

## 01. Introduction

- TTS는 여러 구성 요소를 통해 텍스트의 원시 음성 파형을 합성하는 기술
- 기존 : 2-stage 생성 모델링
  - 1. 텍스트 정규화, 음소화 등의 전처리
    - mel-spectrogram 또는 linguistic 특징과 같은 중간 음성 표현 생성
  - 2. 생성 모델링
    - 전처리 데이터를 기반으로 원시 파형 생성
    - 2단계 모델은 1단계에서 생성된 데이터로 학습되는데, 높은 품질을 위해 sequential training or fine-tuning 요구
    - 미리 정의된 중간 특징에 대한 의존성으로 학습된 숨겨진 표현을 적용하기 어려움.
- NN-based autoregressive TTS
  - 사실적인 음성 합성
  - 순차적 프로세스로 최신 병렬 프로세스를 완전히 활용하기 어려움.
    - 이를 위한 non-autoregressive 기법
      - text-spectrogram 생성 단계에서 pre-trained autoregressive teacher networks에서 attention을 추출
      - 텍스트, 스펙트로그램 간의 정렬 학습 어려움을 감소
    - 최근, likelihood-based 기반 방법
      - target mel-spectrogram의 likelihood 최대화
      - 정렬을 추정하거나 학습하여 외부 정렬 모듈에 대한 의존성 제거
- 최근 여러 연구(FastSpeech 2s, EATS)에서는 전체 파형이 아닌 짧은 오디오 클립에 대한 훈련과 같은 방식 제안
  - mel-spectrogram decoder를 활용하여 텍스트 표현 학습
  - 특수한 spectrogram loss를 사용해 다것 음성과 생성된 음성 간의 불일치 완화
  - 그러나, 학습된 표현을 활용했음에도 품질은 2-stage보다 뒤떨어짐.

## 02. Method

## 01. Variational Inference

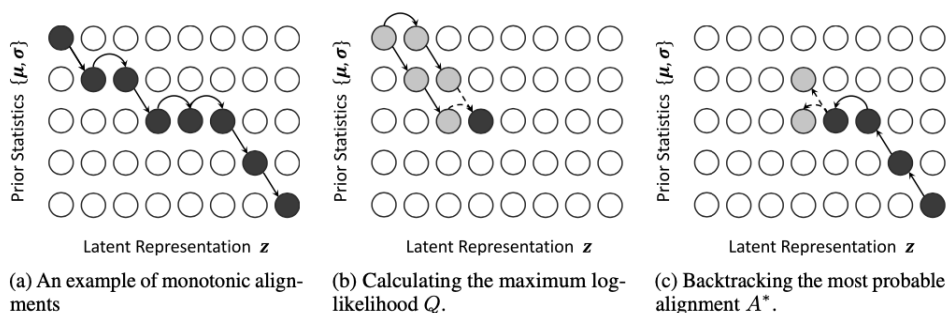
- VITS는 ELBO를 최대화하는 목적을 가진 conditional VAE
- training loss : negative ELBO

## 02. Reconstruction Loss

- melspectrogram 사용
- decoder를 통해 잠재 변수  $z$ 를 waveform domain으로 upsampling
- 이를 mel-spectrogram domain으로 변환
- 라플라스 분포로 가정
  - 상수항을 무시한 max likelihood estimation
- partial sequence를 decoder의 입력으로 사용
  - 효율적인 end-to-end 훈련

### 03. MONOTONIC ALIGNMENT SEARCH(MAS) in Glow-TTS

- latent variable(i.e. input speech)와 prior distribution의 statistics(i.e. text) 사이의 가장 그럴듯한 MONOTONIC ALIGNMENT를 찾는 것
- $Q_{i,j}$ 가 maximum log-likelihood라면, recursive하게  $max(Q_{i-1,j-1}, Q_{i,j-1})$
- backtracking으로 전체 시퀀스를 구할 수 있음. (그림 c)



#### 04. Duration Prediction from Text

- phonemes의 지속 기간 분포를 따르도록 stochastic duration predictor 사용
  - 인간과 같은 음성 리듬을 생성
  - 일반적으로 max likelihood 추정을 통해 훈련
    - 음소의 duration은 연속 정규화 flow를 사용하는 이산 정수
    - 스칼라이기 때문에 직접 적용이 어려움.
  - 이를 위해 variational dequantization, variational data augmentation 적용
- 결과는 phoneme duration의 log-likelihood의 lower bound

