Table of Contents

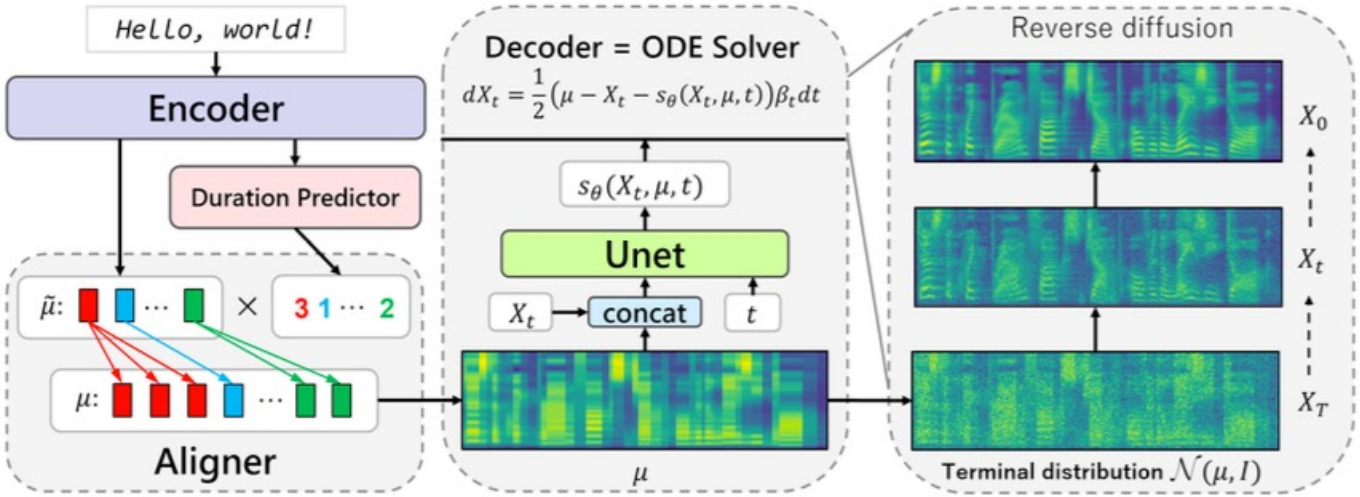
1. Goal

- Getting the hang of the diffusion models
  - DDPM : DiffSinger
  - VPSDE : Grad-TTS
  - VESDE : SGMSE

2. Korean Singing Voice Synthesis with Diffusion

2.1 Method

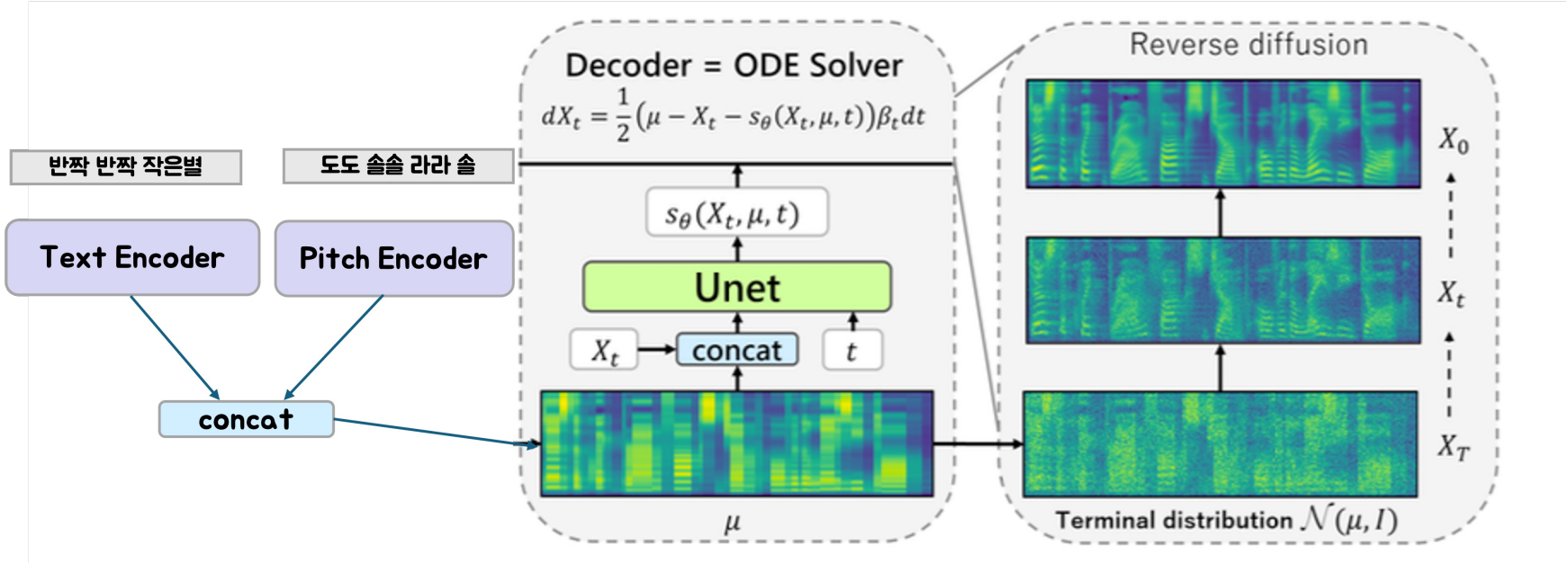
- We have referenced Grad-TTS: A Diffusion Probabilistic Model for Text-to-Speech extensively.
  - Score-based decoder to produce mel-spectrograms of given text.
  - It generalizes the DPM by transforming the forward diffusion over an infinite time horizon, converting all data distributions not to N(0, I) but to a normal distribution N(μ, Σ=I) in forward diffusion process.
  - During sampling, it employs reverse diffusion starting from sampling N(μ, Σ=I).
  - Consists of three modules: encoder, duration predictor, and decoder.



