VALL-E

분류 survey

```
Wang, Chengyi, et al. "Neural codec language models are zero-shot text to speech synthesizers." arXiv preprint arXiv:2301.02111 (2023).
```

- TTS를 위한 언어 모델링 기법(called VALL-E)
- 3초의 음향으로 고품질 맞춤형 음성 합성
  - pipeline : phoneme → discrete code → waveform
  - mel-spectrogram 대체
- 보지 못한 화자에 대한 콘텐츠 생성
  - 텍스트 문장, 등록된 3초 음성, transcription만 제공
  - transcription를 주어진 문장의 시퀀스에 프롬프트로 추가
  - 등록된 음성의 첫 layer 음향 토큰을 prefix로 사용
  - 프롬프트와 음향 prefix를 사용하여 화자의 음성을 복제

01. Introduction

- 지난 수년 간 end-to-end 모델링을 통해 발전
- 현재 cascaded TTS
  - mel-spectrogram을 사용하는 acoustic model + vocoder
  - 단일 또는 여러 화자의 고품질 음성을 합성하려면 고품질 데이터 필요
    - 크롤링 데이터는 성능 저하 야기
  - train dataset이 상대적으로 적어 일반화가 어려움
- zero-shot TTS
  - 보이지 않는 화자의 경우 유사성, 자연성 극적으로 감소
  - 이를 위해 speaker adaptation, speaker encoding 방법 활용
    - speaker adaptation : 몇 개의 복제 샘플을 사용하여 다중 화자 생성 모델을 fine-tuning
    - speaker encoding : 별도의 모델을 훈련하여 새로운 화자를 직접 추론
- VALL-E
  - 개인화 된 음성 합성(zero-shot TTS)
    - 3초 간 녹음된 음성과 프롬프트의 토콘에서 제한된 음향 토큰 생성
    - 생성된 토큰은 코덱 디코더에서 최종 파형을 합성하는 데 사용
  - 개별 음향 토큰을 사용하면 conditional codec language modeling으로 처리 가능
    - GPT 같은 모델을 TTS 작업에 활용 가능
    - 샘플링을 통해 다양한 합성 결과 생성 가능
  - 영어 음성으로 구성된 corpus LibriLight를 사용하여 훈련
    - 원본 데이터는 음성이므로, SER 모델을 사용하여 transcription 생성
  - 동일한 텍스트 및 타겟 speaker를 사용하여 다양한 출력 합성 가능
  - 음향 환경과 감정 유지 가능

02. Related Work

01. zero-shot TTS

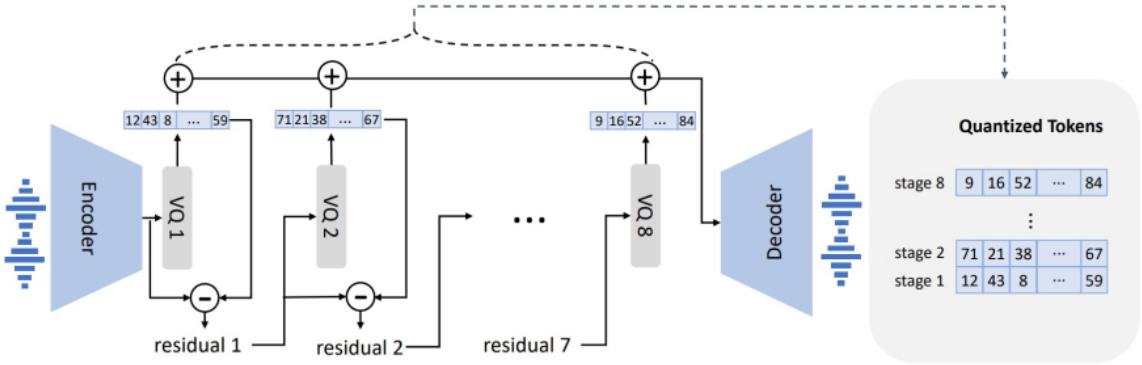
- Current TTS method
  - Cascaded TTS : mel-spectrogram을 중간 표현으로 사용하는 음향 모델과 vocoder
  - end-to-end TTS : 음향 모델과 보코더를 공동으로 최적화
- Speaker encoding based TTS
  - speaker encoder(*) + TTS componet
    - *은 화자 검증 작업에 대해 pre-trained 될 수 있음.
- Diffusion model based TTS
  - 예측된 노이즈를 점진적으로 변환하고 텍스트 입력에 맞춰 오디오 생성

