NeurIPS2024

Paralinguistics-Aware Speech-Empowered Large Language Models for Natural Conversation_분석

Heeseung Kim1 Soonshin Seo2 Kyeongseok Jeong2 Ohsung Kwon2 Soyoon Kim2, Jungwhan Kim2 Jaehong Lee2, Eunwoo Song2,4 Myungwoo Oh2 Jung-Woo Ha2,3Sungroh Yoon1,4,5* Kang Min Yoo2,3,4

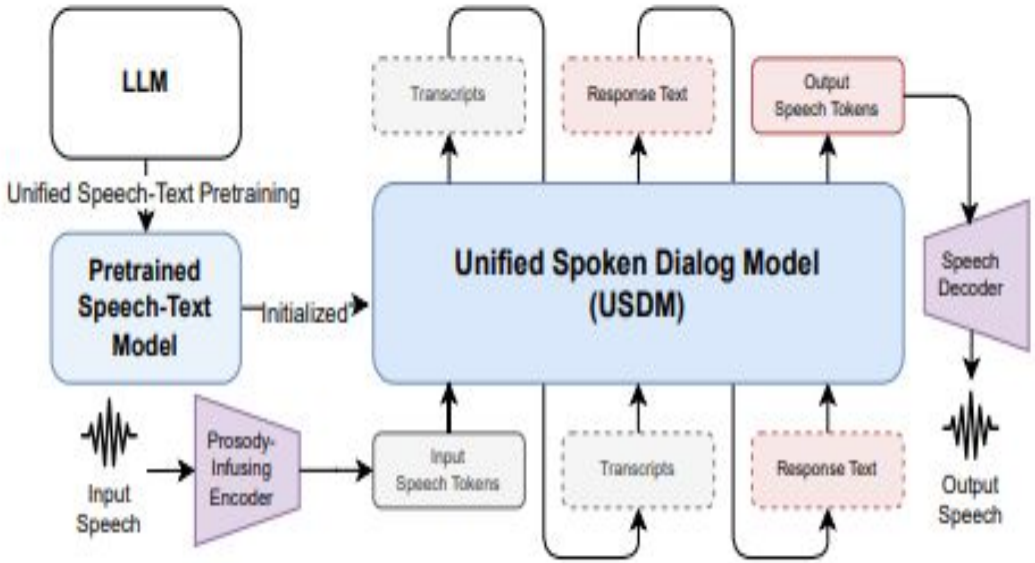
1Data Science and AI Lab, Department of ECE, Seoul National University, 2NAVER Cloud, 3NAVER AI Lab, 4Artificial Intelligence Institute, Seoul National University, 5ASRI, INMC, ISRC, and Interdisciplinary Program in AI, Seoul National University

2026.01.03.
HyorinJung

Index

- Review
- Points
- Experiments
- Proposal

Review



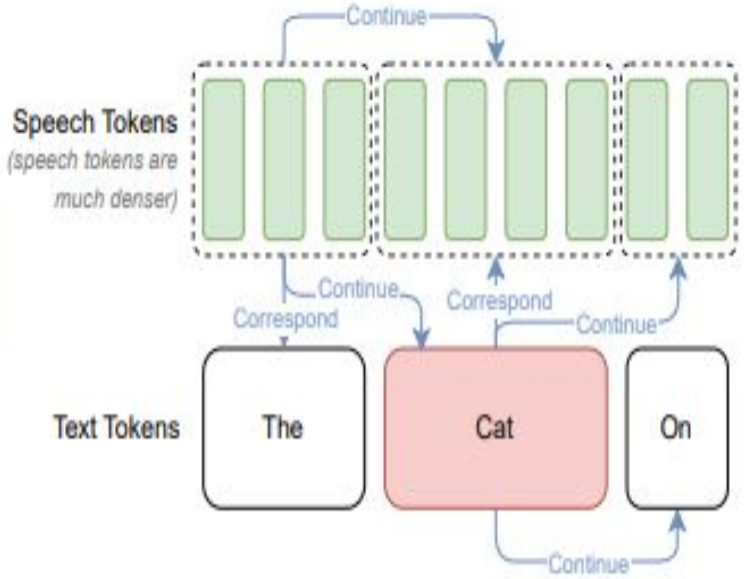


Figure 1: Overview of our spoken dialog modeling approach (Left). All possible self-supervised learning objectives from our speech-text pretraining scheme. (Right)

