Align-SLM: Textless Spoken Language Models with Reinforcement Learning from AI Feedback

Guan-Ting Lin[1,2*] **Prashanth Gurunath Shivakumar**[1] **Aditya Gourav**[1]
Yile Gu[1] **Ankur Gandhe**[1] **Hung-yi Lee**[2] **Ivan Bulyko**[1]

[1]Amazon AGI, USA
[2]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan

Abstract

While textless Spoken Language Models (SLMs) have shown potential in end-to-end speech-to-speech modeling, they still lag behind text-based Large Language Models (LLMs) in terms of semantic coherence and relevance. This work introduces the **Align-SLM** framework, which leverages preference optimization inspired by Reinforcement Learning with AI Feedback (RLAIF) to enhance the semantic understanding of SLMs. Our approach generates multiple speech continuations from a given prompt and uses semantic metrics to create preference data for Direct Preference Optimization (DPO). We evaluate the framework using ZeroSpeech 2021 benchmarks for lexical and syntactic modeling, the spoken version of the StoryCloze dataset for semantic coherence, and other speech generation metrics, including the GPT4-o score and human evaluation. Experimental results show that our method achieves state-of-the-art performance for SLMs on most benchmarks, highlighting the importance of preference optimization to improve the semantics of SLMs.

1 Introduction

Significant strides have been made in Large Language Models (LLMs) by training decoder-only transformer models on vast amounts of text data. In speech processing, Textless NLP (Lakhotia et al., 2021; Kharitonov et al., 2022b; Nguyen et al., 2023; Lin et al., 2022) employs discrete speech units to train Spoken Language Models (SLMs) through next speech unit prediction. This approach is particularly promising, as SLMs are end-to-end speech-to-speech models that bypass the traditional cascaded pipeline of Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems, enabling joint optimization and real-time human-computer interaction. Furthermore, SLMs are applicable to all spoken languages, including those without written scripts, as they only require unlabeled speech data, thus promoting inclusivity in speech technology.

Despite increasing efforts to develop and improve SLMs—through text model initialization (Hassid et al., 2024; Shih et al., 2024), speech tokenizer design (Lakhotia et al., 2021; Hassid et al., 2024; Baade et al., 2024), text & speech token interleaving (Chou et al., 2023; Nguyen et al., 2024), scaling data and model (Hassid et al., 2024; Cuervo and Marxer, 2024)— a substantial gap remains between the understanding capabilities of text-based LLMs and SLMs. Current SLMs, when prompted, often produce speech continuations characterized by repetitive phrases, grammatical inaccuracies, and low relevance. Zhang et al. (2023); Nachmani et al. propose predicting text during intermediate decoding steps in a chain that mimics the ASR, LM, and TTS tasks within a single model. While these intermediate text steps improve the semantics of the generated speech, they still rely on text tokens as conditions to guide speech generation, and the additional decoding steps introduce latency, preventing real-time interactive SLMs. The question of ***whether textless SLMs can generate semantically relevant speech*** remains under-explored.

Most research on SLMs has relied *exclusively on next-speech-token prediction*. Few studies have explored alternative optimization objectives. Compared to text subwords, which on average carry more information, speech tokens are finer-grained and less compact. We argue that the next-speech-token prediction task may overlook long-term semantics, as loosely compressed speech units exhibit significant variability along spectral and temporal dimensions. Consequently, SLMs require a better training objective to effectively capture long-range semantics.

Our motivation stems from the observation that SLMs produce inconsistent results, sometimes gen-

[*]Work done during internship with Amazon AGI.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20395–20411
July 27 - August 1, 2025

erating high-quality speech continuations, while at other times producing suboptimal ones. ***Can we train SLMs to consistently generate better speech continuations while avoiding failures?*** Drawing inspiration from Reinforcement Learning with Human Feedback (RLHF) for text LLM alignment (Ouyang et al., 2022; Rafailov et al., 2024), we propose **Align-SLM**, the first framework that enhances the semantics of SLMs through RL. Starting with a pre-trained SLM (the open-sourced TWIST (Hassid et al., 2024) model), we generate multiple speech continuations from a given speech prompt. The next step is to create preference data (`prompt, chosen, rejected`) for preference optimization. Since collecting human preferences by listening is costly and time-consuming, following the concept of Reinforcement Learning from AI Feedback (RLAIF), we propose an automatic preference data selection strategy with LLM-guided semantic feedback. After preparing the preference data, Direct Preference Optimization (DPO) (Rafailov et al., 2024) is applied to learn from the feedback. Additionally, we couple the proposed technique with curriculum learning and demonstrate further improvements. The proposed framework is **pure speech-to-speech, data efficient**, and does **not require text injection** (Nguyen et al., 2024; Chou et al., 2023) or **text-to-speech synthesized speech** (Zhang et al., 2023).

We evaluate the SLM's performance using the sWUGGY and sBLIMP from ZeroSpeech 2021 (Nguyen et al., 2020) for lexical and syntactic modeling, and Spoken-StoryCloze and Topic-StoryCloze (Hassid et al., 2024) for textual nuances and continuation coherence. Additionally, we perform generative evaluations for speech continuation using (i) human listening tests and (ii) GPT-4 as a proxy for assessing semantic coherence and relevance. The results show that the proposed method achieves superior performance in semantic understanding and speech generation. The contributions can be summarized as follows:

- We propose the first preference optimization framework for textless SLMs, demonstrating that preference optimization is crucial for improving the semantics of SLMs.
- We develop an automated preference data selection strategy by designing effective semantic metrics to score preference data pairs.
- We couple DPO with curriculum learning by iteratively opting for higher criterion of preference data to further enhance performance.
- Align-SLM achieves the ***state-of-the-art*** performance for end-to-end spoken language models on Zerospeech and StoryCloze benchmark (**77.9%** on sWUGGY, **61.1%** on S-StoryCloze, and **86.8%** on T-StoryCloze) and achieves superior Meaningfulness Mean opinion scores with human evaluations.

2 Related Work

2.1 Spoken Language Models (SLMs)

Recent advancements in self-supervised representation learning and acoustic unit discovery convert continuous speech signals into discrete speech tokens (Polyak et al., 2021a). SLMs are end-to-end language models with discrete speech tokens, enabling speech continuation given a speech prompt. GSLM (Lakhotia et al., 2021) utilizes speech tokens to train a decoder-only language model and synthesize speech waveforms using a unit-based vocoder. pGSLM (Kharitonov et al., 2022b) injects prosodic tokens to enhance expressiveness. dGSLM (Nguyen et al., 2023) adopts a dual-tower model for two-channel spoken dialogue modeling.

