NeurIPS2024

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

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

Introduction

Motivation & Problem

- Large Language Models (LLMs) show emergent abilities such as reasoning and instruction following
- Current LLM-based agents are mostly text-based
- Speech interaction is essential for natural and effective human–AI communication
- Conventional systems rely on ASR → LLM → TTS pipelines
- Linguistic mismatch between speech and text causes inefficiency
- Paralinguistic cues (emotion, prosody) are often lost

Introduction

How about visions?

- Recent trends explore Large Foundational Models (LFMs) across modalities
- Vision-language models show that LLMs can model non-text tokens
- Autoregressive language models can model non-text modalities
- Speech modeling with LLMs remains underexplored

Introduction

Speech Modeling with LLMs: Prior Work.

- Early work focused on text-less speech modeling
- Recent studies incorporate LLMs for speech understanding or synthesis
- Most approaches remain limited to text outputs

Challenge.

- It remains unclear whether LLMs can generate and understand speech
- Incorporating paralinguistic information in spoken dialog is challenging

Introduction

Unified Spoken Dialog Model (USDM)

- End-to-end LLM-based spoken dialog modeling
- Directly integrates speech modality into LLMs

Speech–Text Pretraining Strategy

- Novel speech-text pretraining scheme
- Models cross-modal distributional semantics
- Multiple training objectives from speech-text pairs

Spoken Dialog Modeling & Prosody

- Speech-to-speech dialog broken into intermediate steps
- Prosody-infused discrete speech tokenization
- Tokens encode both semantics and prosody

Background

Discrete Speech Representations.

- To construct spoken language models (SLM), prior work uses various discrete speech representations
- These are primarily categorized into two types:
  - Tokens based on speech self-supervised representations
  - Neural audio codecs

Background

Self-supervised Tokens (Acoustic Units)

- A discrete token is obtained by k-means clustering intermediate representations from a speech self-supervised model
- Often called acoustic units, typically encoded at 25–50 Hz
- Speech information depends on the number of clusters k
- With relatively small k, many works preserve semantic information to construct SLMs

Background

Neural Audio Codecs (Semantic + Paralinguistic)

- Neural audio codecs capture both semantic and paralinguistic information
- Encoder–decoder autoencoder with a residual vector quantizer for encoder outputs
- Includes most perceptual information of audio
- Widely used for audio synthesis

Background

Spoken Language & Dialog Models: Task Landscape

- Recent studies explore spoken language modeling for tasks involving speech and text
- Speech input → text output: ASR, spoken QA, speech-to-text translation
- Text input → speech output: TTS, speech synthesis
- Early SLMs are trained **solely** on speech without language models
- With LLMs, studies extend language models to handle **both speech input and output** (pretraining or specific tasks)

Background

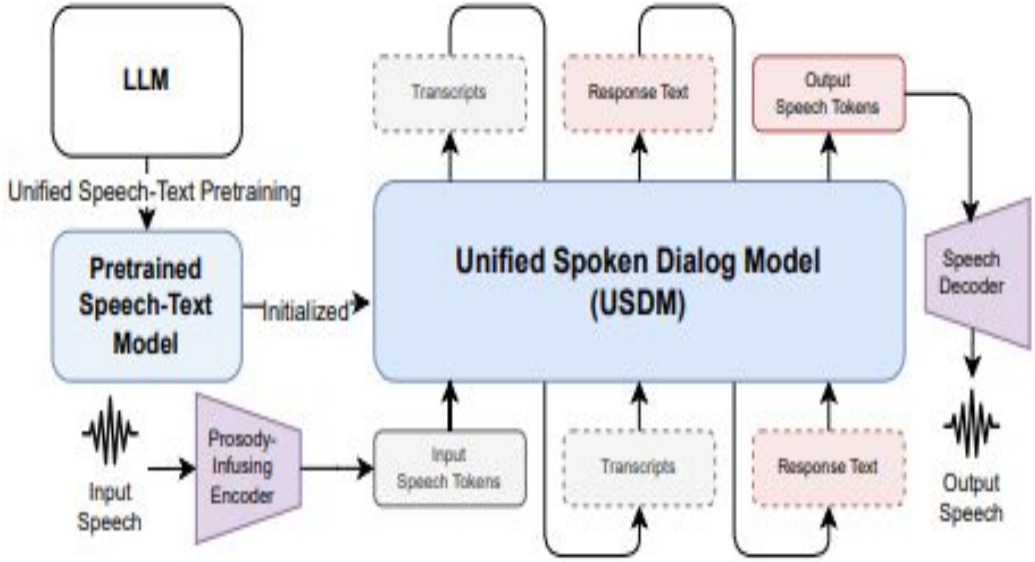
Prior Work for Spoken Dialog Modeling (Speech In/Out)