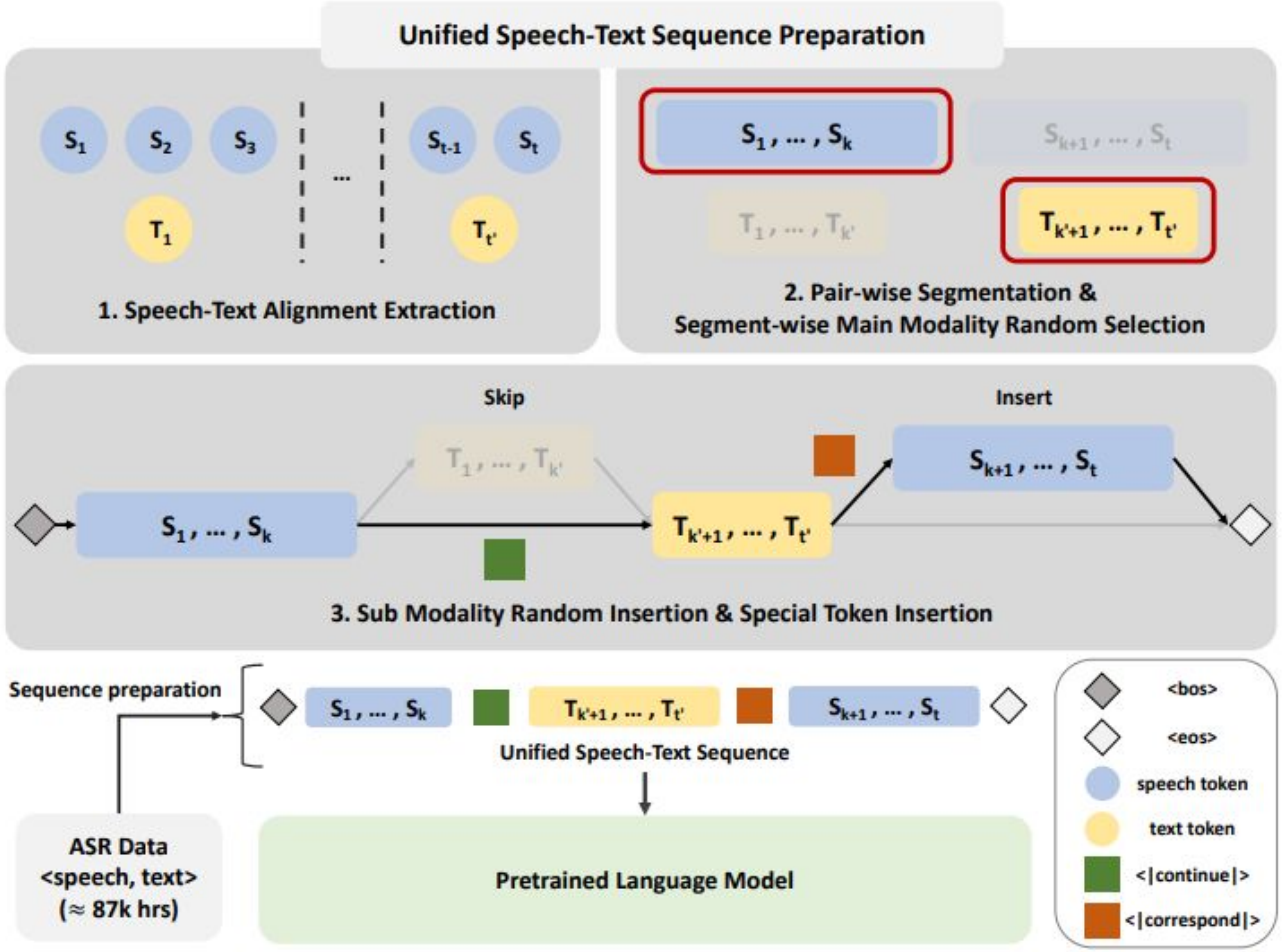


Figure 3: The overall speech-text pretraining scheme.

