# Controllable and Generalizable Speech Generation

## via Explicitly and Implicitly Disentangled Speech Representations

**Wei-Ning Hsu <wnhsu@meta.com>**
Meta FAIR / Research Scientist
2023/08/20 @ The VoxSRC Workshop 2023

Meta AI

# About myself

- Research scientist @ Meta FAIR (2020 - Now). Lead of the audio generation team
  - Research intern @ FAIR (2019), Google Brain (2018), MERL (2016)
  - PhD/SM @ MIT (2015-2020), BS @ National Taiwan University (2010-2014)

# About myself

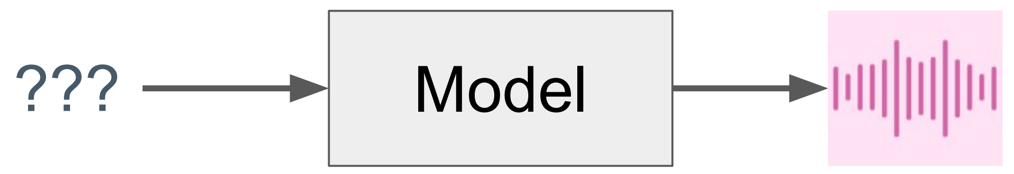
- Research scientist @ Meta FAIR (2020 - Now). Lead of the audio generation team
  - Research intern @ FAIR (2019), Google Brain (2018), MERL (2016)
  - PhD/SM @ MIT (2015-2020), BS @ National Taiwan University (2010-2014)

- Research focus: speech processing & machine learning
  - Unimodal/multimodal speech SSL: HuBERT, data2vec 1 & 2, AV-HuBERT, ResDAVENet, FHVAE
  - SSL-based applications: TextlessNLP, S2ST for the unwritten, unsupervised ASR
  - Speech generation: Voicebox, ReVISE, Unit-HiFiGAN, GMVAE-Tacotron