- Works for spoken dialog modeling with **speech input and output** include:
  - Nguyen et al.: decoder-only transformer trained from scratch for two-speaker conversations
  - Lin et al.: cascaded pipeline (ASR + emotion-aware text dialog LLM + emotional TTS)
  - Zhang et al.: SLMs on top of a pretrained LLM with objectives for ASR and TTS

Limitations of Prior Approaches

- End-to-end pipelines that rely on speech-only training or simple cross-modal objectives fail to fully utilize pretrained language models
- Cascaded models require explicit labels to incorporate paralinguistic features
  - Label dependency makes data collection challenging
  - Limits non-verbal cues to label-definable ones
- Error propagation in cascaded pipelines increases susceptibility to compounded errors

👉 Existing spoken dialog models suffer from limitations in jointly modeling semantic content, paralinguistic cues, and effectively leveraging LLMs



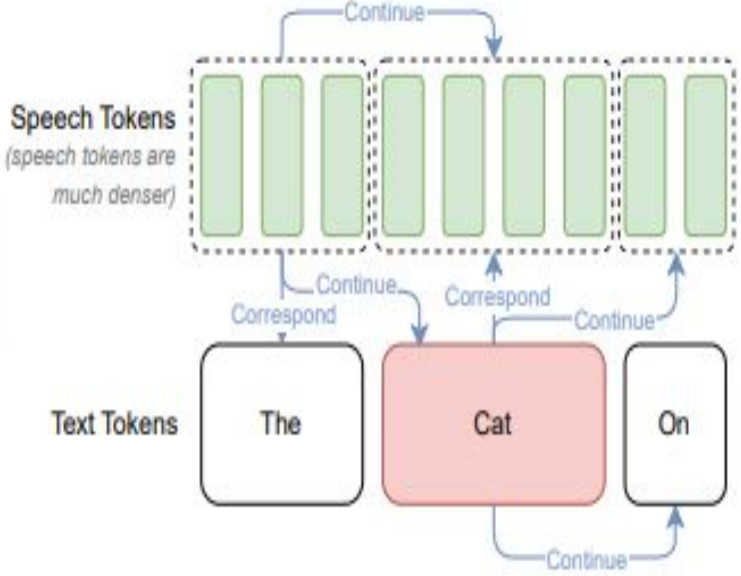


Figure 1: Overview of our spoken dialog modeling approach (Left). All possible self-supervised learning objectives from our speech-text pretraining scheme. (Right)

Methods : Speech-to-Unit Encoder

Speech-to-Unit Encoder: Motivation

- Natural spoken conversations require modeling both speech content and paralinguistic features >> Adopt acoustic units as discrete speech token

What are Acoustic Units?

- Acoustic units are derived from k-means clustering of intermediate representations
- Extracted from a self-supervised speech model
- Known to predominantly capture content and pronunciation

Why Choose This Tokenization Scheme (k = 10,000)?

- Information in acoustic units depends on the number of clusters **k**
- Larger **k** encodes more speech information

Methods : Speech-to-Unit Encoder

Unit Extraction Pipeline (SeamlessM4T)

- Speech is resampled to 16 kHz
- Processed by XLS-R to obtain 50 Hz intermediate representations
- Representations are clustered into 10,000 units

>> speech is converted into a token sequence with a fixed vocabulary.

Do Acoustic Units Contain Paralinguistic Information?