2. Korean Singing Voice Synthesis with Diffusion

2.1 Method

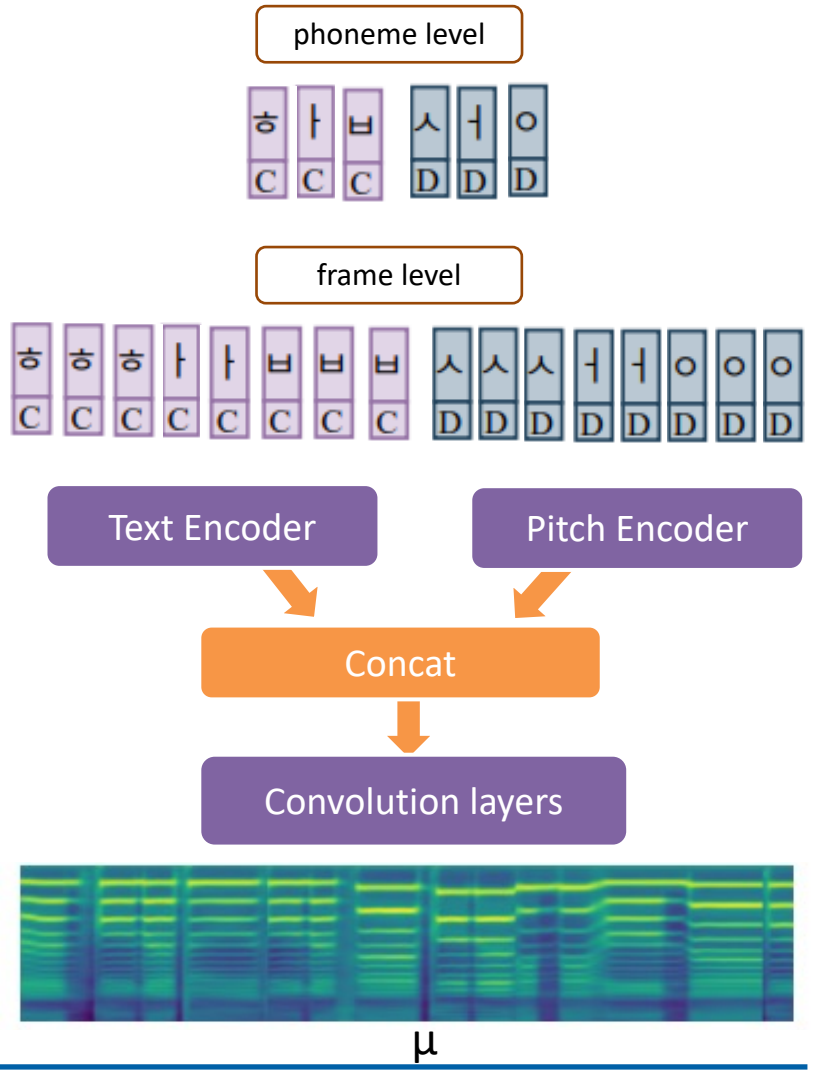
- Unlike TTS, the duration of each syllable is predetermined, so a duration predictor is not needed.
- Therefore, the duration predictor is removed from the Grad-TTS model architecture.
- We added text encoder and a pitch encoder to encode the input information.