Although SLMs can generate words and short-term phrases, the long-term semantics of the generated speech are often poor. TWIST (Hassid et al., 2024) proposes using a text-based LLM as initialization with large-scale training data to improve semantics. VoxtLM (Maiti et al., 2024) leverages both paired and unpaired speech in addition to text data for joint training of ASR, TTS, SLM, and Text LM. Another research direction is text token prediction as an intermediate step like SpeechGPT (Zhang et al., 2023) and SPECTRON (Nachmani et al.), which perform a chain of tasks (ASR $\rightarrow$ LM $\rightarrow$ TTS) in a single model. SUTLM (Chou et al., 2023) and SPIRIT-LM (Nguyen et al., 2024) utilize phrase-level interleaving for speech and text tokens. Concurrent works like LLaMA-Omni (Fang et al., 2024), Mini-Omni (Xie and Wu, 2024), and Moshi (Défossez et al., 2024) leverage simultaneous text token prediction to guide the speech generation. SyllableLM (Baade et al., 2024) recently propose to use syllable-level coarse speech tokens to improve SLM semantics. Compared to prior works focusing on multi-tasking in a single model or using text tokens to guide speech generation, this is the first

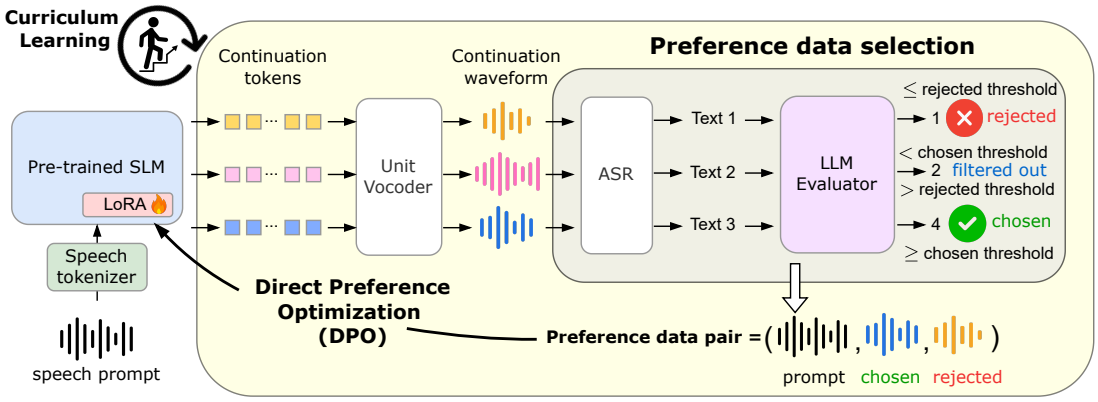

Figure 1: The illustration of the **Align-SLM** framework. Firstly, tokenize the speech prompt into speech tokens using a speech tokenizer. Then, generate and sample multiple speech continuations from the pre-trained SLM. To create the preference data pairs, the framework uses a unit-based vocoder to synthesize the speech continuations back to waveforms, an ASR model to transcribe the waveforms to text, and an LLM evaluator to rate the quality of the semantics. These preference data pairs are then used for direct preference optimization to train the LoRA adapter in the SLM. This alignment process can be coupled with curriculum learning to further improve performance.

work to *improve the long-term semantics of speech-only SLMs through preference optimization*.

2.2 Preference Optimization

Training the language model with *next-token prediction* is effective for learning human knowledge, but this objective might be different from the human preference. RLHF (Christiano et al., 2017; Ouyang et al., 2022) leverages an external reward model, combined with proximal policy optimization (Schulman et al., 2017), to align LLMs based on human feedback. DPO (Rafailov et al., 2024) proposes that LLMs can learn implicit rewards, allowing them to perform preference optimization independently, without the need for an external model. RLAIF (Bai et al., 2022; Lee et al., 2023) demonstrates that using AI as an alternative to human feedback can reduce labor costs and is more scalable, offering performance comparable to RLHF.

These methods show effectiveness in aligning an LLM with human preference. However, *there are limited studies on leveraging preference optimization in the speech and audio processing field*. Recently, Zhang et al. (2024); Chen et al. (2024) adopted preference optimization for the Text-to-Speech (TTS) model to align the quality of speech synthesis with human preference, *but not for enhancing SLMs' semantics*. Liao et al. (2024); Majumder et al. (2024) leverage preference optimization for text-to-audio generation with diffusion model, but the text-to-audio task is very different compared to SLM.

3 Align-SLM Framework

The illustration of the proposed framework is shown in Figure 1.

3.1 Spoken Language Models

The pre-trained SLM used in this work is TWIST (Hassid et al., 2024), a decoder-only transformer model that is trained on the next speech token prediction task with text model initialization. We utilize the TWIST model in two sizes (1.3B and 7B parameters) from the official release[1]. Specifically, the **speech tokenizer** consists of a self-supervised speech model (wen Yang et al., 2021; Lin et al., 2023) and K-means clustering. In this work, HuBERT (Hsu et al., 2021) is used and the cluster number K is set to 500. Notably, when continuous representations are clustered into discrete units, they primarily capture content information, which can be leveraged for modeling and understanding (Polyak et al., 2021b; Lin et al., 2022; Wu et al., 2023). This process first extracts 25Hz frame-level continuous representations from the 11-th layer of the HuBERT model, assigns each frame to its closest cluster index, and then de-duplicates consecutive identical indices to shorten the sequence. The **unit-based vocoder** is a HifiGAN-based (Kong et al., 2020) model that can convert the discrete units back into a continuous

[1]https://github.com/facebookresearch/textlesslib/tree/main/examples/twist (MIT license)

waveform. We use the model checkpoint from the textlesslib (Kharitonov et al., 2022a) library.

3.2 Automatic Preference Data Selection

To prepare the preference data pair (prompt, chosen, rejected), given the speech prompt x, the nucleus sampling (Holtzman et al., 2020) is used to generate N different continuations $y_1, y_2, ..., y_N$. Ideally, humans can listen to the samples and select the desirable and semantically correct one as the chosen continuation y_c, and the semantically incorrect one as the rejected continuation y_r. However, it is costly and time-consuming for human annotators to listen to the samples. Following the idea of RLAIF (Bai et al., 2022; Lee et al., 2023) to simulate human feedback, we propose an automatic preference data selection strategy to create preference data pairs. Since the focus is the semantics of SLMs, which is the content information in the speech, we first use `Whisper-large-v2` (Radford et al., 2023) to transcribe the speech into text, then measure the semantics of the transcribed text. In this work, we explore the two types of AI feedback from a text LLM. The text LLM is the open-sourced Mistral 7B (`instruct-v02`)[2].

3.2.1 Continuation Likelihood: Perplexity

Perplexity (PPL) is a common metric to measure the likelihood of a sentence given a pre-trained language model. In this work, PPL is calculated on the generated transcribed text conditioned on the ground truth text prompt. PPL is used in previous SLM works to evaluate the generation (Lakhotia et al., 2021; Hassid et al., 2024). However, Lakhotia et al. (2021) found out that SLMs sometimes generate *repeated phrases without clear meaning*, and the PPL would be extremely low with naively repeated phrases. To measure this, the auto-BLEU score (a) calculates the n-gram counting within the sentence. Given text sentence t and the set of n-gram $NG(t)$, auto-BLEU score of sentence t is $a_t = \frac{\sum_{s \in NG(t)} \mathbf{1}[s \in (NG(t) \setminus s)]}{|NG(t)|}$. 2-gram is used for auto-BLEU calculation (Lakhotia et al., 2021).