Points

Acoustic unit

Acoustic unit > 모델 > 감정분류	Acoustic Unit > 모델 > 텍스트 생성

생성 방식 점검

USDM → **USDM**
Scratch
Casacade
SpeechGPT

Human Evaluation	Gpt Evaluation

Points

Acoustic unit 생성 (코드 -preporcessing stage1)

XLS-R = self-supervised speech encoder >> 음성 특징을 추출

- 50Hz intermediate representations로 변환(이산값)
- 이것을 clustering하여 acoustic unit으로 생성(연속값)
- 음향 단위 토큰 어휘(vocab) = 10000개

https://anwarvic.github.io/speech-recognition/XLS-R?utm_source

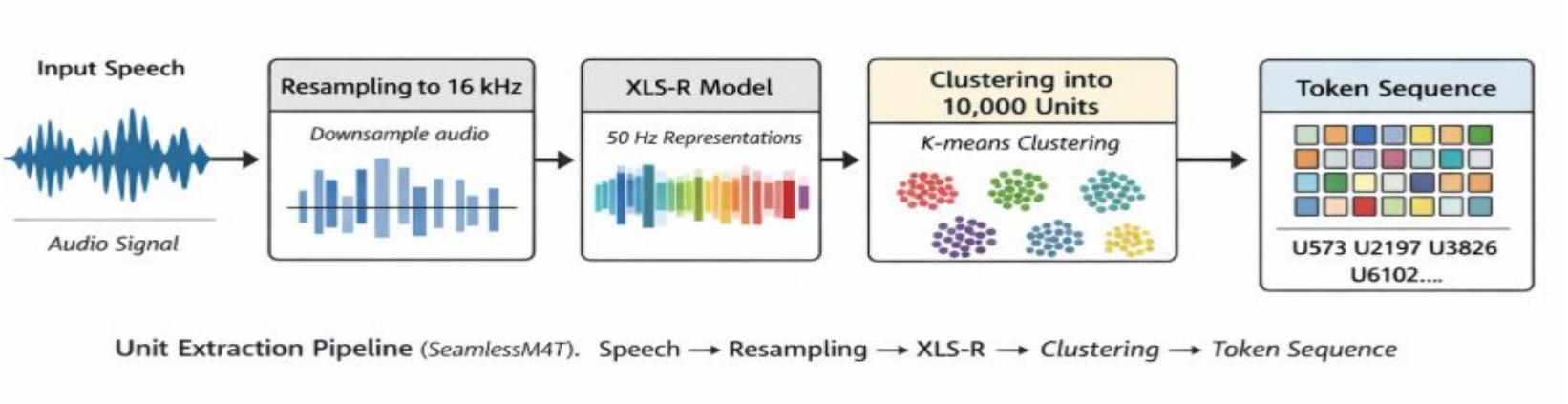
Methods： Speech-to-Unit Encoder

Unit Extraction Pipeline (SeamlessM4T)

speech ➡ resampled to 16 kHz ➡ XLS-R

➡ 50Hz intermediate representations

➡ clustered into 10,000 unit



Unit Extraction Pipeline (*SeamlessM4T*). Speech → Resampling → XLS-R → *Clustering* → *Token Sequence*

Experiment

Unit Emotion Recognition Experiment

- 3-layer transformer emotion classifier trained on **CREMA-D**
- Six emotion categories
- Random guess accuracy: **16.6%** Observed accuracy: **60.8%**