02. Spoken generative pre-trained models

- speech understanding
- speech-to-speech generation
  - 텍스트 없는 환경에서 어떻게 음성을 합성하는가?
  - HuBERT 기반 음성 합성 제안된 연구[Hsu et al., 2021]
  - 오디오 코덱을 사용해 음성합성[AudioLM]
- apply pre-training to the neural TTS
  - mel-spectrogram 예측을 통해 TTS에서 음성 디코더 사전 학습[Chung et al. 2018]
  - 레이블이 지정되지 않은 음성 및 텍스트 데이터를 이용해 TTS 모델 전체 사전 훈련[Ao et al. 2022]
  - 레이블이 지정되지 않은 음성을 개별 토큰으로 양자화하고 토큰-음성 시퀀스를 사용해 학습[VQVAE]
    - 사전 학습 모델이 fine-tuning을 위해 소량의 실제 데이터만 필요
  - mel-spectrogram에 대한 마스크 및 재구성 제안[Baiet al. 2022]

03. Background: Speech Quantization

- 사전학습된 neural audio codec 모델인 Encodec[Dfossez et al., 2022] 토크나이저 사용
  - 입/출력 모두 가변 비트 전송률(24kHz)
  - encoder
    - 24kHz 입력 파형에 대해 75Hz의 임베딩 생성
    - 즉, 샘플링 속도 320배 감소
  - 각 임베딩은 residual vector quantization(RVQ)에 의해 모델링
    - 1024개의 entry가 존재하는 8개의 계층 선택



- discrete representation matrix : 750 X 8 entries
  - 10-second waveform
  - 750 = (24,000*10) / 320
- 높은 비트 전송률은 더 많은 양자화기와 더 나은 재구성 품질
- 모든 양자화기의 개별 코드를 사용하여 decoder는 실수 값 임베딩을 생성하고 24kHz에서 파형 재구성

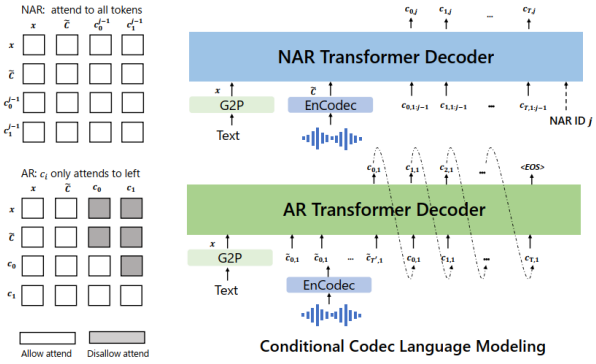
4. Proposed

01. Problem Formulation

- input x(audio sample), y(SER 모델로부터 생성된 transcription)
- pre-trained neural codec 모델을 통해 각 오디오 샘플을 개별 음향 코드로 인코딩
- 양자화 후 neural codec decoder는 파형을 재구성할 수 있음.(y_pred)
- **zero-shot TTS는 보이지 않는 화자에 대해 고품질 음성 합성 모델 필요**

02. conditional codec language modeling

- 어쿠스틱 코드 행렬(C)을 생성하기 위한 언어 모델 학습
  - autogressive(AR) decoder-only language model (first quantizer)
    - conditioned on the phoneme sequence x and the acoustic prompt C
    - 음향 시퀀스 길이 예측에 대한 유연성을 갖춰 생성된 음성의 속도와 녹음된 음성의 속도를 맞추는데 유용함.
  - Non-autograssive language model(second~last quantizer)
    - 화자 ID를 제한하기 위한 음향 프롬프트 행렬
    - 첫 번째 단계의 출력 길이를 따르므로, 시간 복잡도 개선 가능
  - 두 LM의 조합은 음성 품질과 추론 속도 간의 균형 제공
  - **각 양자화기는 이전 양자화기의 잔차를 모델링하도록 훈련**



NAR : Non-autogressive, AR : Autogressive

- 언어 모델이 시퀀스와 프롬프트에서 내용과 화자 정보를 추출하는 방법 학습
- **추론 중에 보이지 않는 화자의 시퀀스와 3초 동안 등록된 녹음이 주어지면, 사전 훈련된 언어 모델에서 해당 내용과 화자의 음성이 포함된 행렬 추정**
- 이후, neural codec decoder가 고품질 음성 합성

05. Experiment

- LibriLight Dataset
- AR model and the NAR model : same transformer architecture
- average length of the waveform in LibriLight is 60 seconds.
  - randomly crop the waveform to a random length 10 ~ 20 seconds.
- AdamW optimizer

baseline

- YourTTS
- VCTK
- LibriTTS
- TTS-Portuguese

Automatic metric

- speaker verification model, **WavLM-TDNN**
- to evaluate the speaker similarity between prompt and synthesized speech.