, \boldsymbol{\epsilon}_s} \left[\left\| \boldsymbol{\epsilon}_\theta \left(\sqrt{\bar{\alpha}_s} \mathbf{x}_0^l + \sqrt{1 - \bar{\alpha}_s} \boldsymbol{\epsilon}_s, \mathbf{H}\mathbf{x}^{l-1}, \mathbf{y}^l, s \right) - \boldsymbol{\epsilon}_s \right\|^2 \right]$$

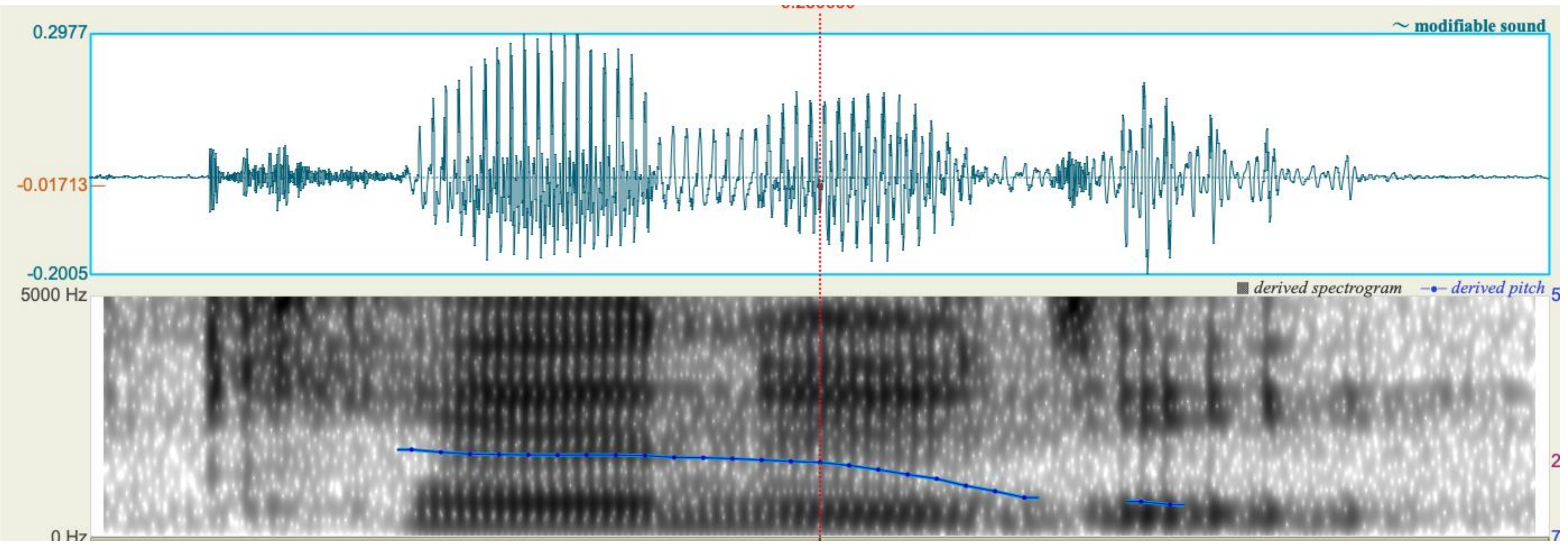
Sampling procedure:

$$\mathbf{x}_s^l = \frac{1}{\sqrt{\bar{\alpha}_s}} \left(\mathbf{x}_{s+1}^l - \frac{1 - \alpha_s}{\sqrt{1 - \bar{\alpha}_s}} \boldsymbol{\epsilon}_\theta \left(\mathbf{x}_{s+1}^l, \mathbf{H}\hat{\mathbf{x}}^{l-1}, \mathbf{y}^l, s \right) \right) + \sigma_s \mathbf{z}_s \ ,$$

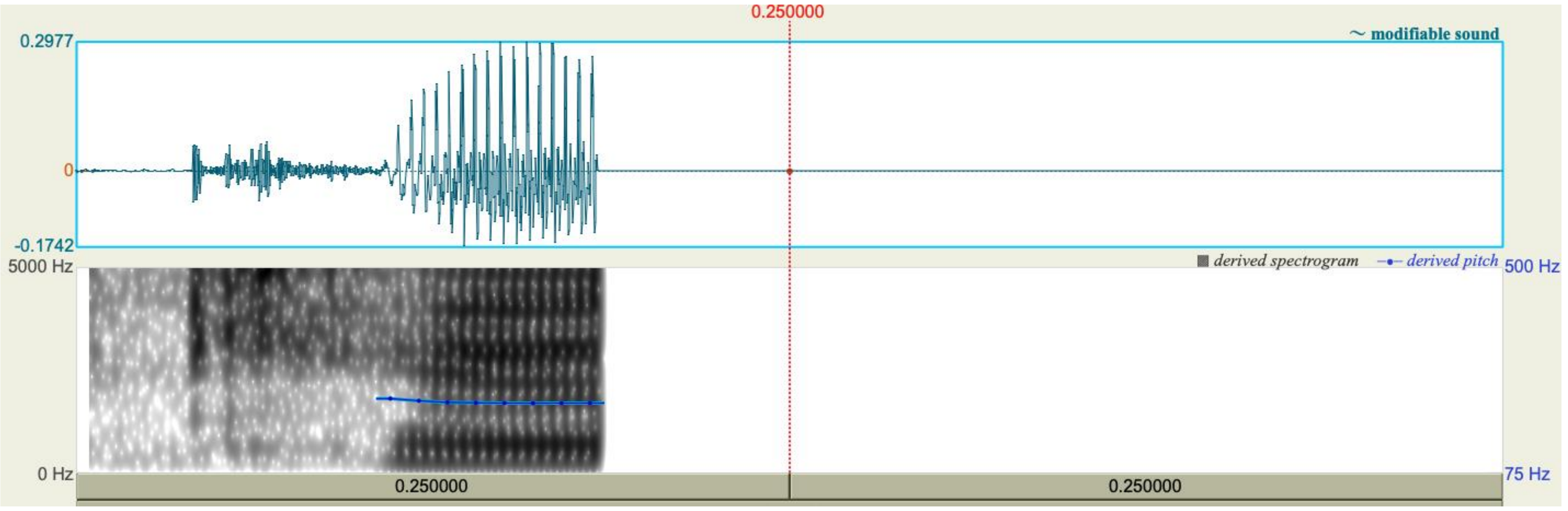




Original audio

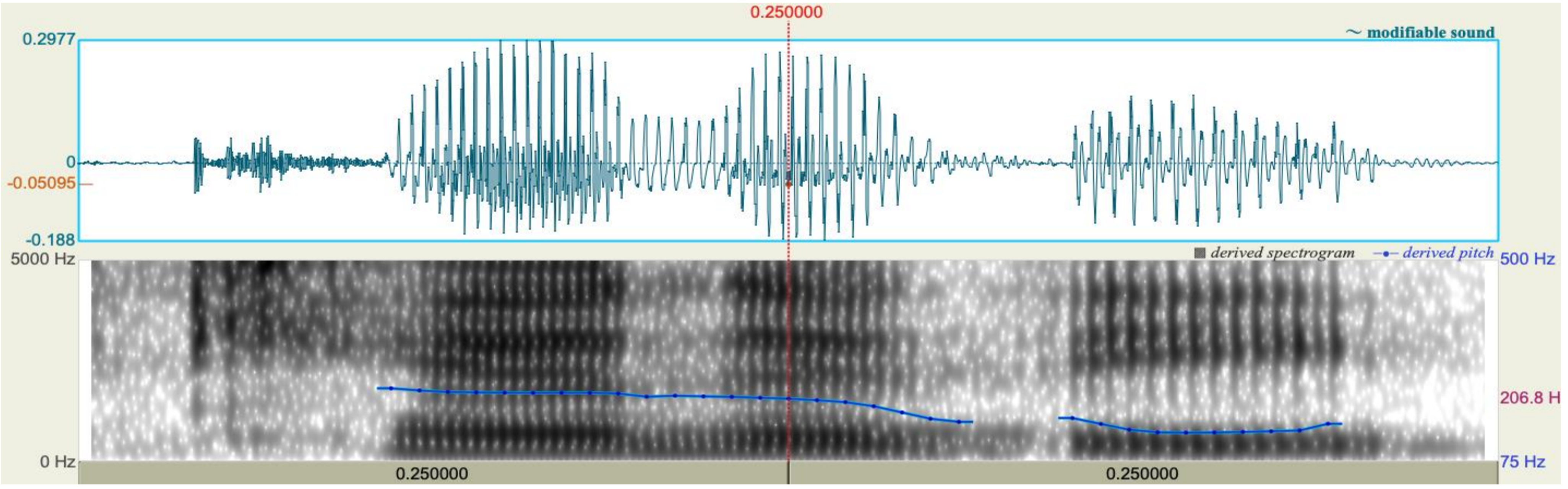


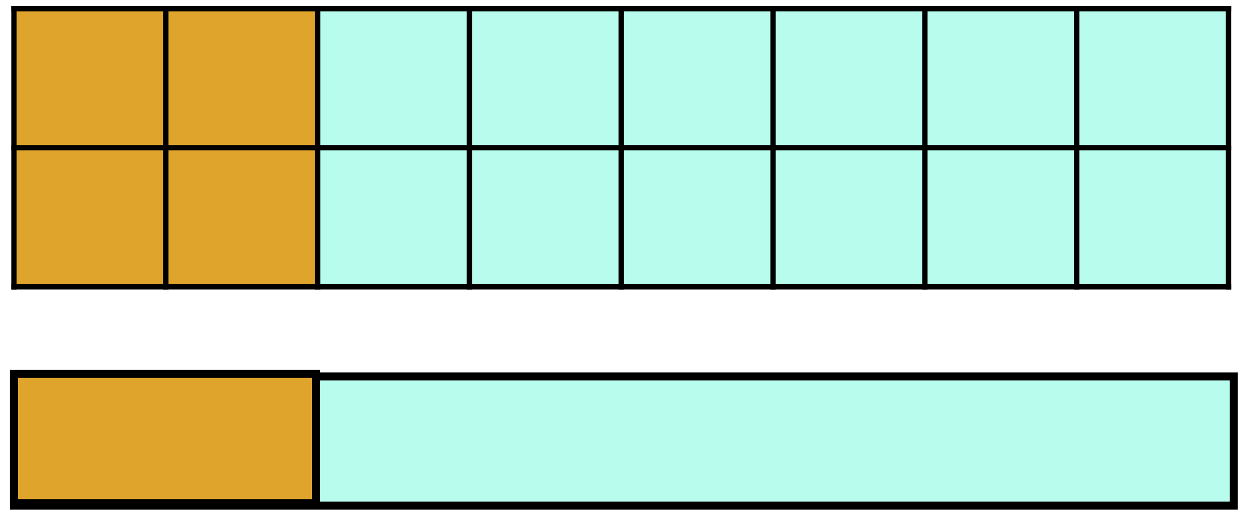
conditioned audio



conditioned phonemes

generated audio





Examples

which had come to rest on a stack of paper.

An examination of certain construction work appearing in the background of this photograph revealed that the picture was taken between March 8

Personal relations

which carry the major responsibility for supplying information about potential threats

who seldom let a session go by without visiting Newgate.


TONE-BYPASS
TONE

Evaluation

Method	↑MOS	↑MOS scaled	↑MUSHRA	↓CER(%)	↓WER(%)
Ground truth	3.98 ± 0.08	4.70 ± 0.09	71.2 ± 2.0	0.89	2.13
WaveGrad 2	3.61 ± 0.09	4.26 ± 0.10	63.8 ± 2.3	3.47	5.75
DiffAR (200 steps)	3.75 ± 0.08	4.43 ± 0.10	65.7 ± 2.2	2.67	6.16
DiffAR (1000 steps)	3.77 ± 0.08	$\mathbf{4.45 \pm 0.09}$	$\mathbf{66.7 \pm 2.2}$	**1.95**	**4.65**

Table 4: VITS (Kim et al., 2021)

Method	↑MUSHRA
Ground truth	74.9 ± 2.2
DiffAR (200 steps)	69.1 ± 2.2
DiffAR (1000 steps)	$\mathbf{71.5 \pm 2.2}$
VITS	69.0 ± 2.3

Table 5: Grad-TTS (Popov et al., 2021)

Method	↑MUSHRA
Ground truth	73.7 ± 2.4
DiffAR (200 steps)	$\mathbf{69.4 \pm 2.5}$
DiffAR (1000 steps)	67.7 ± 2.6
Grad-TTS	68.5 ± 2.5

Table 6: ProDiff (Huang et al., 2022b)

Method	↑MUSHRA
Ground truth	70.0 ± 2.1
DiffAR (200 steps)	66.6 ± 2.4
DiffAR (1000 steps)	$\mathbf{67.5 \pm 2.3}$
ProDiff	64.6 ± 2.4

Table 7: DiffGAN-TTS (Liu et al., 2022)