# Introduction

# What does an ideal speech generation model look like?

include not just TTS, but any model that outputs speech

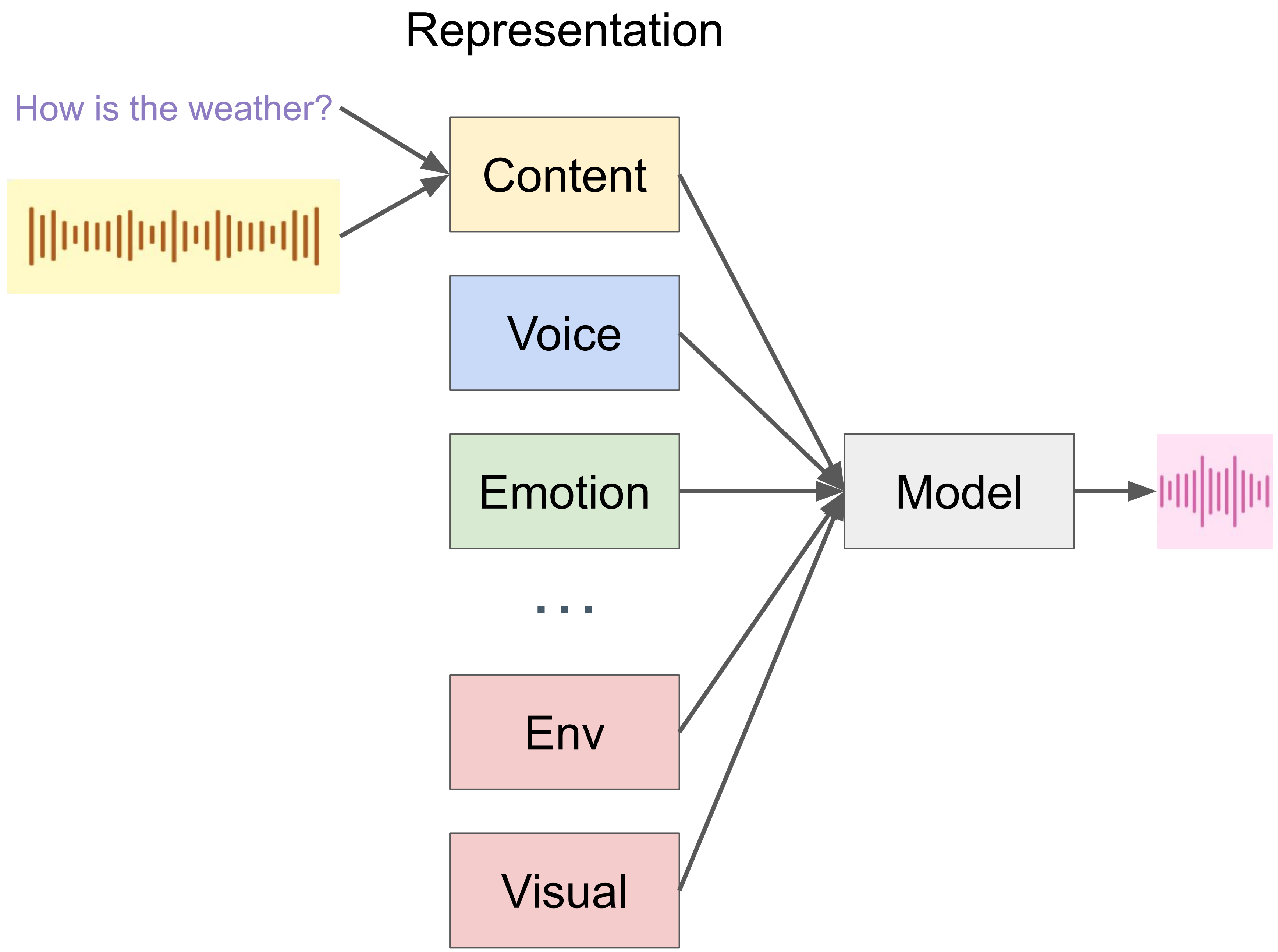


**My personal opinion**

# Controllable, generalizable, and efficient

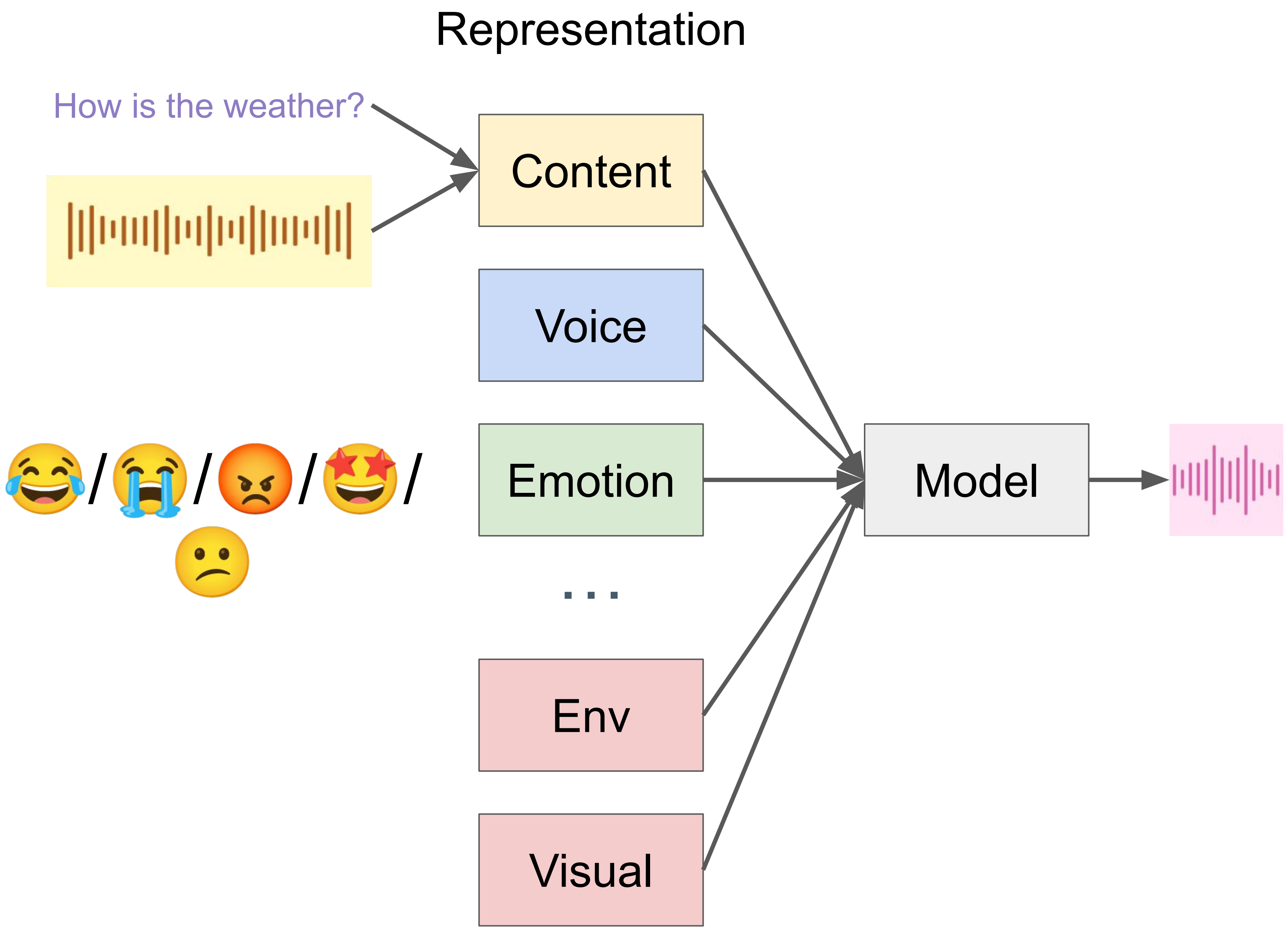
# Controllable

1. How many attributes can we control?
2. What modality can we use to specify each attribute?



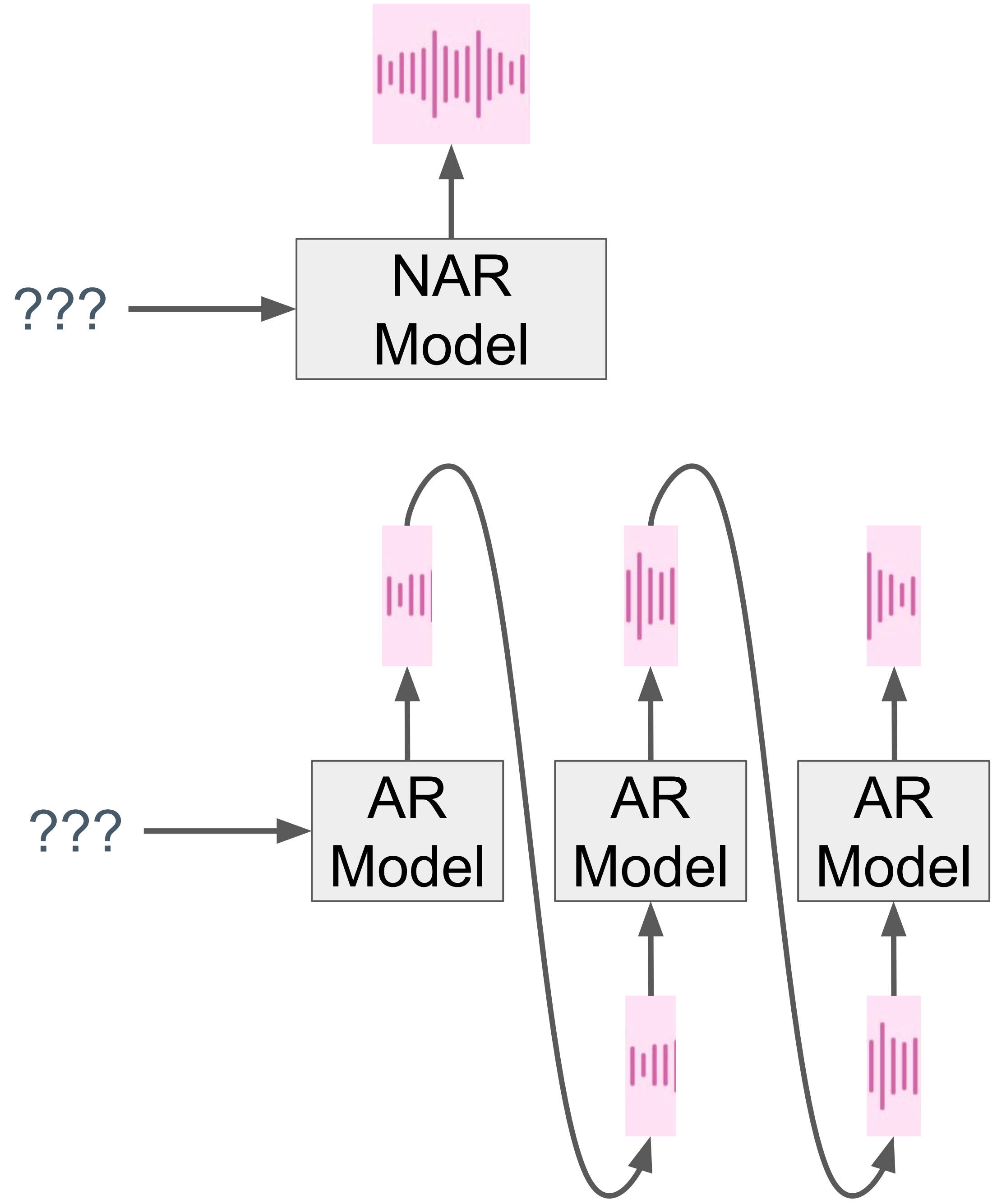
# Generalizable

1. Domain, for example,
   a. How many emotions does it cover if fixed?
   b. Can it generalize to unseen emotion?
2. Task
   a. How many task does it cover?
   b. Can it perform tasks not explicitly trained for?



# Efficient

1. Training efficiency
   a. How much samples do we need?
   b. How fast does the model converge?
2. Inference efficiency
   a. How much time does it take to generate 1 sec?
   b. How much memory does it take?



# Why Speech Generation @ VoxSRC?

- Because representation learning is core to all three criteria
  - Controllability: good representation enables better independent control
  - Generalization: modality agnostic representation enables task generalization
  - Efficiency: model trains faster and requires fewer samples  with pre-trained embedder
- Speaker/voice/accent variations are one of the most important variation to control
  - A focus of this workshop

This talk is about how to use disentangled representation to build the ideal speech generation model

# Are Good Representations All We Need?

NO. We still need the right model and large scale data for generalization

**Research opportunities** : most existing speech generation models are still trained on toy datasets (by today's standard)

Why? Because the right model was not used until very recently

# My Rough Classification on (Model, Data)

| **Model** | Regression | Regression w/<br>low-dimension latent | Generative |
|---|---|---|---|
| **Example** | Fastspeech, Tacotron, HiFi-GAN | GST,GMVAE-Tacotron | VALL-E, NaturalSpeech2, Voicebox |
| **Capability** | Assume deterministic/unimodal mapping between input/output. Low ability to model variation. | Assume unseen variation lies in a low-d manifold. Cannot model high dimensional variation like noise | More powerful generative model that does not have limiting assumptions |
| **Dataset** | LJSpeech, VCTK, Expresso | Blizzard, LibriTTS | Librivox, GigaSpeech |

# Today's Talk

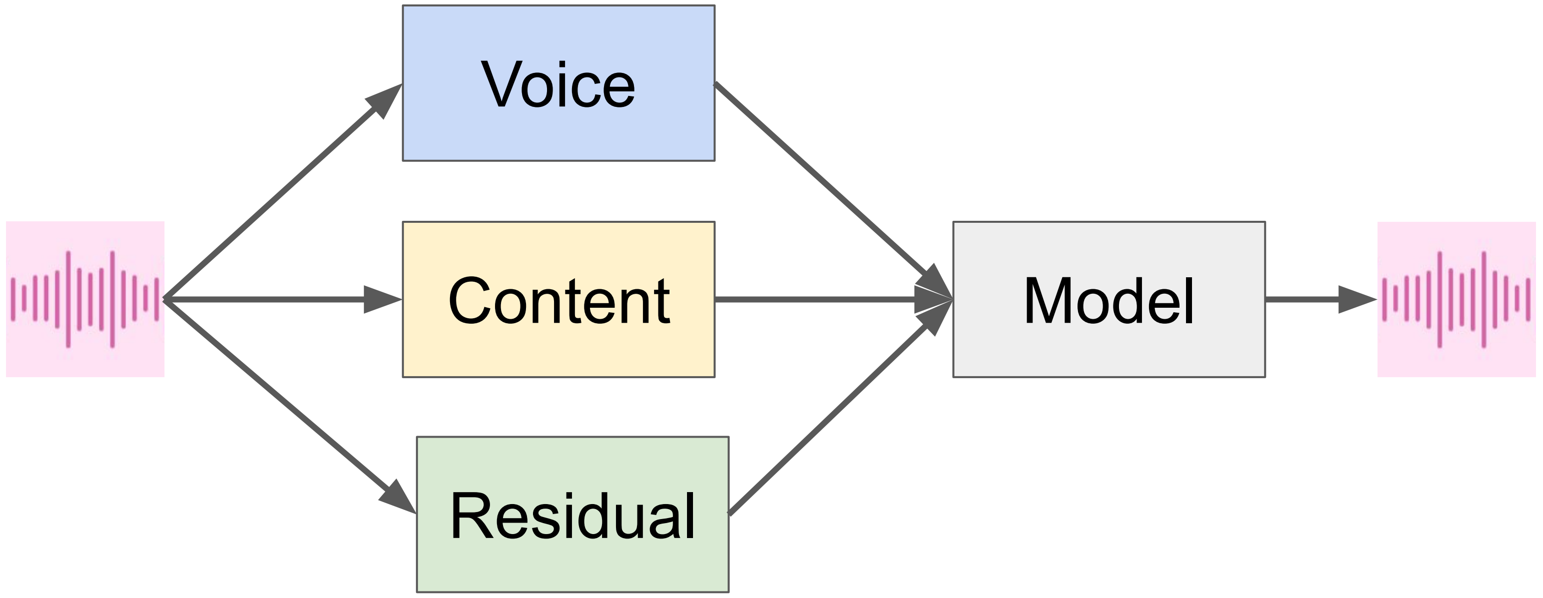
| **Model** | Regression | Regression w/<br>low-dimension latent | Generative |
|---|---|---|---|
| **Example** | Unit-HiFiGAN, ReVISE | | Voicebox |
| **Capability** | 1. Voice conversion<br>2. Generalized audio-visual speech enhancement | | 1. Diverse sampling<br>2. Audio style transfer<br>3. Zero-shot TTS<br>4. Content editing<br>5. Audio infilling |
| **Dataset** | LJSpeech, VCTK, Expresso | | >50K hour multilingual audiobook |