2. Korean Singing Voice Synthesis with Diffusion

2.1 Method

- **Data Representation and Encoder**
  - We followed data representation skim of Tae et al (2021).
  - After passing through the text encoder and pitch encoder, the pitch and text, originally represented at the phoneme level, can be transformed to the mel spectrogram frame level.
    - Following the approach from, we allot 3 frames for the onset(초성) and coda(종성), and n − 6 for the nucleus(중성) per each syllable.
  - The text and pitch embeddings are then concatenated and processed through several convolutional layers to match the dimensions of the original mel-spectrogram, ultimately forming μ.



2. Korean Singing Voice Synthesis with Diffusion

2.2 Experiments and Results

- **Dataset**
  - Children's Song Dataset (Choi et al, 2020)
    - Composed of English and Korean children songs sung by a professional female singer.
    - Each song is accompanied by MIDI and text annotations and sung twice in two different keys.
    - We used 50 Korean songs, which totals approximately two hours in length excluding silence intervals.

2. Korean Singing Voice Synthesis with Diffusion

2.2 Experiments and Results

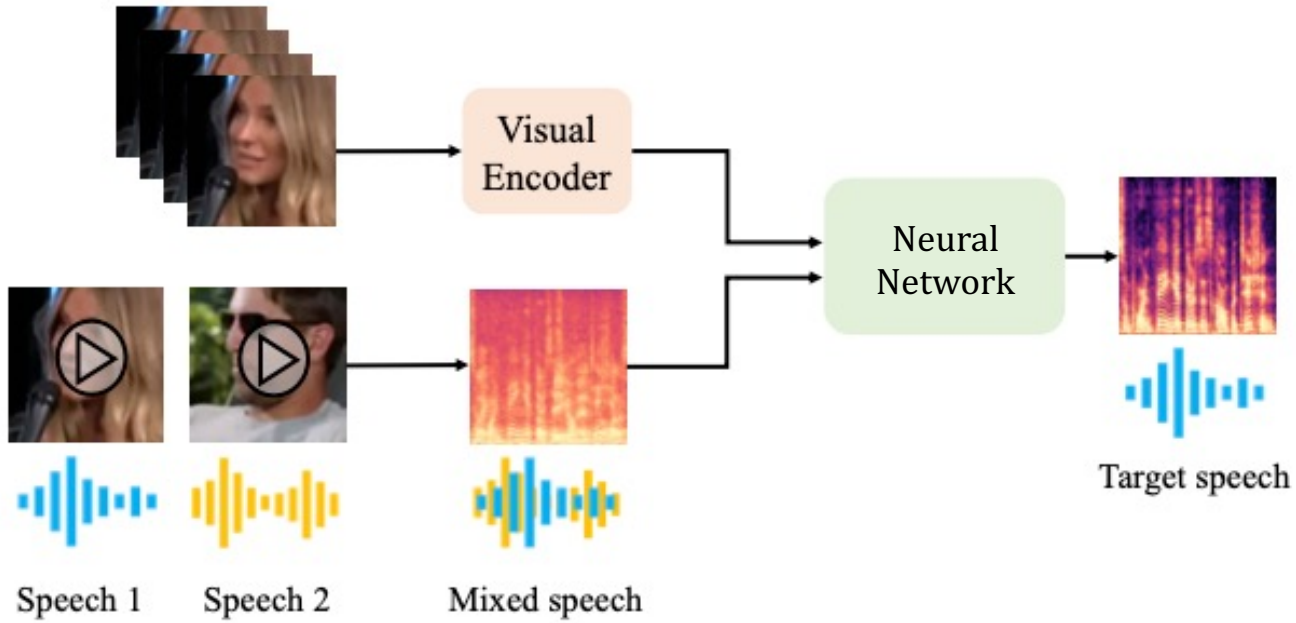
- **Results (left : ground truth, right : generated)**
  - 당신은 누구십니까


