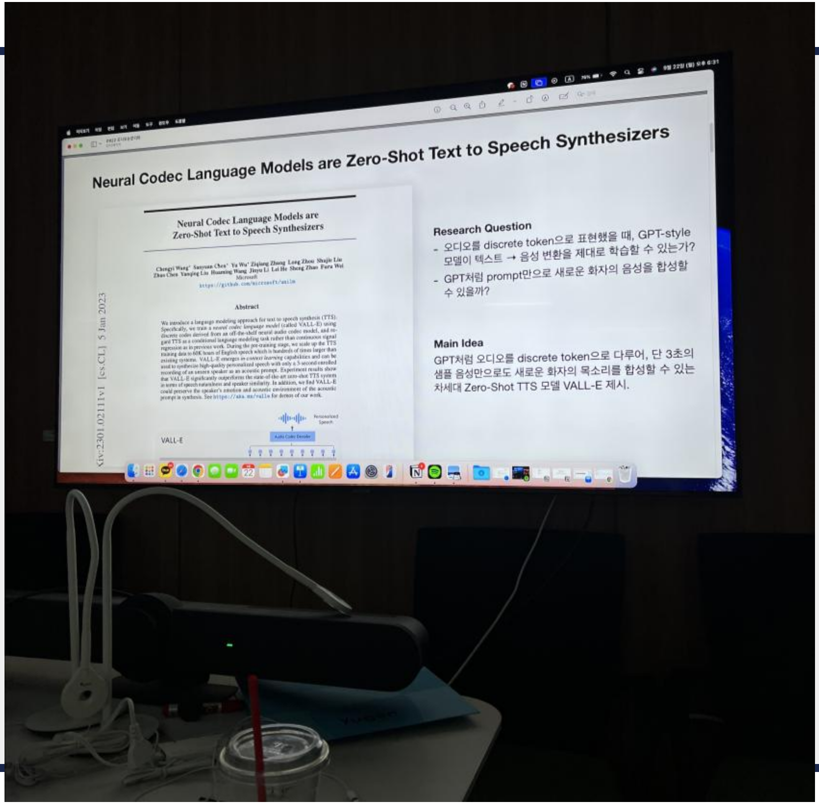
CUAI 딥러닝 논문 리뷰 스터디 (음성 처리)

2025.09.30

발표자 : 김동건

스터디원 소개 및 만남 인증



스터디원 1 : 김동건

스터디원 2 : 양희원

스터디원 3 : 이나현

스터디 방식

월요일 15:00 – 16:30
대면 스터디 진행

각자 공부할 논문 선정 후 간단하게 PPT 제작 후
스터디에서 발표하며 공부한 논문의 내용을 공유

스터디 주제

9월 22일 월요일 15:00 – 16:30 1차 스터디 진행함
각자 공부할 논문 선정 후 간단한 설명 완료.
10월 13일까지 공부 완료하여 10월 13일 2차 스터디에 공부 내용 공유할 예정.

공부할 논문

- **김동건:** ProMode: A Speech Prosody Model Conditioned on Acoustic and Textual Inputs
- **양희원:** Long-Form Speech Generation with Spoken Language Model
- **이나현:** Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

스터디 내용

Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

ProMode: A Speech Prosody Model Conditioned on Acoustic and Textual Inputs

Eray Eren[1], *Qingju Liu*[2], *Hyeongwoo Kim*[3], *Pablo Garrido*[2], *Abeer Alwan*[1]

[1]Dept. of Electrical and Computer Engineering, University of California, Los Angeles, USA
[2]Flawless AI, USA
[3]Imperial College London, UK

erayeren@g.ucla.edu, qingju.liu@flawlessai.com, hyeongwoo.kim@imperial.ac.uk, pablo.garrido@flawlessai.com, alwan@ee.ucla.edu

Abstract

Prosody conveys rich emotional and semantic information of the speech signal as well as individual idiosyncrasies. We propose a stand-alone model that maps text-to-prosodic features such as F0 and energy and can be used in downstream tasks such as TTS. The ProMode encoder takes as input acoustic features and time-aligned textual content, both are partially masked, and obtains a fixed-length latent prosodic embedding. The decoder predicts acoustics in the masked region using both the encoded prosody input and unmasked textual content. Trained on the GigaSpeech dataset, we compare our method with state-of-the-art style encoders. For F0 and energy predictions, we show consistent improvements for our model at different levels of granularity. We also integrate these predicted prosodic features into a TTS system and conduct perceptual tests, which show higher prosody preference compared to the baselines, demonstrating the model's potential in tasks where prosody modeling is important.
Index Terms: speech prosody, speech synthesis, pitch, energy, Perceiver IO