For the PPL of N continuations (PPL_N), we first filter out the auto-BLEU a_i higher than δ, then select the lowest PPL sample as y_c. The threshold of the auto-BLEU score is selected by the score distribution between ground truth continuation and the generated result (Please see the Appendix H for details). The y_r is the continuation with the highest PPL. The (y_c, y_r) is created as below:

$$y_i = \begin{cases} y_c & \text{if } PPL_i = \mathbf{min}(PPL_N) \cap a_i \leq \delta \\ y_r & \text{if } PPL_i = \mathbf{max}(PPL_N) \end{cases} \tag{1}$$

3.2.2 LLM Evaluation: Mistral Score

Instruction-tuned LLMs can follow instructions and understand semantics well (Chung et al., 2024). We propose using an LLM to judge the quality of the speech continuation, which evaluates the entire input and predicts the score. The prompt (see the Appendix for more details) is utilized to instruct the model to provide a score between 1 to 5 (1 denoting bad and 5 denoting good) based on the likelihood and meaningfulness of continuation given text prompt. Since we use the Mistral model for LLM evaluation, we call this "**Mistral score**", denoted as s. To let the model learn to distinguish the preferred and unpreferred continuations, a certain threshold is set for the Mistral score to ensure the difference in quality. s_c is the threshold of the chosen sample, and s_r is the threshold of the rejected sample. s_c should be larger than s_r. The auto-BLEU threshold is also used to recognize the naively repeated samples as rejected. We select the (y_c, y_r) as below:

$$y_i = \begin{cases} y_c & \text{if } s_i \geq s_c \cap a_i \leq \delta \\ y_r & \text{if } s_i \leq s_r \cup a_i > \delta \end{cases} \tag{2}$$

The s_c and s_r values are selected based on a preliminary analysis of the SLM's score distribution. For more details on the distribution of Mistral scores, please see Appendix I. We provide an example to illustrate the preference data selection in the Appendix C section.

3.3 Direct Preference Optimization for SLMs

In our framework, training with online metrics calculation is computationally infeasible due to the chain of models involved (vocoder, ASR, and LLM evaluator) and the computational complexity associated with sampling the SLM multiple times. Instead of calculating the reward online like RLHF, we adopt DPO, a simplified version of RLHF with implicit reward modeling, for preference optimization. The preference data pairs can be prepared offline, making the training more efficient. Additionally, DPO training is stable, simple, and does not require training an external reward model. The

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

Evaluation	Metrics	Pre-trained	Fine-tuned (Diff)	Align-SLM w/PPL (Diff)	Align-SLM w/Mistral score (Diff)
Proxy Metric	auto-BLEU ↓	2.80	2.43 (-0.37)	2.58 (-0.22)	**2.20** (-0.60)
	PPL ↓	114.9	116.3 (+1.4)	**52.1** (-62.8)	100.5 (-14.4)
	Mistral score ↑	1.66	1.70 (+0.04)	1.88 (+0.22)	**2.17** (+0.51)
Zerospeech	sBLIMP ↑	56.8	56.9 (+0.1)	**58.9** (+2.1)	58.1 (+1.3)
	sWUGGY ↑	71.8	70.9 (-0.9)	71.9 (+0.1)	**72.2** (+0.4)
StoryCloze	S-StoryCloze ↑	52.7	53.0 (+0.3)	52.6 (-0.1)	**54.3** (+1.6)
	T-StoryCloze ↑	69.7	70.7 (+1.0)	67.7 (-2.0)	**74.2** (+4.5)
Continuation	GPT4-o ↑	1.82	1.83 (+0.01)	1.85 (+0.03)	**2.06** (+0.24)
	MOSnet ↑	3.99	**4.08** (+0.09)	4.00 (+0.01)	3.98 (-0.01)

Table 1: Comparison of pre-trained model, fine-tuned model with next token prediction, Align-SLM with PPL, and Align-SLM with Mistral score on Zerospeech, StoryCloze, and speech continuation task using Librispeech test-clean. "Diff" means the value difference compared to the Pre-trained model. The number in darkgreen indicates improvement over pre-trained models' performance, while red color stands for performance degradation.

DPO training objective is

$$\mathcal{L}_{DPO} = -\mathbb{E}_{(x,y_c,y_r)\sim D}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_c|x)}{\pi_{ref}(y_c|x)} - \beta\log\frac{\pi_\theta(y_r|x)}{\pi_{ref}(y_r|x)}\right)\right] \quad (3)$$

where π_{ref} is the reference model with the pre-trained model's parameters, which is kept frozen. π_θ is the policy model trained with LoRA (Hu et al., 2022) adapter, while the parameters of the backbone pre-trained model are fixed. β controls the deviation from the reference model π_{ref}.

3.4 Coupling with Curriculum Learning

Curriculum Learning (CL) is a machine learning approach where models are trained by gradually increasing the complexity of tasks, allowing them to learn simpler concepts first before tackling more difficult ones (Bengio et al., 2009). In this work, we propose to couple DPO with curriculum learning to iteratively improve automated preference data selection. We iteratively raise the difficulty in discerning the preference data by tuning the thresholds s_c and s_r in Equation 2. Specifically for the Mistral score, we raise the s_c from 3 to 4 for chosen samples and s_r from 1 to 2 for rejected samples. With Curriculum learning, we expect the model to iteratively improve, given better feedback data.

4 Experiments

4.1 Dataset

We use **LibriSpeech** (Panayotov et al., 2015) as our primary dataset. We use the official training, dev-clean, and test-clean set. To further expand our dataset size, we leverage the English subset of the **Multilingual Librispeech (MLS)** (Pratap et al., 2020) as an additional training set, which is *around 3 times larger than the Librispeech training set*. We use a subset of the MLS data comprising 673K utterances in this work for data scaling, denoted as ***mls***. We apply the following data pre-processing steps to create the final training data:

1) Speech prompt segment selection using word alignment: Since our task involves speech continuation, the speech prompt should contain a sufficient amount of contextual information. We filter out samples shorter than 6 seconds. Unlike previous works that directly split the first 3 seconds as the prompt (Lakhotia et al., 2021; Hassid et al., 2024), we use forced alignment to select the closest word boundary around 3 seconds. This *avoids cutting off spoken words in the middle*, which could cause ASR errors in the speech continuations generated by the model. This potentially leads to poor perplexity or LLM evaluation scores.

2) Filtering out unsuitable chosen/rejected pairs: We apply a second layer of filtering over Mistral score annotations of ASR transcripts by thresholding chosen and rejection scores to ensure separability. Some samples fail to create preference data pairs due to these thresholds. When multiple continuations have the same lowest or highest score, we on-the-fly randomly choose among them. The number of preference data samples for different setups is listed in Table 6 in the Appendix.

4.2 Objective Evaluation

4.2.1 Zerospeech 2021 Benchmark

sWUGGY and sBLIMP metrics evaluate SLMs' lexical and syntactic modeling on pure speech input (Nguyen et al., 2020). **sWUGGY** tests if models prefer real words over phonetically similar non-words. We typically use the "in-vocab" split for reporting results, following the standard practice