- We analyze whether acoustic units encode **non-verbal information**
- Two experiments are conducted:
  - Unit-based emotion recognition
  - Unit-to-speech reconstruction

Methods : Speech-to-Unit Encoder

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**

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

Methods： Speech-to-Unit Encoder

Unit-to-Speech Reconstruction Experiment

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

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

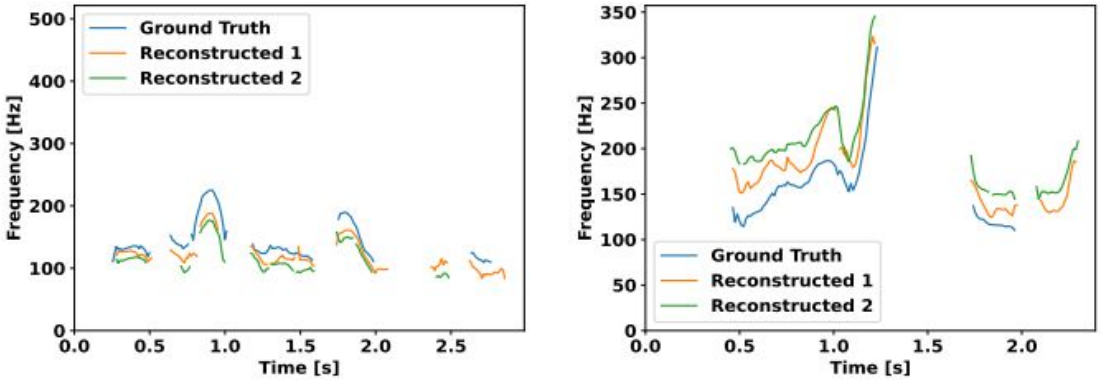


Figure 2: Pitch contour of the original audio and the audio reconstructed from extracted acoustic units. Due to the stochastic nature of the reconstruction model, we attempt reconstruction twice, demonstrating that the pitch variation closely mirrors the ground truth.

Methods : Unified Speech-Text Pretraining

Models

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

Methods : Unified Speech-Text Pretraining

Speech-Text Interleaved Pretraining Data

- Each ⟨speech, text⟩ pair is converted into an interleaved sequence
- An interleaved sample: $\mathcal{I}_j = i_{1,j}, i_{2,j}, ..., i_{||\mathcal{I}_j||,j}$
- Each token can be:
  - an acoustic unit token
  - a text token
  - a special token

The number of tokens in the j-th training example.

$$\mathcal{L}(\theta) = -\Sigma_{j=1}^{||\mathcal{D}||}\Sigma_{k=1}^{||\mathcal{I}_j||}\log p(i_{k,j}|i_{<k,j};\theta),$$

Pretraining Objective

- Dataset: $\mathcal{D} = \{\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_{||\mathcal{D}||}\}$
- Pretraining objective is defined over interleaved sequences
- θ includes:
  - LLM parameters
  - embedding weights of newly added tokens