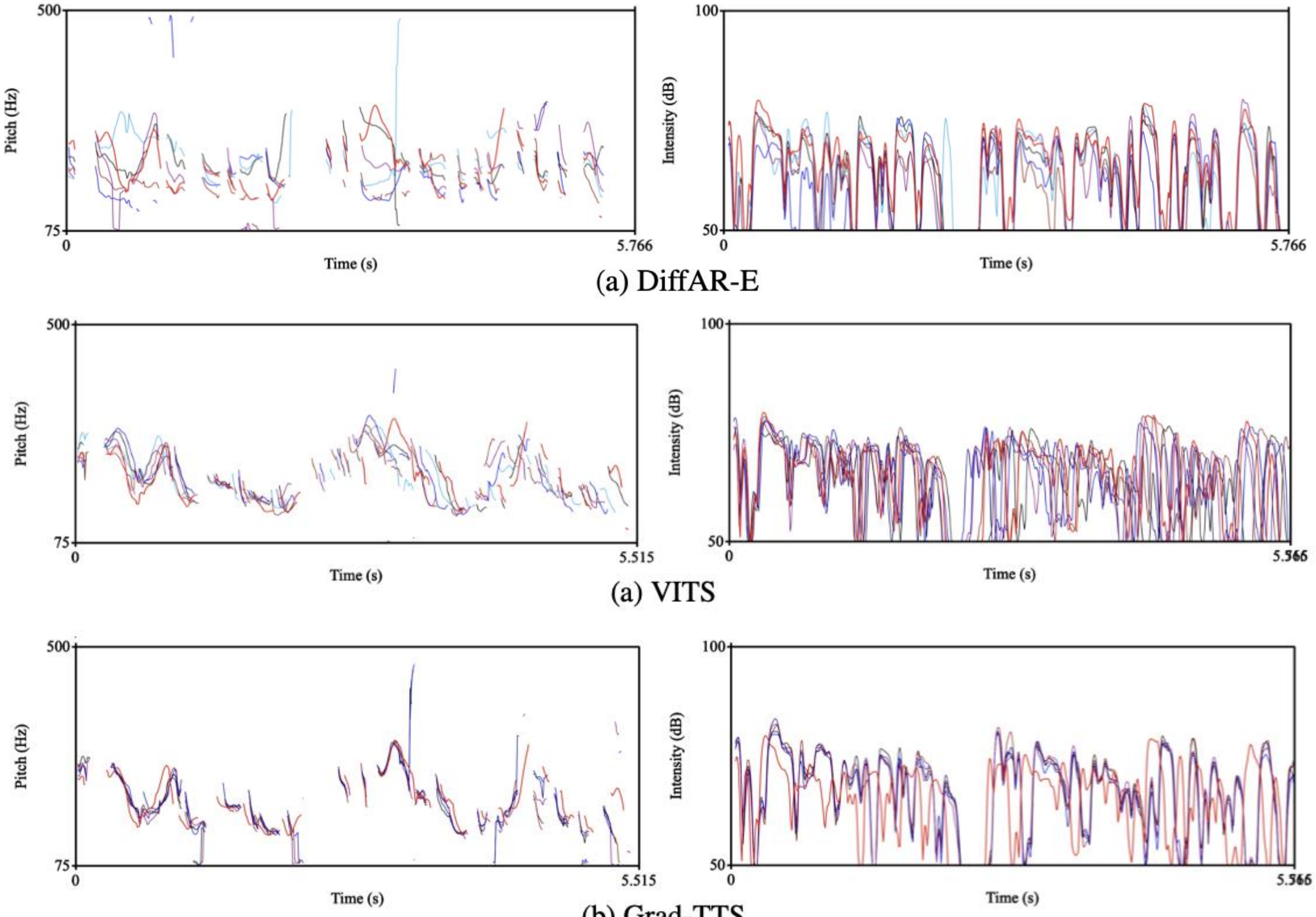
Method	↑MUSHRA
Ground truth	71.2 ± 2.0
DiffAR (200 steps)	$\mathbf{69.5 \pm 2.1}$
DiffAR (1000 steps)	68.4 ± 2.2
DiffGAN-TTS	68.0 ± 2.2

