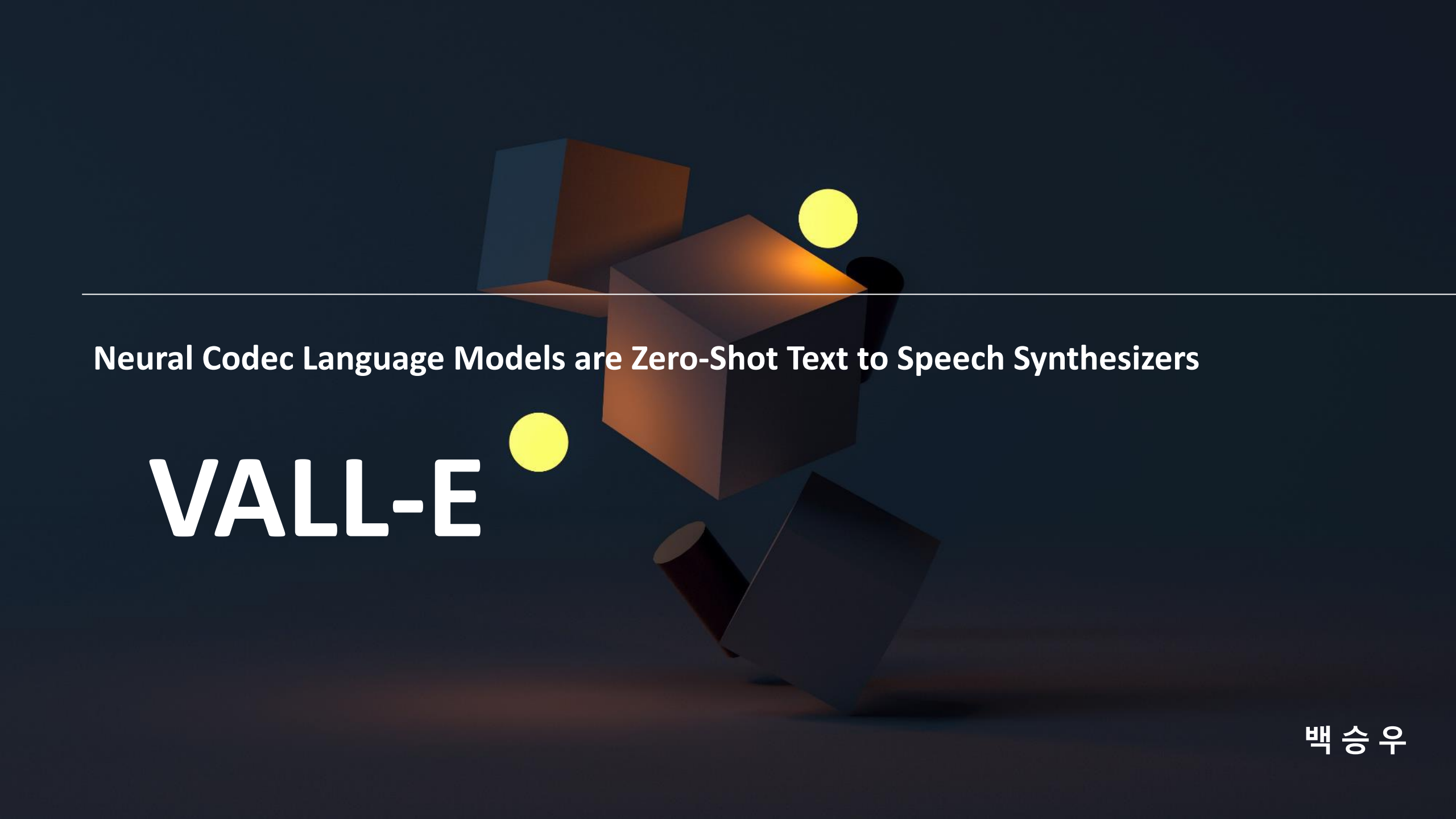