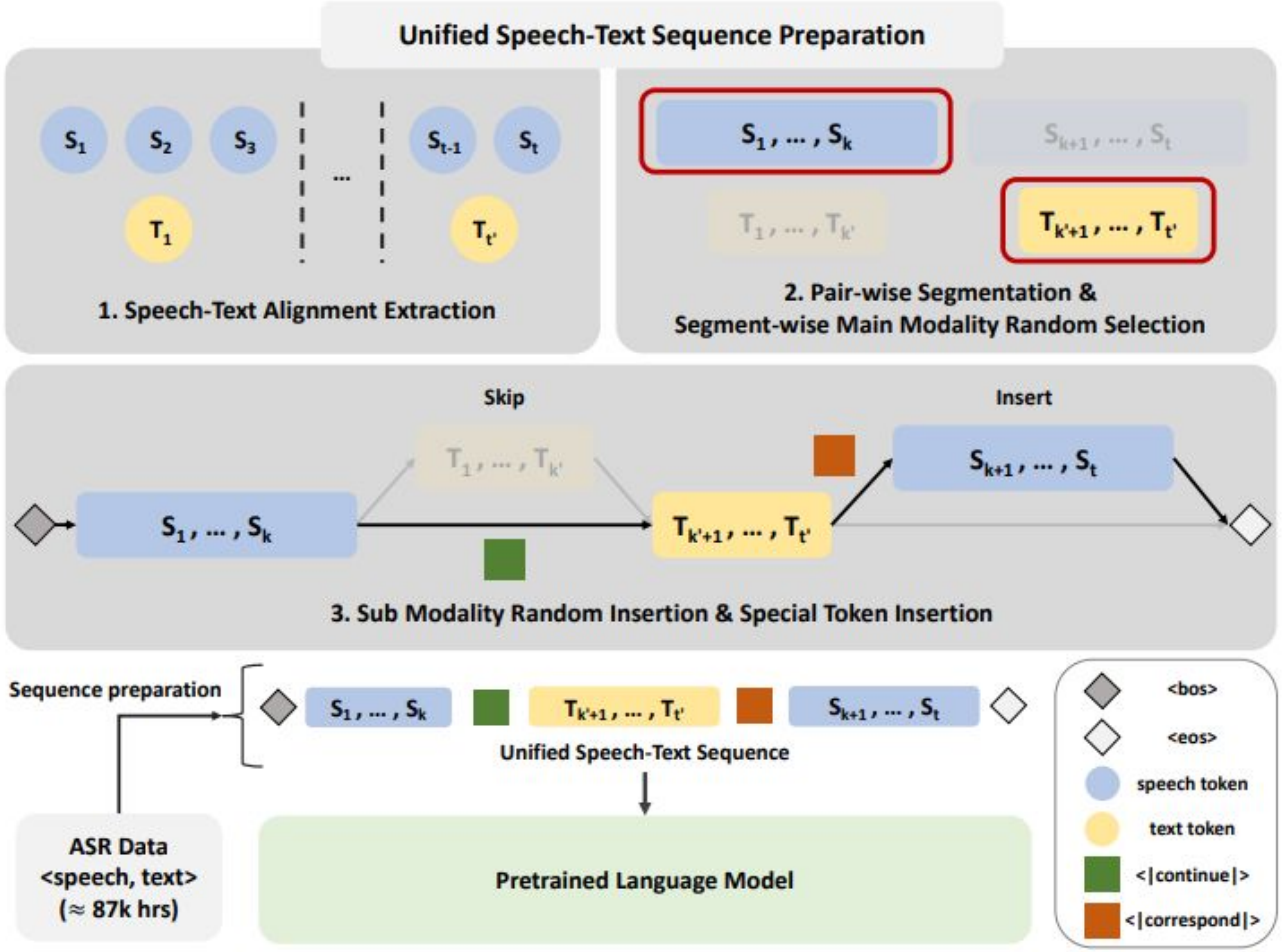


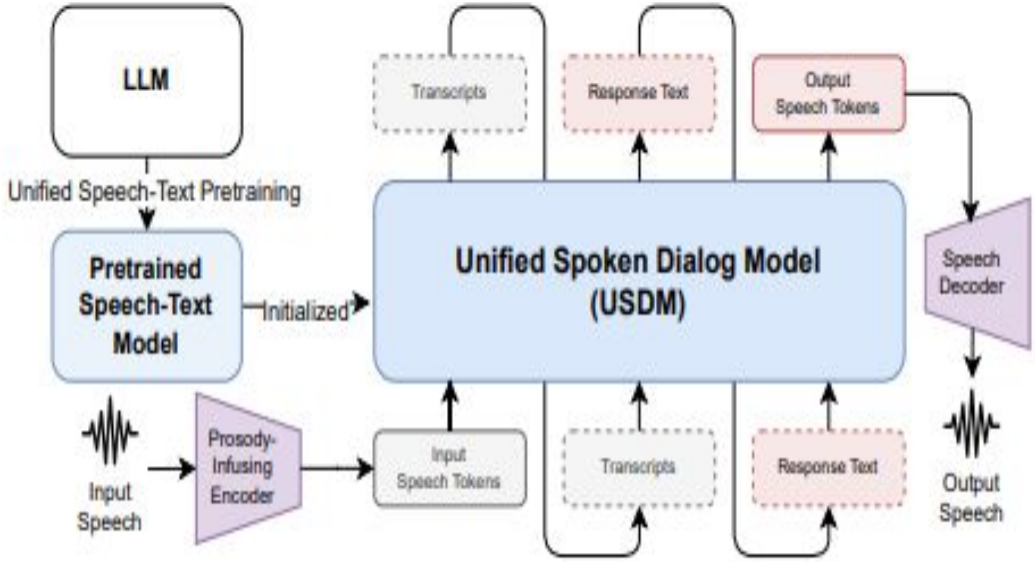
Figure 3: The overall speech-text pretraining scheme.