1. Introduction

Speech prosody is manifested by changes in pitch (fundamental frequency), energy, and phoneme or word durations. These changes in pitch/energy/duration can help convey the meaning of a sentence, and emotion of the speaker. Hence, modeling prosody accurately can, for example, make synthetic speech

pitch/duration predictor in [9], where a prompt speech encoder implicitly models the prosodic information. Instead of predicting prosodic features of pitch or energy, MegaTTS [10] applies an autoregressive (AR) prosody large language model (LLM) to the quantized prosody codes aggregated at the phoneme level.

The aforementioned (sub-)prosody models do facilitate their related tasks, either SER or TTS; however, they are formulated differently, and are not task-agnostic prosody models. An exception is [11], which is a prosody model that also targets TTS applications. The authors proposed to make use of several speech prosody inputs such as the normalized cross-correlation function, F0, energy, and Mel-spectrogram below 500 Hz. After processing the speech waveform, they utilized a BERT encoder to process these prosodic inputs to generate discrete prosody codes. They used these prosody codes to train and improve an existing TTS system. However, their prosody model lacks textual information, which can also contain prosodic information. Their method also requires training a TTS system to produce prosody-aware speech from text.

To address these limitations, we propose a zero-shot (for both prosody and speaker) and stand-alone speech prosody model utilizing the Perceiver IO structure [12], hereafter referred to as ProMode. The model's input consists of masked acoustic and textual features, and the model encodes them into latent prosody embeddings via a Perceiver-based encoder. These prosody embeddings reconstruct prosody of the masked region given contexts in the unmasked regions, with two decoders, one conditional and one unconditional to the unmasked

TTS 합성 시 음성의 자연스러움을 결정하는 Prosody에 대한 임베딩을 미리 학습하여, SER, TTS 등 Downstream Task의 종류에 상관없이 Prosody의 정확도를 높이는 모듈에 대한 내용

스터디 내용

Long-Form Speech Generation with Spoken Language Models

Se Jin Park * [1,2] Julian Salazar * [1] Aren Jansen [1] Keisuke Kinoshita [1] Yong Man Ro [2] RJ Skerry-Ryan [1]

Abstract

We consider the generative modeling of speech over multiple minutes, a requirement for long-form multimedia generation and audio-native voice assistants. However, textless spoken language models struggle to generate plausible speech past tens of seconds, due to high temporal resolution of speech tokens causing loss of coherence, architectural issues with long-sequence training or extrapolation, and memory costs at inference time. From these considerations we derive **SpeechSSM**, the first speech language model family to learn from and sample long-form spoken audio (e.g., 16 minutes of read or extemporaneous speech) in a single decoding session without text intermediates. SpeechSSMs leverage recent advances in linear-time sequence modeling to greatly surpass current Transformer spoken LMs in coherence and efficiency on multi-minute generations while still matching them at the utterance level. As we found current spoken language evaluations uninformative, especially in this new long-form setting, we also introduce: **LibriSpeech-Long**, a benchmark for long-form speech evaluation; new embedding-based and LLM-judged metrics; and quality measurements over length and time. Speech samples, the LibriSpeech-Long dataset, and any future code or model releases can be found at https://google.github.io/tacotron/publications/speechssm/.

