



# Chain-of-Thought Training for Open E2E Spoken Dialogue Systems

Siddhant Arora<sup>1</sup>, Jinchuan Tian<sup>1</sup>, Hayato Futami<sup>2</sup>, Jee-weon Jung<sup>†1</sup>, Jiatong Shi<sup>1</sup>, Yosuke Kashiwagi<sup>2</sup>, Emiru Tsunoo<sup>2</sup>, Shinji Watanabe<sup>1</sup>

<sup>1</sup>Language Technologies Institute, Carnegie Mellon University, USA

<sup>2</sup>Sony Group Corporation, Japan

siddhana@andrew.cmu.edu

## Abstract

Unlike traditional cascaded pipelines, end-to-end (E2E) spoken dialogue systems preserve full differentiability and capture non-phonemic information, making them well-suited for modeling spoken interactions. However, existing E2E approaches often require large-scale training data and generates responses lacking semantic coherence. We propose a simple yet effective strategy leveraging a chain-of-thought (CoT) formulation, ensuring that training on conversational data remains closely aligned with the multimodal language model (LM)’s pre-training on speech recognition (ASR), text-to-speech synthesis (TTS), and text LM tasks. Our method achieves over 1.5 ROUGE-1 improvement over the baseline, successfully training spoken dialogue systems on publicly available human-human conversation datasets, while being compute-efficient enough to train on just 300 hours of public human-human conversation data, such as the Switchboard. We will publicly release our models and training code.

**Index Terms:** spoken dialog systems, speech foundation models, chain-of-thought

## 1. Introduction

Spoken dialogue systems [1, 2] are designed to engage in natural and interactive conversations with end users, playing a critical role in voice assistants and intelligent home devices. Despite their growing importance, building effective spoken dialogue systems remains a challenging task due to the complexity of human communication. Traditionally, spoken dialogue systems [3, 4] comprise multiple modules, including voice activity detection (VAD) [5], automatic speech recognition (ASR) [6, 7], natural language understanding (NLU) [8], natural language generation (NLG) [9], and text-to-speech (TTS) synthesis [10]. Each of these components presents unique challenges. For instance, the ASR module must accurately process shorter, spontaneous speech with disfluencies and filler words [11]. Additionally, spoken dialogue systems must be capable of understanding [12–14], and generating [15, 16] non-phonemic information, like emotions, to create natural interactions.

Recently, E2E spoken dialogue systems<sup>1</sup> [17, 18] have been proposed to process the user’s input speech utterance and generate an appropriate speech response within a single unified architecture. This approach avoids error propagation and can better capture non-phonemic cues such as emotion. Most existing E2E spoken dialogue systems [19] rely on complex architectures that significantly diverge from the model’s pre-training setup,

requiring extensive training compute (7 million hours of unlabelled audio data for multi-stream post-training and 20K hours of speech conversation data). Additionally, the lack of structured reasoning in E2E systems can often lead to less coherent responses, which we will reveal in our experiments (Sec. 5).

In this work, we ① introduce a chain-of-thought (CoT) [20, 21] training approach, restructuring post-training into a composition of ASR, text response generation, and TTS sequences. We apply CoT post-training to an open multimodal LLM [22] pre-trained on ASR, TTS, and Text LM tasks, aligning with its pre-training process to enhance compatibility and enable efficient adaptation to future architectures. ② Our experiments show that this efficient setup builds an effective conversational system using publicly available datasets, achieving strong results with as little as 300 hours of human-human conversation dataset Switchboard [11]. The proposed post-training approach delivers faster convergence, lower compute costs, and better data efficiency, which is crucial for low-resource domains like healthcare, where large datasets are difficult to collect. ③ We compare our method with cascaded baselines, where the same pre-trained SpeechLM is fine-tuned separately for each sub-task. Results show that our CoT-based E2E model achieves comparable performance while being 3× more parameter-efficient. ④ An ablation study confirms that CoT-based training improves both semantic coherence and synthesized speech quality compared to traditional E2E systems. Additionally, our model better preserves emotional nuances in generated speech, showing higher emotional similarity to human references. ⑤ Finally, we will release an open-source framework for training and evaluating cascaded and E2E dialogue systems using open source SpeechLMs on public datasets.

## 2. Related studies

Recently, “Chat” SpeechLMs [23] have gained attention for their ability to engage in natural and interactive conversations. Early efforts [24–26] developed speech-aware LMs that generated text responses from spoken input but relied on external TTS and VAD systems for spoken dialogue. More recently, speech-to-speech LMs [27–31] have emerged, capable of both understanding and generating speech within a single architecture.