👉 This is strong evidence that acoustic units **contain clear emotional cue**

실험 구성

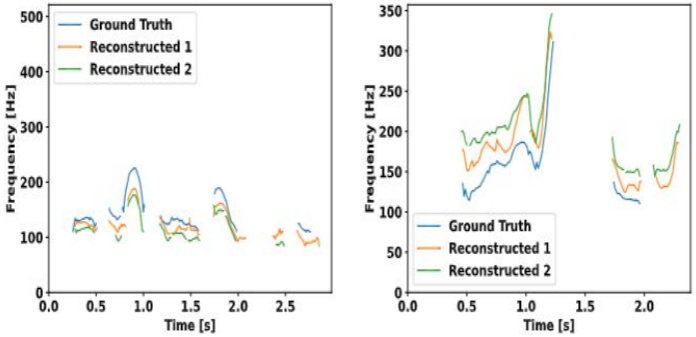
- 모델 구성 : 3-layer Transformer-based emotion classifier
- 손실 함수 : Cross-entropy loss
- 데이터셋 : CREMA-D
  - 총 7,442개 오디오
  - 91명 배우
  - 6개 감정 범주
  - Anger / Disgust / Fear / Happy / Neutral / Sad
- 데이터 분할 : Train / Validation / Test = 70% / 15% / 15%

Methods： Speech-to-Unit Encoder

Unit-to-Speech Reconstruction Experiment

- Train a unit-to-speech reconstruction model
- Uses Voicebox architecture
- Trained on 54,000 hours of speech

>>> Generates speech solely from unit sequences



Results.

- Timbre(음색) and absolute pitch(음높이) differ from original speech
- BUT Pitch variation(억양 변화 패턴) shows a similar trend

👉 This indicates preservation of **non-verbal characteristics**

Unit-to-Speech Reconstruction Experiment

Model	**Training Data**
• Voicebox: Flow-Matching 기반 zero-shot TTS • 출력: Mel-spectrogram • Vocoder: BigVGAN (22,050 Hz)	• Multilingual LibriSpeech (EN subset) • GigaSpeech • 총 54K hours ASR data
Alignment & Duration	**Inference Configuration**
• Montreal Forced Aligner (MFA) 사용 • 음소-멜 스펙트로그램 alignment 추출 • 음소별 duration 계산 • 학습 시: 음소 시퀀스를 멜 길이에 맞게 확장(ex) 안안안녕녕녕녕) • 추론 시: Duration Predictor 사용	• 50 timesteps(50번) • Classifier-free guidance scale = 1(억지로 생성 x, 자연스럽게) • Mel-spectrogram resolution: ~86 Hz • BigVGAN 공식 22,050 Hz 설정과 동일

Unit-to-Speech Reconstruction Experiment

Core Voicebox(기본)

- Flow-Matching 기반 zero-shot TTS 모델
- 입력: 텍스트 + 기준 음성 (reference speech)
- 출력: Mel-spectrogram → BigVGAN으로 음성 변환
- MFA로 음소-음성 정렬
- Duration predictor로 발음 길이 예측

Unit-Voicebox for USDM

- 역할: Speech decoder
  - 모델이 만든 unit sequence → 음성 복원
- 특징:
  - Reference speech 사용 (학습 & 추론) → Zero-shot 음성 복원 가능
- 멀티턴 대화:
  - 이전 턴 음성을 다음 턴의 기준 음성으로 사용 → 화자 음성 일관성 유지
- Duration predictor 불필요 (unit자체가 이미 시간순으로 배열)