# Part 1.1: Speech Resynthesis from Discrete Disentangled Self-Supervised Representations

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux

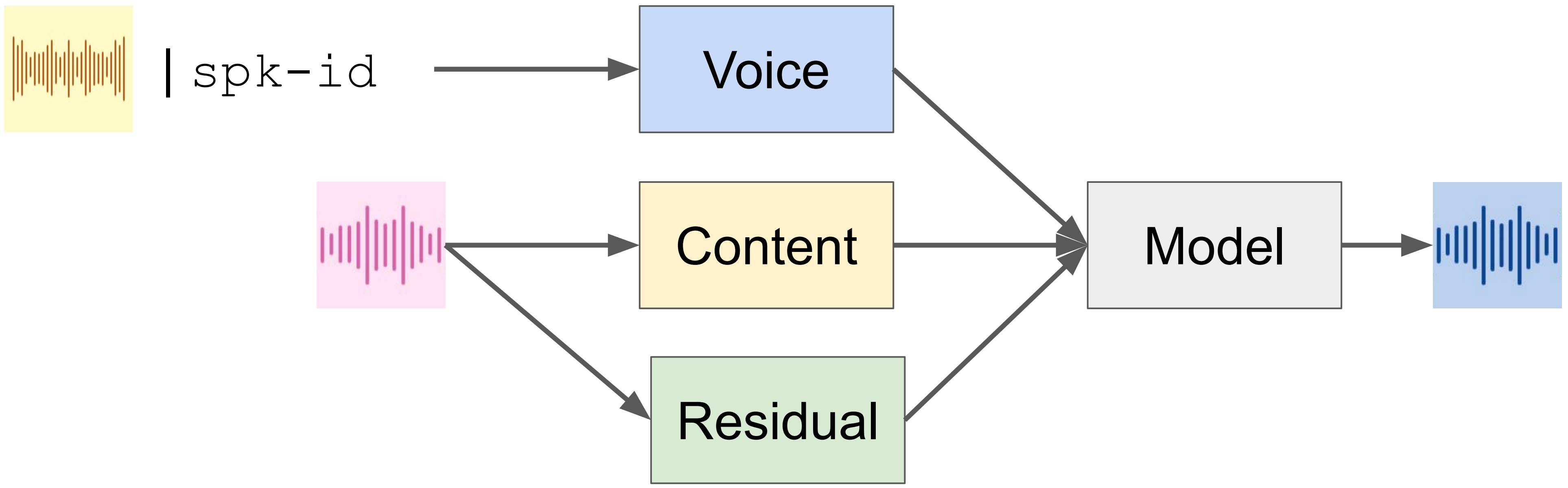
# Goal

- Speech codec: low-bitrate encoding for speech



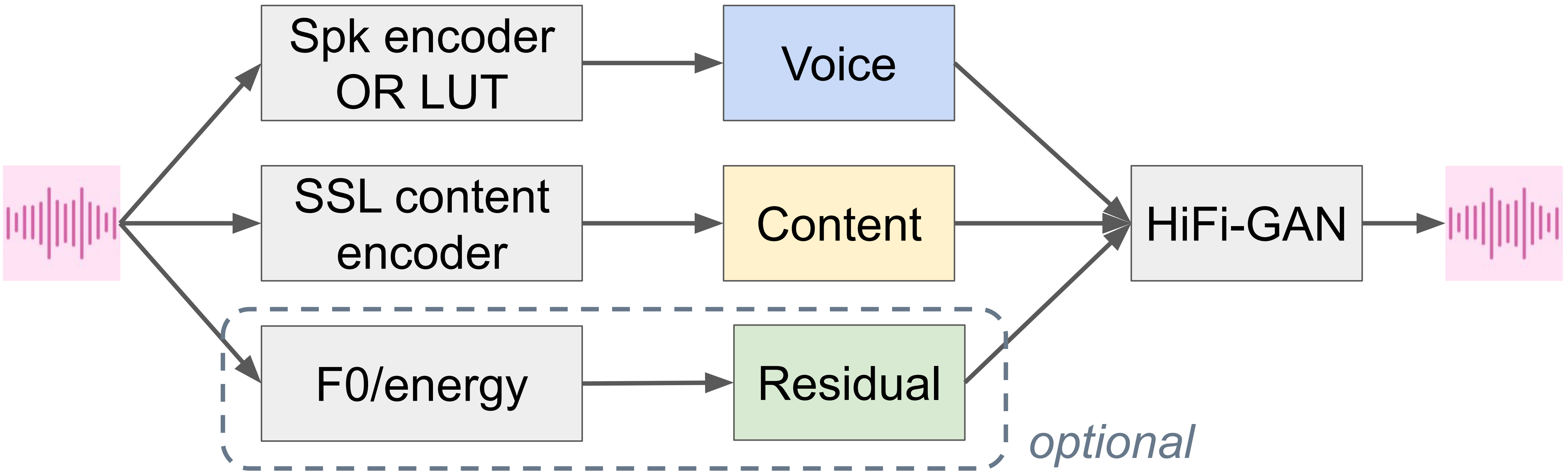
# Goal

- Speech codec: low-bitrate encoding for speech
- Voice conversion: change the voice of source speech while keeping the rest factors
- Voice anonymization: a special case of voice conversion