Innovativeness

Original audio

Synthesized audio

pitch amplitude



(a) DiffAR-E

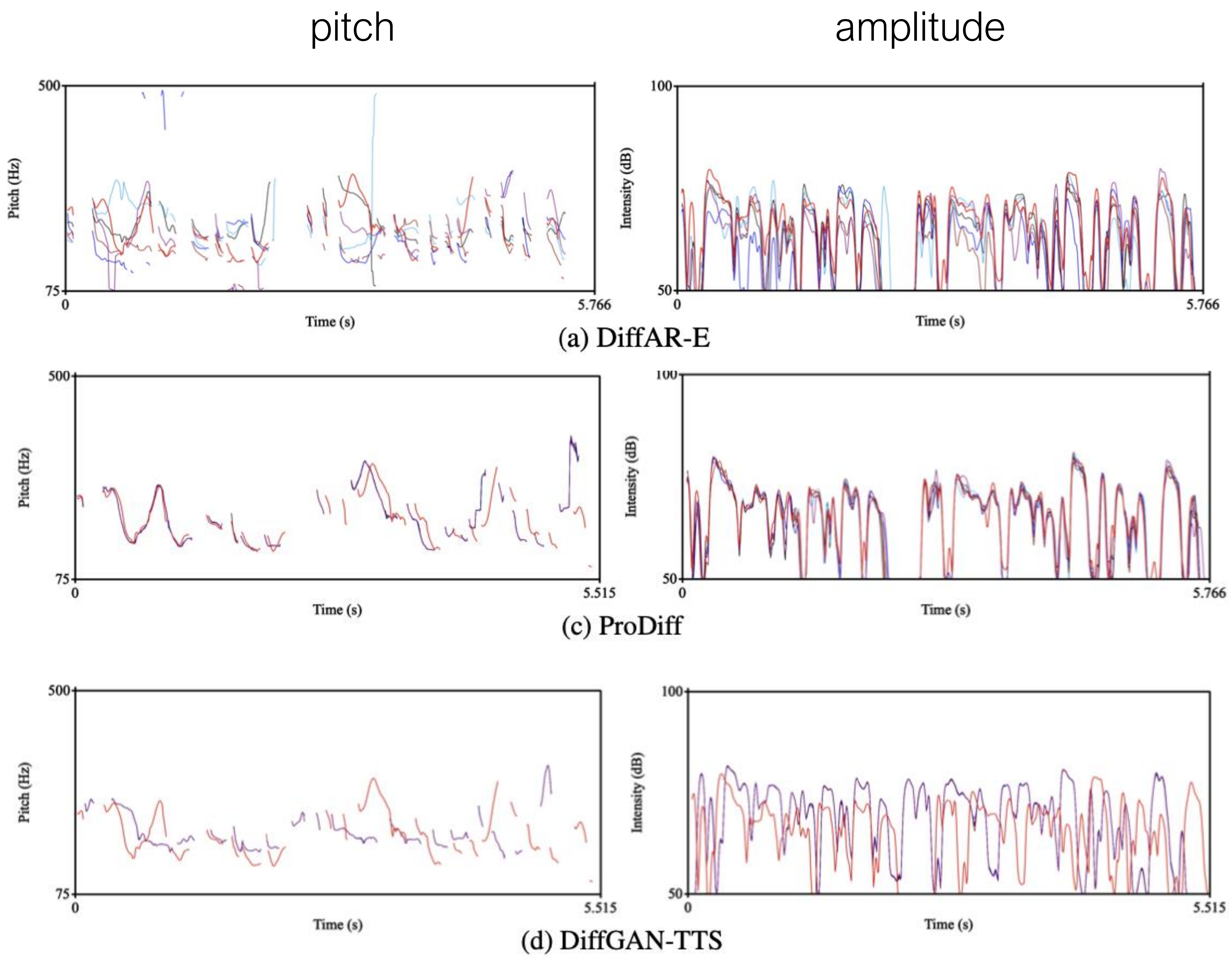
(a) VITS

(b) Grad-TTS

Innovativeness

Original audio

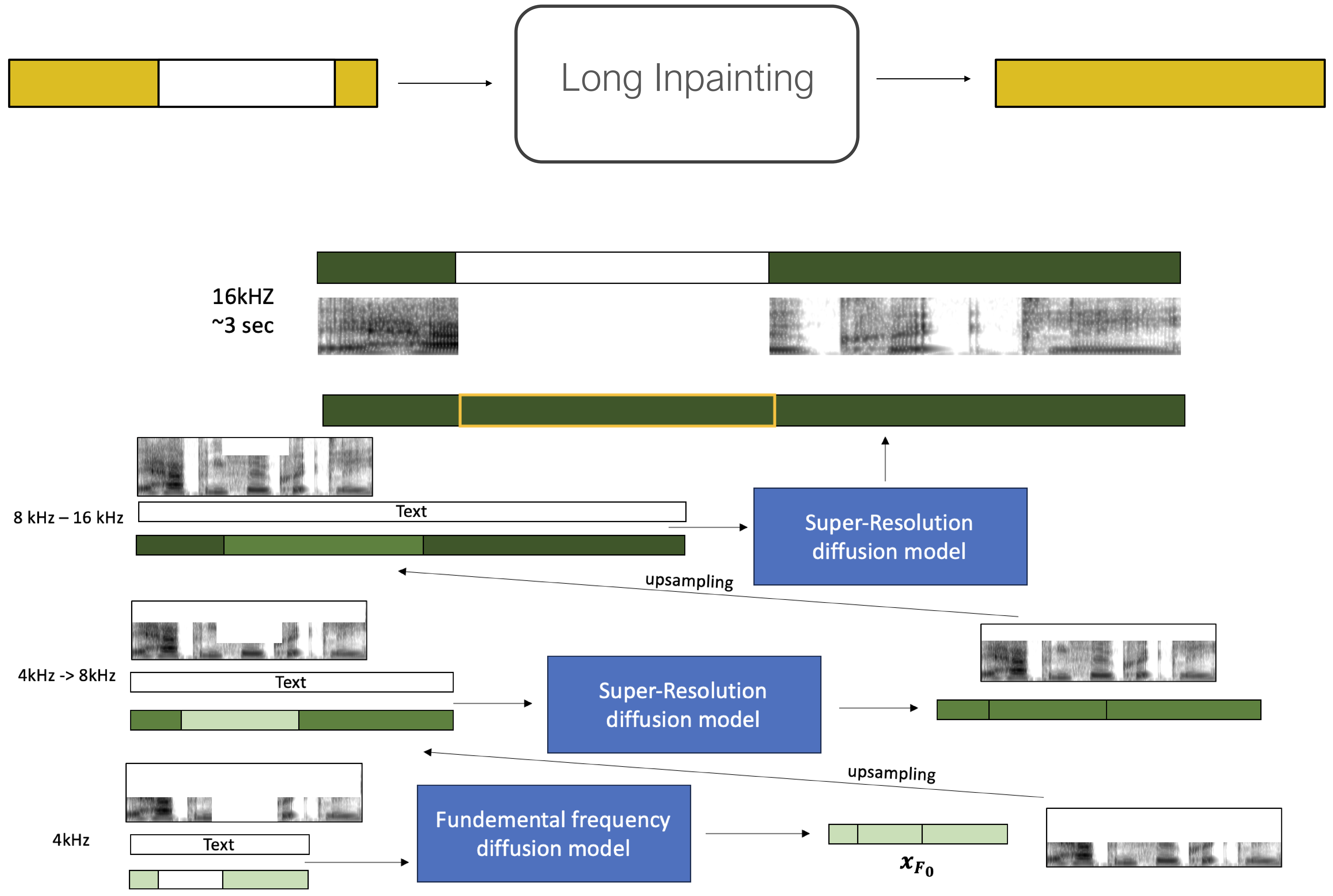
Synthesized audio



Spectral Analysis of Diffusion Models

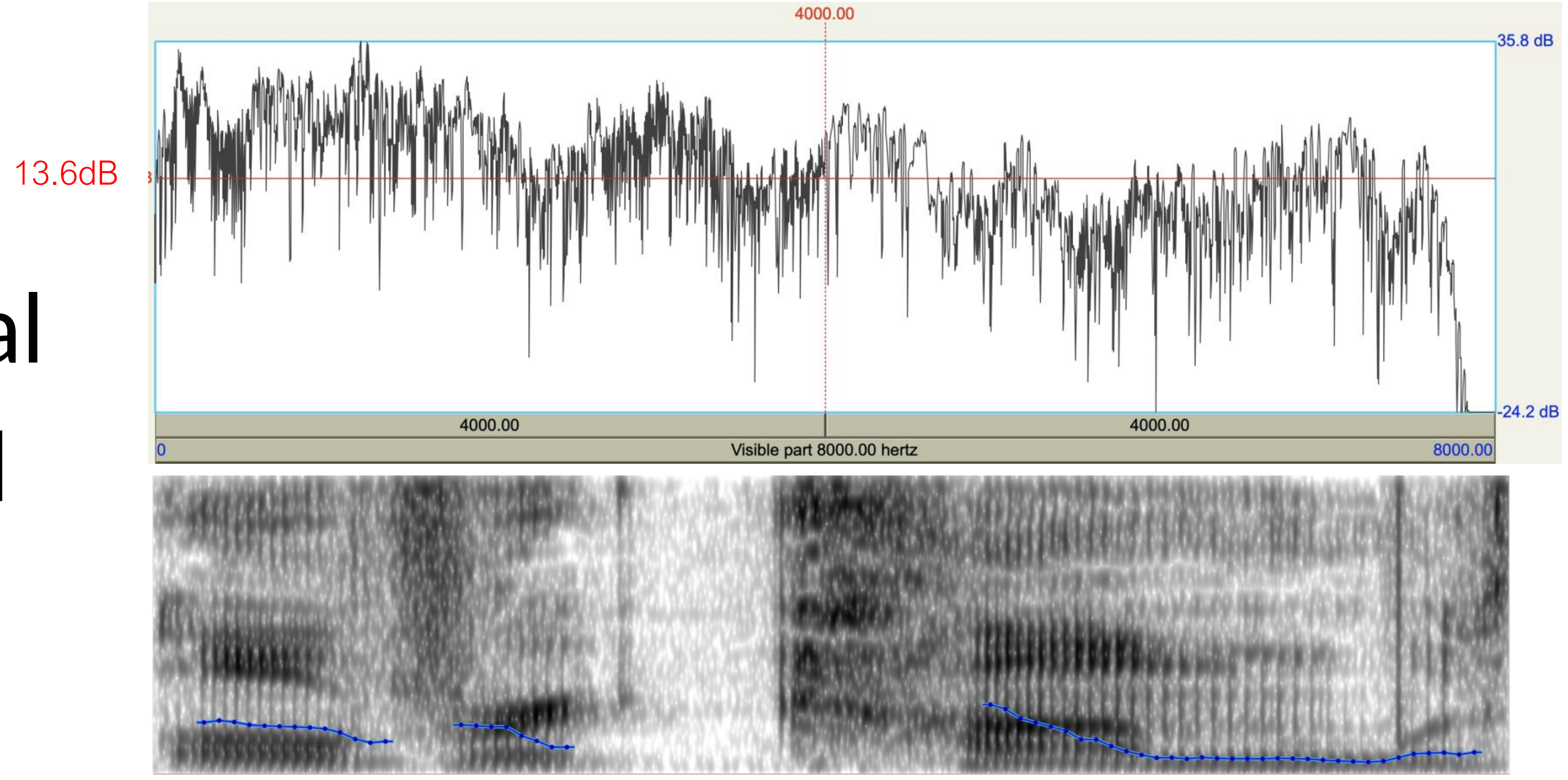
Benita, Elad, and Keshet, 2024

Next: Long Inpainting

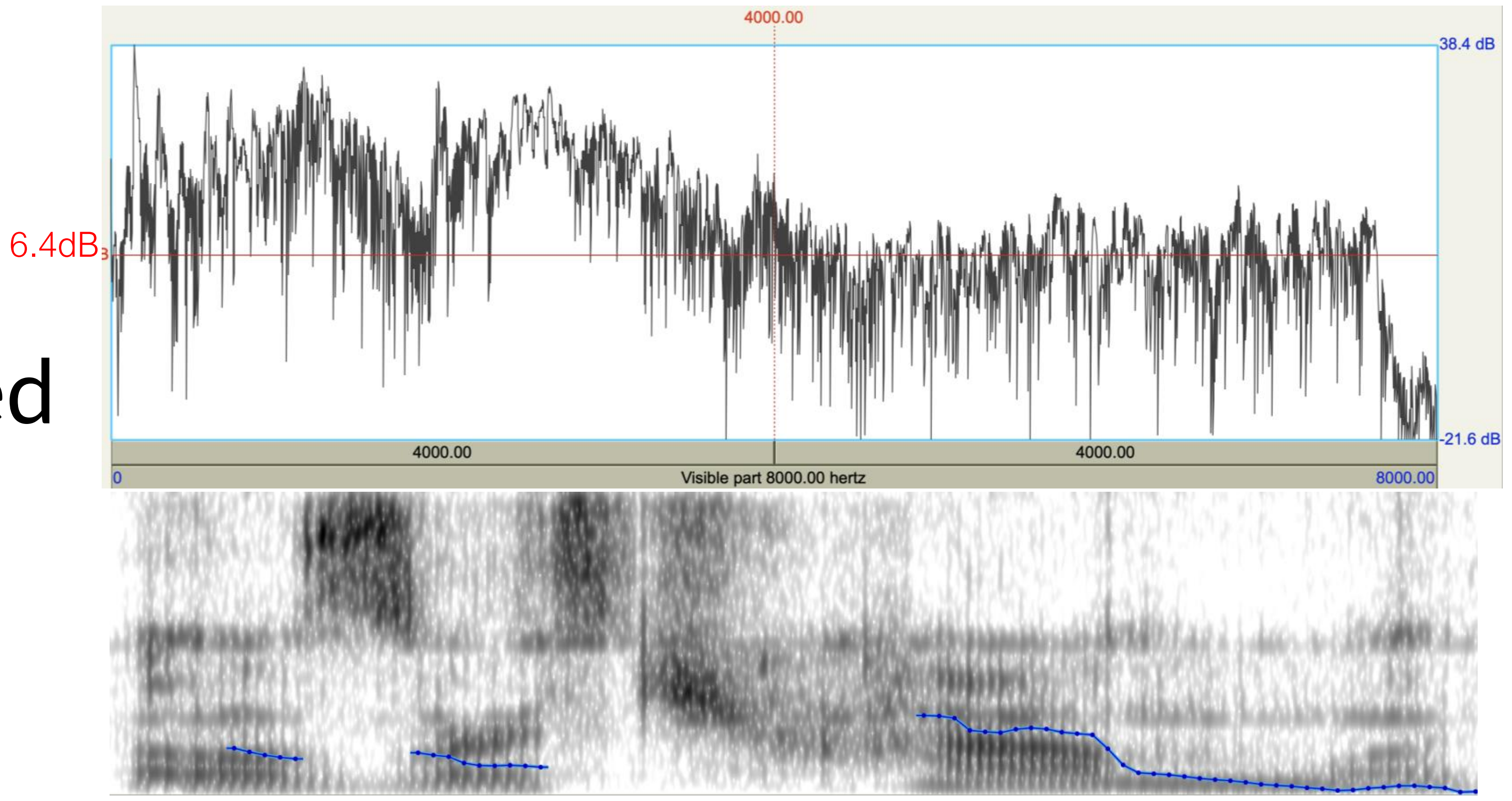


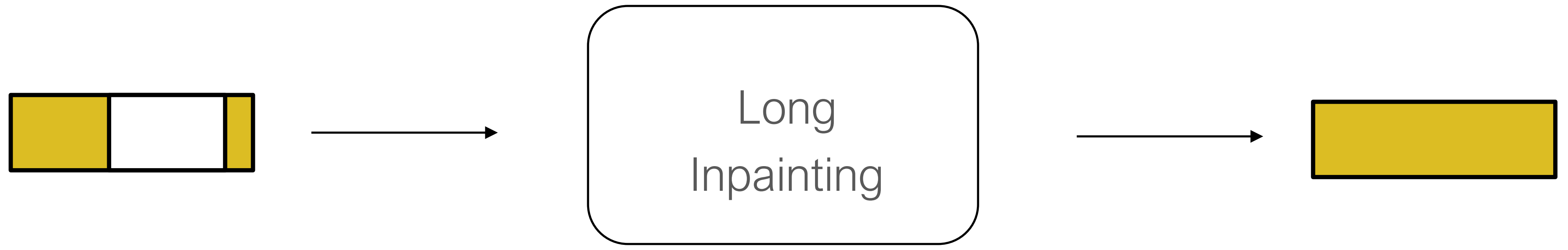
Next: Long Inpainting

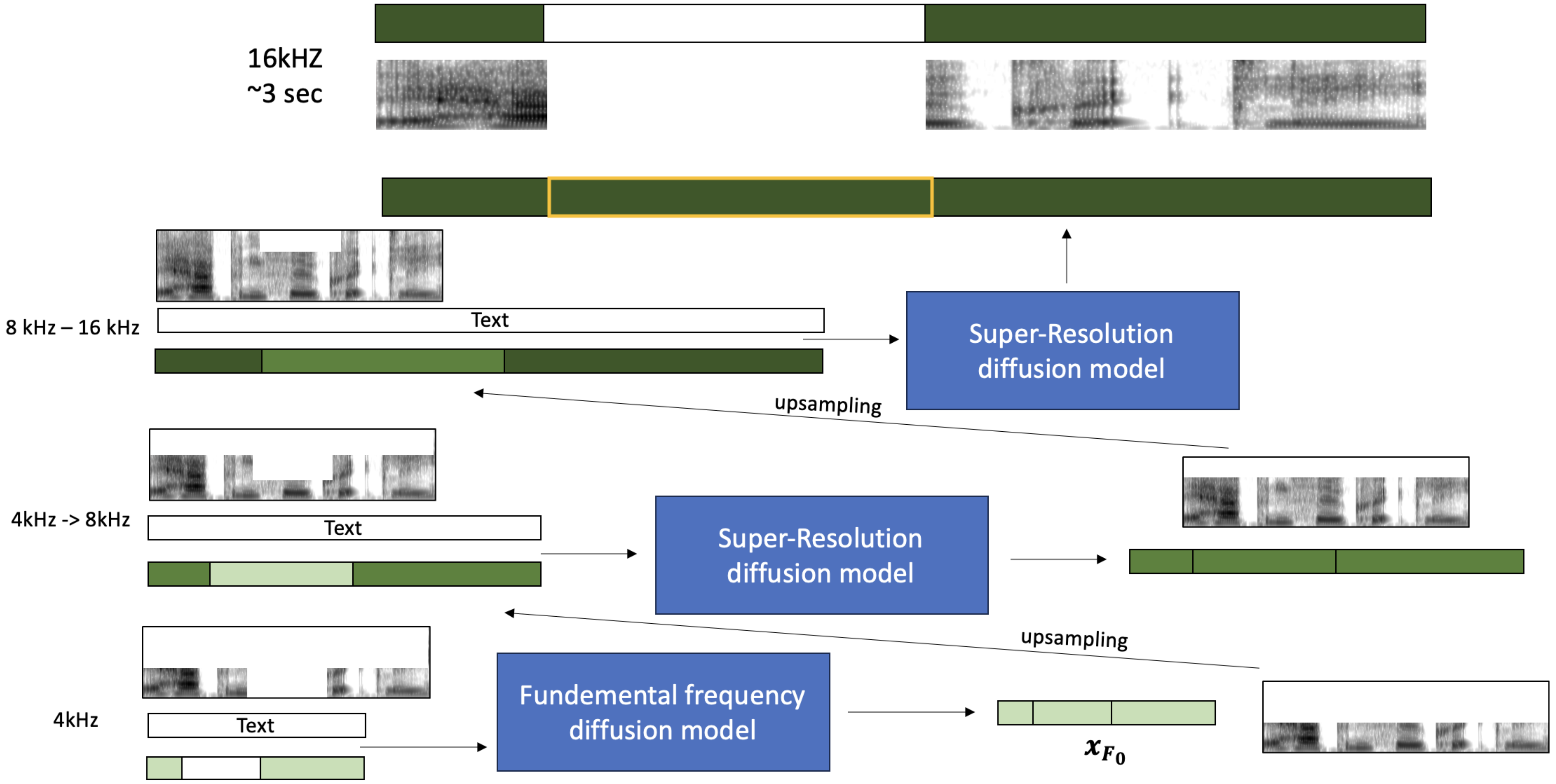
original signal



synthesized signal







Diffusion Process and Frequencies

Intriguing properties of synthetic images: from generative adversarial networks to diffusion models

Riccardo Corvi[1] Davide Cozzolino[1] Giovanni Poggi[1] Koki Nagano[2] Luisa Verdoliva[1]
[1]University Federico II of Naples [2] NVIDIA

Diffusion Probabilistic Model Made Slim

Xingyi Yang[1] Daquan Zhou[2] Jiashi Feng[2] Xinchao Wang[1]
National University of Singapore[1] ByteDance Inc.[2]

TIME SERIES DIFFUSION IN THE FREQUENCY DOMAIN

Jonathan Crabbé,* Nicolas Huynh,* Jan Stanczuk, Mihaela van der Schaar
DAMTP
University of Cambridge