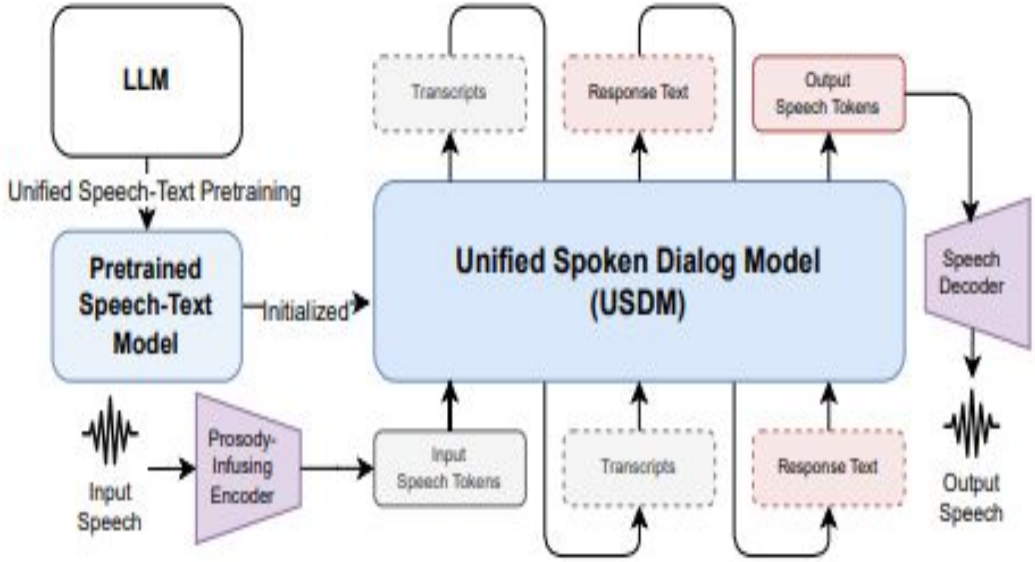
Voicebox for Cascaded Baseline

- 목적: 비교용 베이스라인 TTS
- 설정: Unit-Voicebox와 동일한 데이터·학습 방식 → 공정한 비교
- 차이점:
  - 입력: 텍스트
  - 별도의 duration predictor 학습 필요

Methods : Unified Speech-Text Pretraining

Models

- Backbone: Mistral-7B as the pretrained LLM
- Add 10,000 acoustic unit tokens + 2 special tokens(corresponse+continue) Only the embedding weights of newly added tokens are reinitialized
- Pretraining data: ~87,000 hours of English ASR data



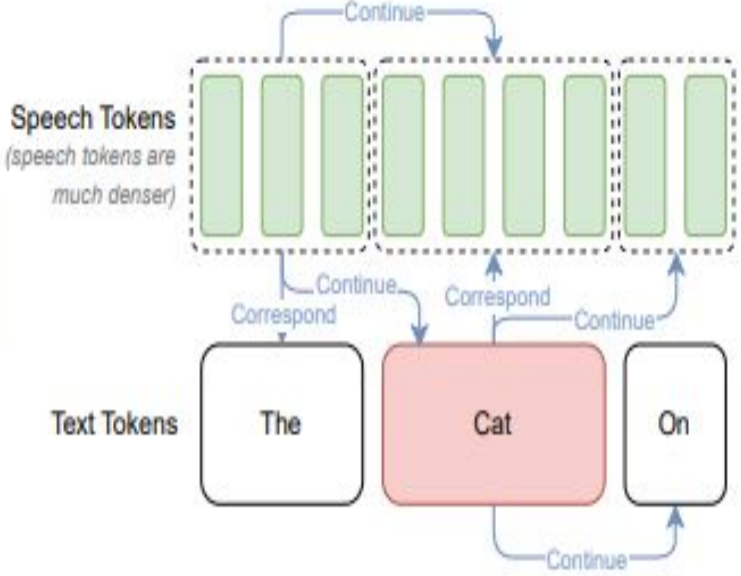


Figure 1: Overview of our spoken dialog modeling approach (Left). All possible self-supervised learning objectives from our speech-text pretraining scheme. (Right)