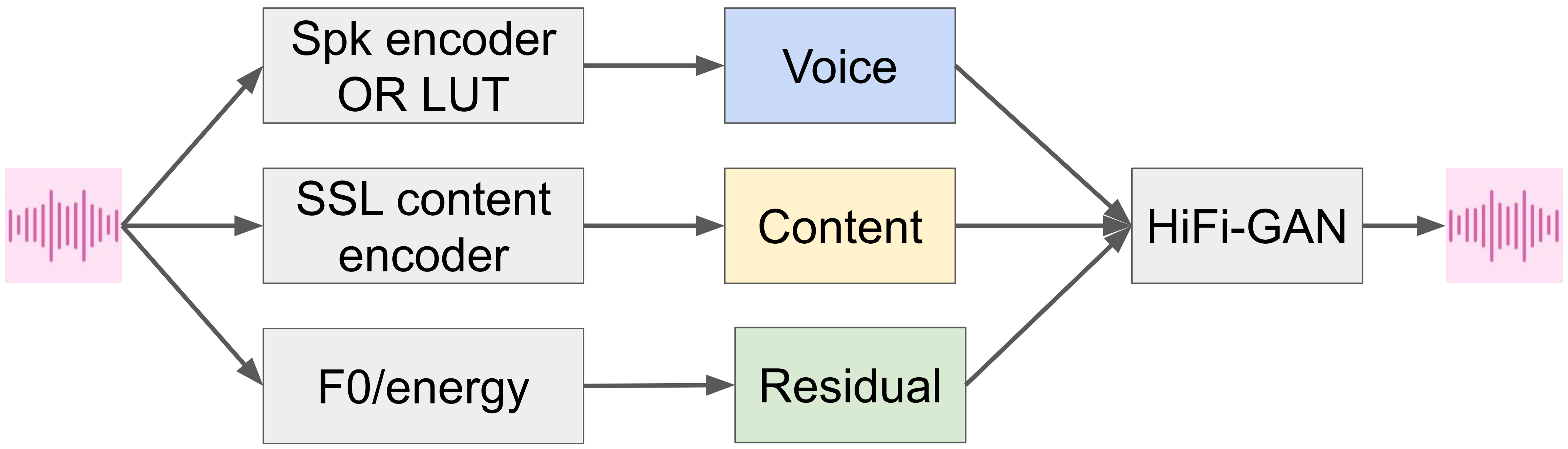
# Method — Unit HiFiGAN

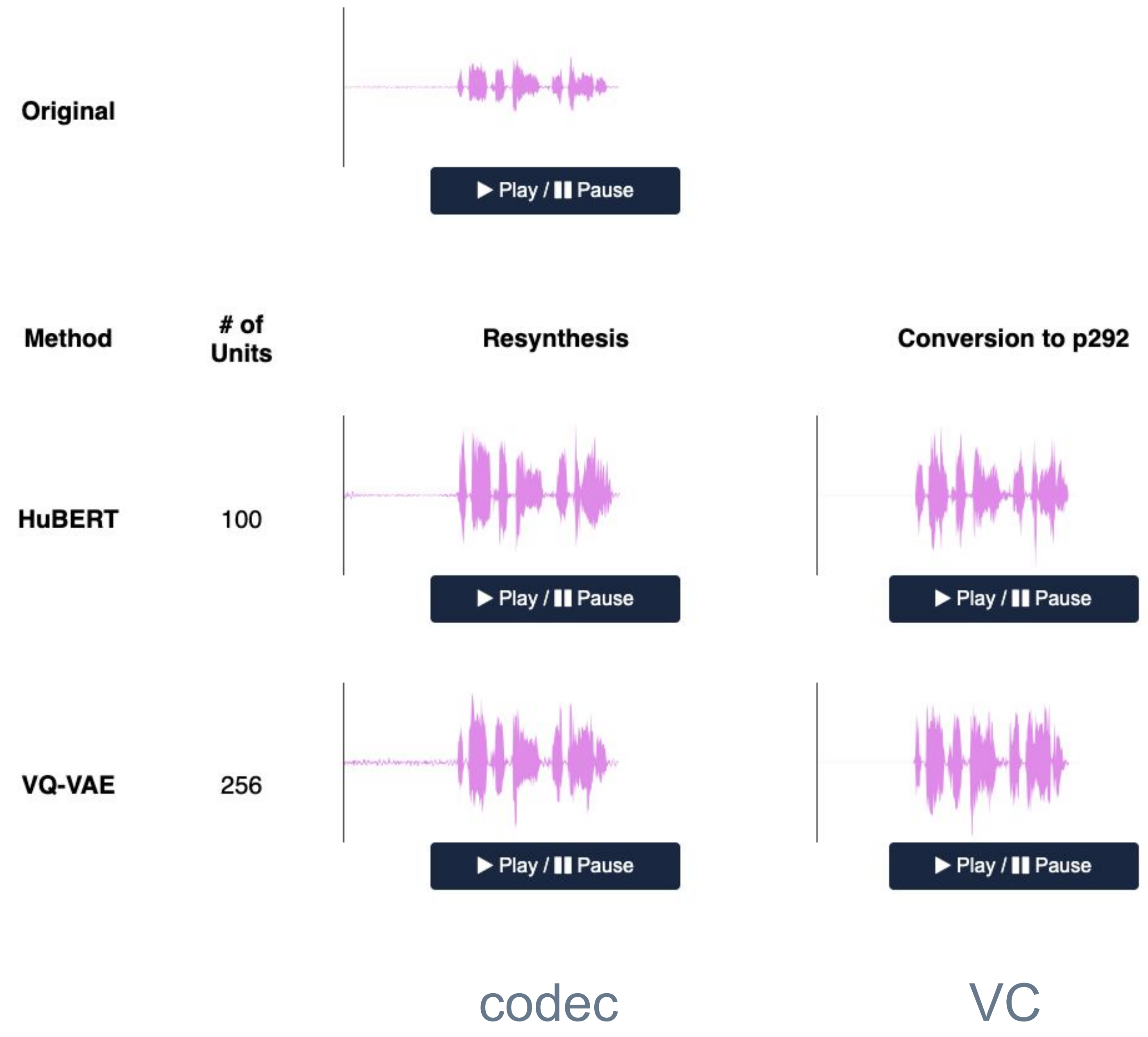
- Use pre-trained disentangled encoders
  - Content: HuBERT (high mutual information with phones) / AE does not work well
    - Why not text or ASR features? Because they drop nonverbal cues (e.g., laughs)
  - Voice: look up table (LUT) or pre-trained speaker embedder
  - Residual: optional if little residual variation
- Backbone: HiFi-GAN (regression + adversarial loss). Decent if most variations are specified



# Results [link]

- Train on LJ+ VCTK
  - multispeaker, clean, non-expressive
- Comparing HuBERT and VQ-VAE for content
  - VQ-VAE encodes *everything,* including speaker
  - Model fails to determine where to infer voice



# Results [link]

- Train on LJ+ VCTK
  - multispeaker, clean, non-expressive
- Comparing HuBERT and VQ-VAE for content
  - VQ-VAE encodes *everything,* including speaker
  - Model fails to determine where to infer voice

| Dataset | Method | Voice Conversion | | | |
|---|---|---|---|---|---|
| | | PER ↓ | WER ↓ | EER ↓ | MOS ↑ |
| VCTK | GT | 17.16 | 4.32 | 3.25 | 4.11±0.29 |
| VCTK | CPC | 23.58 | 15.98 | **4.83** | 3.42 ± 0.24 |
| | HuBERT | **20.85** | **12.72** | 6.01 | **3.58** ± **0.28** |
| | VQ-VAE | 36.88 | 29.44 | 11.56 | 3.08 ± 0.34 |

# Part 1.2: ReVISE: Self-supervised speech resynthesis with visual input for universal and generalized speech enhancement

Wei-Ning Hsu, Tal Remez, Bowen Shi, Jacob Donley, Yossi Adi

# Goal

Tasks for interest

- Lip-to-speech generation
- Audio-visual speech inpainting
- Audio-visual speech enhancement
- Audio-visual source separation

# Goal

Tasks for interest

- Lip-to-speech generation
- Audio-visual speech inpainting
- Audio-visual speech enhancement
- Audio-visual source separation

What are the core requirements?

- Retain textual content
- Improve audio quality

# Goal

Tasks for interest

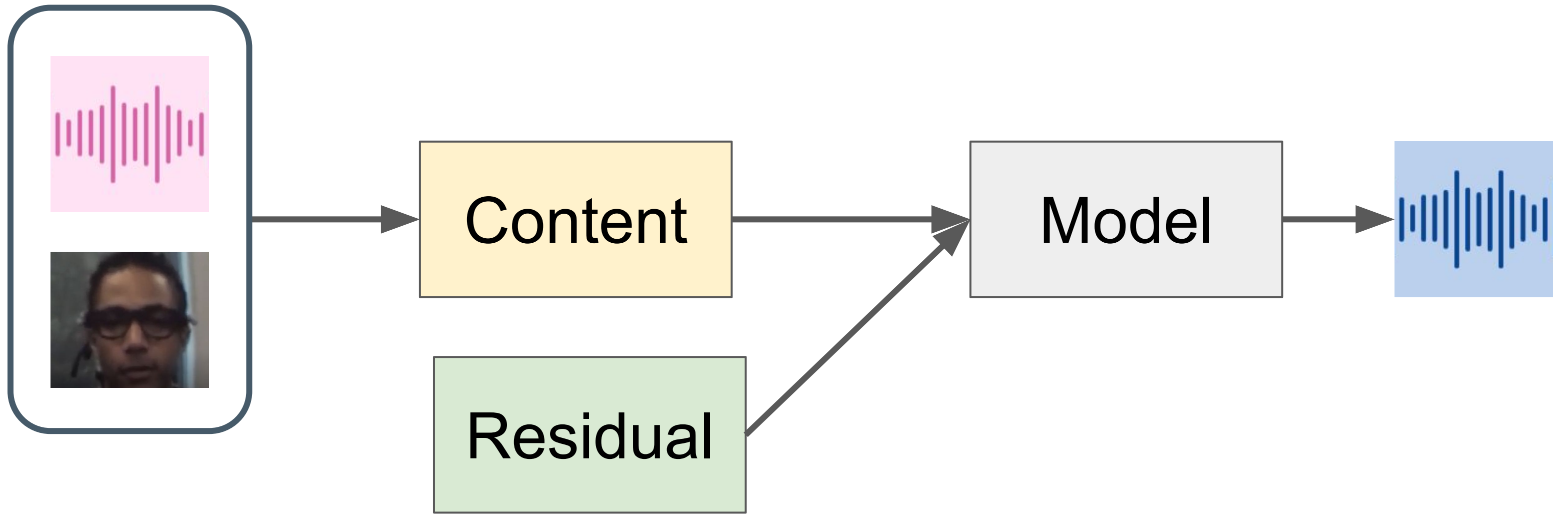
- Lip-to-speech generation
- Audio-visual speech inpainting
- Audio-visual speech enhancement
- Audio-visual source separation

What are the core requirements?

- Retain textual content
- Improve audio quality

Generalized (audio-visual) speech enhancement

- Decompose content, quality, residual
- Focus on improving quality, retaining content, and do not aim to reconstruct the rest



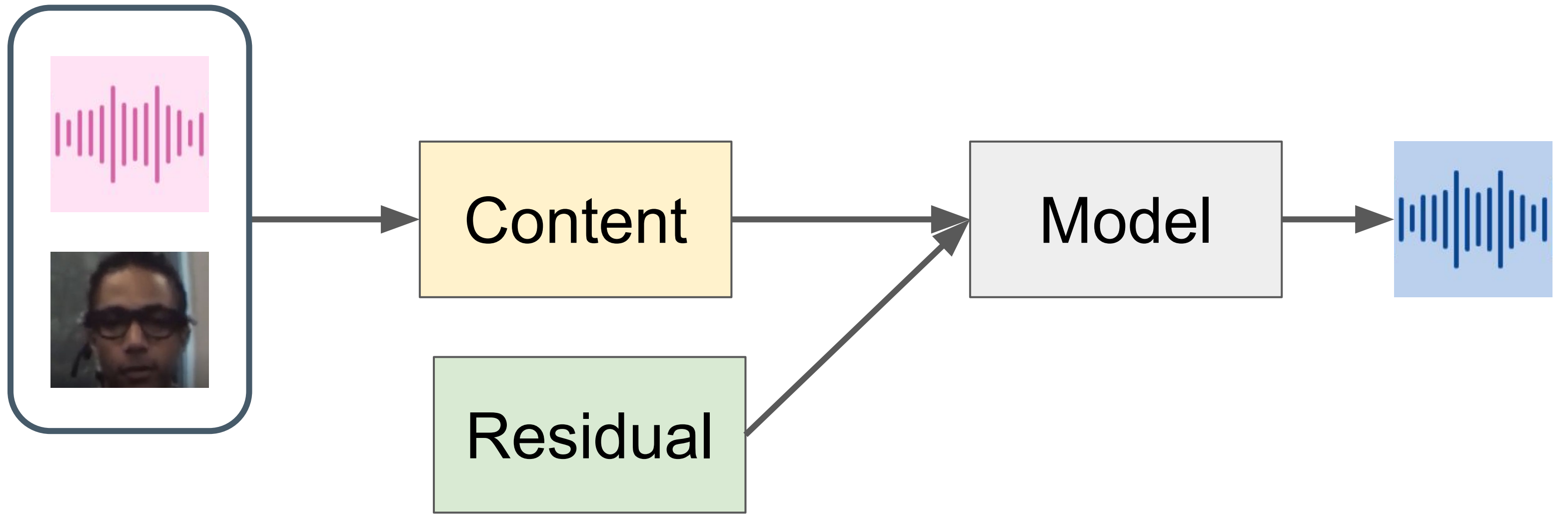
# Goal

Why not aim to reconstruct exactly the original signal?

