https://arxiv.org/pdf/2105.02446

발표자 : 김동규

introduction

- 기존 모델들은 L1, L2 손실값을 사용해서 음향 feature를 재구성
  - unimodal 문제와 오버 스무딩 문제가 발생해 소리가 흐릿해지는 문제가 발생함
  - GAN을 활용해 해결하려 했지만, discriminator의 학습이 제대로 되지 않아 분류가 제대로 되지 않는 문제 발생

- diffusion probabilistic model의 reverse process 활용
  - markrov chain 기법을 활용해 노이즈에서 생성
  - ELBO의 최적화, 악보를 조건으로 ground truth 분포와 강하게 일치하는 mel-spectrogram 생성

diffusion model

Forward Process

$$q(x_{1:T}|x_0) := \prod_{t=1}^{T} q(x_t|x_{t-1})$$

$$q(x_t|x_{t-1}) := \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

$$\alpha_t := 1 - \beta_t, \quad \bar{\alpha}_t := \prod_{s=1}^{t} \alpha_s$$

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I)$$

Reverse Process

제안 기법

- naive version of DiffSinger
  - training process(학습)

    입력
    - t번째 스텝의 mel-spectrogram Mt
    - 스텝 t
    - 악보 x

    학습
    - 추가된 random noise 엡실론 세타 추측

제안 기법

- inference process(추론)

  입력

  - k번째 스텝의 가우시안 노이즈가 추가된 mel-spectrogram Mk
  - 가우시안 노이즈 MT

  T번의 denoising을 통해 결과물 생성

제안 기법

- Shallow diffusion mechanism
  - t=0일때, M은 고조파가 선명하게 보임
  - M~는 비교적 흐리게 보임(오버스무딩)
  - t=50 지점부터 M과 M~의 차이가 거의 없음

제안 기법

- Shallow diffusion mechanism

  M_T에서 M_0로 변형하는 것보다
  M_k에서 M_0로 변환하는 것이 더 효율적임

  보조 디코더(auxiliary decoder)에서 $M_\sim$를 생성하고
  악보 인코더의 출력을 조건으로 L_1 손실 학습
  학습된 내용을 바탕으로 diffusion process에서 k스텝의 샘플을 생성

제안 기법

- Bondary Prediction
  - 교차점 k의 결정 방법
  - 스텝범위 (0,T)에 대하여 M 또는 M~에서 온건지에 대해 cross entropy loss 계산
  - 모든 샘플에 대해서 BP(Mt,t) 와 BP(M~t,t)의 차이가 지정된 threshold보다 작게 되는 k'를 찾음
  - 해당 k'들의 평균을 k로 선택함

model Structure

- Encoder
  - 음소 id를 임베딩 시퀀스로 매핑하는 lyrics encoder, 해당 내용을 언어적 시퀀스로 변경하는 transformer block
  - duration 정보로 언어적 시퀀스를 mel-spectrogram의 길이로 확장하는 length regulator
  - pitch id를 pitch embedding sequence로 매핑하는 pitch encoder

- Step Embedding
  - 디퓨전 스텝 t이 ϵ_θ의 조건으로 주어짐
  - sinusoidal position embedding과 2개의 fc 레이어를 사용해서 c채널의 임베딩 Et를 얻음

model Structure

- Denoiser
  - 디노이저 $\epsilon\theta$는 Mt를 입력받아 추가된 노이즈 ϵ를 예측함
  - Et, Em을 조건으로 받음
    - 현재 모델이서는 non-casual WaveNet을 사용했음
    - (H는 denoiser의 입력 시퀀스)
  - 1. Et에 H의 행렬의 덧셈 시행
  - 2. H를 c 채널에서 2c채널로 변환하는 non casual convolution network
  - 3. Em를 2c채널로 변환하는 1*1 convolution layer
  - 4. 입력과 조건들의 정보를 합치는 gate unit
  - 5. 합쳐진 2c 채널의 정보를 각 c채널의 2개의 branch로 나누는 residual block
- 최종적으로 여러 계층의 feature를 합쳐서 최종 예측 생성

model Structure

- Auxiliary Decoder
  - 간단한 mel-spectrogram 디코더
  - feed forward transformer block
  - fastspeech2의 mel-spectrogram 디코더와 같이 M~를 생성함

- Boundary Predictor
  - Et를 제공하기 위한 step embedding을 시행
  - Et와 스텝 t를 받아서 mel-spectrogram을 입력받아 Mt인지 M~t인지 분류함

Experiment

- dataset
  - PopCS(중국어 팝송), 샘플레이트24kHz, 16비트 샘플링
  - DeepSinger로 문장조각으로 자르고, 노래 조각과 가사 사이를 음소레벨로 맞춰 MFA 모델을 학습
  - pitch는 waveform에서 parselmouth로 추출

- training
  - Warmup stage
    - 악보 인코더와 auxiliary decoder를 16만 스텝동안 학습, 이후 생성된 M~t를 사용해 boundary predictor를 3만 스텝동안 학습시켜 k추출
  - Main stage
    - Diffsinger를 Training process로 수렴할 때까지 학습(16만 스텝)

Experiment

- DiffSinger, Gan-singer가 고조파 사이에서 더 섬세하게 표현되어 있음
- MOS값도 shallow diffusion mechanism 사용 여부와 관계없이 높은 성능을 보여줌
- 또한, shallow diffusion mechanism을 사용했을 때, 초당 0.191초 vs 0.348초의 속도로
- shallow diffusion mechanism을 사용한 경우에 약 45% 빨라짐

감사합니다