Why Not Task-Specific Pretraining?

- Extending LLMs via task-specific objectives may limit model capability
- Examples: ASR, TTS, unimodal modeling, cross-modal continuation
- Such approaches restrict the model to predefined relationships

Redefining Cross-Modal Relationships

- Reinterpret cross-modal relationships as:
  - Continuation ; (speech → speech, text → text)
  - Correspondence ; (speech ↔ text, 발화와 전사 관계)



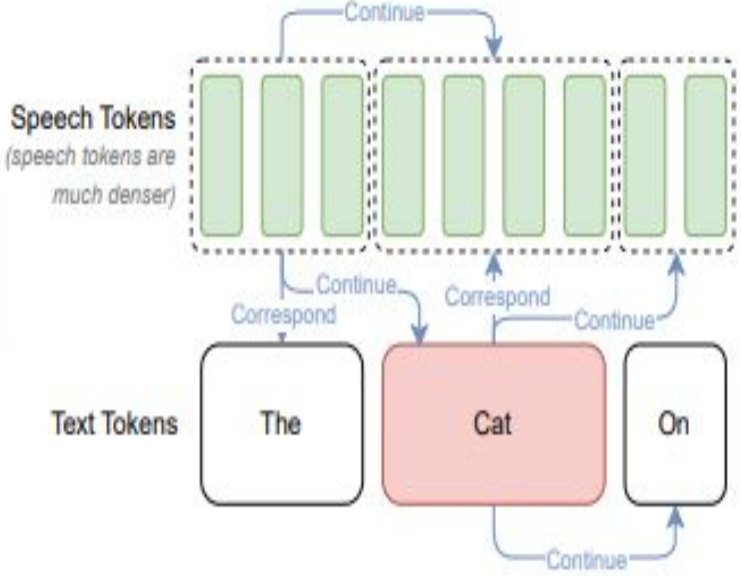


Figure 1: Overview of our spoken dialog modeling approach (Left). All possible self-supervised learning objectives from our speech-text pretraining scheme. (Right)

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

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

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

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

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
	• 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	• Based on SpeechGPT • Pretrained speech-text model • Fine-tuned on DailyTalk for fair comparison

Evaluation Setup

- Spoken responses evaluated on naturalness, prosody, and semantic coherence
- Sampling: top-k = 40, top-p = 0.7, temperature = 0.3
- 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
- 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)
  - TTS WER (50 dialogs, 5 samples each)

Results

- USDM preferred similarly to Ground Truth in human preference tests
- Superior performance over baselines ($p < 0.05$)
- Outperforms baselines in:
  - semantic evaluations
  - P-MOS (prosody naturalness)

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: 평균음질 점수
P-MOS: 운율평균 점수

Results

- From Scratch model shows degraded performance
- Tends to ignore intermediate text representations
- Results in higher STT/TTS WER and lower MOS/P-MOS
- Indicates the necessity of speech-text cross-modal pretraining

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

Ablation

- Unified speech-text pretraining achieves the best overall perplexity (PPL)
- Using only continuation or correspondence leads to limited generalization
- Removing correspondence modeling causes higher STT and TTS WER

Limitations

- Limited exploration of pretraining datasets and backbone LLMs
- Dependency on cross-modal chaining for spoken response generation
- Pretraining relies on large-scale English speech data, limiting multilingual applicability
- Effectiveness on other speech-text tasks remains unexplored
- Future work includes:
  - broader datasets and LLM backbones
  - direct speech-to-speech dialog modeling
  - multilingual extension
  - application to other speech-text tasks

Conclusion

- USDM synthesizes spoken dialog responses with natural prosody
- Introduces a unified speech-text pretraining scheme modeling comprehensive cross-modal relationships
- Acoustic unit tokenization preserves prosodic information
- Outperforms baselines in content, prosody, and naturalness on DailyTalk
- Ablation studies validate the effectiveness of pretraining and fine-tuning
- Demonstrates potential for extending LLM conversational capabilities to voice domains