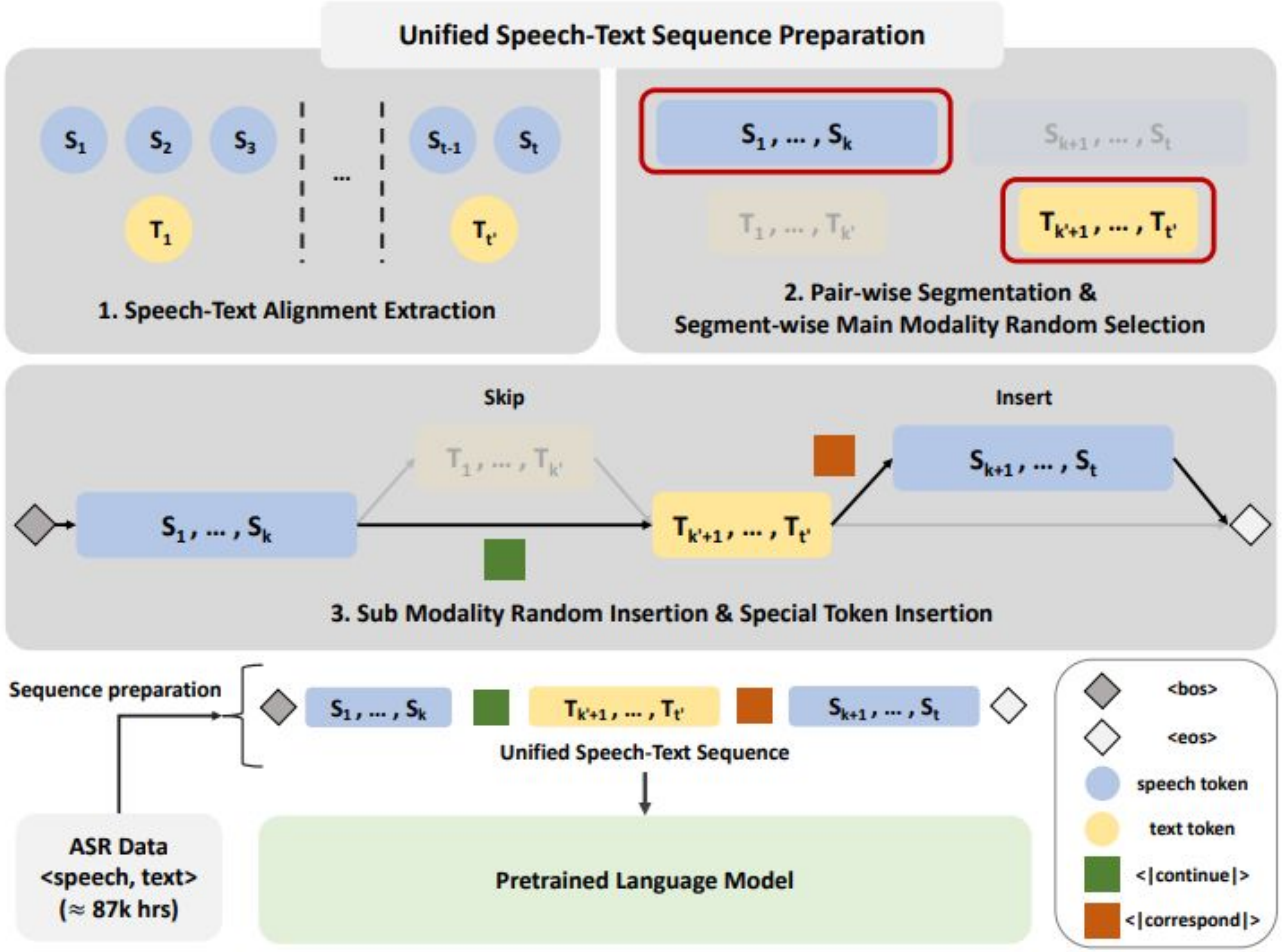


Figure 3: The overall speech-text pretraining scheme.

Methods

Speech-Text Alignment Extraction

- Extract word-level speech-text alignments; Using Montreal Forced Aligner
- Convert speech intervals into unit-index intervals at 50 Hz

Pair-wise Segmentation & Main Modality Selection

- Each speech-text pair is divided into N segments
- From each segment, only one modality is selected
- $N = \lfloor S/10 \rfloor + 1$, where S is speech length (seconds)

Methods

왜 두 곳에서 특수 토큰 추가하는 것? stage2 + train_pt

stage2에서 돌리고 저장안함 > train_pt에서 다시 호출

Sub-Modality Insertion & Special Tokens

- Non-selected modality is inserted with **50% probability**
- Two special tokens are introduced:
  - <|correspond|>
  - <|continue|>
- Special tokens are added **only when modality changes**