SpeechGPT [17] employs a CoT instruction-tuning approach to enhance cross-modal conversational capabilities. However, SpeechGPT is trained solely on large-scale synthetic speech due to the difficulty of preparing speech instruction data. In contrast, we perform CoT training on real conversational data, showcasing its practical applicability and effectiveness in low-resource scenarios. This benefit comes from the explicit incorporation of ASR, text-to-text, and TTS task tokens (Sec. 3.1), enabling a more modular and efficient CoT

<sup>†</sup>Currently at Apple.

<sup>1</sup><https://openai.com/index/hello-gpt-4o/>,  
<https://deepmind.google/technologies/gemini/>

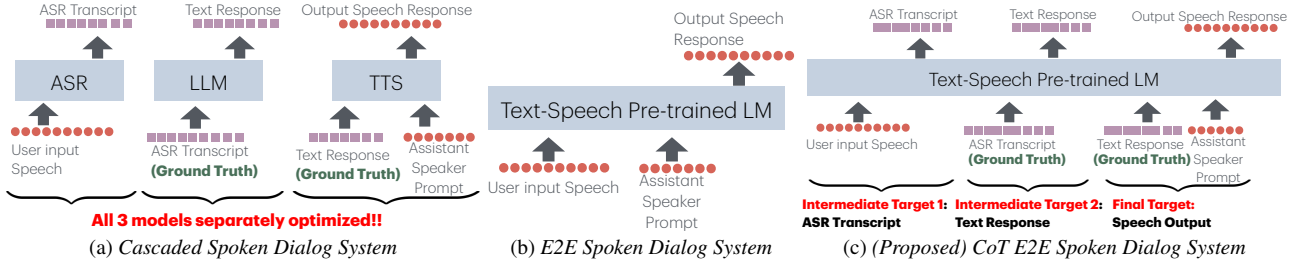


Figure 1: Training schematic for our CoT-based E2E spoken dialogue system, compared with cascaded and conventional E2E systems. Our approach employs multi-stage reasoning, improving semantic coherence and speech quality while preserving E2E differentiability. We use teacher forcing during training, and multi-stage decoding using generated intermediate outputs for inference (Sec. 3.4).

training process. While OMNI-Flatten [32] introduced a half-duplex training approach, where the model sequentially performs ASR, text response generation, and TTS. Although their setup resembles ours, their half-duplex approach serves only as an intermediate training step and lacks CoT reasoning during inference. In contrast, our work presents a novel investigation into the efficacy of CoT E2E spoken dialogue systems, providing a comprehensive comparison with standard cascaded and E2E methodologies using an open-source toolkit.

### 3. Method

We begin our problem formulation with a single-turn spoken dialog system, where  $X$  and  $Y$  represent the user’s and system’s speech feature sequences. Traditional cascaded spoken dialogue systems employ separate modules for sub-tasks, as shown in Fig. 1a. These systems typically consist of an ASR model, which optimizes  $P(S^{\text{asr}}|X)$  to produce the ASR transcript  $\hat{S}^{\text{asr}}$ , a text response generation model based on LLM which optimizes  $P(S^{\text{res}}|\hat{S}^{\text{asr}})$  to estimate the text response  $\hat{S}^{\text{res}}$ ; and a TTS model, which optimizes  $P(Y|\hat{S}^{\text{res}}, X^{\text{spk}})$ , where  $X^{\text{spk}}$  is the TTS speaker prompt. Since each module is separately optimized, the system suffers from error propagation. To address this, E2E spoken dialogue systems [17, 19] maximize the posterior distribution  $P(Y|X, X^{\text{spk}})$  as shown in Fig. 1b. While these systems offer fully E2E differentiable training, they lack structured reasoning, leading to less coherent responses and increased training data requirements.

#### 3.1. CoT formulation of E2E Spoken Dialog Systems

Inspired by the strengths of both schools of thought, our method aims to perform CoT training by explicitly incorporating the prediction of ASR transcript  $S^{\text{asr}}$  and text response  $S^{\text{res}}$  within the E2E spoken dialogue formulation. By regarding the  $S^{\text{asr}}$  and  $S^{\text{res}}$  as a probabilistic variable, we can theoretically incorporate them into the posterior distribution  $P(Y|X, X^{\text{spk}})$  (§ 3) via sum rule. We further use Viterbi approximation and conditional independence (C.I.) assumption to get:

$$\begin{aligned}
 P(Y|X, X^{\text{spk}}) &\approx P(Y|X, X^{\text{spk}}, \hat{S}^{\text{res}}, \hat{S}^{\text{asr}}) \\
 &\text{where } \hat{S}^{\text{res}} = \arg \max_{S^{\text{res}}} p(S^{\text{res}}|X, X^{\text{spk}}, \hat{S}^{\text{asr}}) \\
 &\text{where } \hat{S}^{\text{asr}} = \arg \max_{S^{\text{asr}}} p(S^{\text{asr}}|X, X^{\text{spk}}) \quad (1)
 \end{aligned}$$