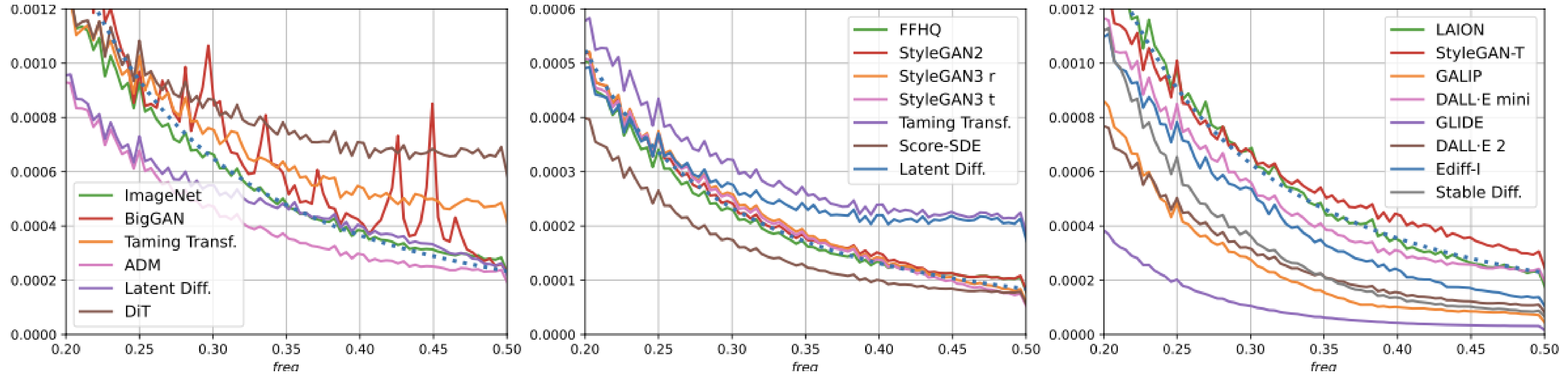
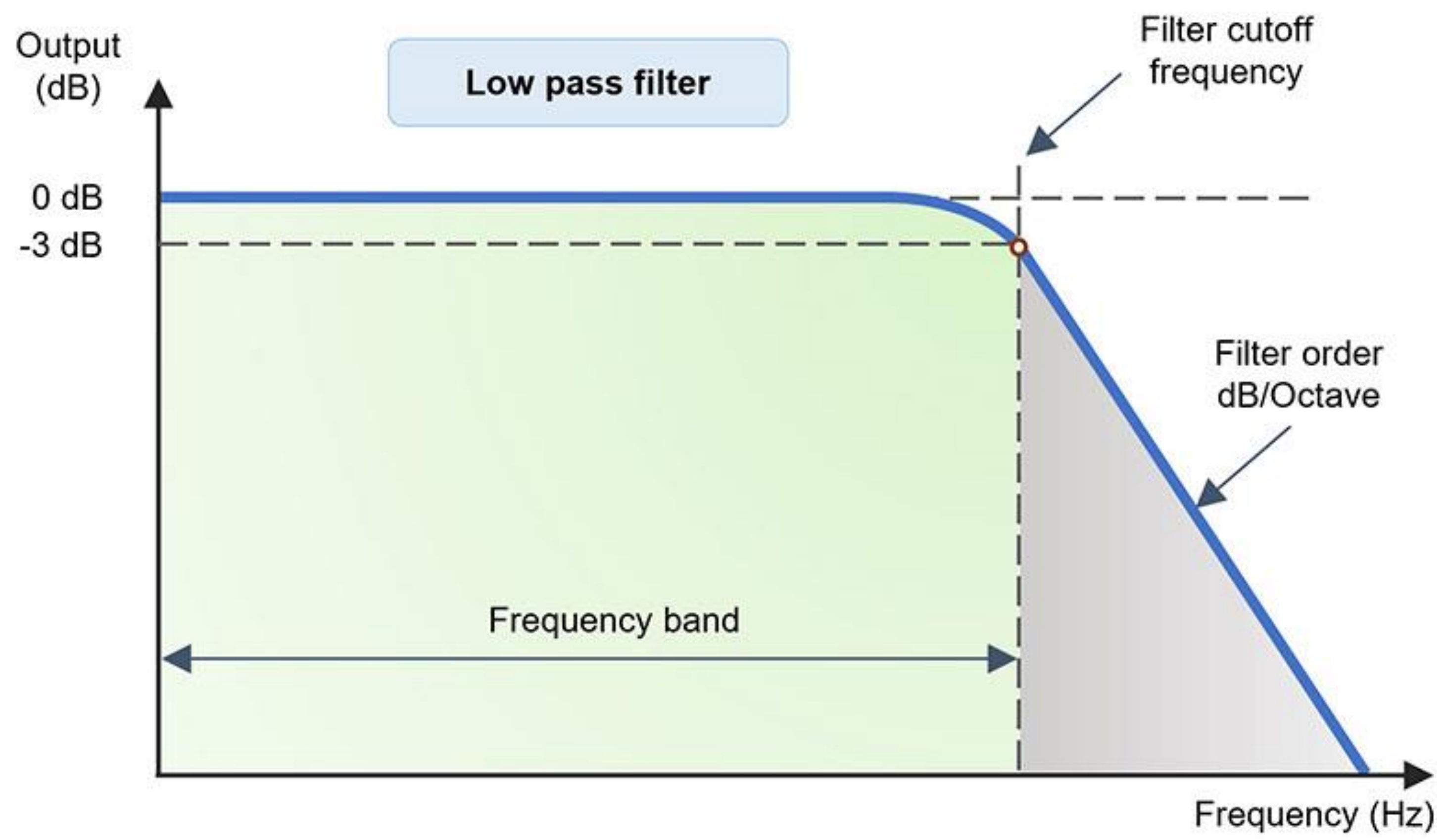


Figure 6. Radial spectrum power density. Synthetic images are compared with the real images used for training the correspondent model. Real images (green) fit very well the expected theoretical curve (dotted).

Spectral Analysis for Diffusion Models

Goal

- Analyze the diffusion model's inference process through a comprehensive frequency response perspective.
- Explore the possibility of **identifying a frequency-domain frequency response the diffusion process**.



The frequency response of a system is the quantitative measure of the magnitude and phase of the output as a function of input frequency.

The Challenge

How to separate the diffusion process from a specific denoiser

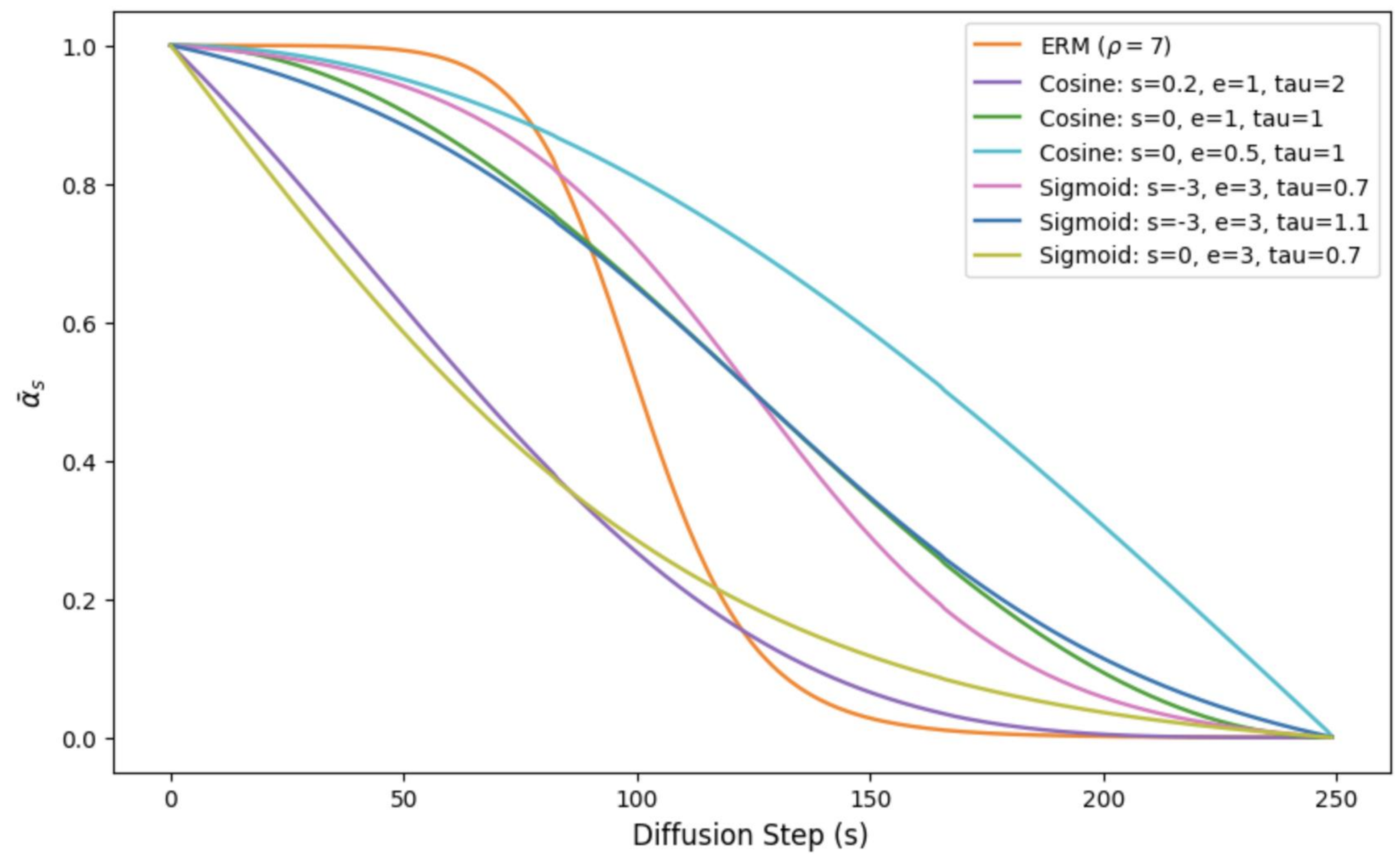
Method

- Assume **a multivariate Gaussian input** and derive **the optimal denoiser analytically**.
- Investigating various setups, including DDPM, DDIM, variance-preserving, variance-exploding, along with the selection of loss functions and additional features like expectancy drift.

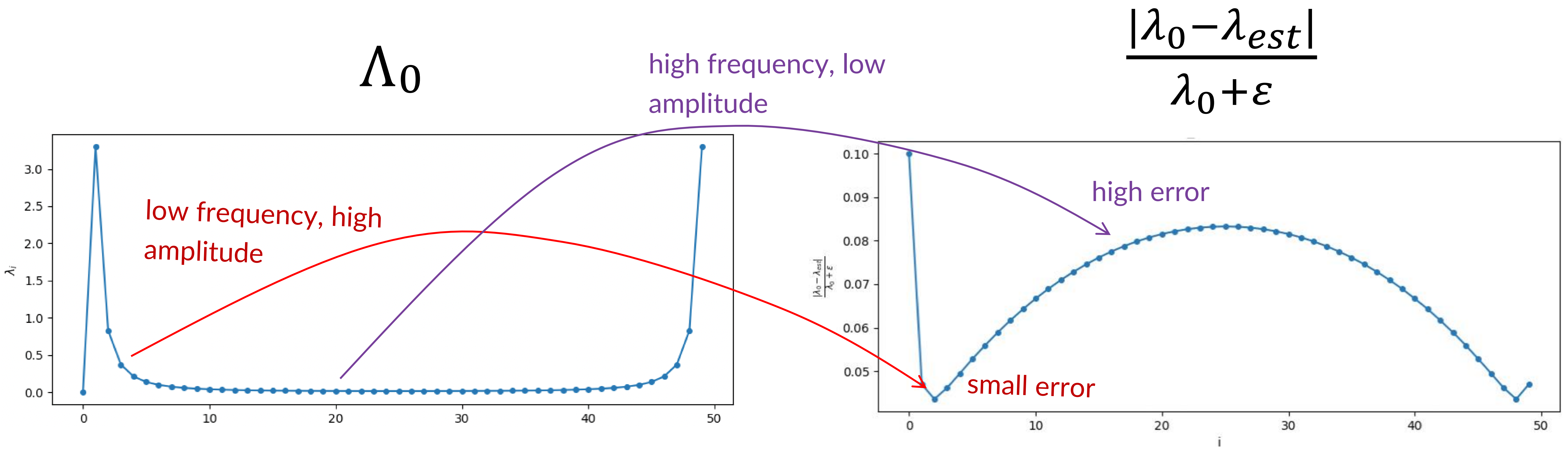
Spectral Analysis for Diffusion Models

Application: Designing Noise Scheduling

- Connection between the noise schedule and the frequency-domain phenomena
- Formulating an optimization problem to determine the optimal noise schedule that aligns with the dataset's characteristics and evaluating it against existing heuristics.



The Error of Each Frequency (Eigenvalue)



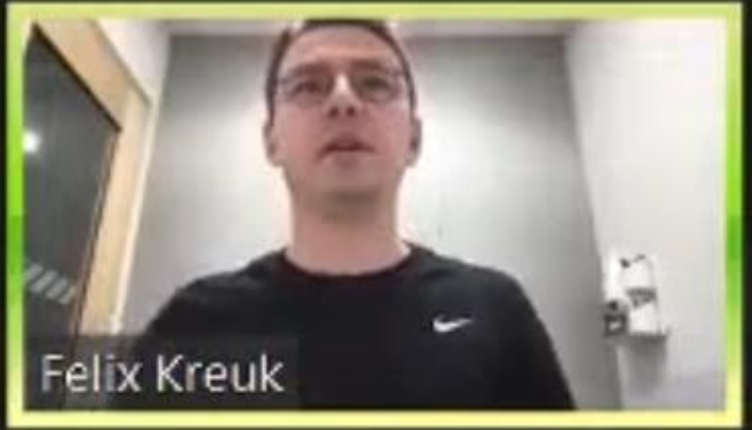
Time-scale modification of speech

Cohen, Kreuk, and Keshet, 2022

Analysis: more data

Table 2: *Analysis of model performance on the TIMIT and Buckeye test sets before and after augmenting them with examples from Librispeech.*

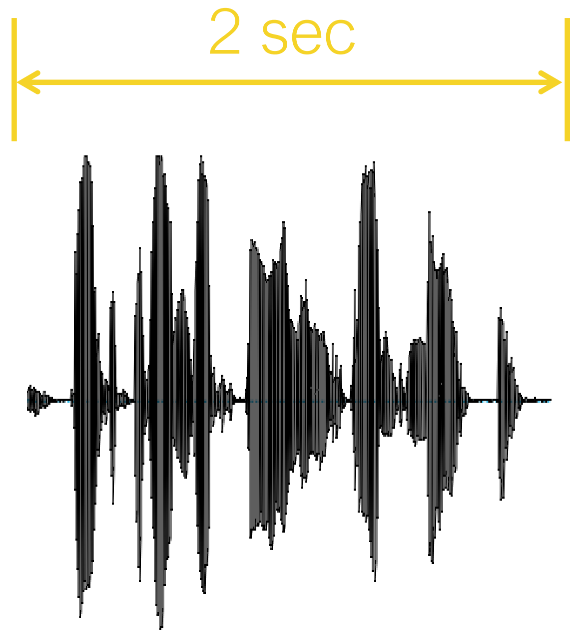
Training set	Test set	P	R	F1	R-val
TIMIT	TIMIT	83.89	83.55	83.71	86.02
TIMIT+	TIMIT	**84.11**	**84.17**	**84.13**	**86.40**
Buckeye	Buckeye	**75.78**	76.86	76.31	79.69
Buckeye+	Buckeye	74.92	**79.41**	**77.09**	**79.82**