#	Method	Zerospeech		StoryCloze		Speech Continuation		
		sBLIMP↑	sWUGGY↑	S-StoryCloze↑	T-StoryCloze↑	Mistral↑	GPT4-o↑	MOSnet↑
<1B								
1	GSLM (Lakhotia et al., 2021)	54.2	64.8	53.3	66.6	∅	∅	∅
2	AudioLM (Borsos et al., 2023)	**64.7**	71.5	∅	∅	∅	∅	∅
3	Cuervo and Marxer (2024)	61.3	∅	56.7	78.0	∅	∅	∅
4	SyllableLM (Baade et al., 2024)	63.7	72.2	∅	75.4	∅	∅	∅
1.3B								
5	VoxtLM (Maiti et al., 2024)	57.1	66.1	∅	∅	∅	∅	∅
6	TWIST (Hassid et al., 2024)	57.0	72.7	52.4	70.6	∅	∅	∅
7	Pre-trained TWIST*	56.8	71.8	52.7	69.7	1.66	1.82	3.99
8	Align-SLM	58.1	72.2	54.3	74.2	2.17	2.06	3.98
9	Align-SLM + CL	58.2	72.2	53.9	76.1	2.35	2.29	4.05
10	Align-SLM-*mls*	59.0	72.7	54.0	76.7	2.37	2.34	4.00
11	Align-SLM-*mls* + CL	59.8	72.7	55.0	80.0	2.50	2.43	3.94
7B								
12	Moshi (Défossez et al., 2024)	58.8	72.6	60.8	83.0	∅	∅	∅
13	SPIRIT-LM (Nguyen et al., 2024)	58.3	69.0	61.0	82.9	∅	∅	∅
14	TWIST (Hassid et al., 2024)	59.0	73.9	55.3	74.1	∅	∅	∅
15	Pre-trained TWIST*	58.8	73.5	55.1	75.4	2.03	2.70	3.80
16	Align-SLM	61.1	75.3	59.1	83.8	2.89	3.50	4.08
17	Align-SLM + CL	61.4	75.5	58.2	85.6	**3.22**	**3.56**	**4.09**
18	Align-SLM-*mls*	62.2	77.5	58.6	85.6	2.92	3.46	4.02
19	Align-SLM-*mls* + CL	62.3	**77.9**	**61.1**	**86.8**	3.11	3.50	3.99
Cascade Topline and Human Performance								
20	ASR+LLM (Nguyen et al., 2024)	71.6	79.2	75.7	94.8	∅	∅	∅
21	Human (Hassid et al., 2024)	∅	∅	79.2	90.2	∅	∅	∅
22	Ground Truth Continuation	∅	∅	∅	∅	2.89	4.25	4.02

Table 2: Performance on Zerospeech, StoryCloze, and Speech Continuation task. The methods marked with underline indicate they are trained with paired speech and text tokens, *not the speech-only model*. "Pre-trained TWIST*" uses open-sourced TWIST model checkpoint but we use `bf16` precision for inference, so the performance is slightly different compared to published "TWIST" results. "ASR+LLM" is using Whisper (Radford et al., 2023) as ASR model with Llama 2 model (Touvron et al., 2023), reported by Nguyen et al. (2024).

established by Lakhotia et al. (2021). **sBLIMP** assesses grammaticality judgments between correct and incorrect sentences. Both metrics compare geometric means of sequence probabilities assigned to paired utterances.

4.2.2 Spoken StoryCloze Benchmark

StoryCloze benchmarks (Mostafazadeh et al., 2017) evaluate the model's ability to identify the more plausible ending among two scenarios given a short story as a prompt. This requires a degree of high-level semantic understanding and common sense. We utilize the spoken version of the original StoryCloze (**S-StoryCloze**) as well as the topic-based StoryCloze (**T-StoryCloze**) created by Hassid et al. (2024). T-StoryCloze uses simpler negative samples that are randomly drawn while S-StoryCloze uses adversarially curated negative samples. The random baseline performance for the above tasks is 50%. We name the spoken version of StoryCloze as "**StoryCloze**" for simplicity, but note that this is different from the text StoryCloze.

4.2.3 Generative Speech Continuation

GPT4-o score: GPT4-o (OpenAI, 2023) has shown remarkable text understanding performance and can serve as alternative human evaluators (Chiang and Lee, 2023a; Liu et al., 2023; Chiang and Lee, 2023b), showing high correlation with human judgments. We leverage GPT4-o as a proxy for human evaluations. Following *llm evaluation* (Chiang and Lee, 2023a), the instruction first analyzes the sentence and then provides the score from 1 to 5, to judge the semantic coherence, meaningfulness, and grammatical correctness of the ASR transcribed continuation given a prompt (1 denoting bad and 5 denoting good). The instruction prompt is shown in Appendix G.

MOSnet score: To measure the audio quality, we utilize the MOSnet (Cooper et al., 2022) to predict the Mean Opinion Score (MOS) of audio quality. Specifically, MOSnet is based on self-supervised wav2vec 2.0 (Baevski et al., 2020), fine-tuned on MOS prediction task[3]. The model has shown a high correlation with human MOS scores and good generalization ability for unseen data.

[3]https://github.com/nii-yamagishilab/mos-finetune-ssl

Method	MMOS↑
Target Re-synthesized	3.50 ± 0.07
Pre-trained TWIST 7B	3.48 ± 0.07
Align-SLM 7B + CL	**3.73 ± 0.06**

Table 3: Meaningfulness MOS score. We report the MMOS score as the mean ± 95% confidence interval of the standard deviation. "Target Re-synthesized" uses the speech tokens from the original continuation and re-synthesize back to the waveform.

4.3 Subjective Evaluation

We conducted human listening evaluations to assess the meaningfulness of the generated speech. We follow Lakhotia et al. (2021); Kharitonov et al. (2022b) to use the **M**eaningfulness **M**ean **O**pinion **S**core (**MMOS**). Specifically, given a speech prompt and generated speech continuations, evaluators listen to audio samples and rate the meaningfulness in terms of relevance, coherence, and grammatical correctness. We randomly sample 100 speech prompts from the Librispeech test-clean set for evaluation. Each sample has 10 evaluators to provide the rating on a scale between 1 to 5 with an increment of 1. The human evaluation template and instruction are shown in Appendix B. CrowdMOS (Ribeiro et al., 2011) package is used for outlier removal (Lakhotia et al., 2021).

4.4 Baselines

We compare Align-SLM with other SLMs on the Zerospeech and StoryCloze benchmarks. The most comparable baselines are *speech-only SLMs*, including GSLM (Lakhotia et al., 2021), AudioLM (Borsos et al., 2023), TWIST (Hassid et al., 2024), SyllableLM (Baade et al., 2024), and the model from Cuervo and Marxer (2024). Among these, the TWIST model is initialized from a text-based LLM. Additionally, we compare the performance against *SLMs leverage text modality* (Table 2 with underline), specifically VoxtLM (Maiti et al., 2024) with multi-task training, SPIRIT-LM (Nguyen et al., 2024) with speech-text interleaving, and Moshi (Défossez et al., 2024), which leverages text-guided speech generation.

5 Results

5.1 Mistral Score Provides Better Semantic Feedback Than Perplexity

To determine which preference data selection strategy is beneficial for SLM's semantics, we first conduct preliminary experiments on Align-SLM with PPL and Mistral score in Table 1 using the TWIST 1.3B model. Additionally, we continually fine-tune the pre-trained model using the same data, which serves as the baseline for the next speech token prediction. "Proxy Metric" refers to the metrics used for preference data selection, while "Zerospeech", "StoryCloze", and "Speech Continuation" are zero-shot speech evaluation metrics.

Align-SLM w/PPL successfully improves the proxy metrics for auto-BLEU and PPL, but the performance on speech continuation is slightly worse than the pre-trained model. As for the Zerospeech and StoryCloze benchmark, Align-SLM w/PPL has marginal improvement on most metrics and degrades on T-StoryCloze. Particularly, perplexity feedback shows a much greater improvement on sBLIMP, which measures grammatical correctness. This finding suggests that optimizing toward perplexity might overly focus on grammar rather than general semantics and relevance.