Resulting Interleaved Sequences

- Generate interleaved sequences {Ij}
- Enable:
  - unimodal modeling
  - comprehensive cross-modal modeling

Unit-to-Speech Decoder

- Unit-to-speech decoder based on the Voicebox architecture
- Voicebox: a zero-shot TTS model using text and reference speech
- Adapted to generate speech from acoustic unit sequences
- Reference speech enables speaker adaptation
- Paralinguistic information in units supports prosodic speech generation >> Referred to as unit-Voicebox

Unified Spoken Dialog Model (USDM)

- Fine-tune a pretrained speech–text model on spoken dialog data to leverage pretrained LLMs.
- Motivated by step-by-step reasoning and text-based speech bridging
- End-to-End Speech–Text–Speech Pipeline:
  - Instead of directly modeling speech-to-speech
  - The model performs:
    - Speech transcription
    - Response text generation
    - Response speech generation
  - All within a single end-to-end pipeline
- Text-related tasks are inserted between speech input and output
  - Enables leveraging pretrained LLM knowledge
  - Supports chained reasoning over the intermediary modality
- Each stage attends to all input and output tokens generated so far >> Joint modeling improves robustness to transcription errors
- Comparison with Cascaded Approaches:
  - More robust than independent cascaded modules
  - Avoids severe error propagation
- Supervised Fine-tuning Template
  - Loss is computed only for:
    - input transcript
    - answer text
    - answer unit tokens

Setting

- Experiments conducted on DailyTalk
- 20 hours of spoken dialog data
- Sampling rate: 22,050 Hz
- Two speakers: one male, one female
- Model comparisons with three baselines
- Compared Models:
  - USDM
  - From Scratch (Evaluate effectiveness of speech-text pretraining)
  - Cascaded (unified vs cascaded)
  - SpeechGPT(Validate advantages)

USDM Configuration	From Scratch Baseline

• Speech-to-unit: XLS-R, k = 10,000 quantizer
• Unit-to-speech: unit-Voicebox
• Vocoder: BigVGAN
• Unified speech-text pretraining:
  ○ ~87,000 hours of English ASR data
  ○ Max sequence length: 8,192
• Spoken dialog fine-tuning:
  ○ 5 epochs
  ○ Global batch size: 64
  ○ max learning rate: 2x10-5
	• Identical to USDM architecture
	• No speech-text pretraining
	• Mistral-7B fine-tuned directly on spoken dialog data

Cascaded Baseline	SpeechGPT Baseline
• Separate modules for each stage • ASR: whisper-large-v3 • Text dialog model: Mistral-7B • TTS: Voicebox • ASR → Text Dialog → TTS pipeline • Identical to USDM architecture	• Based on SpeechGPT • Pretrained speech-text model • Fine-tuned on DailyTalk for fair comparison

Table 4: Models for each component of the USDM and the baselines.

Model	ASR Model	Speech Encoder	LLM	Speech Decoder	TTS Model
USDM	–	XLS-R	Mistral-7B	unit-Voicebox + BigVGAN	–
From Scratch	–	XLS-R	Mistral-7B	unit-Voicebox + BigVGAN	–
SpeechGPT	–	mHuBERT	Llama-7B	unit-HiFi-GAN	–
Cascaded	*whisper-large-v3*	–	Mistral-7B	–	Voicebox + BigVGAN

Evaluation Sampling

- Spoken responses evaluated on naturalness, prosody, and semantic coherence
- Sampling: top-k = 40, top-p = 0.7, temperature = 0.3(SpeechGPT제외)
- All audio resampled to 16 kHz and normalized to −27 dB
- Reference speech: previous turn (Voicebox, unit-Voicebox)

Human Evaluation

- Human preference test via Amazon Mechanical Turk
- 50 randomly selected spoken dialogs
- Criteria: naturalness, prosody, semantic coherence/ 텍스트 내용은 동일
- Additional metrics:
  - MOS (5-scale)
  - P-MOS (5-scale prosody MOS)