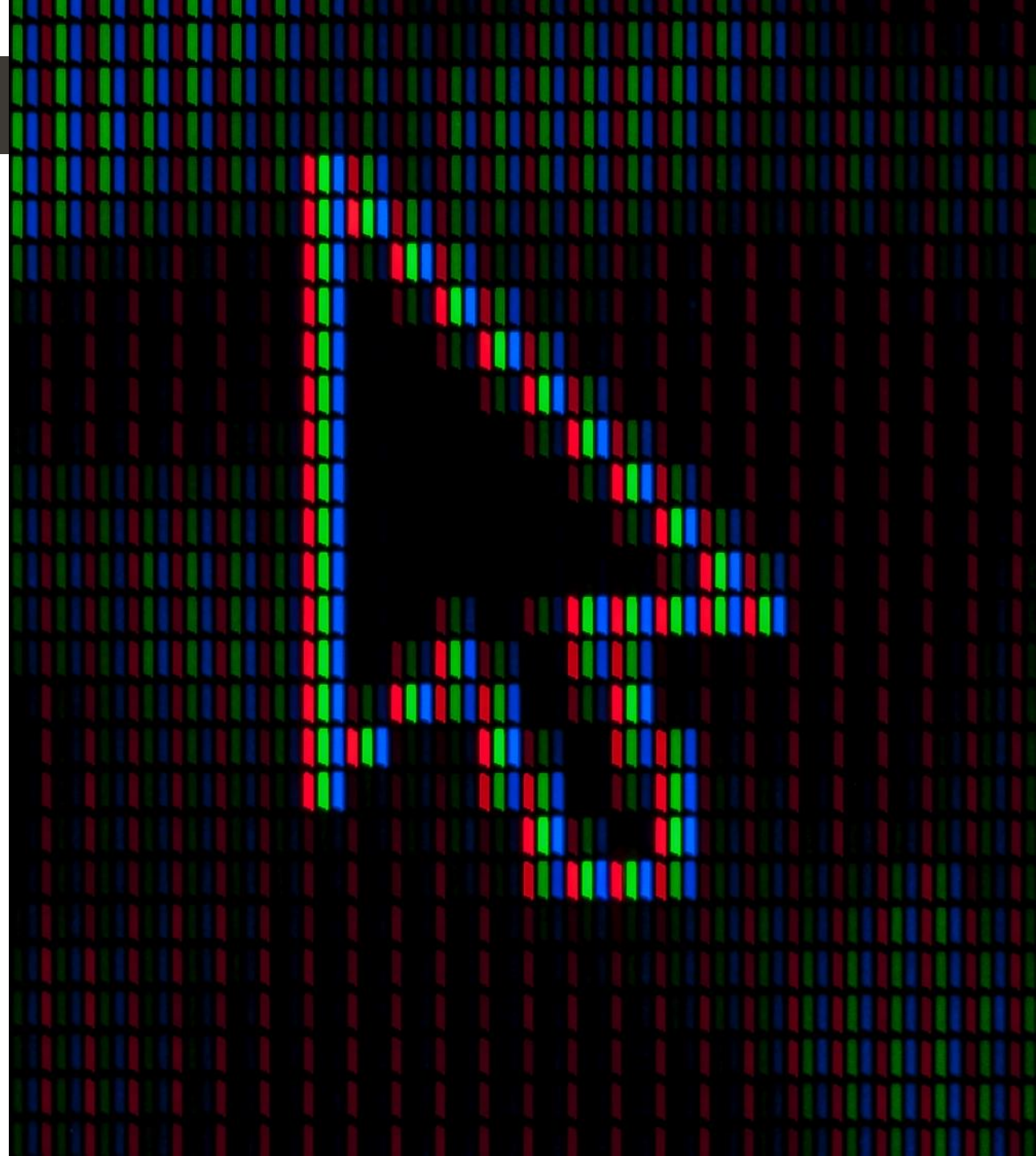


Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers

VALL-E

백 승 우

목차 a table of contents



Part 1 Abstract & Introduction

VALL-E

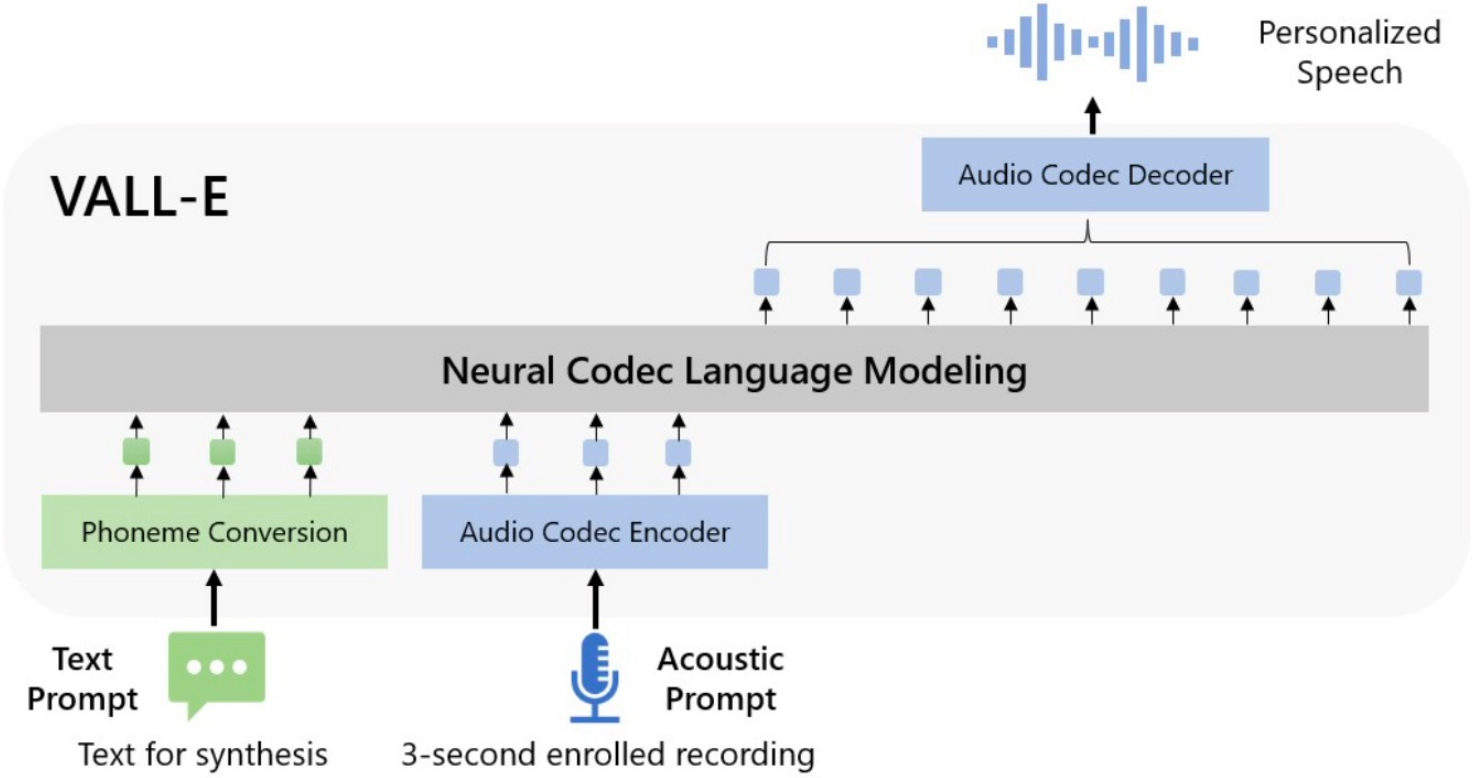
일반적인 음성 합성 모델 :

파형 조작을 통한 음성 합성하는 방법

VALL-E :

텍스트와 음향 프롬프트에서

개별 음성 코덱 코드를 생성하는 구조



ⓒSaebyeol Yu. Saebyeol's PowerPoint

Current TTS System vs VALL-E

Table 1: A comparison between VALL-E and current cascaded TTS systems.

	Current Systems	**VALL-E**
Intermediate representation	mel spectrogram	audio codec code
Objective function	continuous signal regression	language model
Training data	$\leq$ 600 hours	60K hours
In-context learning	✗	✓

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Current TTS System

- 고급 TTS 시스템의 경우, 녹음 스튜디오에서 고품질의 깨끗한 데이터를 요구

- 인터넷에서 검색된 대규모 데이터는 요구 사항을 충족할 수 없으며 항상 성능 저하를 초래

- 훈련 데이터가 상대적으로 작기 때문에 현재의 TTS 시스템은 여전히 일반화가 잘 되지 않음

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Current TTS System

- 제로샷 시나리오에서 보이지 않는 화자에 대한 화자 유사성과 음성 자연성은 극적으로 감소

- 이러한 문제를 해결하기 위해 기존 연구는 스피커 적응과 스피커 인코딩 방법을 활용하여 추가 미세 조정, 복잡한 사전 설계 기능 또는 무거운 구조 엔지니어링이 필요하다.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

VALL-E

- 텍스트 합성 분야의 성공에 동기부여 받아 가능한 한 크고 다양한 데이터로 모델을 훈련 (16GB - > 160GB -> 570GB -> 1TB)

- 크고 다양한 다중 화자 음성 데이터를 활용하는 최초의 언어 모델 기반 TTS 프레임워크

ⓒSaebyeol Yu. Saebyeol's PowerPoint

VALL-E

- 개인화된 음성(예: 제로샷 TTS)을 합성하기 위해, VALL-E는 3초 등록된 녹음의 음향 토큰과 스피커와 콘텐츠 정보를 각각 제한하는 음소 프롬프트에 따라 해당하는 음향 토큰을 생성

- 생성된 음향 토큰은 해당 신경 코덱 디코더와 최종 파형을 합성하는 데 사용

- 오디오 코덱 모델에서 파생된 이산 음향 토큰은 TTS를 조건부 코덱 언어 모델링으로 처리할 수 있게 하며, GPT를 TTS 작업에 활용

ⓒSaebyeol Yu. Saebyeol's PowerPoint

VALL-E

- 7000명 이상의 화자가 있는 60,000시간의 영어 스피치로 구성된 말뭉치인 LibriLight로 훈련

- 더 많은 잡음이 있는 음성과 부정확한 녹음을 포함하지만 다양한 화자와 운율을 제공
  (소음에 강하고, 큰 데이터를 활용하여, 잘 일반화시킬 것이라고 함)