To realize the formulation described in Eq.1, this work proposes a CoT E2E spoken dialog system, as shown in Figure 1c. The architecture is built around a pre-trained decoder-only LM capable of both processing and generating audio and text tokens. The **CoT model** follows a structured decoding process: **ASR Decoding**: Generates the ASR transcript  $\hat{S}^{\text{asr}}$  by modeling  $P(S^{\text{asr}}|X)$ . **Text Response Prediction**: Predicts the text

Table 1: Task templates for ASR, Text Response Generation (T2T), TTS, and both E2E and CoT E2E (Proposed) spoken dialogue models, showing their respective conditions and targets. The CoT E2E model includes intermediate targets, highlighted in blue. Here,  $\langle a\_tk \rangle$ ,  $\langle t\_tk \rangle$ , and  $\langle s\_tk \rangle$  denote the audio, text, and speaker tokenizer identifiers, respectively.

Task	Condition	Target
ASR	$\underline{C}^{\text{asr}} = (\langle \text{asr} \rangle, \langle a\_tk \rangle, \underline{X}, \langle t\_tk \rangle)$	$\underline{S}^{\text{asr}}$
T2T	$\underline{C}^{\text{t2t}} = (\langle t2t \rangle, \langle t\_tk \rangle, \underline{\hat{S}}^{\text{asr}}, \langle t\_tk \rangle)$	$\underline{S}^{\text{res}}$
TTS	$\underline{C}^{\text{tts}} = (\langle \text{tts} \rangle, \langle t\_tk \rangle, \underline{\hat{S}}^{\text{res}}, \langle s\_tk \rangle, \underline{X}^{\text{spk}}, \langle a\_tk \rangle)$	$\underline{Y}$
E2E	$\langle s2s \rangle, \langle a\_tk \rangle, \underline{X}, \langle s\_tk \rangle, \underline{X}^{\text{spk}}, \langle a\_tk \rangle$	$\underline{Y}$
CoT E2E	$\underline{C}^{\text{asr}}, \underline{S}^{\text{asr}}, \underline{C}^{\text{t2t}}, \underline{S}^{\text{res}}, \underline{C}^{\text{tts}}$	$\underline{Y}$

response  $\hat{S}^{\text{res}}$ , optimizing  $P(S^{\text{res}}|X, \hat{S}^{\text{asr}})$ , akin to a cascaded system but additionally conditions on input speech  $X$ . **Final Speech Output Generation**: Generates the speech output  $\hat{Y}$  by modelling  $P(Y|X, X^{\text{spk}}, \hat{S}^{\text{asr}}, \hat{S}^{\text{res}})$ .

Compared to standard E2E systems that model  $P(Y|X, X^{\text{spk}})$ , our CoT model preserves E2E differentiability while improving semantic coherence and speech quality through multi-stage reasoning, conditioning on intermediate outputs  $\hat{S}^{\text{asr}}$  and  $\hat{S}^{\text{res}}$ . Compared to TTS in the cascaded system that model  $P(Y|\hat{S}^{\text{res}}, X^{\text{spk}})$ , our approach reduces error propagation and improves expressive speech synthesis by conditioning on input speech  $X$  and ASR transcript  $\hat{S}^{\text{asr}}$  to generate final speech output. We apply the proposed CoT training as a post-training strategy on an open SpeechLM [22] that has been pre-trained on ASR, TTS, Speech Continuation, and TextLM tasks. By aligning each reasoning stage with the SpeechLM’s pre-training tasks—namely ASR, text LM, and TTS — our approach achieves faster convergence and improved training data efficiency.

#### 3.2. Text and Audio Joint Tokenizer

Our CoT training procedure utilizes autoregressive generation of a joint text/audio token, as defined in [22]. First, speech feature sequence  $X$ , text  $S$ , and special token (token) are tokenized ( $\text{Tok}(\cdot)$ ) to the following discrete representation:

$$\underline{X} = \text{Tok}(X), \underline{S} = \text{Tok}(S), \langle \text{token} \rangle = \text{Tok}(\langle \text{token} \rangle), \quad (2)$$

where the underline  $\underline{\quad}$  indicates the tokenized value.  $\underline{X} = (\underline{x}_1, \underline{x}_2, \dots)$  is a sequence of  $M$ -dimensional discrete audio codec and its  $t$ -th vector is represented as  $\underline{x}_t = [x_t^1, \dots, x_t^M]^T$ .  $x_t^1$  is reserved for a speech semantic token [33]. Similarly,  $\underline{S}$  is composed of  $j$ -th discrete vector  $\underline{s}_j = [s_j, \emptyset, \dots, \emptyset]^T$ , where  $\emptyset$  pads null values from the 2nd to  $M$ th dimension, as text tokens are one-dimensional.  $\langle \text{token} \rangle$  is also converted similarly to text tokens. All tokens belong to a *shared discrete vector space*,