- Ill-posed problem
- Phase can differ while speech sounds the same
- Reference may not be ideal (mild noise, bad mic)

**Noisy audio-visual input**
**(distant, single channel)**

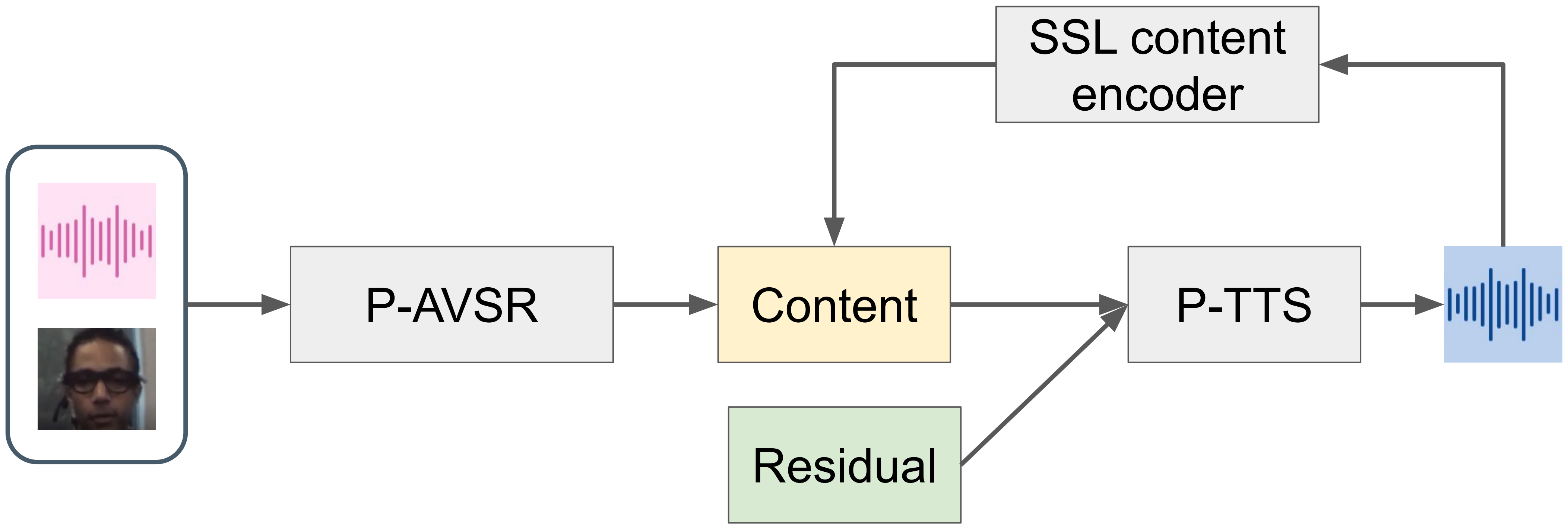
**Reference target**
**(close-talking mic)**



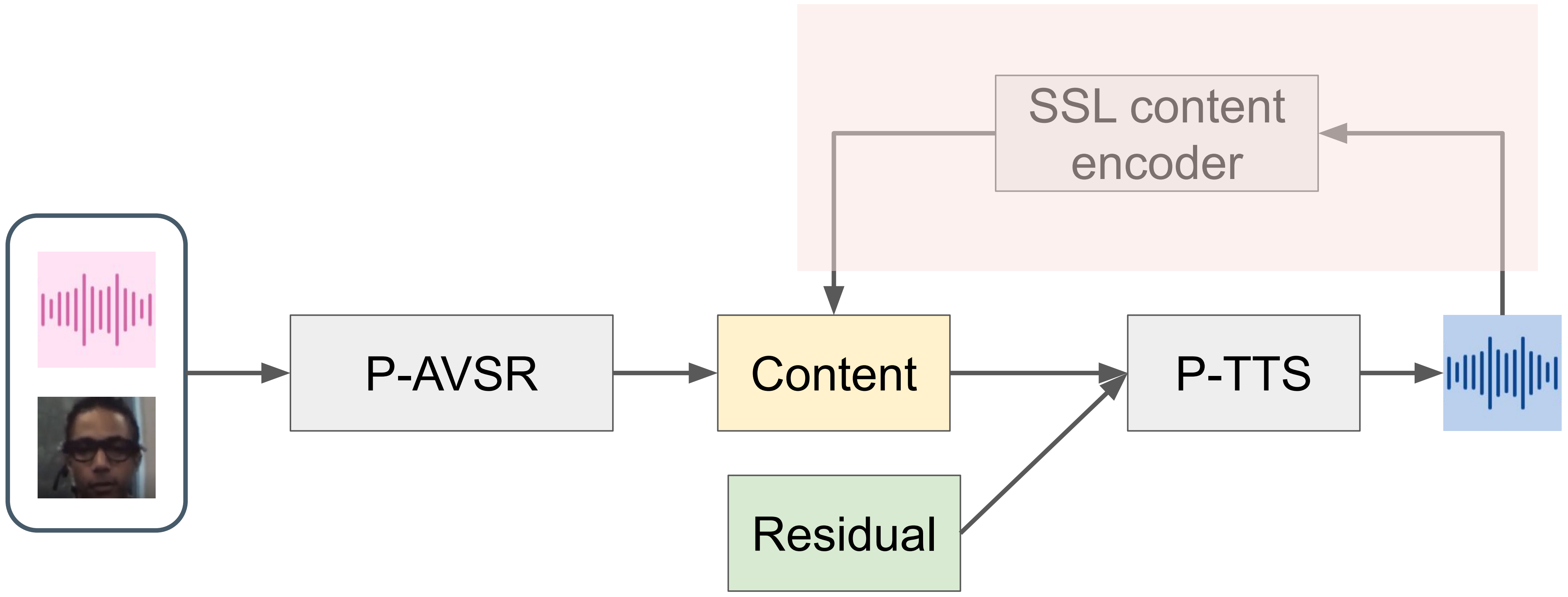
# Method

Treat the problem as pseudo audio-visual speech recognition and pseudo text-to-speech synthesis

- P-AVSR: predict SSL units given audio-visual input
- P-TTS: synthesize clean speech given SSL unit (content) and residual attributes (e.g., speaker)