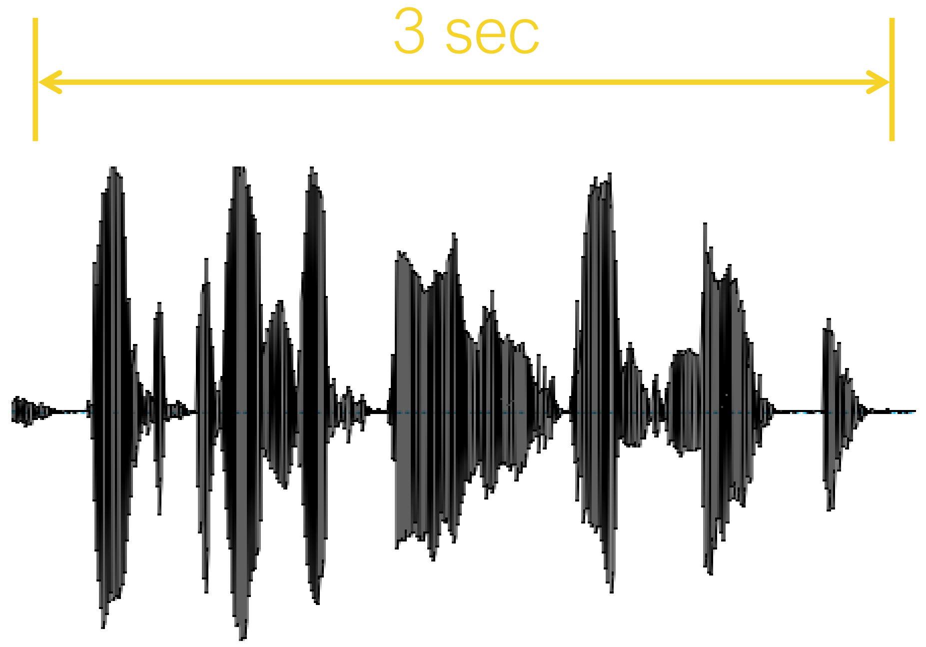
Gene-Ping Yang

Anton Ragni

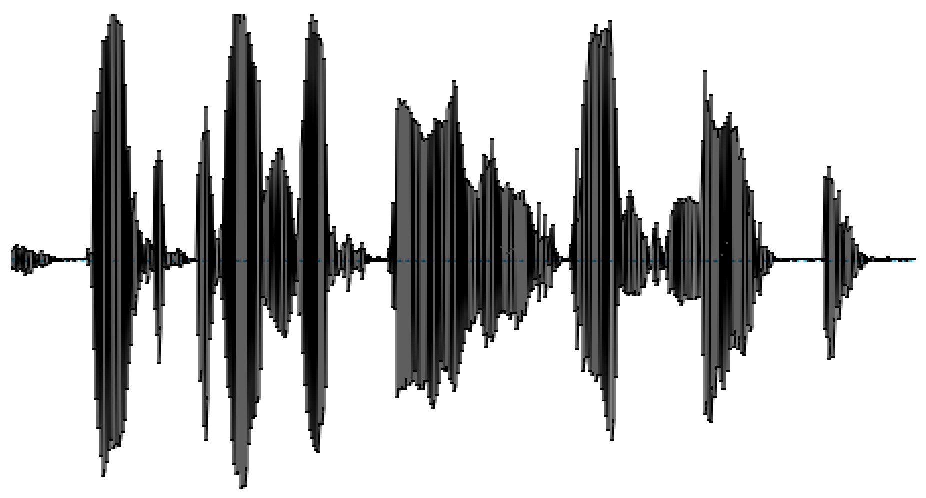
Herman Kamper



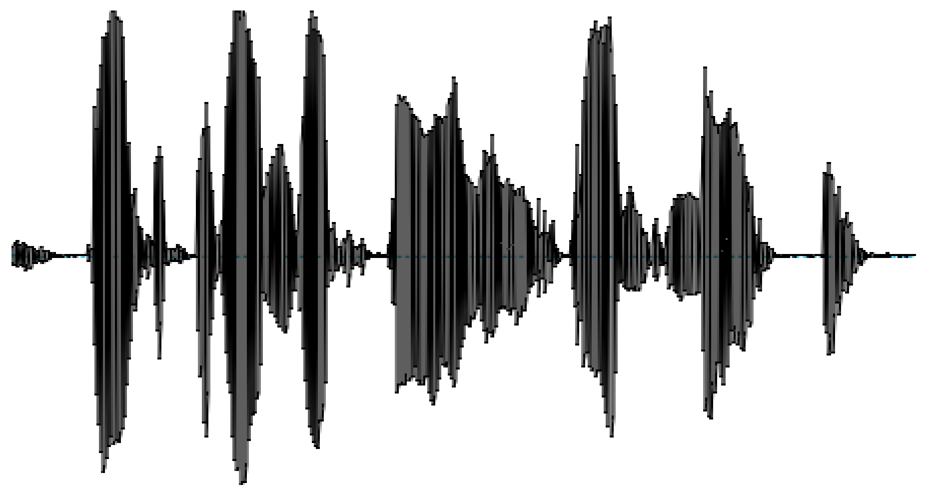
down-sampling by simple decimation *r=1.5*

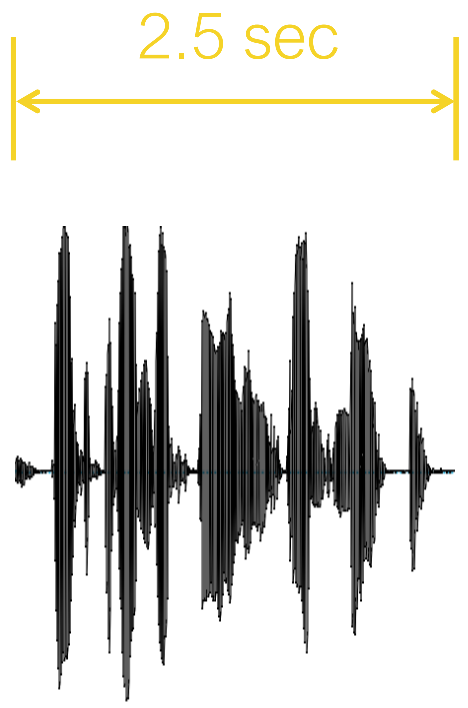


professional tool (Élastique)

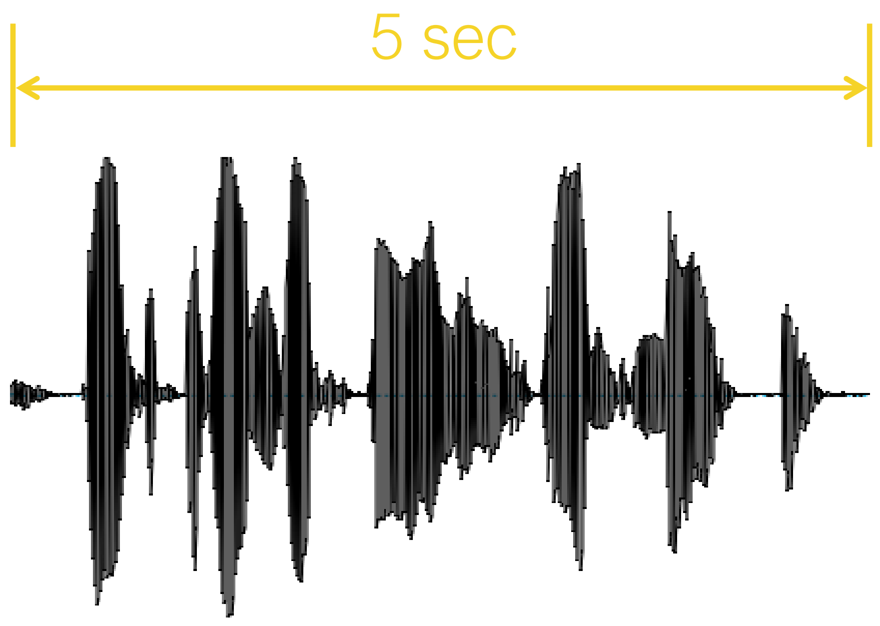


our deep learning model

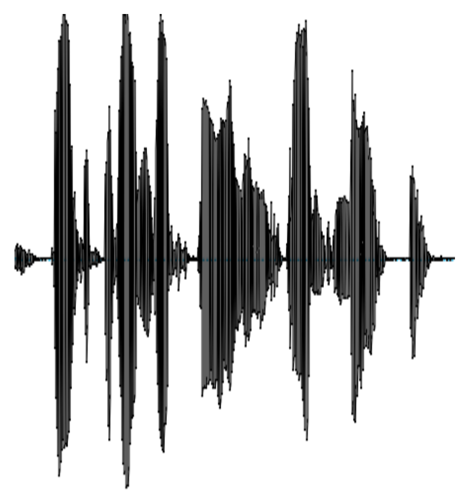




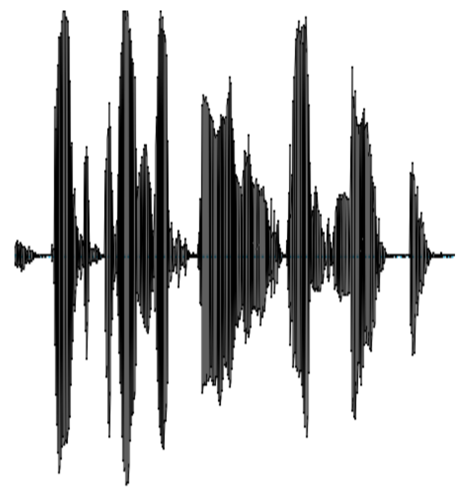
down-sampling by simple decimation *r=0.5*



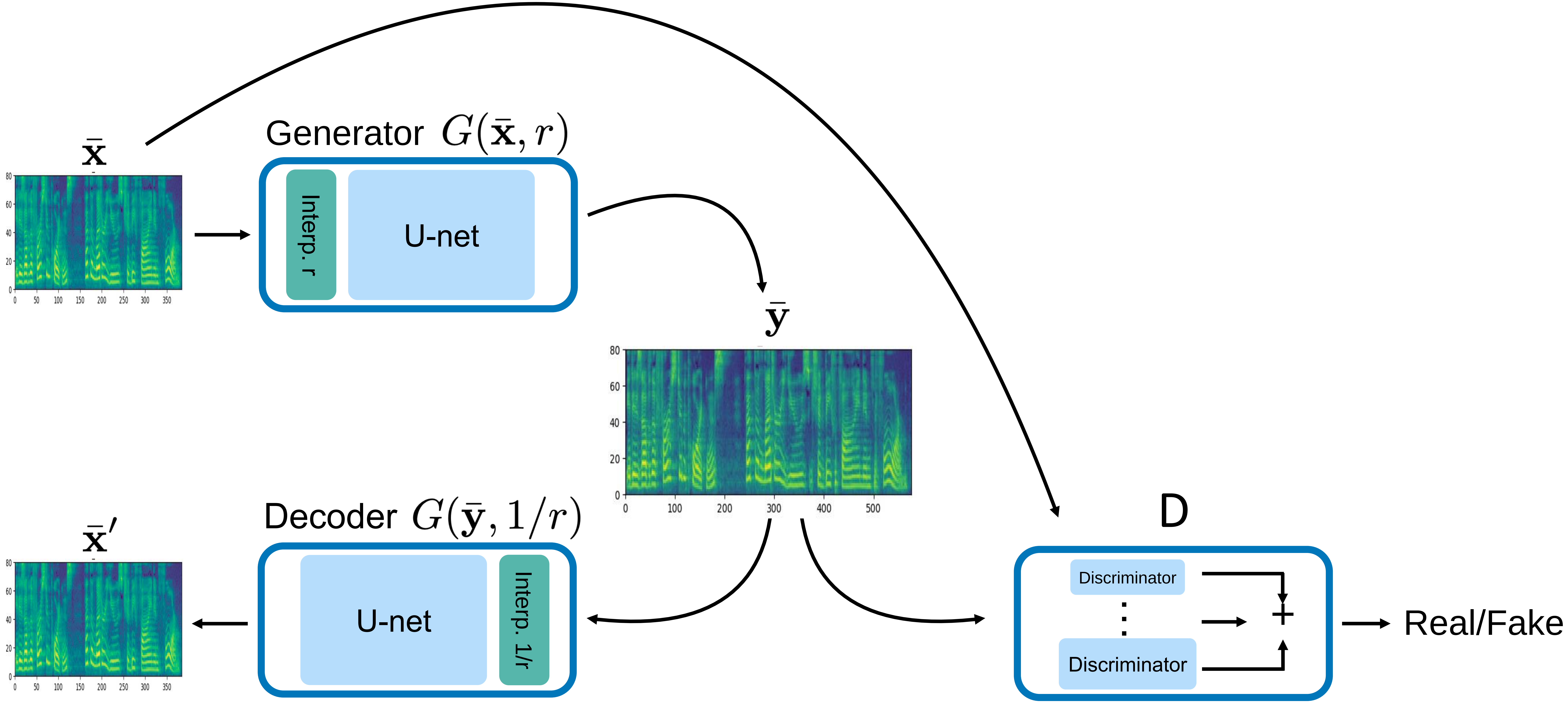
professional tool (Élastique)