Table 2: *Semantic Quality on Switchboard Eval 2000 and Fisher: Our CoT-based E2E model produces more coherent responses ( $p < 0.05$ ) than conventional E2E systems and matches the performance of the state-of-the-art LLM, SmolLM.*

Model	SWBD			Fisher		
	ROUGE-1/2/L (↑)	METEOR (↑)	Perplexity (↓)	ROUGE-1/2/L (↑)	METEOR (↑)	Perplexity (↓)
SmolLM v2-1.7B (GT Transcript)	14.0 / <b>2.3</b> / <b>11.5</b>	12.9	25.4	15.3 / <b>3.4</b> / <b>12.5</b>	13.9	<b>17.9</b>
SpeechLM E2E	10.5 / 0.8 / 8.4	9.7	302.2	11.3 / 1.3 / 8.9	10.1	138.7
SpeechLM CoT E2E						
Text Response ( $\hat{S}^{\text{res}}$ )	14.2 / 1.3 / 10.2	14.2	37.2	<b>17.6</b> / 3.2 / 12.2	17.6	33.0
w/o Post-process ( <i>ablation</i> )	12.3 / 1.3 / 8.6	14.8	24.1	14.6 / 2.9 / 10.0	17.7	16.1
using GT Transcript $\hat{S}^{\text{asr}}$ ( <i>ablation</i> )	12.3 / 1.4 / 8.8	15.0	24.5	15.0 / 3.0 / 10.2	17.8	16.6
Speech Response ( $\hat{Y}$ )	14.2 / 1.3 / 10.3	12.9	28.5	<b>17.6</b> / 3.2 / 12.3	<b>17.7</b>	18.5
SWBD+Fisher train $\hat{Y}$ (§ 4.1)	<b>14.6</b> / 1.5 / 10.4	<b>14.9</b>	<b>22.2</b>	<b>X</b>	<b>X</b>	<b>X</b>

i.e.,  $\mathbf{x}_t, \mathbf{s}_j, \langle \text{token} \rangle \in \mathcal{V}^M$ , where  $\mathcal{V}^M$  comprises the union of text, speech semantic and acoustic token vocabularies, along with special tokens. This formulation enables speech, text, and special tokens to be autoregressively predicted using an LM.

### 3.3. Training Sequence Format

All the training and decoding steps of our model, introduced in Section 3.1, are tokenized into a single sequence and uniformly predicted. Tab. 1 presents our sequence formats for each sub-task in the cascaded pipeline (ASR, text response generation, and TTS) and for E2E and CoT E2E spoken dialogue training. Task tokens are denoted as  $\langle \text{asr} \rangle$ ,  $\langle \text{t2t} \rangle$ ,  $\langle \text{tts} \rangle$ , and  $\langle \text{s2s} \rangle$ , representing ASR, text response generation, TTS, and E2E spoken dialogue tasks, respectively. Notably, the CoT E2E model does not introduce new task tokens and consists of **intermediate targets** that represent outputs from the first and second decoding stages (Sec. 3.1). Prompts for each decoding stage are carefully designed to align closely with SpeechLM’s pre-training objectives: ASR ( $\mathcal{C}^{\text{asr}}$ ), text LM ( $\mathcal{C}^{\text{t2t}}$ ), and TTS ( $\mathcal{C}^{\text{tts}}$ ) as shown in Tab. 1, enabling better training efficiency.<sup>2</sup>

### 3.4. Training and Inference

During CoT-post training, we compute the loss only on the “target” sequences: including intermediate targets  $\hat{S}^{\text{asr}}$ ,  $\hat{S}^{\text{res}}$  and final target  $\hat{Y}$ . During inference, our decoding process consists of three steps: ① The SpeechLM generates the ASR transcript  $\hat{S}^{\text{asr}} = \arg \max P(\hat{S}^{\text{asr}} | \mathcal{C}^{\text{asr}})$ . ② The generated ASR transcript is then incorporated to generate text response, sampled as  $\hat{S}^{\text{res}} \sim P(\hat{S}^{\text{res}} | \mathcal{C}^{\text{asr}}, \hat{S}^{\text{asr}}, \mathcal{C}^{\text{t2t}})$  using top-k sampling. ③ The predicted text response is similarly used to construct prompt (Tab. 1) for sampling the final speech response  $\hat{Y} \sim P(\hat{Y} | \mathcal{C}^{\text{asr}}, \hat{S}^{\text{asr}}, \mathcal{C}^{\text{t2t}}, \hat{S}^{\text{res}}, \mathcal{C}^{\text{tts}})$ . By employing joint tokenization (Sec. 3.2), our approach performs CoT-based inference (Eq. 1) within a standard LLM inference framework.