On the other hand, *Align-SLM w/Mistral score significantly outperforms the pre-trained model across metrics*. Specifically, the performance on the Zerospeech and StoryCloze benchmarks improve significantly (+1.6 on S-StoryCloze and +4.5 on T-StoryCloze). The generated continuations also yield a better GPT4-o score, indicating the generated content is more relevant and coherent to the speech prompt. Regarding proxy metrics, the auto-BLEU and Mistral scores are improved, whereas the PPL is similar to the pre-trained model.

This finding suggests that LLM evaluations, such as the Mistral score, provide general semantic feedback and achieve superior performance across benchmarks. Furthermore, Align-SLM w/Mistral score significantly outperforms the fine-tuned baseline, which only marginally improves performance. Therefore, *in the following experiments for Align-SLM, we use the* ***Mistral score as the AI feedback***.

5.2 Consistently Improves Pre-trained SLMs

Given our findings from Section 5.1, we use the Mistral score as a "proxy metric" for the rest of our experiments. In Table 2, we observe that Align-SLM consistently outperforms the pre-trained model on the Zerospeech, StoryCloze, and speech continuation task for both the 1.3B and 7B models (rows 7 to 8 for 1.3B, row 15 to 16 for 7B). For instance, on T-StoryCloze, training Align-SLM with Librispeech data yields relative improvements of 6.5% and 11.1% for the 1.3B and 7B models,

respectively. Additionally, Table 2 demonstrates Align-SLM 7B performs better in speech continuation, as reflected by the GPT-4 score improving from 2.70 to 3.50.

5.3 Improvement of Curriculum Learning

After training the pre-trained model with the first iteration of Align-SLM, we consider the resulting model a stronger starting point and generate new preference data with more stringent preference data selection criteria. Results in Table 2 indicate that curriculum learning improves most metrics on the Zerospeech and StoryCloze benchmarks (rows 8 to 9 for 1.3B, rows 16 to 17 for 7B), particularly for T-StoryCloze, which requires fine-grained relevance between the speech story prompt and its continuation. Additionally, we observe improvements in speech continuation; for example, the GPT4-o score increases from 2.06 to 2.29 for the 1.3B model. We also experimented with further increasing the number of curriculum learning iterations, which continued to enhance performance (see the discussion in Appendix E).

5.4 Amount of Preference Data

In the previous experiments, we only use the LibriSpeech training data for DPO training. After applying filtering as described in Section 4.1, the number of samples is around 39K and 63K for the 1.3B model and 7B model, respectively. To investigate whether additional data helps the Align-SLM to learn semantics, we scale up the data around three times by including a subset of MLS (*mls*). It is worth noting that the scale of the MLS subset is still much smaller than the pre-training data used in Hassid et al. (2024). Table 2 shows that that with more data, the **model learns significantly better semantics across model sizes and benchmarks** (row 10 to 11 for 1.3B, row 18 to 19 for 7B). Nevertheless, we observe that adding *mls* data for the 7B SLM yields little improvement in GPT4-o score compared to 1.3B model[4].

5.5 Comparison with Baselines

Align-SLM-*mls*-CL achieves **state-of-the-art performance** for textless SLMs in T-StoryCloze (**86.8**), S-StoryCloze (**61.1**), and sWUGGY (**77.9**), **even surpassing *text-guided approaches* (moshi) and *speech-text interleaving methods* (SPIRIT-LM)**. The performance on T-StoryCloze is close to human-level accuracy (90.2). However, for sBLIMP (grammatical correctness), AudioLM and SyllableLM achieve the best performance, possibly due to their superior design of speech tokens. Compared to the cascaded topline (ASR+LLM), end-to-end SLMs still have room for improvement.

5.6 Human Prefer Align-SLM's Generation

Table 3 presents the MMOS scores from the subjective evaluation. We compare the re-synthesized speech of the original continuation, the pre-trained TWIST 7B, and the proposed Align-SLM 7B with CL. The results show that human evaluators perceive Align-SLM as generating more meaningful speech continuations than the pre-trained model, and even surpassing the original continuation. This can be attributed to the fact that the original continuation, derived from audiobook content, may rely on the broader context, whereas Align-SLM learns to generate more relevant and meaningful content based on the speech prompt.

5.7 Impact on Audio Quality

Preference optimization can potentially adversarially exploit the ASR and LLM evaluators, leading the SLM to generate nonsensical or noisy speech with artificially high rewards. To address this concern, we evaluate the audio quality of generated speech using MOSnet (Cooper et al., 2022), which has shown a high correlation with human judgments. Table 2 presents the MOS scores from LibriSpeech dev-clean and test-clean sets. Results indicate that the MOSnet scores for Align-SLM are comparable to or slightly higher than those of the pre-trained SLM, suggesting that the training of Align-SLM preserves audio quality. The proposed framework requires the generated speech to pass through the ASR model for speech-to-text conversion, natural and clear speech is essential to avoid speech recognition errors. Consequently, in some cases, the MOSnet score for Align-SLM is even higher than that of the pre-trained SLM.

6 Conclusion

This work introduces **Align-SLM**, a novel framework that significantly enhances the semantics of SLM via preference optimization. By utilizing LLM-guided semantic feedback and direct preference optimization, Align-SLM achieves state-of-

[4]This can be attributed to the amount of preference data for the 7B model from Librispeech already being sufficient (63K for the first iteration and 71K for the second iteration). In contrast, the 1.3B model only has 39K and 20K preference data pairs by Librispeech.

the-art performance of SLMs across various benchmarks and generative tasks, consistently outperforming pre-trained SLMs. The framework demonstrates superior results with the proposed LLM evaluation feedback and curriculum learning. This work highlights the critical role of preference optimization for SLM and paves the way for better end-to-end speech-to-speech models.

Limitations and Future Works

Limited to Semantics of SLMs: This work investigates only the semantic aspect of SLMs, though other aspects such as speaking styles, paralinguistics, and prosody are also important for spoken dialogue (Lin et al., 2024b,a). Our Align-SLM framework can be generalized to these aspects, but that is beyond the scope of this paper.

Integrating Align-SLM with other SLMs: We integrated the open-source pre-trained TWIST (Hassid et al., 2024) model with Align-SLM, demonstrating significant performance improvements. As the SLM community grows rapidly, more powerful pre-trained SLMs are emerging (Défossez et al., 2024; Baade et al., 2024; Fang et al., 2024). The Align-SLM framework is a general framework that can be extended to integrate with other SLMs.

Language Support: The current model is trained solely on English speech data. Future work can extend the Align-SLM framework to multilingual speech data for more inclusive speech technology. For unwritten languages, instead of using ASR models for text transcription, we can use speech translation models to convert unwritten spoken languages into high-resource languages (Chen et al., 2023), similarly obtaining semantic feedback from the LLM.

Diverse Data and Model Scaling: The dataset used in this work is much smaller than typical pre-training datasets, and the model size is relatively small compared to text-based LLMs. Additionally, the training data is limited to the audiobook domain. Expanding the training data to include more diverse domains in future work could lead to improved model performance.

References

Alan Baade, Puyuan Peng, and David Harwath. 2024. Syllablelm: Learning coarse semantic units for speech language models. *Preprint*, arXiv:2410.04029.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533.