ⓒSaebyeol Yu. Saebyeol's PowerPoint

VALL-E

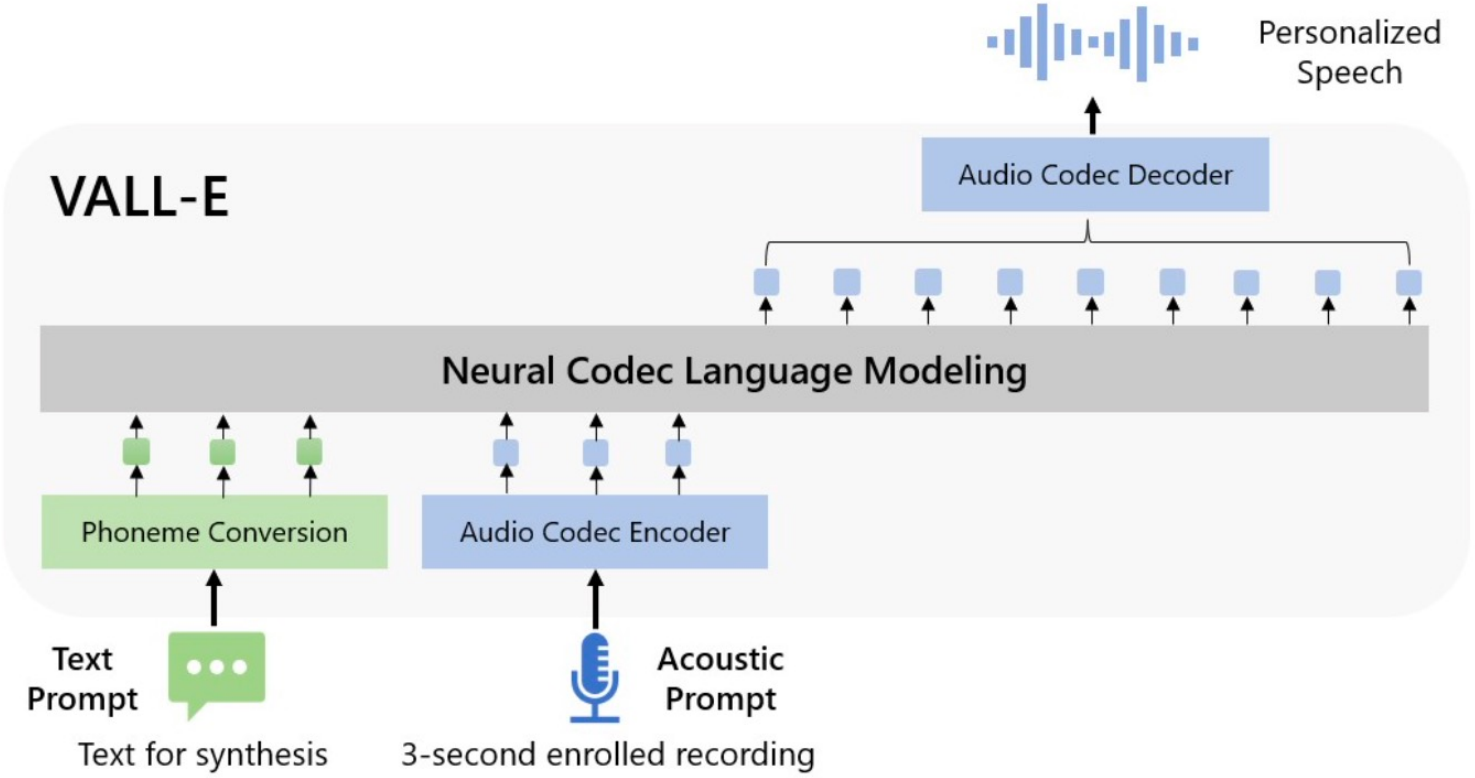
- 동일한 입력 텍스트로 음향 환경과 프롬프트의 화자 감정을 유지하는 등의 다양한 출력을 제공