Automatic Evaluation (Content & Error Rates)

- Content evaluation:
  - METEOR, ROUGE-L (ASR-transcribed text)
  - GPT-4-based preference test
- Error rates:
  - STT WER (entire test set)
    - Cascaded: Whisper-large-v3 outputs
    - End-to-end models: intermediate transcriptions
  - TTS WER (50 dialogs, 5 samples each> 평균사용)
    - Whisper-large-v3 used as ASR

METEOR: 생성된 응답 텍스트와 정답 텍스트 간의 일치도
ROUGE-L: 생성된 응답 텍스트와 정답 텍스트 간의 최장 공통 부분열(LCS)을 기반으로한 일치도

STT: 전사된 텍스트의 정확도
TTS: 모델이 생성한 응답 텍스트를 음성으로 변환한 후, 이 음성을 다시 ASR 모델로 텍스트화했을 때(재전사) 발생하는 오류율

Results

논문 주장: ground truth결과와 win, lose가 거의 동등하게 나옴> > USDM의 높은 성능을 보여줌
의문점: 그렇다면 나머지 모델도 ground truth랑 비교했을 때의 결과를 보여줘야하는 것 아닌지?

The results are presented in Table 1 and 2. In human preference tests that consider comprehensive factors, our model is preferred similarly to the Ground Truth and demonstrates superior preferences

Table 1: Human evaluation results of our model and the baselines. We report the MOS and P-MOS scores with a 95% confidence interval.

Method	Overall			Acoustic	
	win	*tie*	*lose*	MOS	P-MOS
Ground Truth	45.9%	8.0%	46.1%	4.51 ± 0.05	4.35 ± 0.05
USDM				4.31 ± 0.07	4.31 ± 0.06
Cascaded	55.3%	4.9%	39.8%	4.26 ± 0.07	4.22 ± 0.07
From Scratch	53.3%	7.6%	39.1%	3.71 ± 0.11	3.65 ± 0.10
SpeechGPT [25]	53.8%	6.9%	39.3%	4.08 ± 0.09	4.04 ± 0.08

MOS: 평균음향 점수
P-MOS: 운율평균 점수

Results

논문 주장: Scratch모델과 비교했을 때 pretrain되지 않으면 성능 낮아진다 > > USDM의 높은 성능을 보여줌

의문점:

1) ground truth와의 win/lose비율이 비슷하다로 usdm의 높은 성능을 강조한 앞 결과로는 현재 결과 설득x

2) casacaded방식이 STT,TTS 성능 훨씬 잘나옴.
앞의 결과도 오차범위 내,
>> 그렇다면 casacaded방식이 더 우수한 것 아닌가?

Table 2: GPT-4 evaluation and quantitative results of our model and the baselines.

Method	Semantic					WER	
	win	*tie*	*lose*	METEOR	ROUGE-L	STT	TTS
Ground Truth	32.7%	19.6%	47.7%	–	–		2.2%
USDM	–	–	–	13.1	15.7	7.4%	2.0%
Cascaded	42.7%	24.6%	32.7%	12.5	15.0	3.8%	1.3%
From Scratch	79.7%	10.1%	10.2%	8.6	10.6	58.1%	64.0%
SpeechGPT [25]	61.0%	13.1%	25.9%	9.9	11.8	12.4%	23.2%

METEOR: 생성된 응답 텍스트와 정답 텍스트 간의 일치도
ROUGE-L: 생성된 응답 텍스트와 정답 텍스트 간의 최장 공통 부분열(LCS)을 기반으로한 일치도
STT: 전사된 텍스트의 정확도
TTS: 모델이 생성한 응답 텍스트를 음성으로 변환한 후, 이 음성을 다시 ASR 모델로 텍스트화했을 때(재전사) 발생하는 오류율

[input tokens (text + speech)]

↓

CustomMistralModel (Transformer stack)

↓

hidden states [H]

↓

lm_head (Linear)

↓

logits over (text + speech vocab)

↓

softmax → next token