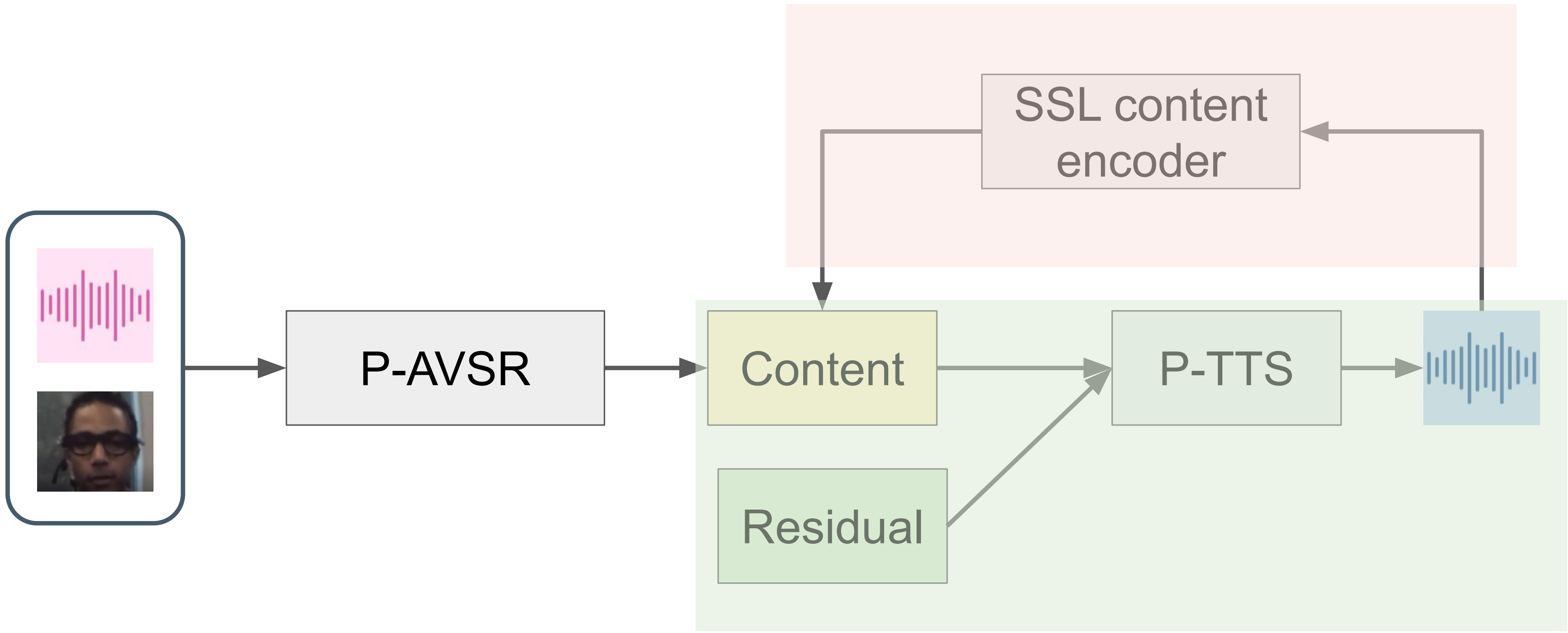


# Method




- Pre-trained HuBERT
- Frozen

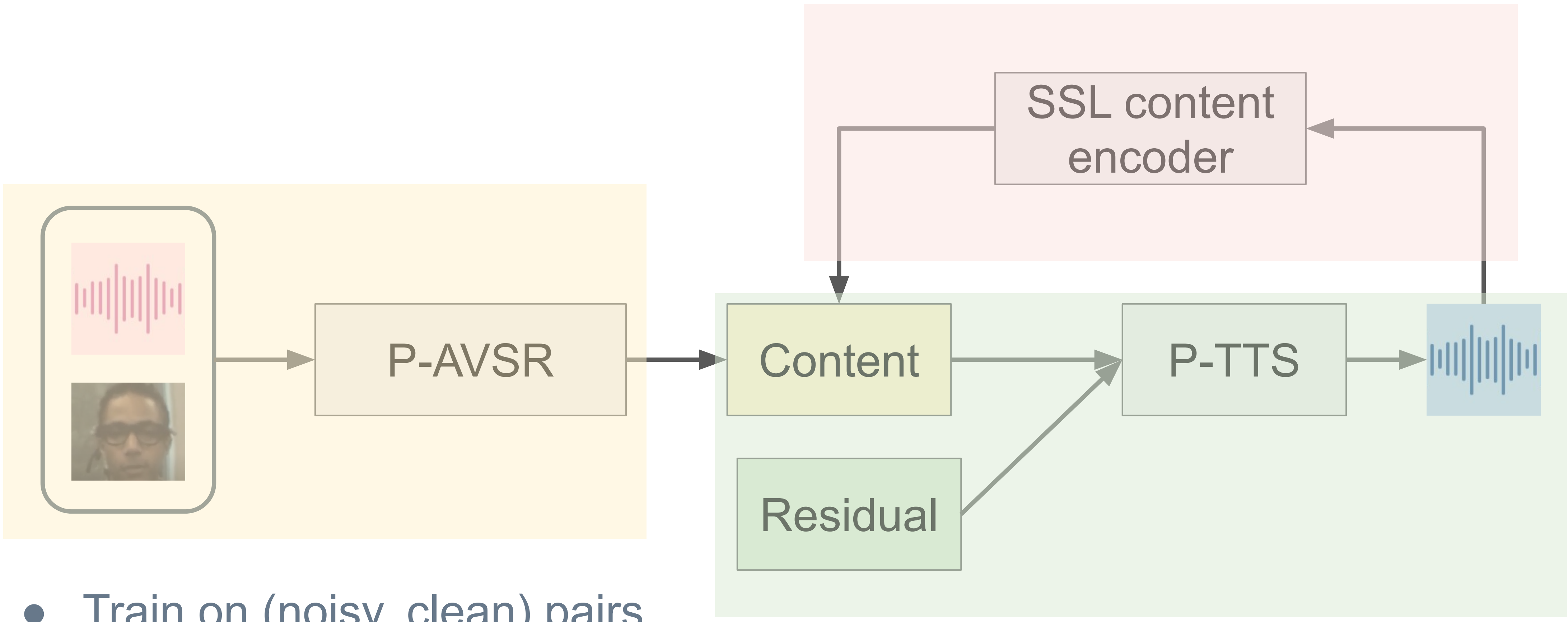
# Method




- Pre-trained HuBERT
- Frozen

- Unit HiFiGAN from previous part
- Train on single-speaker unlabeled clean data
- Does not reconstruct voice, but can easily be extended to preserve voice

# Method




- Pre-trained HuBERT
- Frozen

- Train on (noisy, clean) pairs
- Initialize with AV-HuBERT

- Unit HiFiGAN from previous part
- Train on single-speaker unlabeled clean data
- Does not reconstruct voice, but can easily be extended to preserve voice

# Results [link]

- Effective even in real-world low-SNR low-resource case (EasyCom: 2.2h)
- Better quality than reference audio

**Noisy audio-visual input (distant, single channel)**

**Our model output (beamform +ReVISE)**

**Reference target (close-talking mic)**

# Results [link]

- Effective even in real-world low-SNR low-resource case (EasyCom: 2.2h)
- Better quality than reference audio
- A single model works for all 4 tasks

Noisy audio-visual input
(distant, single channel)

Our model output
(beamform +ReVISE)

Reference target
(close-talking mic)

Package loss

ReVISE output

Reference target

Silent video

ReVISE output

Reference target

# Results [link]

- Effective even in real-world low-SNR low-resource case (EasyCom: 2.2h)
- Better quality than reference audio
- A single model works for all 4 tasks

Disentangled representation

1. reduces labeled data needed
2. enables better modularity/controllability

**Noisy audio-visual input (distant, single channel)**

**Our model output (beamform +ReVISE)**

**Reference target (close-talking mic)**

**Package loss**

**ReVISE output**

**Reference target**

**Silent video**

**ReVISE output**

**Reference target**

# Part 2: Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale

Matthew Le*, Apoorv Vyas*, Bowen Shi*, Brian Karrer*, Leda Sari, Rashel Moritz,
Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, Wei-Ning Hsu*

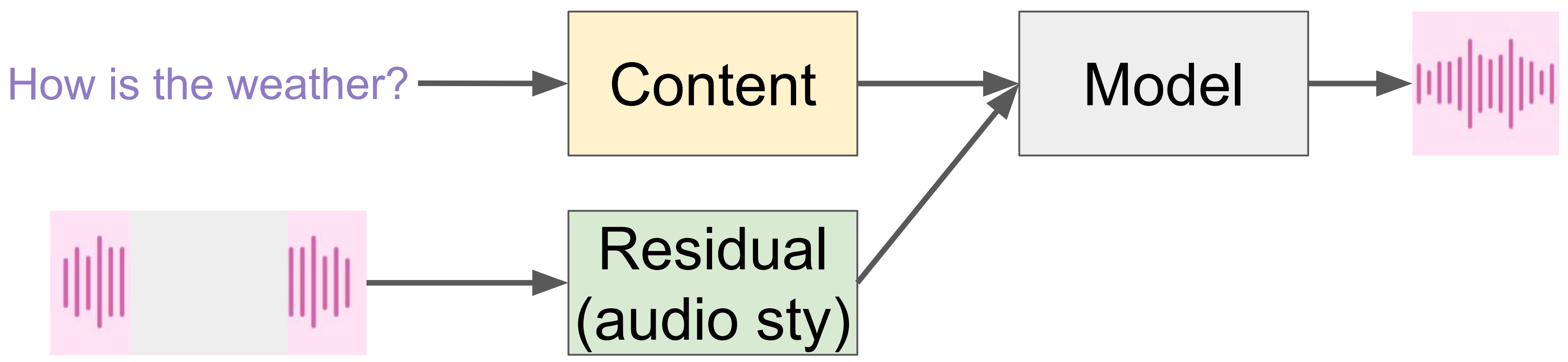
# Key Limitations of Prior Studies

- Limited ability to model stochastic mapping
  - Require input to capture most variation (more deterministic)
  - Use supervised and simple data (less variation)
  - Popular AE/VAE-based models tries to tackle this
    - Still has the assumption that unseen variation lies in low-D manifold
- Case 1: HiFi-GAN with unseen emotion variation
- Case 2: Global style token with unseen noise variation

**In order to scale data, we need to find **a right model** and **a right way to control****

# What is Voicebox?

- Flow-Matching Model with the Optimal Transport probability path
  - Non-autoregressive. Based on ODE and estimate gradient
  - Similar to score-matching diffusion models but with fast training and inference
- We train the model with a text-guided masked infilling task
  - A generalization of next token / chunk prediction. Future context is taken into account
  - We sometimes drop the entire context
  - One model for duration, one model for audio
- How do we control the model?
  - Content: text
  - Audio style (voice, noise, emo, etc.): audio context
  - Implicitly disentangled
- Trained on >50K hours of in-the-wild data in 6 languages