- 제로샷 시나리오에서 프롬프트를 표시하여 스피커 유사성이 높은 자연 음성을 합성하는지 확인

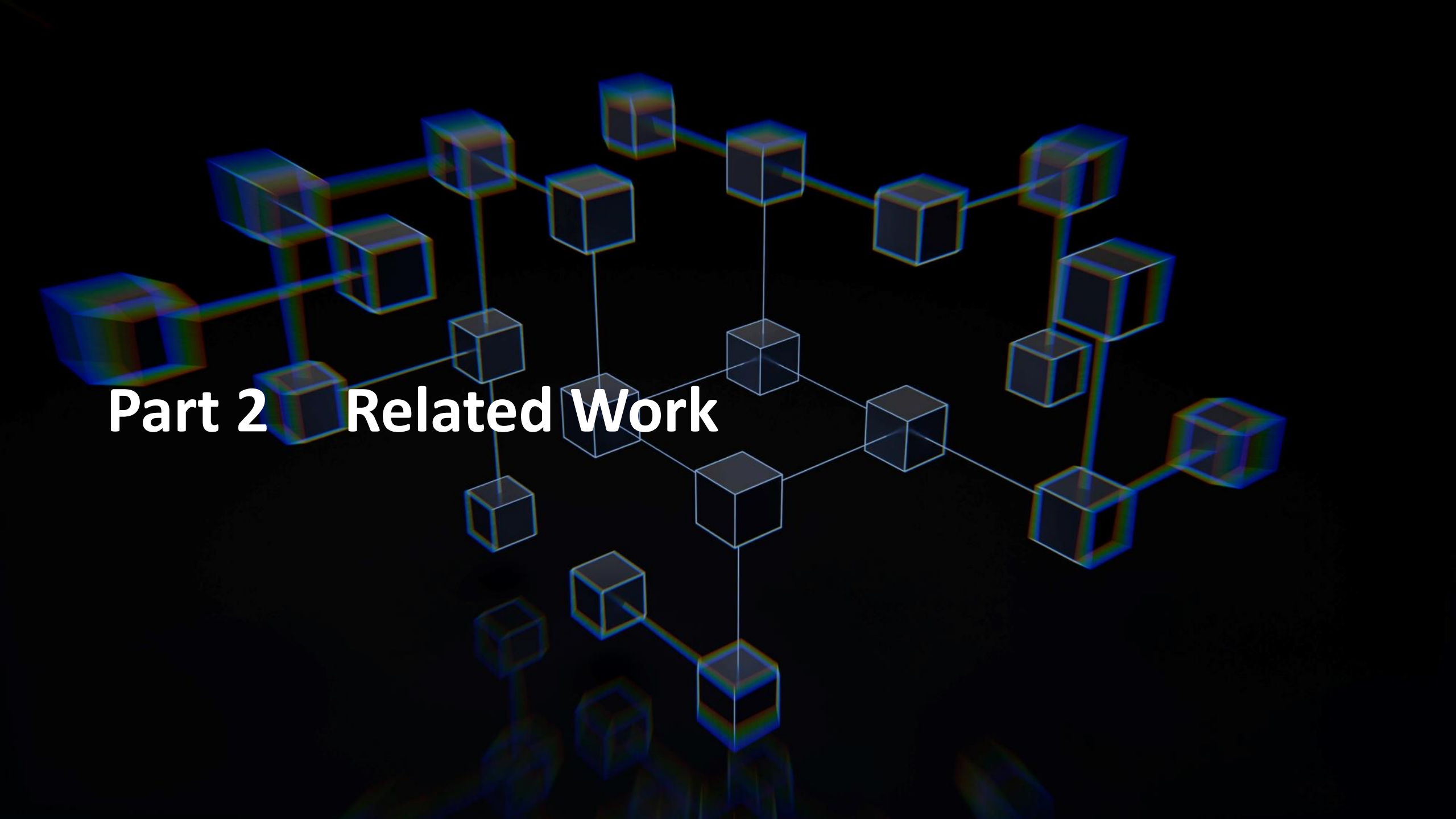
ⓒSaebyeol Yu. Saebyeol's PowerPoint

VALL-E

https://valle-demo.github.io/



ⓒSaebyeol Yu. Saebyeol's PowerPoint

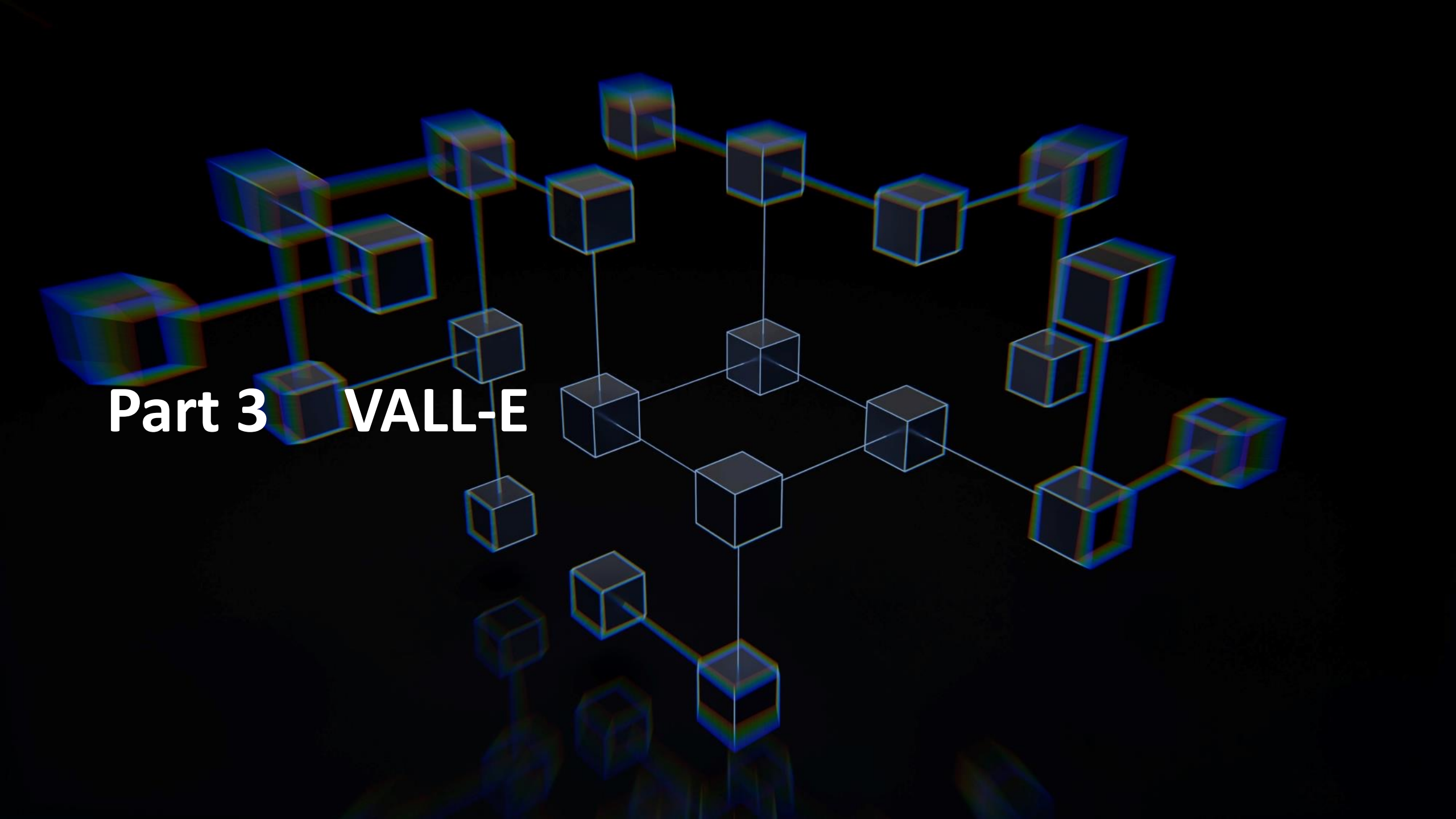


Part 2 Related Work

Zero-Shot TTS

Spoken generative pre-trained models

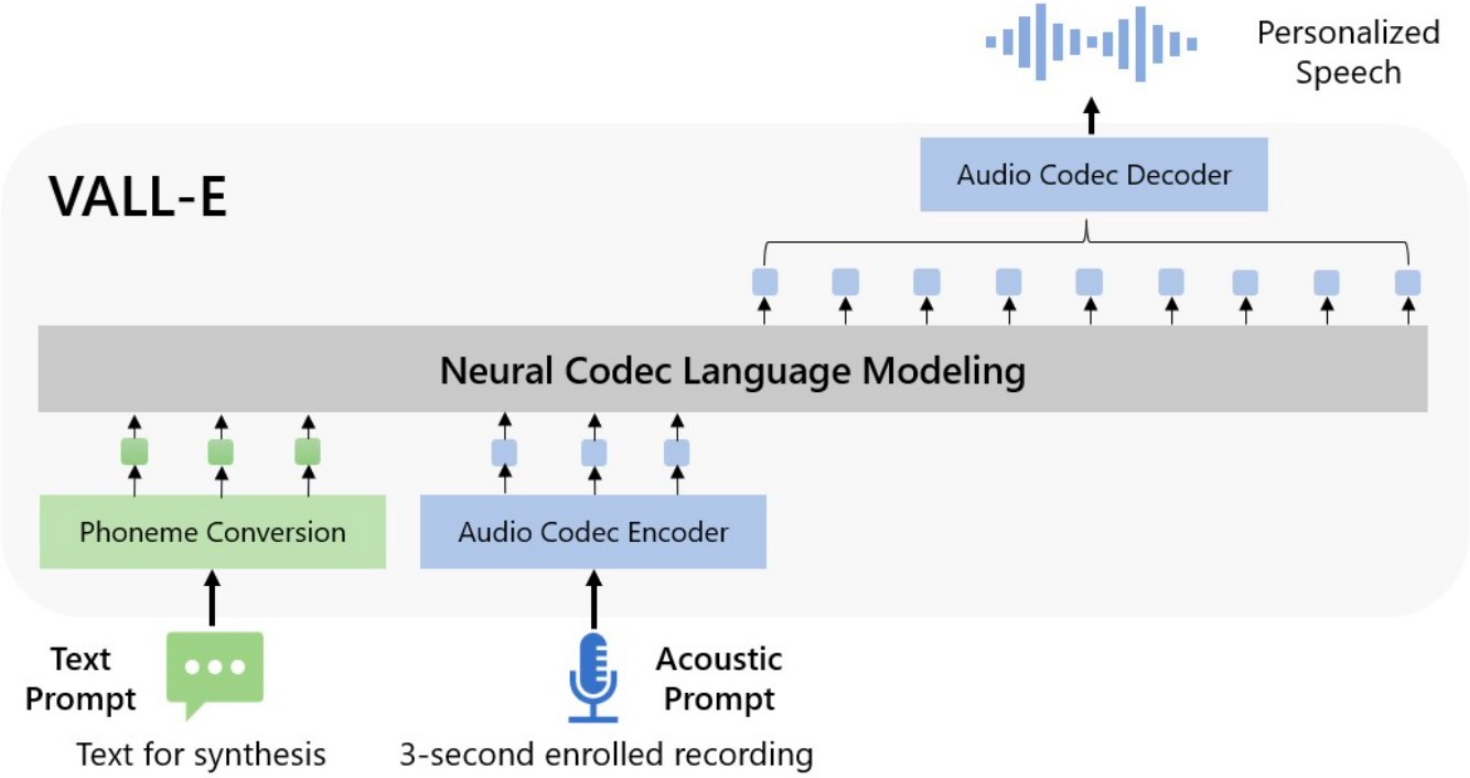
ⓒSaebyeol Yu. Saebyeol's PowerPoint



Part 3 VALL-E

Part 3

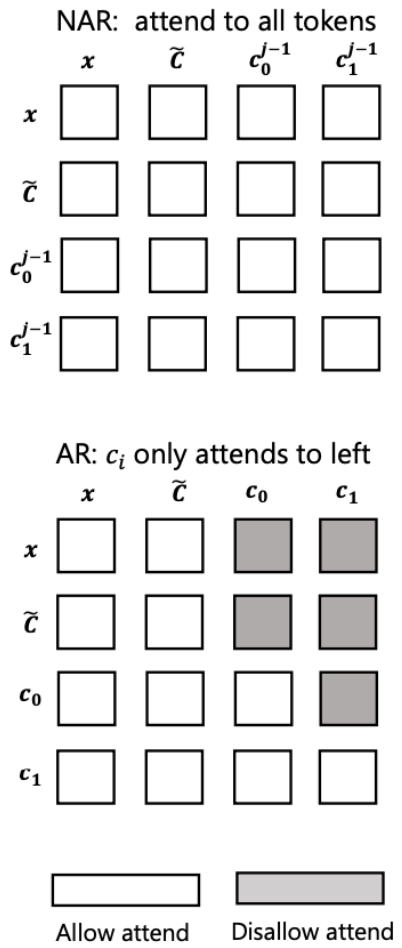
VALL-E

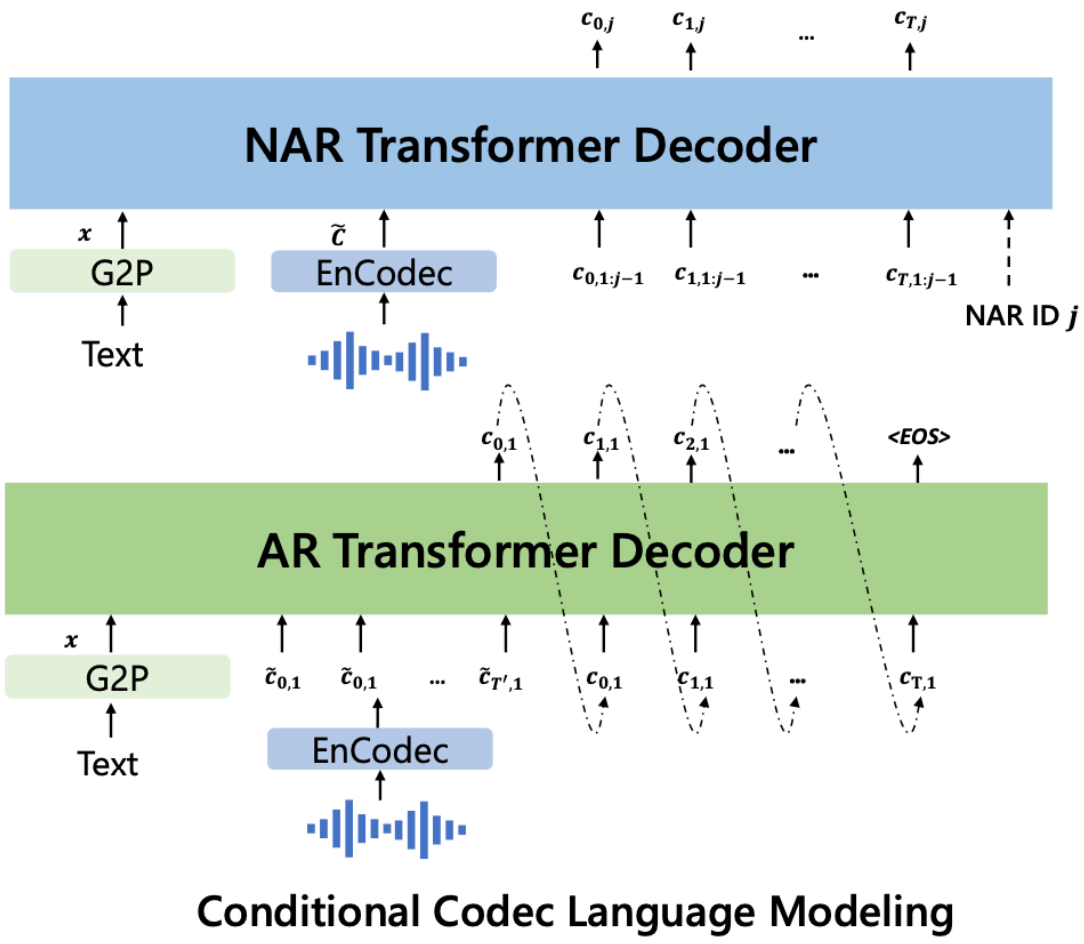


ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 3

VALL-E





ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 4 Experiment

Experiment - Setup

Dataset

- The EnCodec model is used to generate the acoustic code matrix for the 60K hours of data.

Baseline

- SOTA zero-shot TTS model YourTTS

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Model