Chen Chen, Yuchen Hu, Wen Wu, Helin Wang, Eng Siong Chng, and Chao Zhang. 2024. Enhancing zero-shot text-to-speech synthesis with human feedback. *arXiv preprint arXiv:2406.00654*.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, et al. 2023. Speech-to-speech translation for a real-world unwritten language. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4969–4983.

Cheng-Han Chiang and Hung-Yi Lee. 2023a. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631.

Cheng-Han Chiang and Hung-yi Lee. 2023b. A closer look into using large language models for automatic evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8928–8942, Singapore. Association for Computational Linguistics.

Ju-Chieh Chou, Chung-Ming Chien, Wei-Ning Hsu, Karen Livescu, Arun Babu, Alexis Conneau, Alexei Baevski, and Michael Auli. 2023. Toward joint language modeling for speech units and text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6582–6593, Singapore. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi

Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Erica Cooper, Wen-Chin Huang, Tomoki Toda, and Junichi Yamagishi. 2022. Generalization ability of mos prediction networks. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8442–8446. IEEE.

Santiago Cuervo and Ricard Marxer. 2024. Scaling properties of speech language models. *arXiv preprint arXiv:2404.00685*.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. Technical report, Kyutai.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2024. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Eugene Kharitonov, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Paden Tomasello, Ann Lee, Ali Elkahky, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2022a. textless-lib: a library for textless spoken language processing.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, et al. 2022b. Text-free prosody-aware generative spoken language modeling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8666–8681.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, et al. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Huan Liao, Haonan Han, Kai Yang, Tianjiao Du, Rui Yang, Zunnan Xu, Qinmei Xu, Jingquan Liu, Jiasheng Lu, and Xiu Li. 2024. Baton: Aligning text-to-audio model with human preference feedback. *arXiv preprint arXiv:2402.00744*.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. 2024a. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6626–6642, Bangkok, Thailand. Association for Computational Linguistics.

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung yi Lee, and Lin shan Lee. 2022. DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering. In *Proc. Interspeech 2022*, pages 5165–5169.

Guan-Ting Lin, Chi-Luen Feng, Wei-Ping Huang, Yuan Tseng, Tzu-Han Lin, Chen-An Li, Hung-yi Lee, and Nigel G. Ward. 2023. On the utility of self-supervised models for prosody-related tasks. In *Proc. IEEE SLT*, pages 1104–1111.

Guan-Ting Lin and Hung-yi Lee. 2024. Can llms understand the implication of emphasized sentences in dialogue? *arXiv preprint arXiv:2406.11065*.

Guan-Ting Lin, Prashanth Gurunath Shivakumar, Ankur Gandhe, Chao-Han Huck Yang, Yile Gu, Shalini Ghosh, Andreas Stolcke, Hung-Yi Lee, and Ivan Bulyko. 2024b. Paralinguistics-enhanced large language modeling of spoken dialogue. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10316–10320.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval:

Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. 2024. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. *arXiv preprint arXiv:2404.09956*.

Nasrin Mostafazadeh, Michael Roth, Annie Louis, Nathanael Chambers, and James Allen. 2017. LSDSem 2017 shared task: The story cloze test. In *Proceedings of the 2nd Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pages 46–51, Valencia, Spain. Association for Computational Linguistics.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. In *The Twelfth International Conference on Learning Representations*.

Tu Anh Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Evgeny Kharitonov, Alexei Baevski, Ewan Dunbar, and Emmanuel Dupoux. 2020. The zero resource speech benchmark 2021: Metrics and baselines for unsupervised spoken language modeling. In *NeuRIPS Workshop on Self-Supervised Learning for Speech and Audio Processing*.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, et al. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, et al. 2024. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*.

OpenAI. 2023. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021a. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*, pages 3615–3619.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021b. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. 2011. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2416–2419.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Min-Han Shih, Ho-Lam Chung, Yu-Chi Pai, Ming-Hao Hsu, Guan-Ting Lin, Shang-Wen Li, and Hung yi Lee. 2024. Gsqa: An end-to-end model for generative spoken question answering. In *Interspeech 2024*, pages 2970–2974.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shu wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Guan-Wei Wu, Guan-Ting Lin, Shang-Wen Li, and Hung yi Lee. 2023. Improving Textless Spoken Language Understanding with Discrete Units as Intermediate Target. In *Proc. INTERSPEECH 2023*, pages 1503–1507.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics.

Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2024. Speechalign: Aligning speech generation to human preferences. *arXiv preprint arXiv:2404.05600*.

Appendix

A Implementation Details

For sampling the speech continuations, we use a decoding temperature of 0.8, following the original sampling setup in Hassid et al. (2024). The number of sampled continuations, N, is set to 5. The hyperparameters of the LoRA adapter are set with a rank of 32 and an alpha value of 8. For training on LibriSpeech, models are trained with a batch size of 512, using a peak learning rate of 1e-6 that linearly decays over 100k steps with 500 warmup steps. For *mls* data, since there are more training samples, we further increase the training steps to 300k. For curriculum learning, the peak learning rate is 1e-6 and linearly decays without the warmup step. We utilize 64 NVIDIA A100 GPUs for training. The model checkpoint is selected by the highest reward accuracy on the Librispeech dev-clean set.

B Details of Subjective Evaluation

We used the Amazon Mechanical Turk crowdsourcing platform for subjective evaluations. The evaluators, all based in the United States, had HIT approval rates above 98%, as the speech samples were in American English. Each evaluation task contained 10 samples, and annotators were paid 0.50 USD per task. On average, each task took about 2 minutes to complete.

Prompt: *a man was looking in from the corridor behind*

#	Sampled speech continuation	Score
1	He was seen, he stopped before the man. He said, *I have a book you read.*	3
2	When they drew near and received their letter, and looking down, showed them something of the startling effect of the diamond.	1
3	But every time he reached the open window he saw the little man pressing in a silk sash.	2
4	When he heard the dog talking to him. Said, *My father, did you hear the dog?* Said he, *did you...*	1
5	And the door opened, and he heard the words. And there they all began, Boyce said to himself.	3

Table 4: Prompt, sampled continuations, and their MISTRAL scores.

C Preference-based selection illustration

Given the speech prompt "***a man was looking in from the corridor behind***", we draw five candidate continuations from the model and assign each a Mistral score (higher is better). Table 4 lists the continuations and their scores. We keep a continuation if $s_i \geq s_c = 3$, reject it if $s_i \leq s_r = 1$, and otherwise mark it as *filtered* (neither accepted nor explicitly rejected). In this example, continuations (1) and (5) are kept; continuations (2) and (4) are rejected due to clear irrelevance; and continuation (3) is filtered. Notably, the lowest-scoring continuations (score 1) tend to be semantically unrelated to the prompt, underscoring the value of the LLM feedback.

D Speech Continuation Examples

We present some qualitative examples of speech continuations in Table 5. We observe that the pre-trained TWIST 7B model occasionally generates nonsensical speech based on the given speech prompt. For instance, in the first example, the TWIST continuation includes the repetitive phrase "only, easy, only, and" at the end.