#### 05. Adversarial Training

- decoder G에 의해 생성된 출력과 판별기 D 추가
- least-squares loss for adversarial training
- additional featurematching loss for training the generator

### 03. Model Architecture


#### 01. Posterior Encoder

- non-causal WaveNet residual blocks used in **WaveGlow, Glow-TTS**
- WaveNet residual block : dilated CNN, gated activation unit, skip-conn
- Linear projection layer : produces the mean and variance of the normal posterior distribution
- multi-speaker의 경우, residual block의 global conditioning 사용

#### 02. Prior Encoder

- consist of
  - text encoder : input text, normalizing flow
    - flexibility of the prior distribution
    - **text encoder : transformer encoder**
      - uses **relative positional representation** instead of absolute positional encoding
    - **above text encoder**
      - **text encoder, linear projection layer**를 통해 은닉 표현 추출
      - 사전 분포를 구성하는 데 사용되는 평균과 분산 생성
    - **normalizing flow**
      - stack of affine coupling layers
      - consisting of a stack of WaveNet residual blocks

#### 03. Decoder

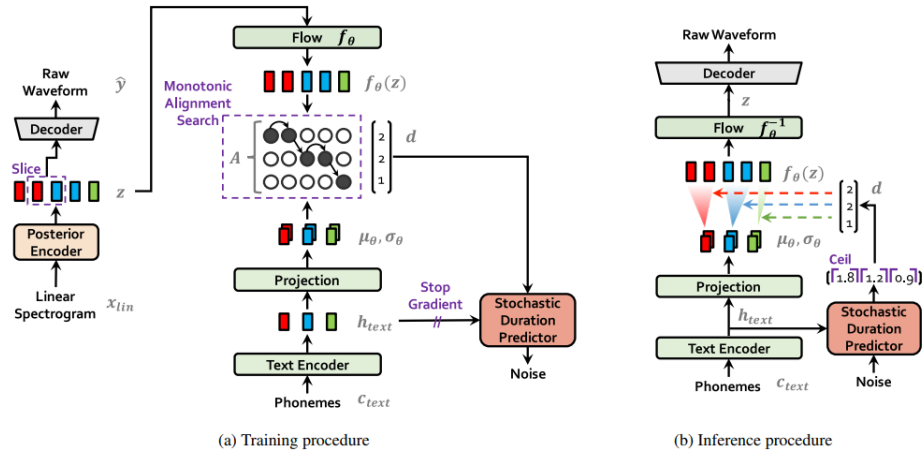
- 기본적으로 HiFi-GAN V1 generator
  - transposed conv로 구성
  - 출력 : 다른 RF 크기를 가지는 residual block 출력의 합
  - For multi-speaker setting
    - add linear layer for transforms speaker embedding
    - add input latent variables 

#### 04. Discriminator

- follow multi-period discriminator proposed in HiFi-GAN
- input waveform의 다양한 주기적인 패턴에서 작동

#### 05. Stochastic Duration Predictor(SDP)

- conditional input(text)로부터 phoneme duration의 분포 추정
- SDP의 효율적인 매개변수화를 위해 dilated, depth-separable conv를 stacking
- invertible nonlinear transformation을 취하기 위해 **neural spline flow** 적용
  - normalizing flow는 단순 기본 밀도의 역변환으로 복소 확률 밀도 모델링
    - 밀도 평가 및 샘플링을 제공
    - 그러나, 쉽게 반전 가능한 요소별 변환의 매개변수화에 의존
    - 선택에 따라 이러한 모델의 유연성 결정
  - monotonic rational-quadratic splines
    - 일반적인 affine coupling layer와 유사한 매개변수의 수를 가짐.
    - 변환 표현력 향상
- for multi-speaker, add linear layer that transforms speaker embedding



### 03. Experiments

#### 01. Datasets

- LJ Speech dataset (Ito, 2017)
  - consists of 13,100 short audio clips of a single speaker
  - total length of approximately 24 hours.
  - 16-bit PCM with a sample rate of 22kHz
  - without any manipulation
- VCTK dataset (Veaux et al., 2017)
  - approximately 44,000 short audio clips
  - 109 native English speakers with various accents.
  - 16-bit PCM with a sample rate of 44kHz
    - reduced the sample rate to 22 kHz
- split dataset to train, val, test

#### 02. Preprocessing

- Short-time Fourier transform(STFT) as input of the posterior encoder.
- FFT/window/hop size : 1024, 1024, 256
- 80 bands mel-scale spectrogram
- use International Phonetic Alphabet (IPA) sequences as input to the prior encoder.