our deep learning model

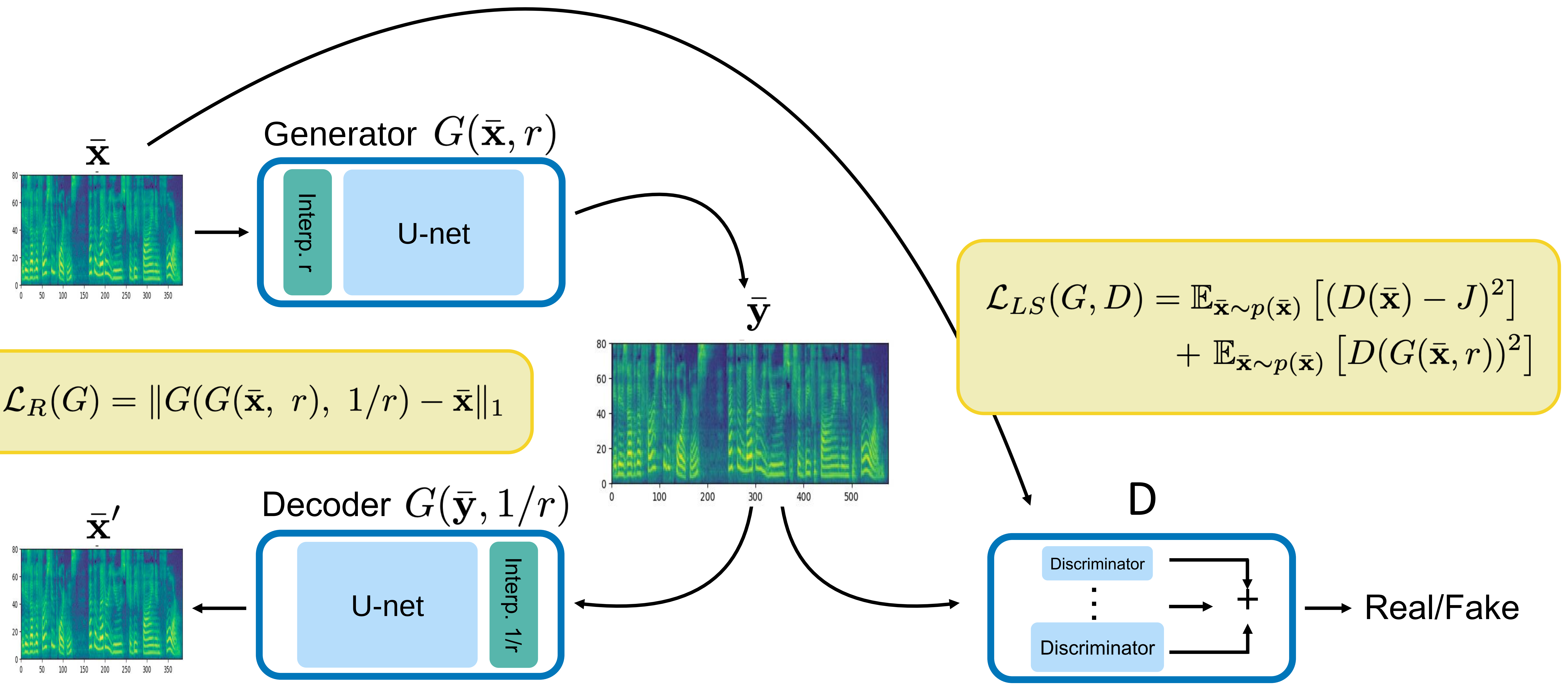


Training



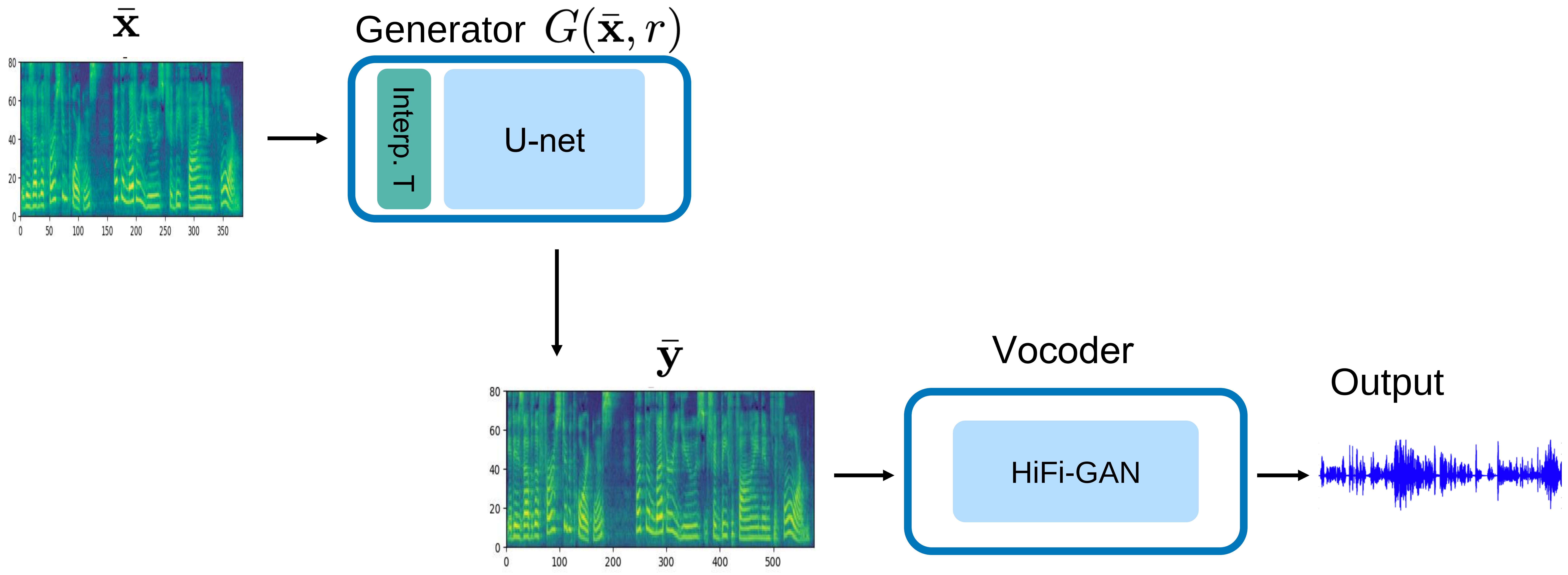
$\bar{\mathbf{x}}$
Generator $G(\bar{\mathbf{x}}, r)$
Interp. r
U-net
$\bar{\mathbf{y}}$
Decoder $G(\bar{\mathbf{y}}, 1/r)$
U-net
Interp. 1/r
$\bar{\mathbf{x}}'$
D
Discriminator
Discriminator
+
Real/Fake

Training



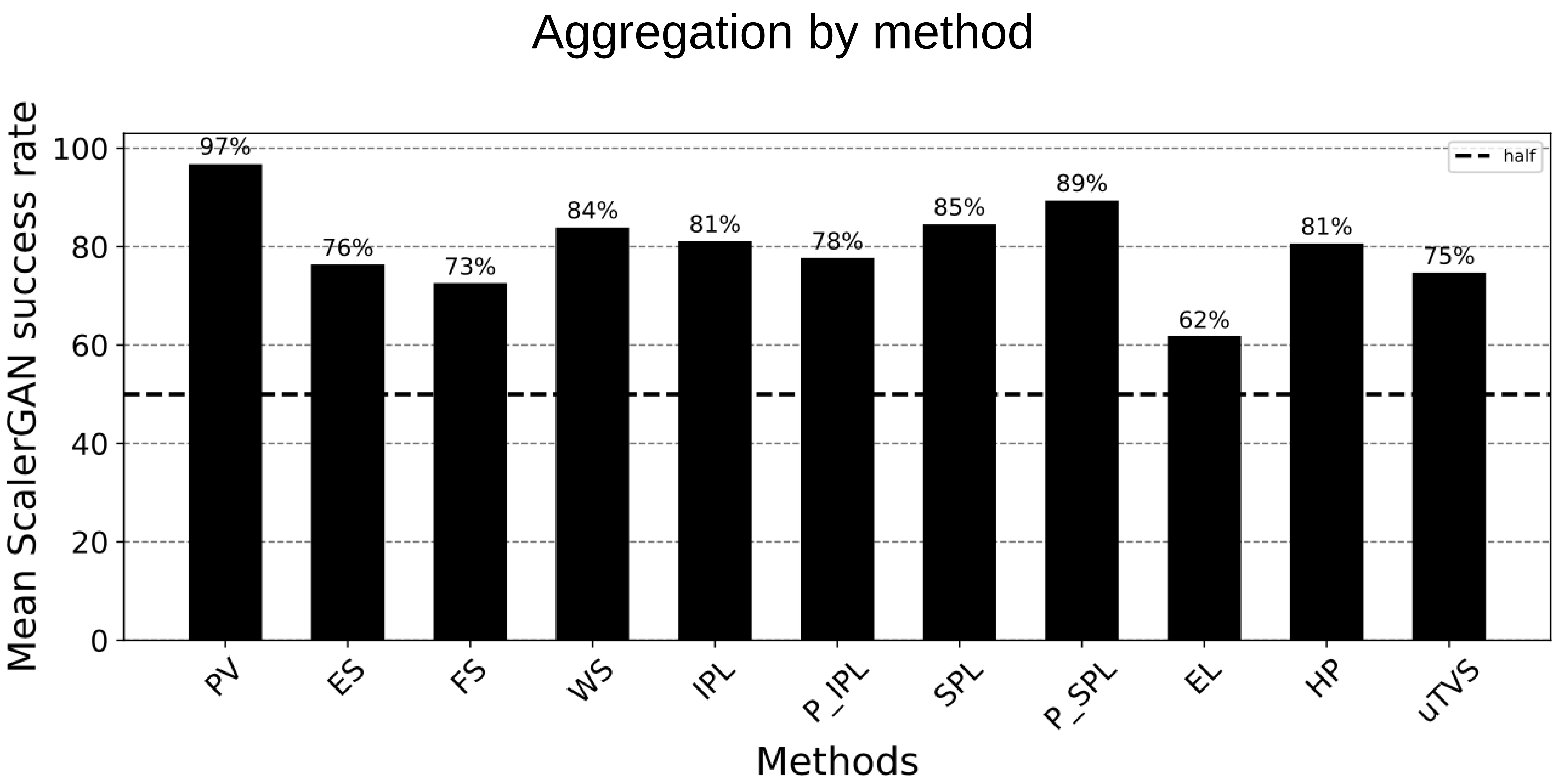
$\bar{\mathbf{x}}$
Generator $G(\bar{\mathbf{x}}, r)$
Interp. r
U-net
$\bar{\mathbf{y}}$
$\mathcal{L}_R(G) = \|G(G(\bar{\mathbf{x}},\ r),\ 1/r) - \bar{\mathbf{x}}\|_1$
$\mathcal{L}_{LS}(G, D) = \mathbb{E}_{\bar{\mathbf{x}} \sim p(\bar{\mathbf{x}})}\left[(D(\bar{\mathbf{x}}) - J)^2\right] + \mathbb{E}_{\bar{\mathbf{x}} \sim p(\bar{\mathbf{x}})}\left[D(G(\bar{\mathbf{x}}, r))^2\right]$
$\bar{\mathbf{x}}'$
Decoder $G(\bar{\mathbf{y}}, 1/r)$
U-net
Interp. 1/r
D
Discriminator
Discriminator
Real/Fake

Inference



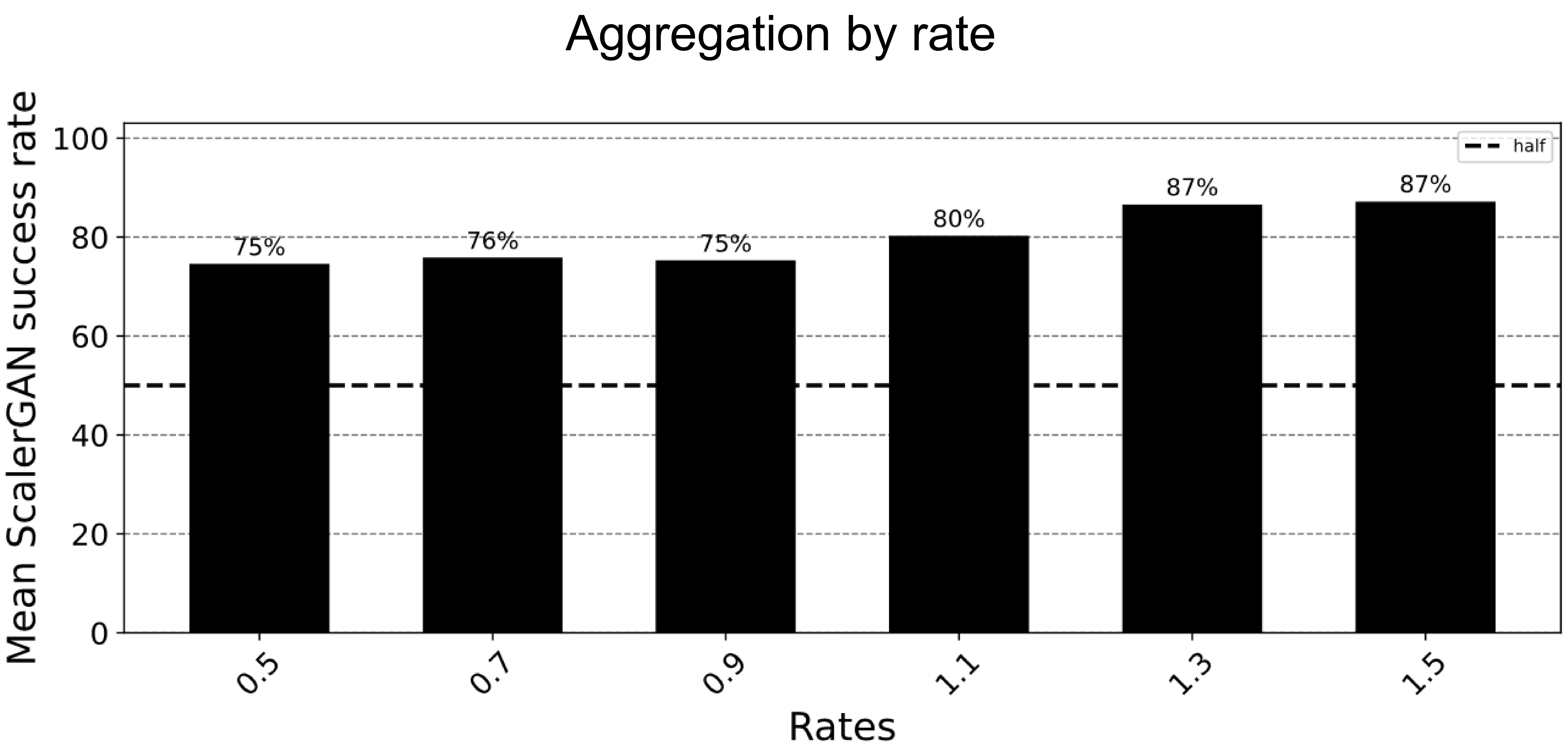
$\bar{\mathbf{x}}$
Generator $G(\bar{\mathbf{x}}, r)$
Interp. T
U-net
$\bar{\mathbf{y}}$
Vocoder
HiFi-GAN
Output