- Both the AR model and the NAR model have the same transformer architecture with 12 layers
- 16 attention heads
- Embedding dimension of 1024
- Feed-forward layer dimension of 4096
- Dropout of 0.1
- The average length of the waveform in LibriLight is 60 seconds
- Random length between 10 seconds and 20 seconds
- 16 NVIDIA TESLA V100 32GB GPUs
- Batch size of 6k acoustic tokens per GPU for 800k steps.

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Automatic metrics

- We employ the SOTA speaker verification model
- We also evaluate the synthesis robustness of our model
- We perform ASR on the generated audio and calculate the WER with respect to the original transcriptions

Human evaluation

- 12 and 6 native speakers are invited as CMOS and SMOS contributors
- CMOS is an indicator of speech naturalness, and SMOS measures whether the speech is similar to the original speaker’s voice

ⓒSaebyeol Yu. Saebyeol’s PowerPoint

Part 5

Experiment - Result

model	WER	SPK
GroundTruth	2.2	0.754
Speech-to-Speech Systems		
GSLM	12.4	0.126
AudioLM*	6.0	-
TTS Systems		
YourTTS	7.7	0.337
VALL-E	5.9	**0.580**
VALL-E-continual	**3.8**	0.508

	SMOS	CMOS (v.s. VALL-E)
YourTTS	$3.45_{\pm 0.09}$	-0.12
VALL-E	$4.38_{\pm 0.10}$	0.00
GroundTruth	$4.5_{\pm 0.10}$	+0.17

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 5

Experiment - Result

Table 5: Ablation study of the AR model.

	WER	SPK
VALL-E	**5.9**	**0.585**
w/o acoustic prompt	**5.9**	0.236

Table 6: Automatic evaluation of speaker similarity with 108 speakers on VCTK. *YourTTS has observed 97 speakers during training, while VALL-E observed none of them.

	3s prompt	5s prompt	10s prompt
108 full speakers			
YourTTS*	0.357	0.377	0.394
VALL-E	0.382	0.423	**0.484**
GroundTruth	0.546	0.591	0.620
11 unseen speakers			
YourTTS	0.331	0.337	0.344
VALL-E	0.389	0.380	**0.414**