E Number of Iteration for Curriculum Learning

In this section, we examine how the number of curriculum learning iterations influences performance. Figure 2 presents the results from the pre-trained

Speech Prompt	Pretrained TWIST Continuation	Align-SLM Continuation
He came down to earth lived among men suffered	for a friend, she would have been glad of the last word, the pointed word, possible only, easy, only, and.	died, and yet somehow he felt that he lived more than they, and he came to know men of his own who loved him still.
Then there were three or four leading men of the community	and the majority of these were hirelings, hard-headed and stolid.	with them at the table, and they spoke together about the measures they were preparing for the day's work. But one among them was speaking feelingly.
In the early morning of the next day may twenty ninth	W. H. Barker, the prisoner before the magistrate and the charge of the prosecution,	we went to our rooms and sat down and had a long talk over all the present circumstances of our journey to the place where we had spent so much time.

Table 5: Continuation examples from Pretrained TWIST 7B model and Align-SLM 7B + CL. The transcribed texts are shown in the table.

7B model after one to three iterations of Align-SLM with curriculum learning using Librispeech data. For the first, second, and third iterations, (s_r, s_c) is $(1, 3)$, $(2, 4)$, and $(3, 5)$, respectively. We observe that increasing the number of curriculum learning iterations gradually improves the SLM's performance across all benchmarks. While the T-StoryCloze performance saturates after the second iteration, other metrics show their best performance in the third iteration.

F Instruction for the Mistral LLM for evaluation

The [text_prompt] and [generated_transcription] are the placeholder for different input. The entire instruction is shown as following:

```
Rate the text continuation based on how likelihood is the text continuation given the text prompt. You should also consider whether the meaning of the text continuation is making sense. Don't be too strict. The text prompt is [text_prompt], and the text continuation is [generated_transcription].
You must give an overall rating from 1 to 5. The rating guideline is
1: Very Unlikely and Irrelevant;
2: Unlikely and Marginally Relevant;
3: Moderately Likely and Relevant;
4: Likely and Relevant;
5: Very Likely and Highly Relevant.
Output format is: I would rate the score as [NUMBER]
```

G Instruction for the GPT4-o for evaluation

```
The task is evaluating the relevance and likelihood of the predicted text continuation, given the text prompt. You should also consider whether the meaning of the text continuation is making sense. The text prompt is: [text_prompt], and the text continuation is :[generated_audio_transcription].

You must give an overall rating from 1 to 5. The rating guideline is as below:
1: The text continuation is very unlikely and irrelevant to the text prompt.
2: The text continuation is unlikely and marginally relevant to the text prompt.
3: The text continuation is moderately likely and relevant to the text prompt.
4: The text continuation is likely and relevant to the text prompt. 5: The text continuation is very likely and highly relevant.
You should take the following steps to provide the score:
First: briefly analyze the sample with the above definition.
Second: MUST follow the output format as: I would rate the score as _
```

H auto-BLEU Score Distribution

Table 5 presents the auto-BLEU score distribution on of the TWIST 1.3B model and the target continuation. We observe that most target continuations have an auto-BLEU score lower than 0.1,

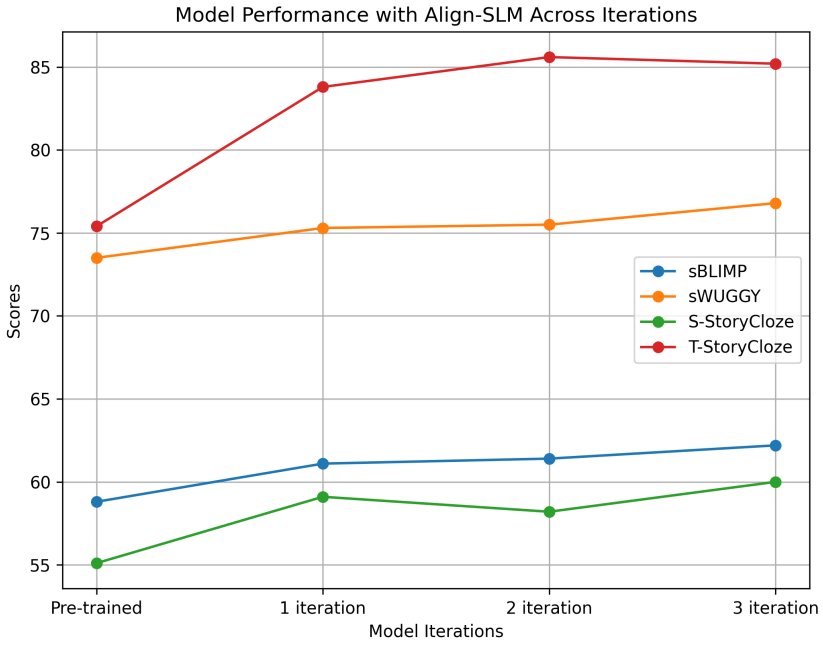

Figure 2: The 7B model undergoes one, two, and up to three iterations of curriculum learning with Align-SLM using Librispeech data.

while TWIST's generated continuations show auto-BLEU scores ranging from 0.1 to 0.3 (or even higher). This difference suggests that an auto-BLEU score higher than 0.1 may not indicate a good continuation. Therefore, we set the auto-BLEU score threshold (δ) at 0.1 to distinguish the chosen and rejected continuations.

I Mistral Score Distribution

Given N generated speech continuations, selecting an appropriate scoring threshold is crucial for determining which samples are chosen or rejected. Thresholds that are too high or too low fail to create effective preference data pairs. In this work, we first analyze the score distribution of the pre-trained SLMs, as shown in Figures 3 and 4. We observe that the Mistral score is 2 with more than a 50% probability for both the 1.3B and 7B pre-trained SLMs. Therefore, a score of 2 is considered the norm, a score of 1 represents rejected cases, and scores of 3 or higher are considered chosen samples. After the initial Align-SLM DPO training, the score distribution shifts towards higher scores, reducing the proportion of low scores. Increasing the s_c and s_r values in the second iteration can further improve the overall ratings.

Dataset	Iter 1	Iter 2
1.3B		
Librispeech	39,566	20,942
+ *mls*	131,071	113,243
7B		
Librispeech	63,107	71,944
+ *mls*	234,195	262,929

Table 6: Number of data samples after filtering. Note that different models have different filtered samples since the filtering depends on the quality of the generated speech. The number of samples of original Librispeech and *mls* are 247K and 673K, respectively.

J Correlation Between GPT4-o score and MMOS

We calculate the Pearson's correlation coefficient between the GPT4-o score and MMOS score. Pearson's coefficient is 0.51, suggesting that the GP4-o score correlates well with the human rating. The finding is aligned with previous works that the rating from GPT4 can be alternative human evaluators (Chiang and Lee, 2023a; Liu et al., 2023; Lin and Lee, 2024).