## 4. Experiments

### 4.1. Datasets

For our experiments, we focus exclusively on *real* human-human conversation datasets, selecting 2 widely used corpora: Switchboard [11] ( $\approx 300$  hours) and Fisher [34] ( $\approx 2000$  hours). For Switchboard, we use the Eval2000 dataset for evaluation, while for Fisher, we follow Dialog GSLM [28], splitting the dataset into 98:1:1 for train, dev, and evaluation, respectively. Additionally, we train the model in a *combined* setting (SWBD+Fisher train) using the **entire** Fisher dataset and the Switchboard training set and evaluate it on the Eval2000 dataset. Before training, we apply 3 preprocessing steps: ① We merge silence-separated utterances within a single speaker’s turn to preserve conversational coherence. ② We remove very short utterances (fewer than five words) to improve linguistic

<sup>2</sup>Other prompt designs resulted in sub-optimal performance.

Table 3: *Audio Quality and Intelligibility: Our CoT E2E model, trained on Fisher and Switchboard, matches single-speaker TTS (LJSpeech VITS) performance with a high-quality speaker prompt (“Spk prompt”). “GT” indicates decoding with ground-truth responses  $\underline{S}^{\text{res}}$  and (abl) denote ablation study.*

Model	SWBD		Fisher	
	WER (↓)	UTMOS (↑)	WER (↓)	UTMOS (↑)
LJSpeech VITS (GT)	8.5	4.19	9.5	4.14
SpeechLM				
Pre-train (GT)	47.8	2.08	55.7	1.93
Finetune (GT)	30.5	2.21	32.6	2.14
E2E	<b>X</b>	2.03	<b>X</b>	2.02
CoT E2E	13.6	3.55	12.5	3.32
w/o Spk prompt ( <i>abl</i> )	15.1	2.05	15.5	2.04
w/ GT ( <i>abl</i> )	15.5	2.25	13.2	2.23
w/o Post-process ( <i>abl</i> )	42.0	2.10	36.5	2.10
SWBD+Fisher train (§ 4.1)	9.1	3.40	<b>X</b>	<b>X</b>

context and response quality. ③ We truncate all utterances to a maximum length of 30 seconds.

We evaluate spoken dialogue systems across semantic quality and audio quality and intelligibility of response. For semantic quality, we report ROUGE [35] and METEOR [36] scores, using human references as ground truth, alongside perplexity [37], computed using GPT-2 [38]. For E2E models, semantic quality is evaluated by transcribing synthesized speech  $\hat{Y}$  using Whisper large [6]. For our CoT model, we also report metrics on the intermediate text response  $\hat{S}^{\text{res}}$  (Eq. 1). We utilize the VERSA toolkit [39], measuring intelligibility similarly through Whisper hypotheses and evaluating audio quality using UTMOS [40]. We evaluate conversation-level performance by extracting emotion vectors from both synthesized outputs and ground-truth responses using Emo2Vec [41], then measure their cosine similarity. We rank all spoken dialogue systems based on their emotional alignment and compute the average rank (“Emotion Rank”) across all utterances. Finally, we report the model sizes for both cascaded and E2E dialogue systems.

### 4.2. Baseline and Experimental Setups

We compare our CoT E2E system on response quality against strong task-specific baselines<sup>3</sup>: SmolLM v2 1.7B-Instruct [42] for text generation and the single-speaker TTS model (**LJSpeech VITS**) [43] from ESPnet-TTS. Our system applies the CoT-based post-training strategy to a pre-trained open-source SpeechLM [22]. For TTS, we also evaluate the pre-trained SpeechLM in two settings: zero-shot (**SpeechLM Pre-train**) and fine-tuned (**SpeechLM Fine-tune**).<sup>4</sup> We further compare with cascaded and E2E spoken dialogue systems using conversation-level metrics. The cascaded systems combine: SpeechLM (ASR), SmolLM v2 (text response), and SpeechLM (TTS), evaluated in both zero-shot (**SpeechLM Cascaded (Pre-train)**) and fine-tuned (**SpeechLM Cascaded**

<sup>3</sup>Baselines are not fine-tuned on spoken dialogue datasets.

<sup>4</sup>SpeechLM is not instruction fine-tuned on conversational data, so we exclude it from the text response evaluation.

Table 4: *Conversation Level Statistics on Switchboard: our CoT E2E model better captures emotion and is parameter efficient.*

Model	Emotion Rank ( $\downarrow$ )	Model Size ( $\downarrow$ )
SpeechLM		
Cascaded (Pre-train)	3.08	3.4B
Cascaded (Fine-tune)	2.57	5.1B
E2E	2.59	<b>1.7B</b>
CoT E2E (SWBD + Fisher)	<b>1.77</b>	<b>1.7B</b>

(**Fine-tune**) modes. We also compare with a traditional E2E system (**SpeechLM E2E**, Sec. 3), where SpeechLM is trained end-to-end to directly predict speech outputs.