  - 창밖을 보라


3. Audio-Visual Speech-Separation with Diffusion

3.1 Problem Definition

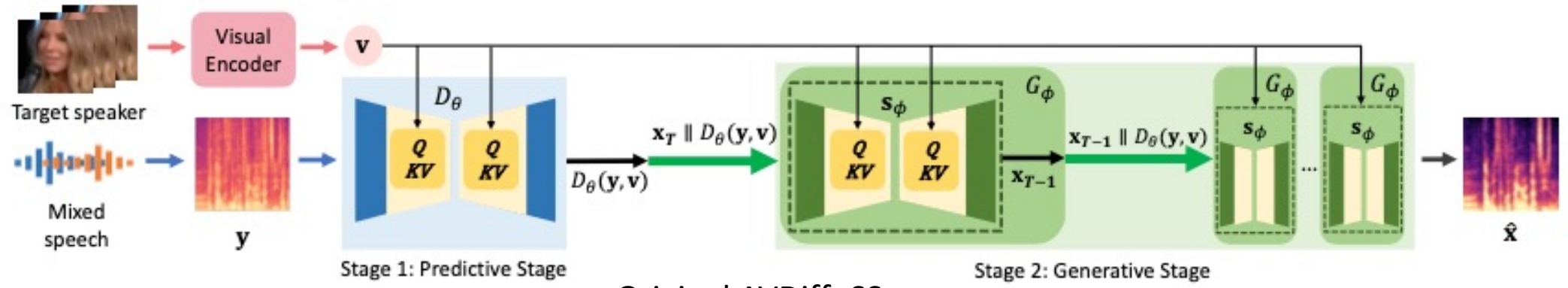
- Audio-visual speech separation (AVSS) : leverages both audio and visual cues to separate speech signals from a mixture of speeches.



3. Audio-Visual Speech-Separation with Diffusion

3.2 Approach 1 : Generator Only

- Attempted to implement AVDiffuSS (Lee et al., 2023) in a more compact form with few modifications.

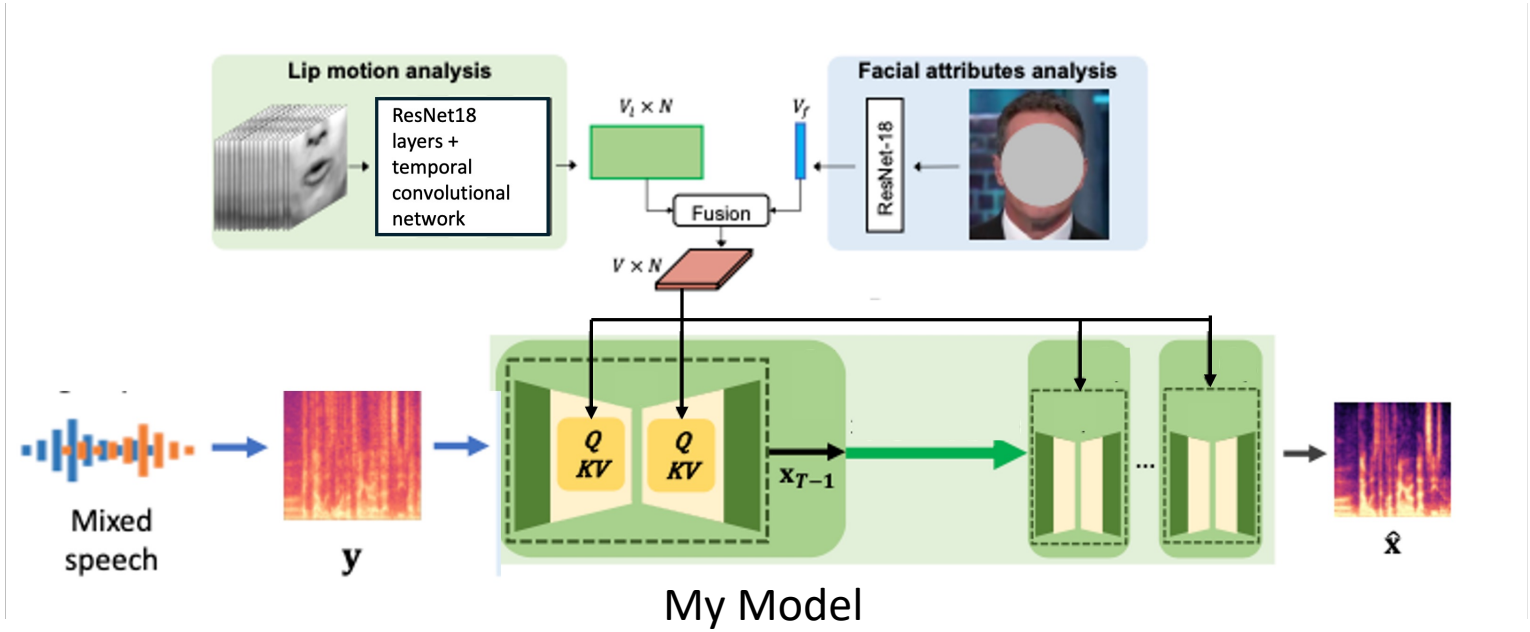


Origianl AVDiffuSS

3. Audio-Visual Speech-Separation with Diffusion

3.2 Approach 1 : Generator Only

- Unlike AVDiffuSS,
  - I separated lip motion encoder and facial attributes encoder for encoding purpose.
  - Remove predictor for simplicity.