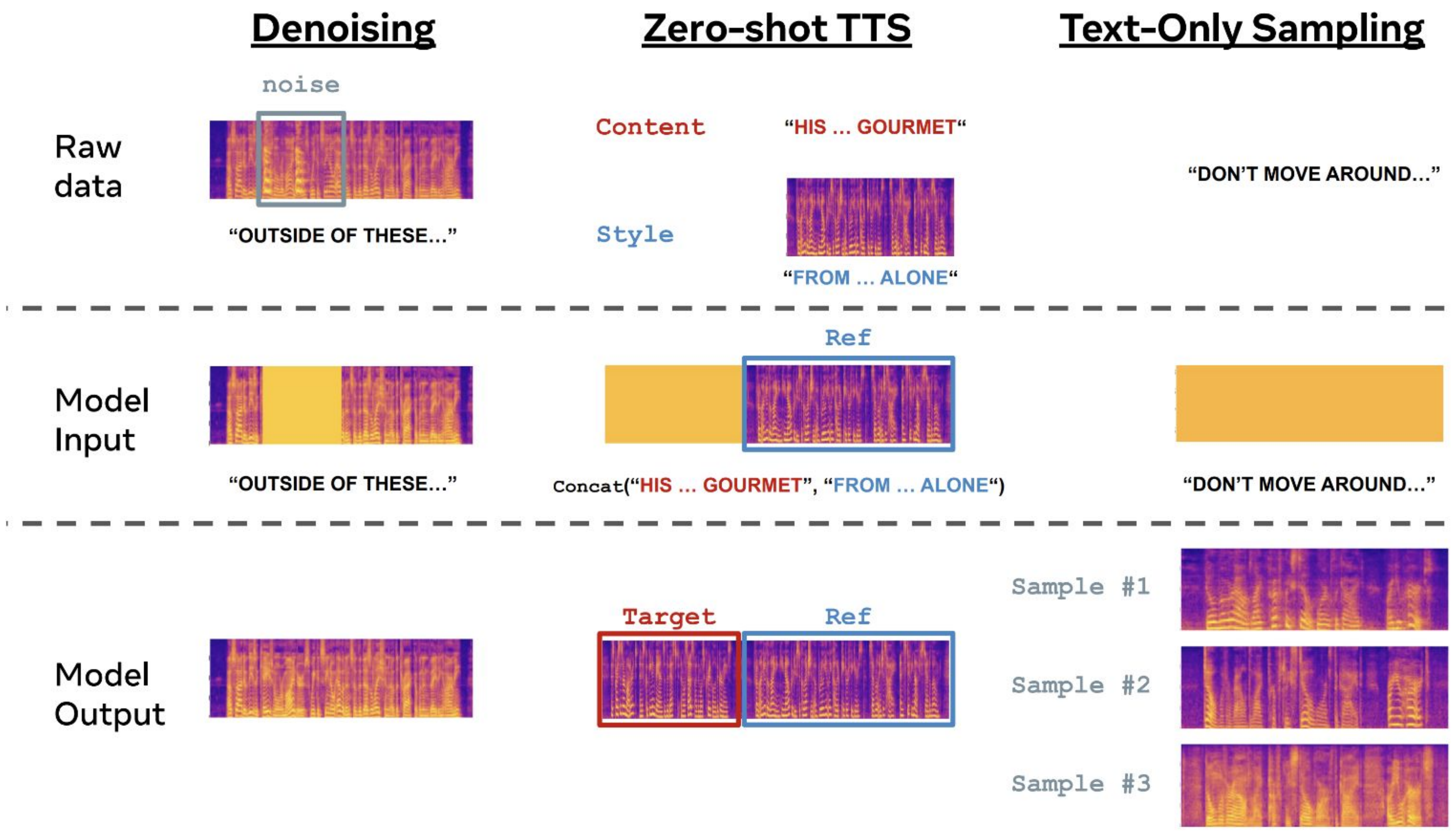


# What Can Voicebox Do?

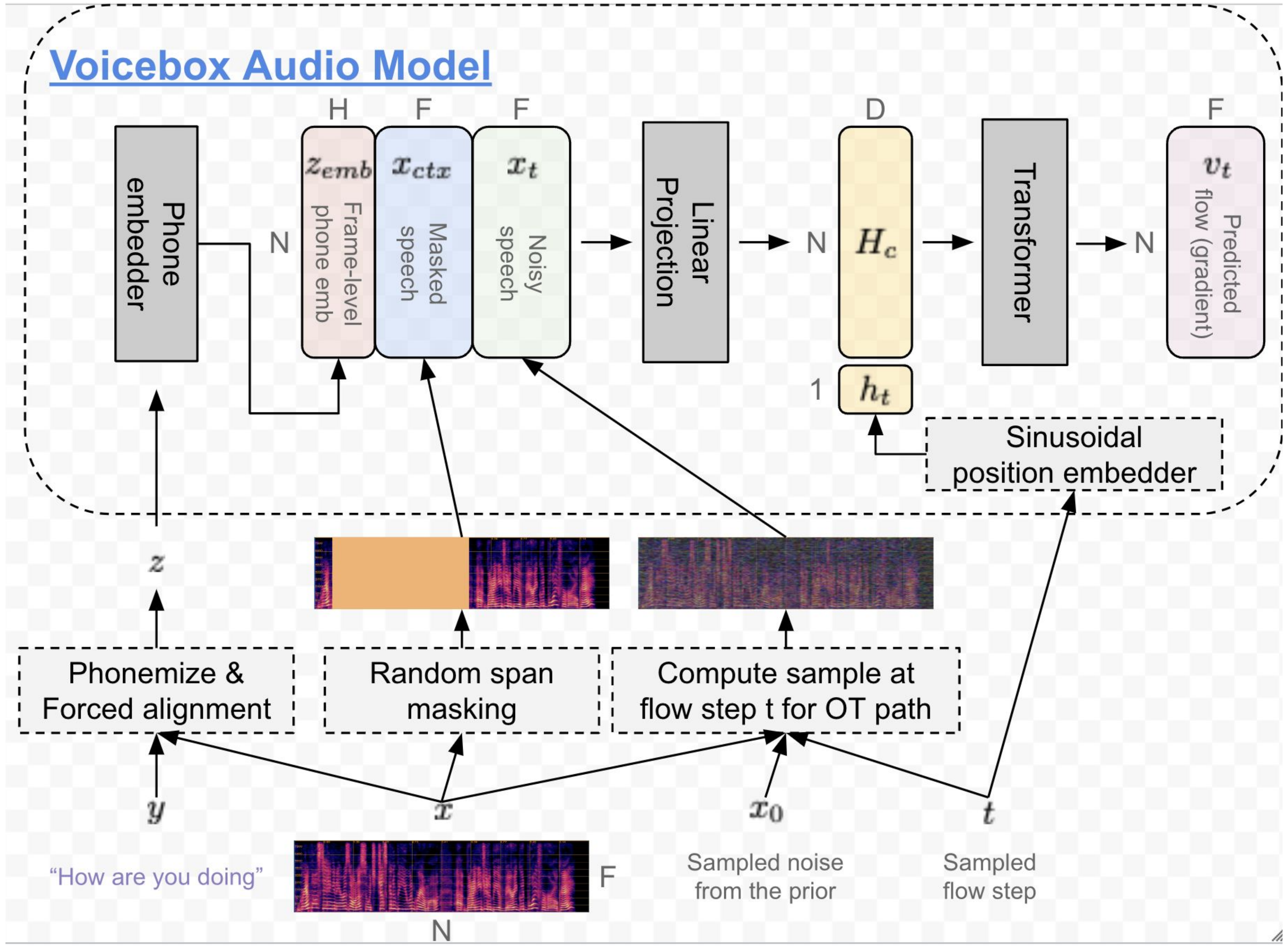
Text-guided speech infilling is powerful, because it subsumes many task

- Transient noise removal through infilling
- Speech content editing
- Voice/emotion/noise/… conversion by example
- Monolingual/cross-lingual zero-shot TTS
- Diverse speech generation for data augmentation

All we need is forming the input differently



# Voicebox Training




- Randomly mask audio
- Sample t ~ [0, 1] and noise from N(0, 1), then compute the x_t and gradient v_t according to the chosen probability path (OT)
- Predict v_t conditioned on (aligned text, masked audio feature, noisified audio feature)