Our models are implemented in PyTorch, with all experiments conducted using the ESPnet [44, 45] toolkit. The pre-trained SpeechLM leverages the SmoLLM2 1.7B text LLM for initialization. We adopt the delay interleave architecture [46] for multi-stream language modeling. For audio tokenization, we concatenate codec and SSL tokens frame-by-frame. Specifically, we utilize ESPnet-Codec [47]<sup>5</sup> for codec tokenization and XEUS<sup>6</sup> [48] for SSL tokenization. For decoding (Sec. 3.4), ASR uses greedy search followed by post-processing to remove hallucinations, while text response generation employs top-k sampling ( $k=30$ , temperature = 0.8), followed by post-processing to remove hallucinations and constrain response length. For speech response generation, we apply top-k sampling (same as text response), and further post-process the outputs, computing top-10 samples and selecting the speech with the highest intelligibility to generated text response  $\hat{S}^{\text{res}}$  (Eq. 1). We conducted Wilcoxon signed-rank and Signed Paired Comparison tests for statistical significance. Models are trained using 4 NVIDIA H200 GPUs. We will release data processing, training and inference details as part of ESPnet [44, 49] toolkit.

## 5. Results and Discussion

**Semantic Coherence and Audio Quality Results:** Tab. 2 presents the semantic quality of responses. Even state-of-the-art LLM SmoLLM2 struggles to achieve high ROUGE and METEOR scores (Tab. 2), reflecting the spontaneity and unpredictability of human-human conversations. Despite this, our CoT E2E model generates semantically coherent responses, performing on par with SmoLLM2, even when SmoLLM2 uses ground-truth transcripts. Whisper hypotheses produced from final speech response  $\hat{Y}$  exhibit similar semantic quality to the intermediate text response  $\hat{S}^{\text{res}}$ .<sup>7</sup> We observe no significant performance drop when using generated ASR transcripts  $\hat{S}^{\text{asr}}$  in place of ground-truth transcripts ( $\hat{S}^{\text{asr}}$  in CoT E2E prompt (Tab. 1)). Conventional E2E models (“SpeechLM E2E”) show poor overlap with human references and generate semantically incoherent sentences, as shown by their high perplexity scores.

Next, we evaluated audio quality performance in Tab. 3. The pre-trained SpeechLM performs significantly worse in both intelligibility and speech quality against a single-speaker VITS model, likely due to poor generalization on disfluent, conversational text and the low-quality audio in Switchboard, as the model uses speaker prompts  $\underline{X}^{\text{spk}}$  (TTS prompt in Tab. 1) from the corresponding speaker. Fine-tuning on conversational datasets improves both speech quality and intelligibility. Our CoT post-training achieves similar performance improvements,

<sup>5</sup>[https://huggingface.co/ftshijt/espnet\\_codec\\_dac\\_large\\_v1.4\\_360epoch](https://huggingface.co/ftshijt/espnet_codec_dac_large_v1.4_360epoch)

<sup>6</sup><https://huggingface.co/espnet/xeus>, K-means tokenizer trained on the last-layer representation with 5k clusters

<sup>7</sup>While our post-processing increases perplexity, it prevents very long responses.

Table 5: *Ablation Study on intermediate task ASR: While ASR performance slightly decreases with CoT training, the quality of text responses improves (Tab. 2).*

Model	SWBD		Fisher	
	WER ( $\downarrow$ )	CER ( $\downarrow$ )	WER ( $\downarrow$ )	CER ( $\downarrow$ )
OWSM CTC (3.2)	17.2	12.7	12.9	8.6
SpeechLM				
Pre-train	18.3	13.3	14.3	9.6
Finetune	17.8	13.1	13.4	9.0
CoT E2E ( $\hat{S}^{\text{asr}}$ )	17.6	13.1	19.5	14.1
SWBD + Fisher Train	22.2	17.0	$\times$	$\times$

further boosting TTS quality through the post-processing step (Sec. 4.2, “Post-process” in Tab. 3). Additionally, replacing Switchboard speaker prompts with high-quality prompts from Librispeech (UTMOS Score  $\approx 4.5$ ) during inference (“Spk prompt” in Tab. 3) results in substantial gains in both intelligibility and audio quality. For the E2E SpeechLM, we cannot compute intelligibility due to the absence of ground-truth text references. However, we observe that its synthesized audio quality is inferior to that of the CoT-based E2E model, further underscoring the advantages of multi-stage reasoning. Finally, our CoT-based E2E model, trained on a combined dataset of Switchboard and Fisher, significantly ( $p < 0.05$ ) outperforms all baselines and matches performance of the single-speaker VITS model, with added flexibility for multi-speaker speech.