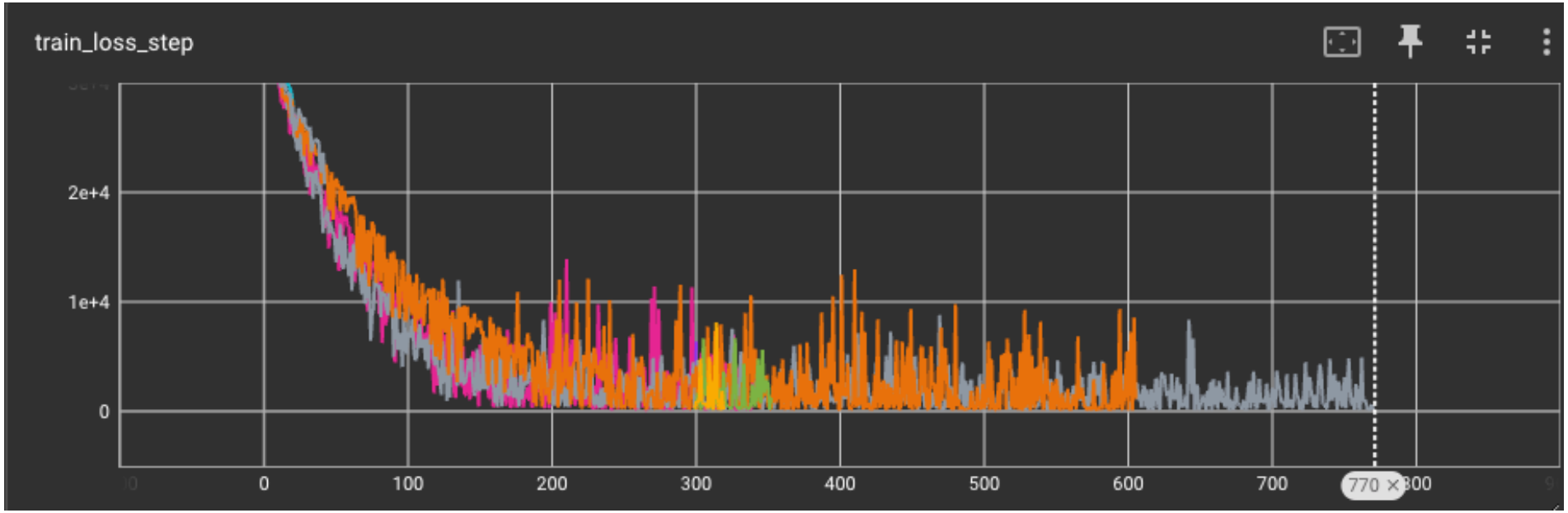


My Model

3. Audio-Visual Speech-Separation with Diffusion

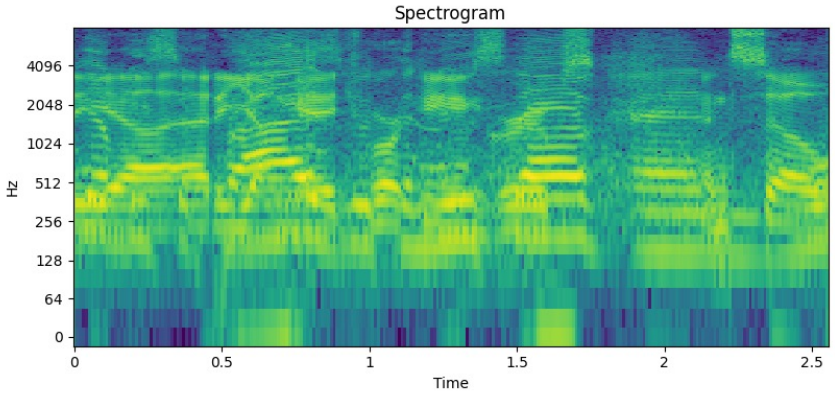
3.2 Approach 1 : Generator Only

- Error stop decreasing at some point.