1. Introduction

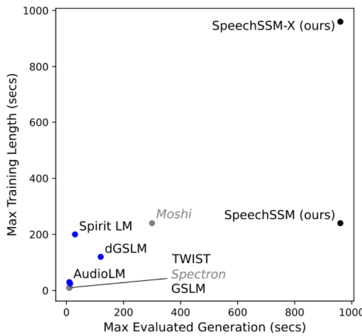


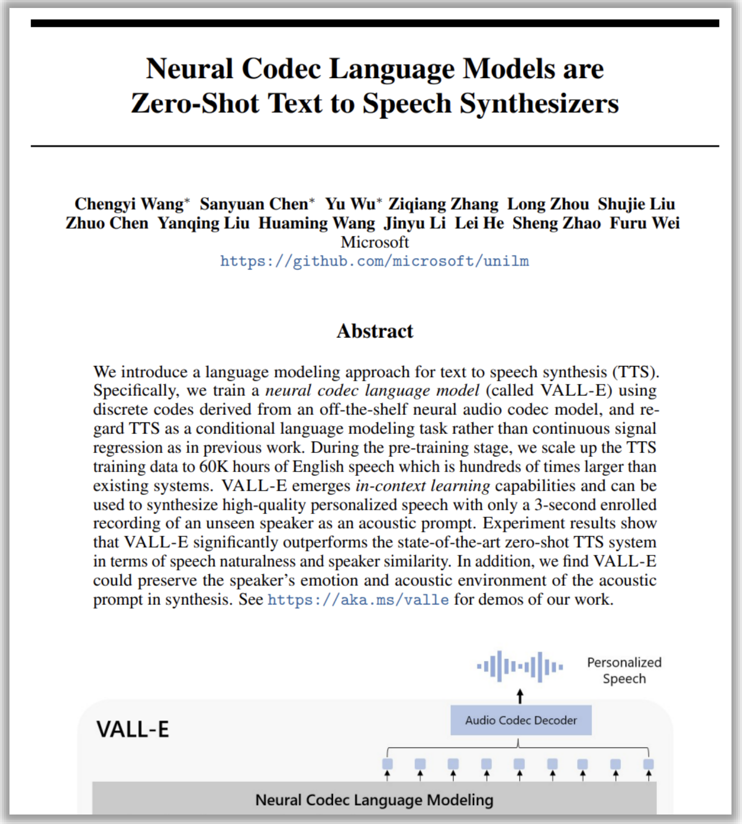
Figure 1. Maximum sequence lengths considered by various spoken LMs. *Italicized* models used text intermediates at generation time. Our models can generate indefinitely due to their constant memory footprint, but we cap our evaluations to 16 minutes.

its paralinguistic aspects, such as prosody (Kharitonov et al., 2022) and turn-taking (Nguyen et al., 2023b). These capabilities make speech-native language models (LMs) promising for applications like media understanding and co-creation, audio-native voice assistants, and textless NLP. However, real-world use-cases of spoken LMs require the ability to both understand and generate long-form speech. For example, voice interactions can last many minutes, requiring a model to maintain a growing conversational history in real time, and expressive media like audiobooks and podcasts can require semantic, paralinguistic, and speaker coherence

ICML 2025 Oral

수십분의 긴 음성 생성을 목표로, SSM을
활용해 텍스트를 거치지 않고 직접
음성을 생성하는 모델을 제안함.

스터디 내용



Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

Chengyi Wang* Sanyuan Chen* Yu Wu* Ziqiang Zhang Long Zhou Shujie Liu
Zhuo Chen Yanqing Liu Huaming Wang Jinyu Li Lei He Sheng Zhao Furu Wei
Microsoft
https://github.com/microsoft/unilm

Abstract

We introduce a language modeling approach for text to speech synthesis (TTS). Specifically, we train a *neural codec language model* (called VALL-E) using discrete codes derived from an off-the-shelf neural audio codec model, and regard TTS as a conditional language modeling task rather than continuous signal regression as in previous work. During the pre-training stage, we scale up the TTS training data to 60K hours of English speech which is hundreds of times larger than existing systems. VALL-E emerges *in-context learning* capabilities and can be used to synthesize high-quality personalized speech with only a 3-second enrolled recording of an unseen speaker as an acoustic prompt. Experiment results show that VALL-E significantly outperforms the state-of-the-art zero-shot TTS system in terms of speech naturalness and speaker similarity. In addition, we find VALL-E could preserve the speaker's emotion and acoustic environment of the acoustic prompt in synthesis. See https://aka.ms/valle for demos of our work.

Microsoft 2023

텍스트를 음성 코드로 바로 변환해 Zero-Shot TTS 를 가능하게 하고, 짧은 음성 샘플만으로 화자의 음색을 모방할 수 있고, 별도 학습 없이도 새로운 화자와 문장을 자연스럽게 합성할 수 있는 VALL-E를 제안함