**Conversation Level Analysis:** Tab. 4 shows a conversation-level analysis of our CoT-based E2E model, comparing its performance with various cascaded and E2E spoken dialogue systems. Our analysis reveals that CoT modeling enhances dialogue expressiveness, as shown by higher emotion similarity with ground-truth human responses. Additionally, our CoT model demonstrates strong parameter efficiency while delivering performance comparable to the modules in cascaded systems (Tab. 2, 3), showcasing its applicability to on-device scenarios. While its latency remains similar to cascaded systems due to multi-stage decoding, future work will focus on reducing latency through techniques like quantization.

**Ablation Study: Intermediate ASR task:** Tab. 5 report the ASR performance of our CoT spoken dialogue model, comparing it with the OWSM 3.1 [7] and pre-trained SpeechLM. The pre-trained SpeechLM performs competitively with OWSM, with fine-tuning further reducing WER. While CoT post-training slightly impacts ASR performance due to hallucinations, note that ASR serves only as an intermediate task. Importantly, Tab. 2 demonstrates that the response quality remains robust against ASR hallucinations.

## 6. Conclusion

We propose a CoT-based formulation for E2E spoken dialogue systems, achieving competitive performance with task-specific baselines, while producing more coherent responses with superior audio quality than conventional E2E models. Our CoT-based approach also surpasses cascaded systems in parameter efficiency and emotional expressiveness. Future work will explore methods to enable real-time interaction and support “speaking while listening” [18, 19].

## 7. Acknowledgement

Experiments of this work used the Bridges2 system at PSC and Delta system at NCSA through allocations CIS210014 and IRI120008P from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