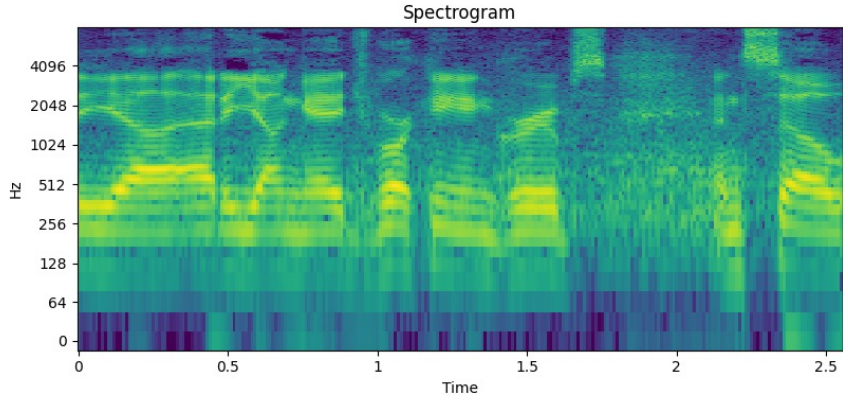


3. Audio-Visual Speech-Separation with Diffusion

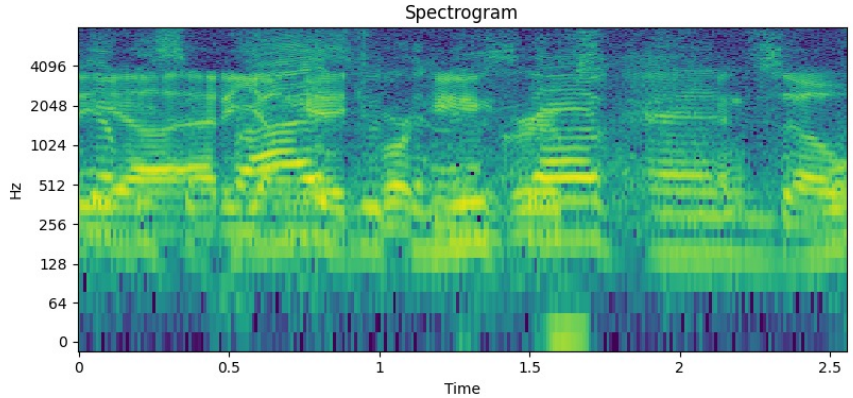
3.2 Approach 1 : Generator Only



Mixed Speech



Separated Speech GroundTruth



Result

3. Audio-Visual Speech-Separation with Diffusion

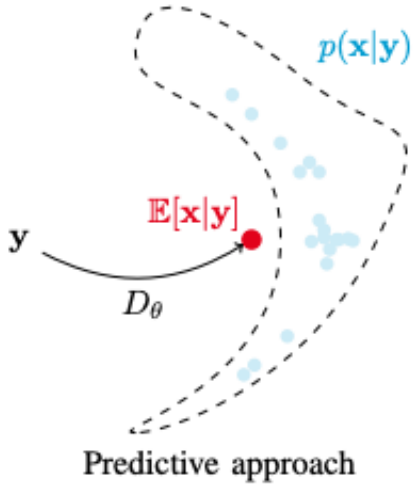
3.2 Approach 1 : Generator Only

- Approaches Tried
  - Switching to a Pre-trained Encoder
  - Increasing Model Size
  - Reducing Learning Rate
  - Modifying Cross Attention Code
  - Multi-phase Training
    - Begin training with easy samples (mixed speech with different gender speakers) and then gradually generalize to more complex cases.
  - Adding Predictor as in the Original Paper -> Approach 2

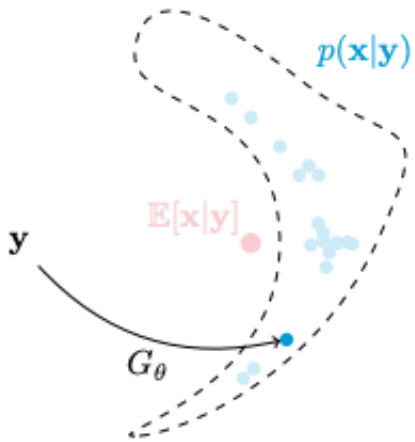
3. Audio-Visual Speech-Separation with Diffusion

3.2 Approach 2 : Generator with Predictor

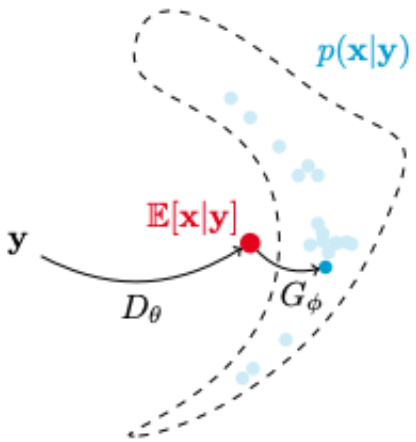
- In the predictive stage, the same Unet model used in the generator is employed to directly predict the spectrogram.
  - The predicted spectrogram is then optimized with MSE loss.
- The output of the predictive stage is then fed into the generator, which employs a diffusion-based model.