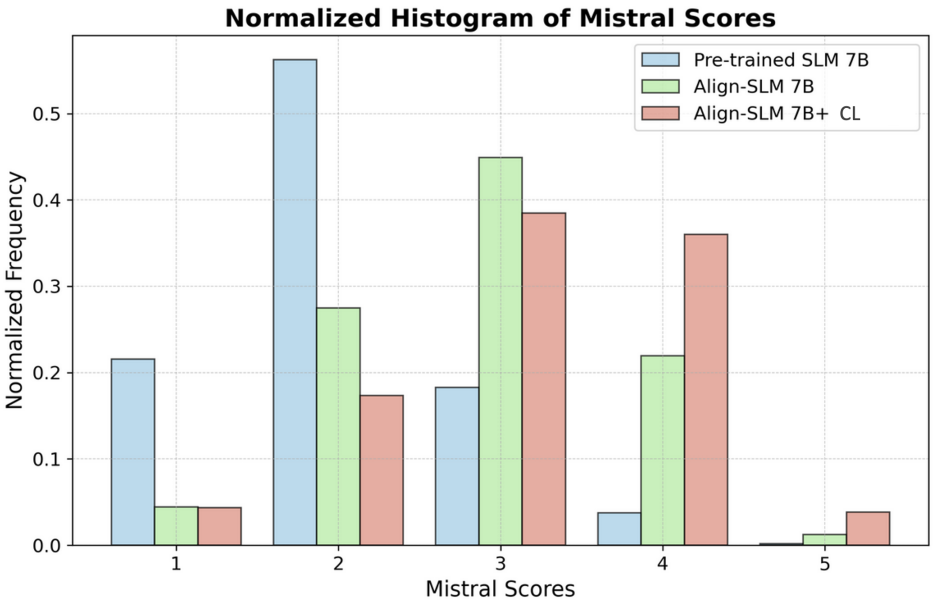

Figure 3: The normalized histogram of Mistral score distribution on Librispeech dev-clean with 7B model.

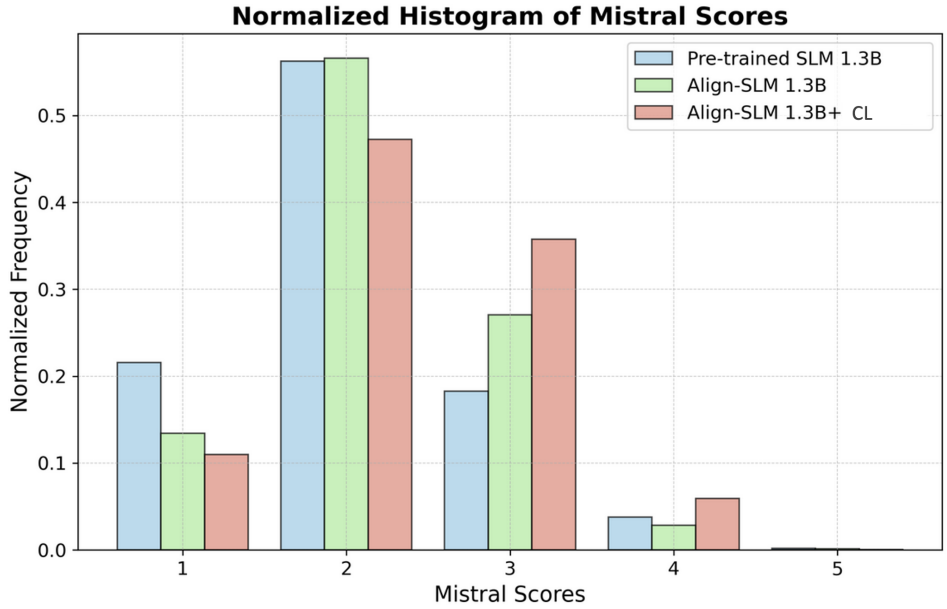

Figure 4: The normalized histogram of Mistral score distribution on Librispeech dev-clean with 1.3B model.

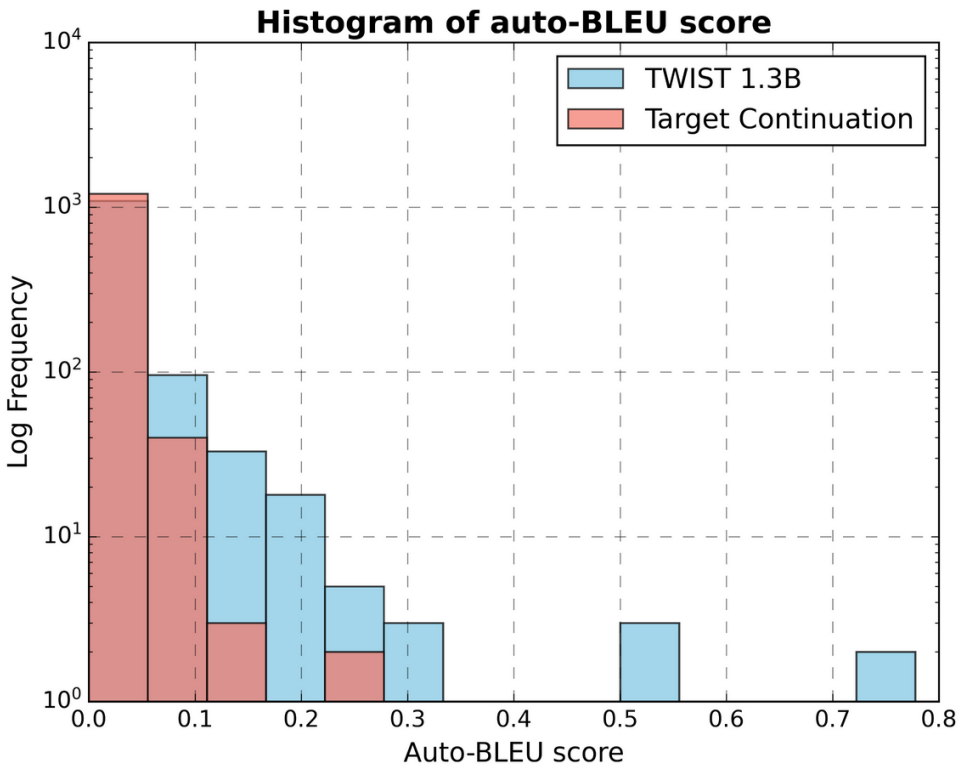

Figure 5: The log normalized histogram of auto-BLEU score distribution on Librispeech dev-clean with 1.3B TWIST model.

Instructions for speech meaningfulness evaluation

(This is the listening evaluation. Please read through the instruction carefully, and wearing the headphone when doing the evaluation.)

Disclaimer: If you are not the native English speaker, or you don't have headphone to listen carefully, please skip this task.

Introduction and Instruction

In this task, you will hear samples of speech recordings, composed of the [PROMPT] and [CONTINUATION].
- [PROMPT] is a sentence beginning.
- [CONTINUATION] is a few more words that continue this prompt.

The [CONTINUATION] has been generated by a computer and your task will be to concentrate and evaluate the **relevance, coherence, meaningfulness and grammatical correctness** of the [CONTINUATION] with respect to the [PROMPT] on a 1 to 5 scale, (**irrespective of audio quality, sound clarity or intonation**).

You should give a score according to the following scale:

Score	relevance, coherence, meaningfulness and grammatical correctness
5	Excellent
4	Good
3	Fair
2	Poor
1	Bad

IMPORTANT:

- Use the **headphone** to listen carefully to the content in speech recording.

- You can repeatedly play the audio recording until you are sure about your decision.

- Note that some of the [CONTINUATION] sentences **are truncated unexpectedly in the end, which is OK and should not affect the score.**

- We will reject the submission that is obviously spamming or not following the instructions and guidelines.

[Sample 1]

PROMPT:

CONTINUATION:

How **relevance, coherence, meaningfulness and grammatical correctness** of the [CONTINUATION] with respect to the [PROMPT] (**irrespective of audio quality, sound clarity or intonation**)

- ○ (1) Bad
- ○ (2) Poor
- ○ (3) Fair
- ○ (4) Good
- ○ (5) Excellent

Figure 6: The template and instruction of the subjective evaluation of the MMOS score.