## 8. References

- [1] K. Jokinen *et al.*, *Spoken dialogue systems*. Morgan & Claypool Publishers, 2009.
- [2] C. Breazeal *et al.*, “Social robots that interact with people,” *Springer handbook of robotics*, pp. 1349–1369, 2008.
- [3] J. Glass, “Challenges for spoken dialogue systems,” in *Proceedings of the 1999 IEEE ASRU Workshop*, MIT Laboratory for Computer Science Cambridge, vol. 696, 1999.
- [4] R. Huang *et al.*, “Audiogpt: Understanding and generating speech, music, sound, and talking head,” in *Proceedings of the AAAI*, 2024, pp. 23 802–23 804.
- [5] J. Wiseman, *Py-webrtcvad*, Accessed: 2024-12-10, 2024.
- [6] A. Radford *et al.*, “Robust speech recognition via large-scale weak supervision,” in *Proc. ICML*, 2023.
- [7] Y. Peng *et al.*, “Reproducing whisper-style training using an open-source toolkit and publicly available data,” in *Proc. ASRU*, 2023.
- [8] S. Mehri *et al.*, “Dialogue: A natural language understanding benchmark for task-oriented dialogue,” *arXiv preprint arXiv:2009.13570*, 2020.
- [9] B. Peng *et al.*, “Few-shot natural language generation for task-oriented dialog,” in *Findings of EMNLP 2020*, T. Cohn *et al.*, Eds., Online: Association for Computational Linguistics, Nov. 2020, pp. 172–182.
- [10] Coqui, *Introducing open-xts: An open-source toolkit for tts*, Accessed: 2024-12-10, 2024.
- [11] J. Godfrey *et al.*, “Switchboard: Telephone speech corpus for research and development,” in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, 517–520 vol.1.
- [12] H. Rashkin, “Towards empathetic open-domain conversation models: A new benchmark and dataset,” *arXiv preprint arXiv:1811.00207*, 2018.
- [13] K. Hara *et al.*, “Prediction of turn-taking using multitask learning with prediction of backchannels and fillers,” *Listener*, vol. 162, p. 364, 2018.
- [14] N. Ward *et al.*, “Prosodic features which cue back-channel responses in english and japanese,” *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.
- [15] S. Sundaram *et al.*, “Automatic acoustic synthesis of human-like laughter,” *The Journal of the Acoustical Society of America*, vol. 121, no. 1, pp. 527–535, 2007.
- [16] S. Fujie *et al.*, “A conversation robot with back-channel feedback function based on linguistic and nonlinguistic information,” in *Proc. ICARA Int. Conference on Autonomous Robots and Agents*, Citeseer, 2004, pp. 379–384.
- [17] D. Zhang *et al.*, “Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities,” *arXiv preprint arXiv:2305.11000*, 2023.
- [18] Z. Xie *et al.*, *Mini-omni: Language models can hear, talk while thinking in streaming*, 2024.
- [19] A. Défossez *et al.*, “Moshi: A speech-text foundation model for real-time dialogue,” *Kyutai, Tech. Rep.*, 2024.
- [20] J. Wei *et al.*, “Chain-of-thought prompting elicits reasoning in large language models,” *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022.
- [21] Z. Zhang *et al.*, “Automatic chain of thought prompting in large language models,” *arXiv preprint arXiv:2210.03493*, 2022.
- [22] J. Tian *et al.*, “ESPnet-SpeechLM: An open speech language model toolkit,” *arXiv*, 2024.
- [23] S. Ji *et al.*, *Wavchat: A survey of spoken dialogue models*, 2024.
- [24] Y. Chu *et al.*, “Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models,” *arXiv preprint arXiv:2311.07919*, 2023.
- [25] C. Fu *et al.*, “Vita: Towards open-source interactive omni-modal llm,” *arXiv preprint arXiv:2408.05211*, 2024.
- [26] W. Held *et al.*, “Distilling an end-to-end voice assistant without instruction training data,” *arXiv:2410.02678*, 2024.
- [27] D. Zhang *et al.*, “Speechgpt-gen: Scaling chain-of-information speech generation,” *arXiv preprint arXiv:2401.13527*, 2024.
- [28] T. A. Nguyen *et al.*, “Generative spoken dialogue language modeling,” *Trans. Assoc. Comput. Linguistics*, vol. 11, pp. 250–266, 2023.
- [29] B. Veluri *et al.*, “Beyond turn-based interfaces: Synchronous llms as full-duplex dialogue agents,” *arXiv preprint arXiv:2409.15594*, 2024.
- [30] Q. Fang *et al.*, “Llama-omni: Seamless speech interaction with large language models,” *arXiv preprint arXiv:2409.06666*, 2024.
- [31] Z. Meng *et al.*, “Parrot: Autoregressive spoken dialogue language modeling with decoder-only transformers,” in *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*.
- [32] Q. Zhang *et al.*, “Omniflatten: An end-to-end gpt model for seamless voice conversation,” *arXiv preprint arXiv:2410.17799*, 2024.
- [33] Z. Borsos *et al.*, “Audiolm: A language modeling approach to audio generation,” *IEEE/ACM TASLP*, vol. 31, pp. 2523–2533, 2023.
- [34] C. Cieri *et al.*, “The fisher corpus: A resource for the next generations of speech-to-text,” in *LREC*, vol. 4, 2004, pp. 69–71.
- [35] C.-Y. Lin, “Rouge: A package for automatic evaluation of summaries,” in *Text summarization branches out*, 2004, pp. 74–81.
- [36] S. Banerjee *et al.*, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in *ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
- [37] F. Jelinek *et al.*, “Perplexity—a measure of the difficulty of speech recognition tasks,” *The Journal of the Acoustical Society of America*, vol. 62, no. S1, S63–S63, 1977.
- [38] A. Radford *et al.*, “Language models are unsupervised multitask learners,” *OpenAI*, 2019.
- [39] J. Shi *et al.*, “Versa: A versatile evaluation toolkit for speech, audio, and music,” *arXiv preprint arXiv:2412.17667*, 2024.
- [40] T. Saeki *et al.*, “UTMOS: UTokyo-SaruLab system for voice-MOS challenge 2022,” in *Interspeech*, 2022, pp. 4521–4525.
- [41] Z. Ma *et al.*, “Emotion2vec: Self-supervised pre-training for speech emotion representation,” *Proc. ACL 2024 Findings*, 2024.
- [42] L. B. Allal *et al.*, *Smollm2: When smol goes big – data-centric training of a small language model*, 2025.
- [43] T. Hayashi *et al.*, “Espnet-tts: Unified, reproducible, and integratable open source end-to-end text-to-speech toolkit,” in *ICASSP, IEEE*, 2020, pp. 7654–7658.
- [44] S. Watanabe *et al.*, “ESPnet: End-to-end speech processing toolkit,” in *Proceedings of Interspeech*, 2018, pp. 2207–2211.
- [45] S. Arora *et al.*, “Espnet-slu: Advancing spoken language understanding through espnet,” in *ICASSP, IEEE*, 2022, pp. 7167–7171.
- [46] J. Copet *et al.*, “Simple and controllable music generation,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [47] J. Shi *et al.*, “Espnet-codec: Comprehensive training and evaluation of neural codecs for audio, music, and speech,” *arXiv preprint arXiv:2409.15897*, 2024.
- [48] W. Chen *et al.*, “Towards robust speech representation learning for thousands of languages,” *arXiv preprint arXiv:2407.00837*, 2024.
- [49] S. Arora *et al.*, “ESPnet-SDS: Unified toolkit and demo for spoken dialogue systems,” in *NAACL (System Demonstrations)*, Apr. 2025, pp. 248–259.