GroundTruth	0.528	0.556	0.586

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Part 5 Limitations and Future Work

Synthesis robustness

- 일부 단어가 음성 합성에서 불분명하거나 누락되거나 중복될 수 있음

- phoneme-to-acoustic language part가 무질서한 정렬이 존재하고, 문제 해결에 제약이 없는 자기 회귀 모델

- (바닐라 트랜스포머 기반 TTS에서도 관찰)
  비자동 회귀 모델을 적용하거나 모델링에서 주의 메커니즘을 수정하여 해결

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Data coverage

- 여전히 모든 사람들의 목소리, 특히 억양을 커버할 수 없음

- LibriLight의 대부분이 책 읽듯이 말하는 데이터 세트로, 말하기 스타일의 다양성은 충분하지 않음

- 훈련 데이터의 확장과 더불어, 제로샷 TTS 작업은 모델 및 데이터 스케일업을 사용해 거의 해결될 것임

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Model Structure

- 현재, 두 가지의 다른 양자화기(아날로그->디지털)를 이용한 예측 모델 사용

- 큰 보편적 모델로 예측하는 것

- NAR 모델을 사용하여 프레임워크에서 모델 추론 속도를 높이는 것

ⓒSaebyeol Yu. Saebyeol's PowerPoint

Broader impacts

- 음성 식별을 통해 속이거나 사칭하는 등 모델을 오용할 때 잠재적 위험을 수반할 수 있다.

- 오디오 클립이 VALL-E에 의해 합성되었는지 여부를 구별하는 탐지 모델을 구축

ⓒSaebyeol Yu. Saebyeol's PowerPoint