Empirical evaluation



PhaseVocoder (Laroche & Dolson, 1999), ESOLA (Rudresh et al. 2018), FESOLA (Roberts & Paliwal, 2019), WSOLA (Verhelst & Roelands,1993), IPL (Laroche & Dolson, 1999), Phavorit IPL (Karrer et al., 2006), SPL (Laroche & Dolson, 1999), Phavorit SPL (Karrer et al., 2006), Élastique, HPTSM (Driedger et al., 2013), and μTVS (Sharma et al., 2017).

Empirical evaluation

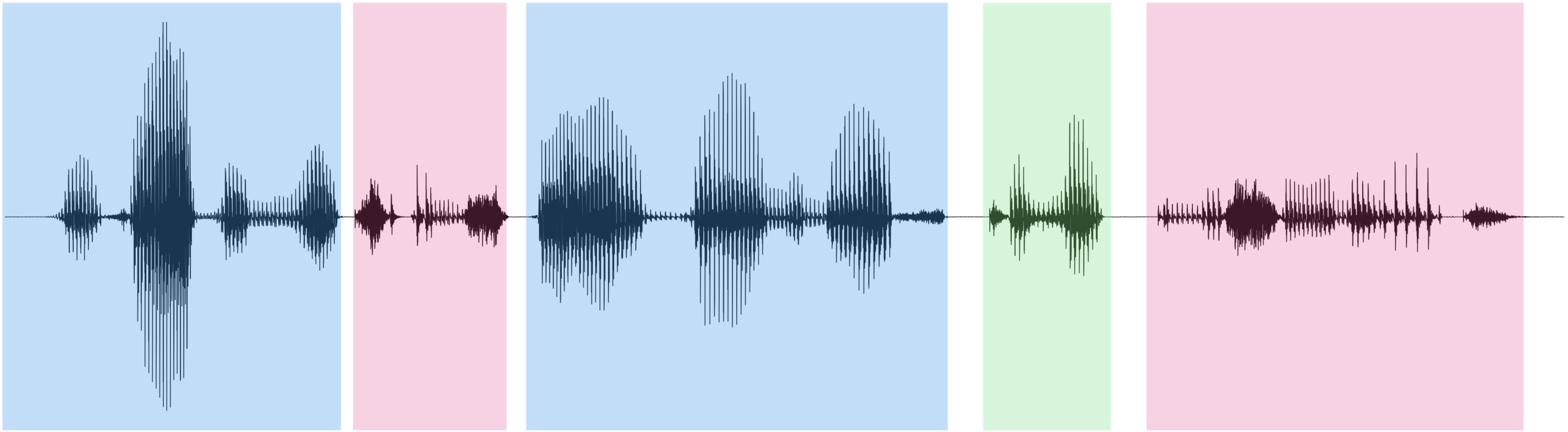


PhaseVocoder (Laroche & Dolson, 1999), ESOLA (Rudresh et al. 2018), FESOLA (Roberts & Paliwal, 2019), WSOLA (Verhelst & Roelands,1993), IPL (Laroche & Dolson, 1999), Phavorit IPL (Karrer et al., 2006), SPL (Laroche & Dolson, 1999), Phavorit SPL (Karrer et al., 2006), Élastique, HPTSM (Driedger et al., 2013), and μTVS (Sharma et al., 2017).

Self-Supervised Speaker Diarization

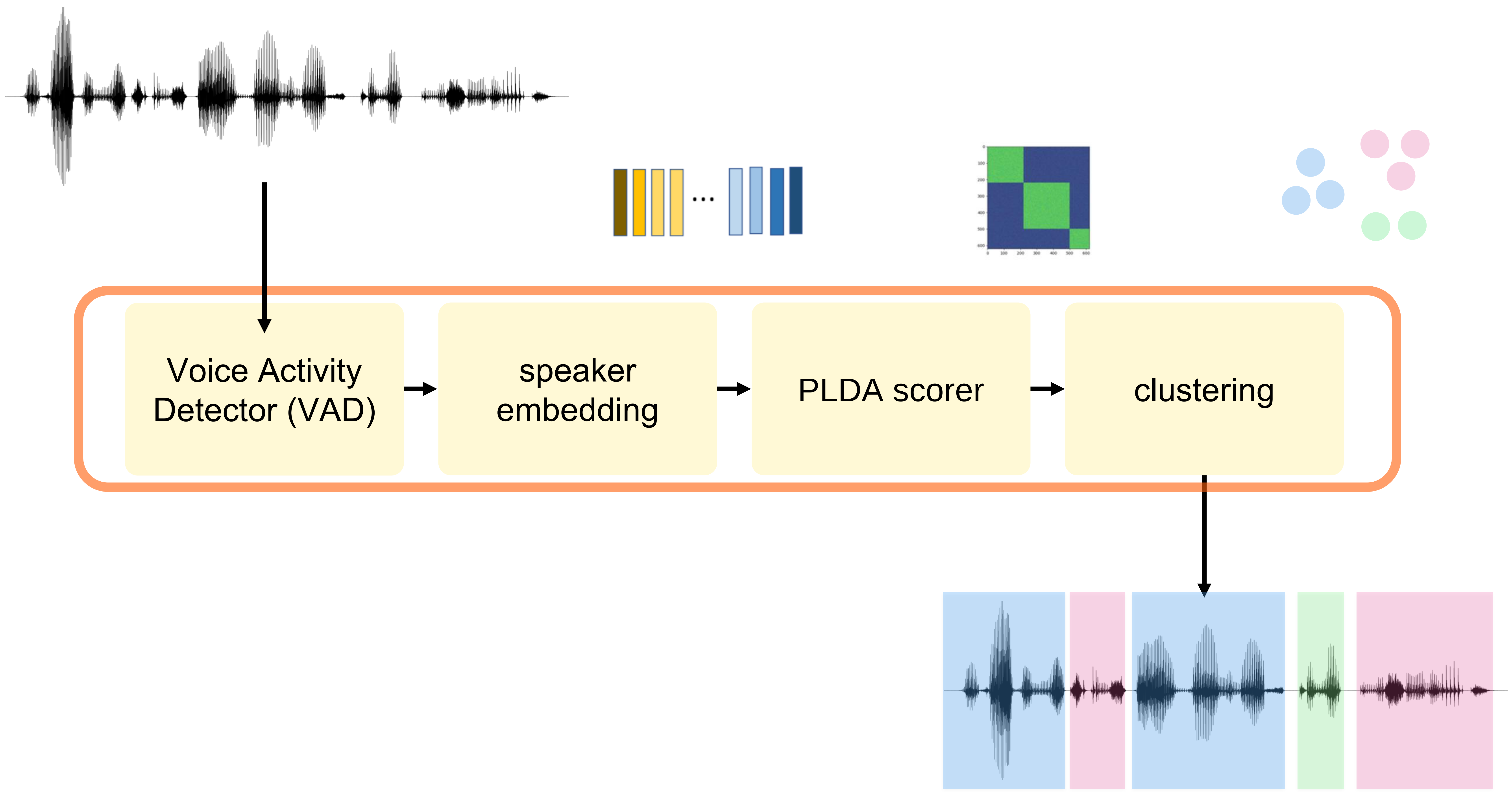
Dissen, Kreuk, and Keshet, 2022

Speaker diarization: Who spoke when?

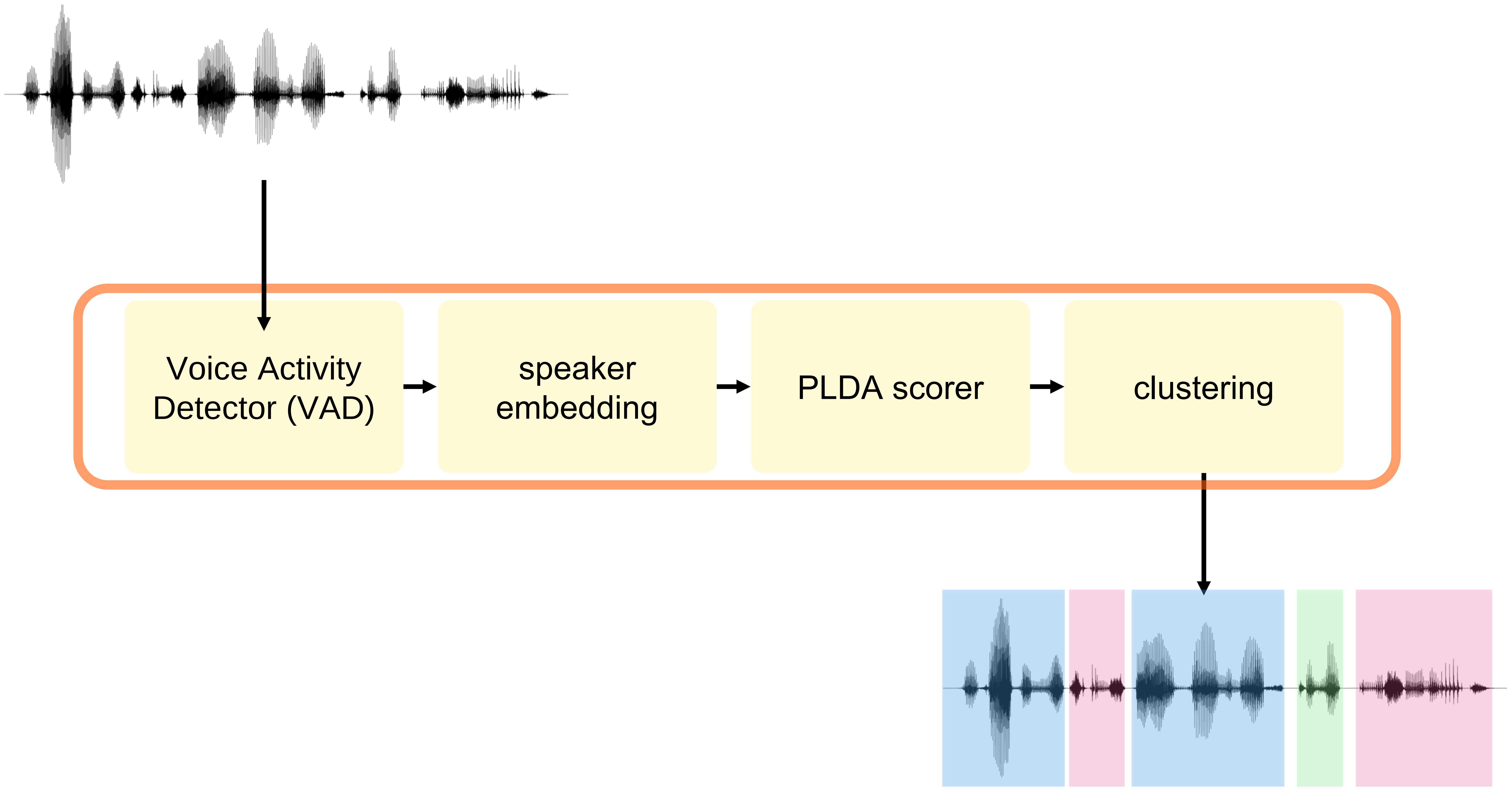


Speaker diarization: Who spoke when?