Predictive approach



Generative model



StoRM

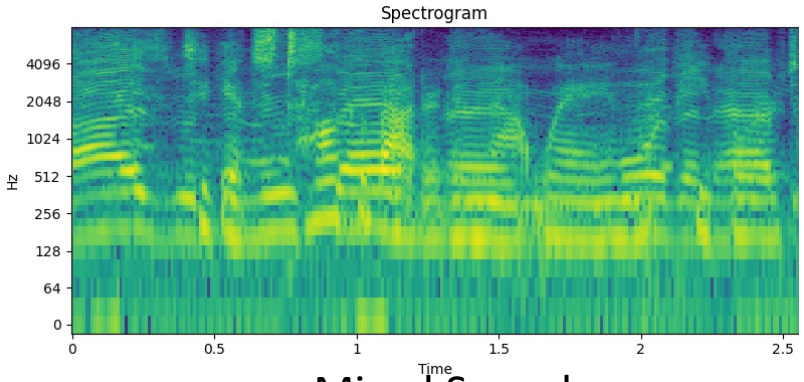
$$L_{pred} = \mathbb{E}\left[\|\mathbf{x} - D_\theta(\mathbf{y}, \mathbf{v})\|_2^2\right],$$

$$L_{diff} = \mathbb{E}\left[\|\mathbf{s}_\phi(\mathbf{x}_\tau, \mathbf{y}, \mathbf{v}, \tau) + \frac{\mathbf{x}_\tau - \mathbf{x}}{\sigma_\tau}\|_2^2\right],$$

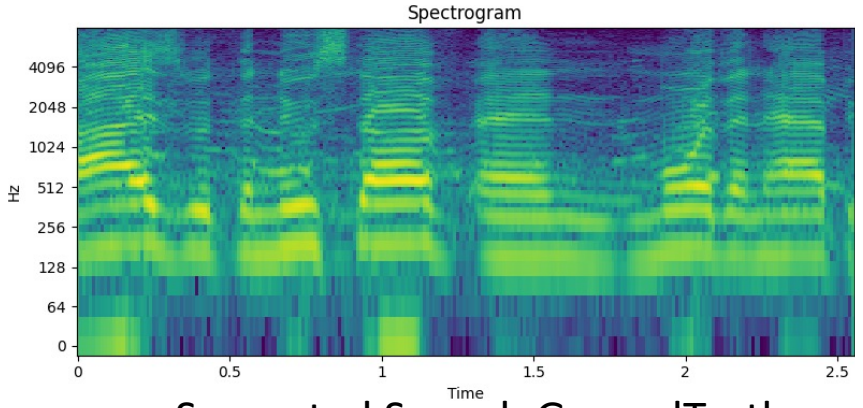
$$L_{total} = \lambda_1 * L_{pred} + \lambda_2 * L_{diff}.$$

3. Audio-Visual Speech-Separation with Diffusion

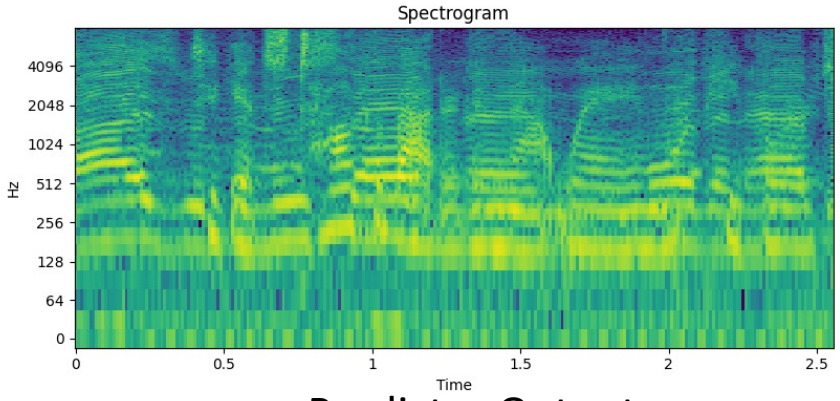
3.2 Approach 2 : Generator with Predictor