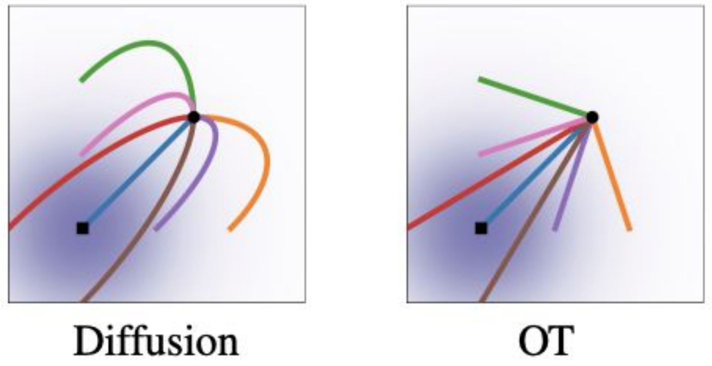
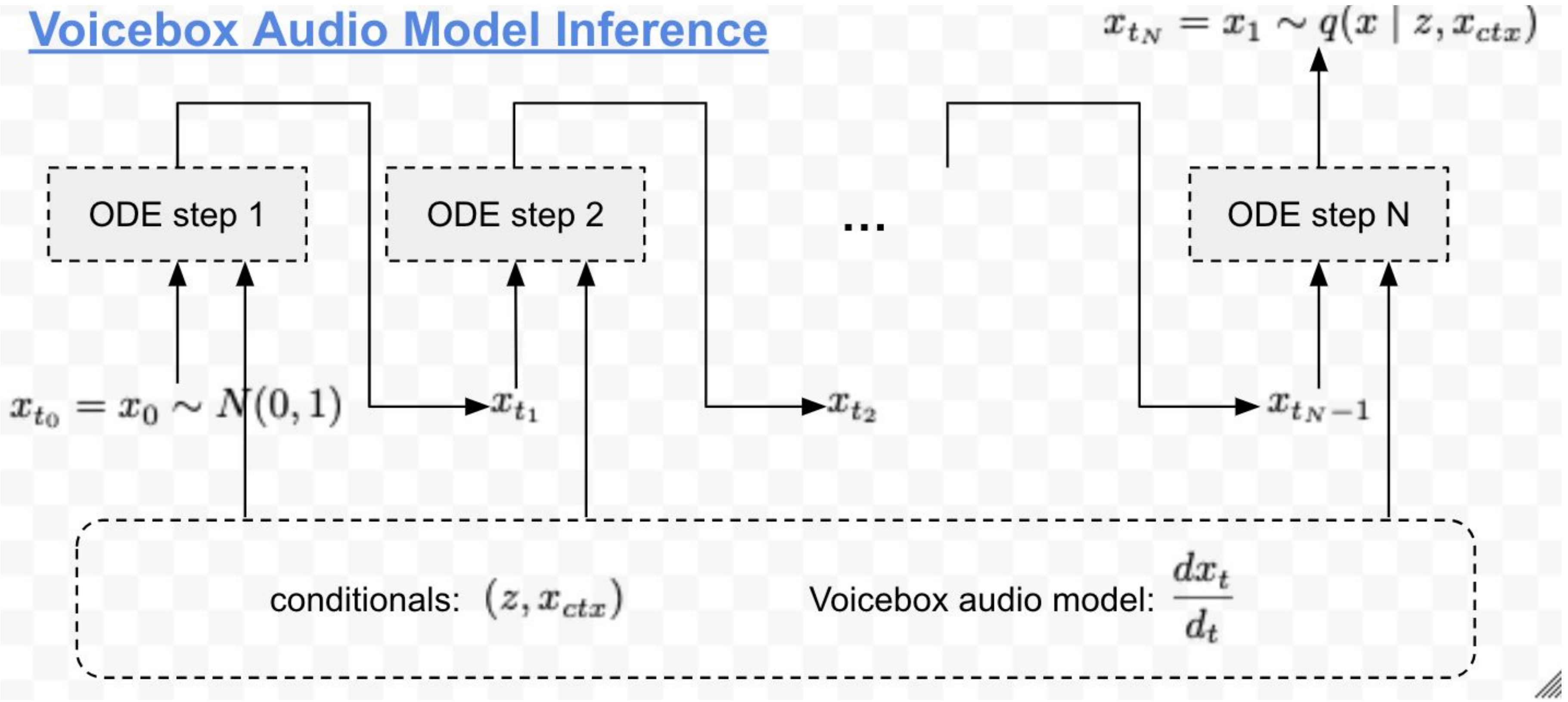



Figure 3: Diffusion and OT trajectories.

# Voicebox Inference




- Use and ODE solver
  - The trained model parameterize dx / dt
  - Sample an initial noise x_0 from N(0, 1)
  - Compute x_1 by doing integration
- Inference speed depends on #ODE steps
  - Configurable
  - Fixed-step and adaptive-step solver

# Demo

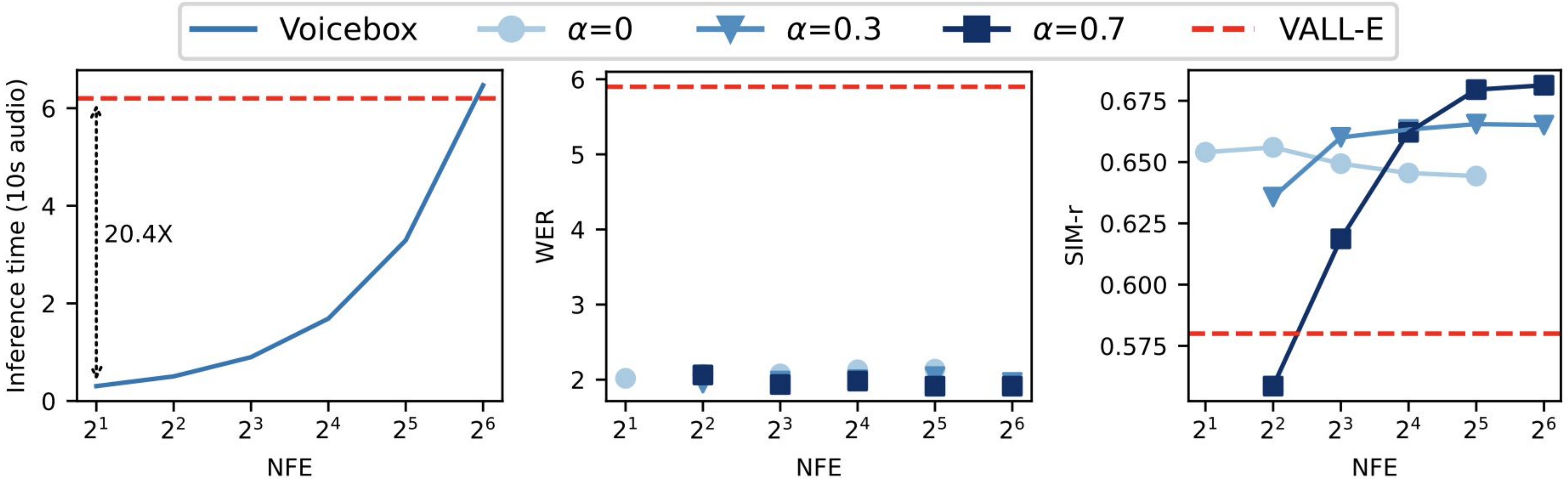
All prompts are recorded by Meta employees (out-of-domain!)

- Denoising/Editing: Acoustic condition (e.g., static noise) is transferred
- ZS-TTS: Accent is also transferred
- Cross-lingual ZS TTS: Only 11 Polish speakers in training data
- Diverse speech sampling: obvious prosody, accent, voice, quality variation
    - Can be effectively used for ASR data creation

| | **WER on real data** | | | |
|---|---|---|---|---|
| | **No LM** | | **4-gram LM** | |
| **ASR training data** | test-c | test-o | test-c | test-o |
| Real audio (100hr) | 9.0 | 21.5 | 6.1 | 16.2 |
| Real audio (960hr) | 2.6 | 6.3 | 2.2 | 5.0 |
| VITS-LJ | 58.0 | 81.2 | 51.6 | 78.1 |
| VITS-VCTK | 33.8 | 55.5 | 30.2 | 53.1 |
| YourTTS (ref=LS train) | 25.0 | 54.6 | 20.4 | 51.2 |
| VB-En ($\alpha = 0$, dur=regr) | 7.1 | 17.6 | 6.5 | 14.6 |
| VB-En ($\alpha = 0$, dur=FM, $\alpha_{dur} = 0$) | 3.1 | 8.3 | 2.6 | 6.7 |

# Efficient Inference [link]

- The model can work fine even with only 2 diffusion steps
  - Take 0.3 seconds to generate 10 second audio
  - 20.4x faster than VALL-E (Token-based LM)




(a) NFE vs Inference Time (b) NFE vs WER (Zero-shot TTS) (c) NFE vs SIM-r (Zero-shot TTS)

# Final Remark

# What’s Next?

- Better controllability for large scale speech generative model
  - Can we independently control speaker while changing other factors?
  - How to better disentangle factors within speaker representation?
- Generalize to more task
  - Global speech enhancement, source separation, translation, …
  - More ways to specify input (audio, video, image, …)
- Scaling law for speech generative models
  - Can we predict how model improves with data?
  - Can we improve scaling law?

# Acknowledgement

Yossi Adi, Robin Algayres, Michael Auli, Arun Babu, Alexei Baevski, Benjamin Bolte, Peng-Jen Chen, Alexis Conneau, Jade Copet, Jacob Donley, Emmanuel Dupoux, Paul-Ambroise Duquenne, Ali Elkahky, Maryam Fazel-Zarandi, Hongyu Gong, Jiatao Gu, Qing He, Brian Karrer, Eugene Kharitonov, Felix Kreuk, Ann Lee, Matthew Le, Kushal Lakhotia, Xutai Ma, Vimal Manohar, Jay Mahadeokar, Rashel Moritz, Abdelrahman Mohamed, Tu-Anh Nguyen, Juan Pino, Adam Polyak, Sravya Popuri, Tal Remez, Morgane Rivière, Leda Sari, Benoit Sagot, Ruslan Salakhutdinov, Holger Schwenk, Bowen Shi, Paden Tomasello, Yun Tang, Yao-Hung Tsai, Apoorv Vyas, Changhan Wang, Mary Williamson, Qiantong Xu, Seamless Team

# Reference


[1] Ren, Yi, et al. "Fastspeech 2: Fast and high-quality end-to-end text to speech."

[2] Shen, Jonathan, et al. "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions."

[3] Kong, Jungil, Jaehyeon Kim, and Jaekyoung Bae. "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis."

[4] Wang, Yuxuan, et al. "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis."

[5] Hsu, Wei-Ning, et al. "Hierarchical generative modeling for controllable speech synthesis."

[6] Wang, Chengyi, et al. "Neural codec language models are zero-shot text to speech synthesizers."

[7] Shen, Kai, et al. "Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers."

[8] Le, Matthew, et al. "Voicebox: Text-Guided Multilingual Universal Speech Generation at Scale."

[9] Polyak, Adam, et al. "Speech resynthesis from discrete disentangled self-supervised representations."

[10] Hsu, Wei-Ning, et al. "Revise: Self-supervised speech resynthesis with visual input for universal and generalized speech enhancement."

Meta AI

Meta AI