Traditional pipe:



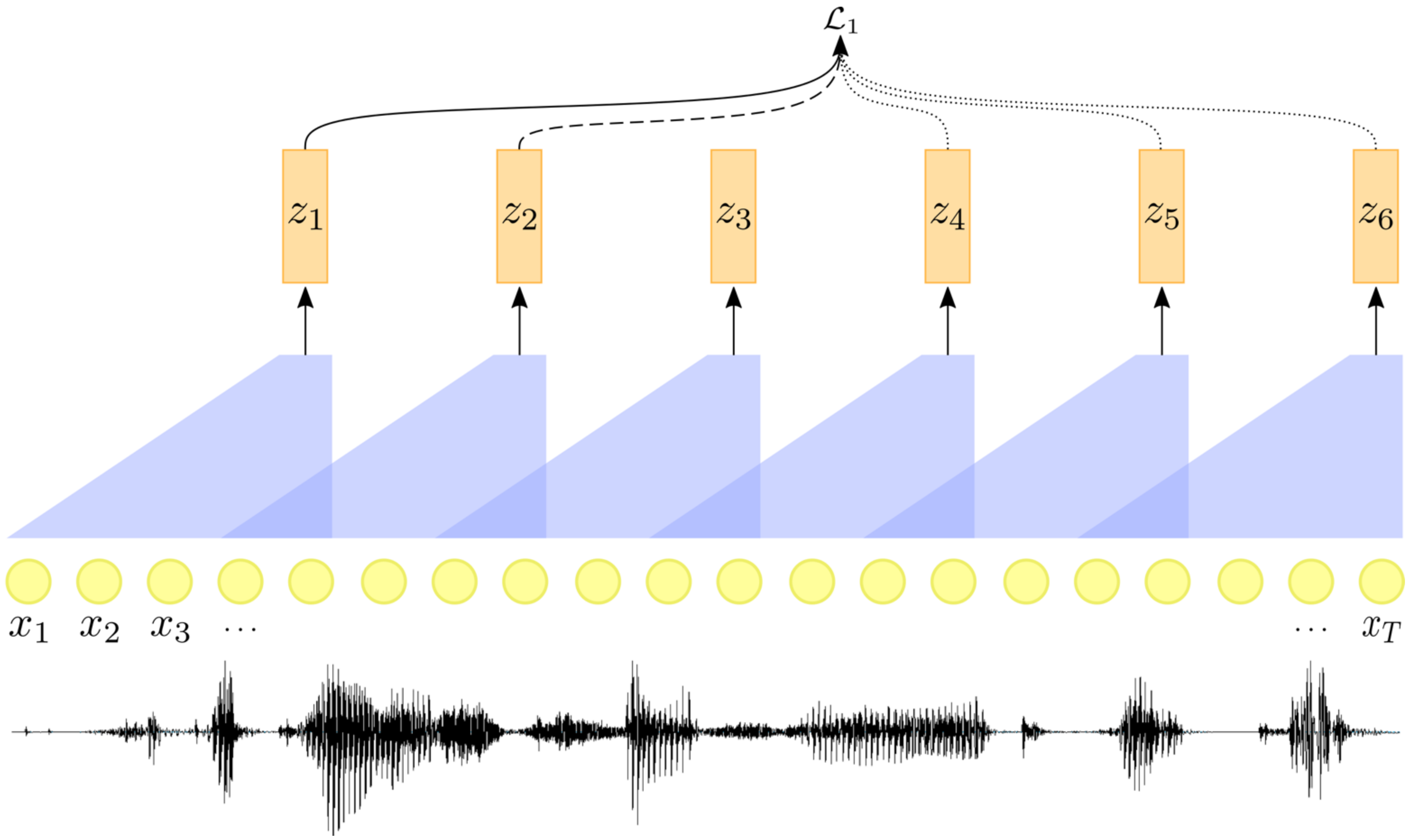
Goal



Propose a complete pipeline for speaker diarization training **with no annotated data**.

Speaker embedding

- Proposed: Contrastive learning.
  - Learn a metric by which positive pairs are similar and negative pairs are dissimilar. \pause
  - **Positive** pairs: Assume close frames are of the **same speaker**.
  - **Negative** pairs: Assume frames from different files are of **different speakers**.



- Problem: Using different files can introduce unwanted learned artifacts such as acoustic environment.
- Solution: **use only positive examples**.

Speaker embedding

Our self-supervised loss function is $L_{BT}(\mathbf{Z}_t, \mathbf{Z}_\tau) = \|\mathbf{R}_{\mathbf{Z}_t\mathbf{Z}_\tau} - \mathbf{I}\|_{\mathcal{F}}^2$

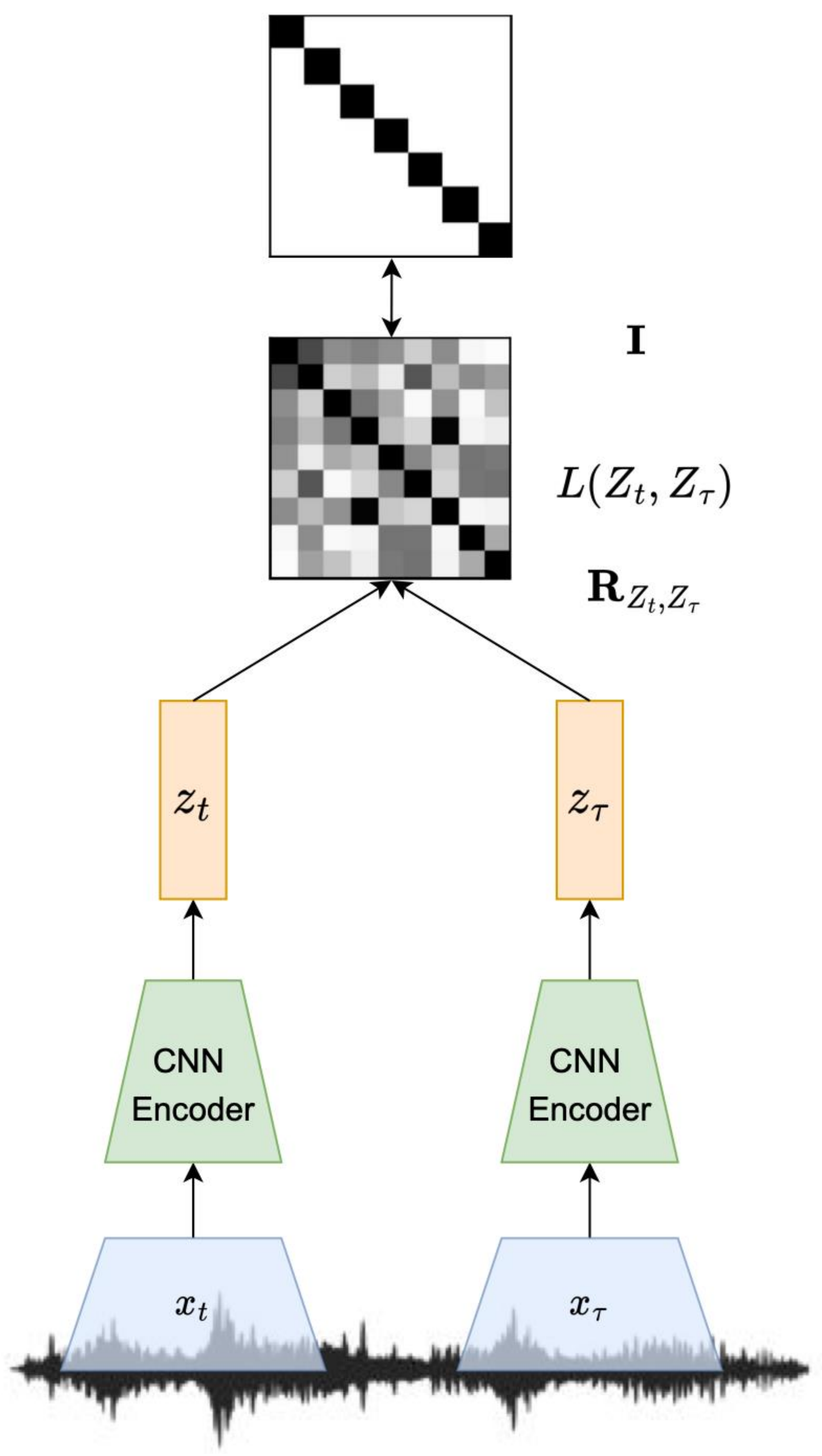
makes the cross-correlation matrix as close as possible to the identity matrix

Cross-correlation matrix between the two embeddings: $\mathbf{R}_{\mathbf{Z}_t\mathbf{Z}_\tau} = \mathbb{E}\left[\mathbf{z}_t^\top \mathbf{z}_\tau\right]$

CNN-based encoder: $\mathbf{z}_t = f_\theta(\mathbf{x}_t) \quad \mathbf{z}_\tau = f_\theta(\mathbf{x}_\tau)$

We work on two raw waveform segment of T samples:

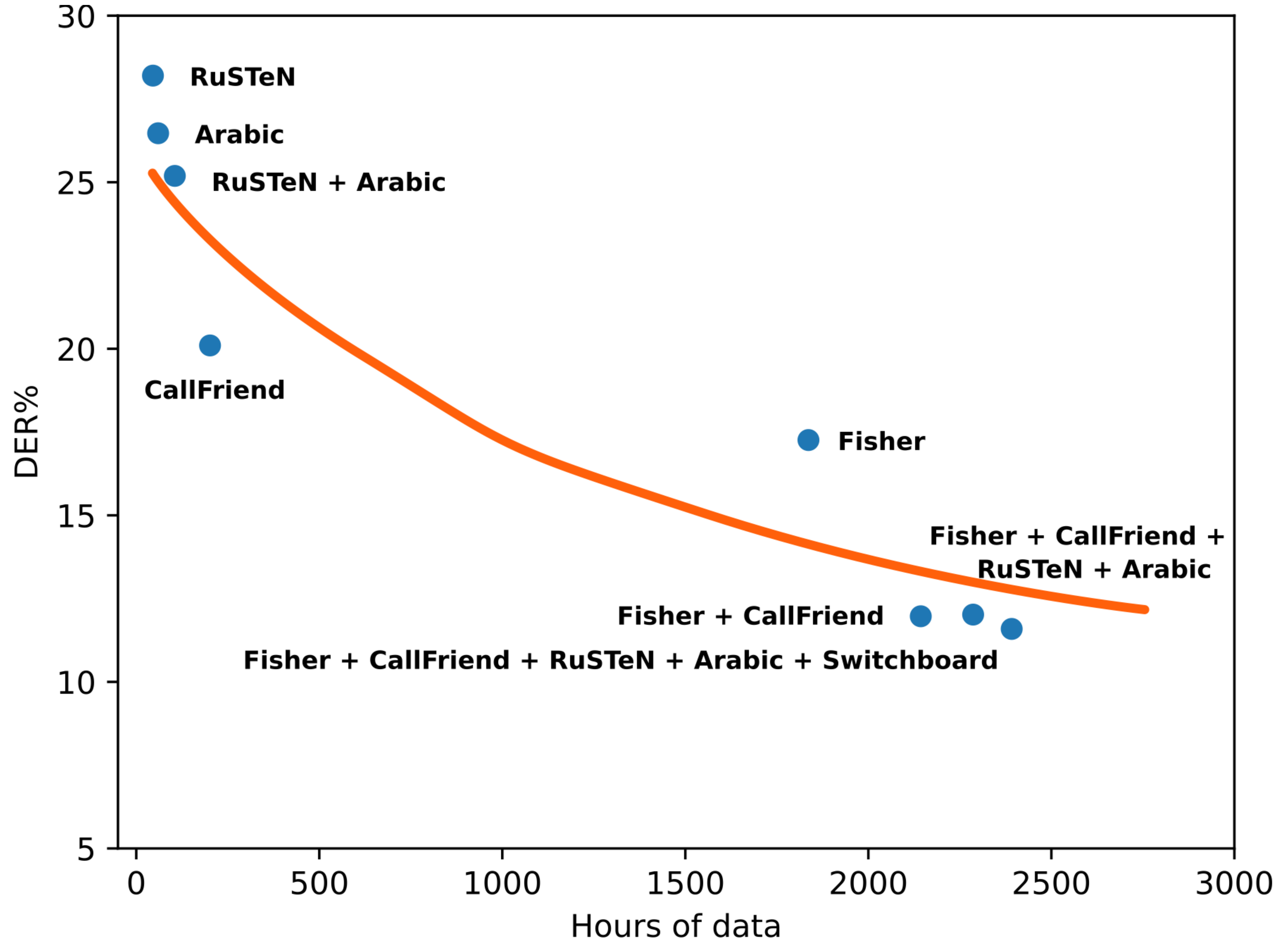
$\mathbf{x}_t = (x_{t-T}, x_{t-T+1}, \ldots, x_{t+T-1}) \qquad \mathbf{x}_\tau = (x_{\tau-T}, \ldots, x_{\tau+T-1})$



Inspired by Barlow Twins (Zbontar, Jing, Misra, LeCun, and Deny, 2021) and VICReg (Bardes, Ponce, and LeCun, 2021).

Speaker embedding

The more data the better the embeddings:



Empirical evaluation

Diarization error rate (DER) in % on the test set of CallHome compared with recent SOTA supervised works

Model	DER
UIS-RNN V1 [25]	10.6
UIS-RNN V2 [25]	9.6
UIS-RNN V3 [25]	7.6
x-vector + LSTM (oracle VAD) [5]	6.6
DIVE [17]	5.9
Ours (unsupervised, oracle VAD)	6.6
Ours (unsupervised)	9.1

Keyword Spotting and Automatic Speech Recognition

aiOla KWS Demo

English

aiOla & OpenAI Whisper

Word Error Rate:

Keyword List

OpenAI Whisper

Word Error Rate:

Original Text

The patient is taking the following medications: Levetiracetam, SGLT2 Inhibitors, Isavuconazonium Sulfate and Artemether. He is also having trouble with his blood pressure and will need to be prescribed

Compare

aiOla’s Jargonic V2 Demo

Japanese

aiOla

Execution Time:

Word Error Rate:

aiOla Keyword Spotting:

Cloud ASR

Execution Time:

Word Error Rate:

Original Text

モーターに問題があります。以下は圧力と温度の値で 期待される範囲から外れています
電圧・電流・周波数を計測しました。

Compare

Thanks!