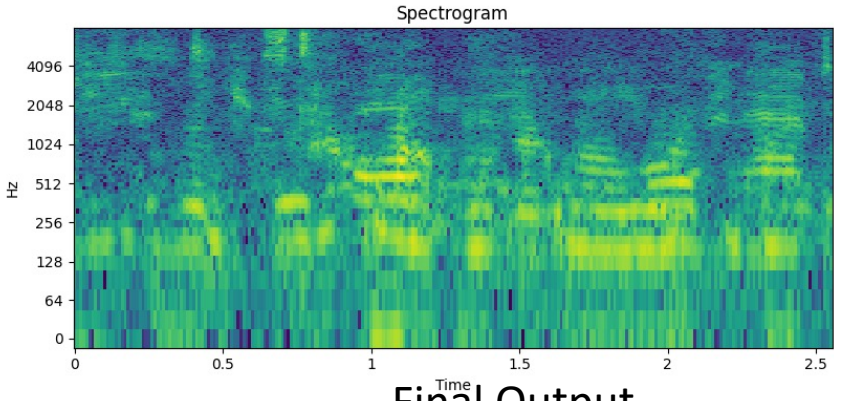
Mixed Speech



Separated Speech GroundTruth



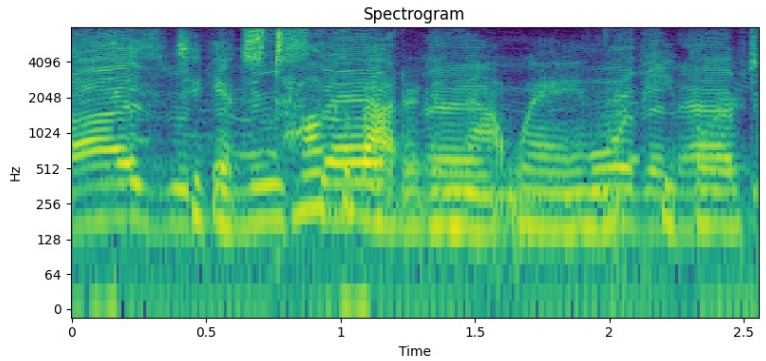
Predictor Output



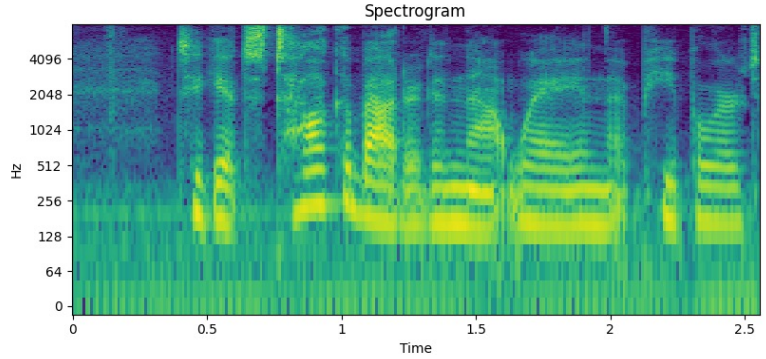
Final Output

3. Audio-Visual Speech-Separation with Diffusion

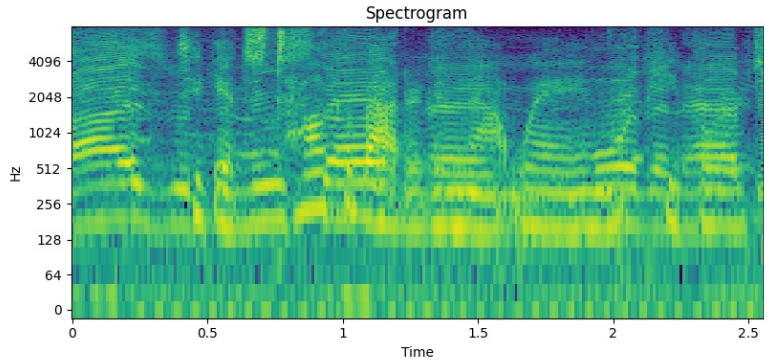
3.2 Approach 2 : Generator with Predictor



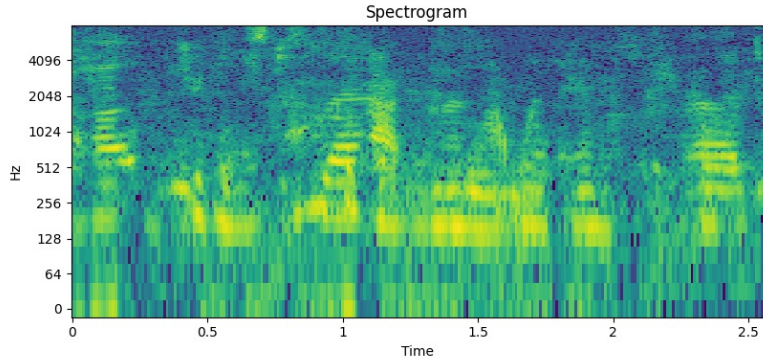
Mixed Speech



Separated Speech GroundTruth



Predictor Output



Final Output

3. Audio-Visual Speech-Separation with Diffusion

- Possible Reasons for Failure
  - Insufficient training time: The original AVDiffuSS paper trained the model for 12 days.
  - Inadequate hyperparameter tuning
  - Suboptimal model architecture