SPOKEN QUESTION ANSWERING AND SPEECH CONTINUATION USING SPECTROGRAM-POWERED LLM

Eliya Nachmani[1,*], **Alon Levkovitch**[1,3,*,†], **Roy Hirsch**[2] , **Julian Salazar**[1], **Chulayuth Asawaroengchai**[1], **Soroosh Mariooryad**[1], **Ehud Rivlin**[2] **RJ Skerry-Ryan**[1], **Michelle Tadmor Ramanovich**[1]
[1]Google Research, [2]Verily AI, [3] Tel-Aviv Univeristy
{eliyn, alevkovitch, royhirsch}@google.com

ABSTRACT

We present a novel approach to adapting pre-trained large language models (LLMs) to perform question answering (QA) and speech continuation. By endowing the LLM with a pre-trained speech encoder, our model becomes able to take speech inputs and generate speech outputs. The entire system is trained end-to-end and operates directly on spectrograms, simplifying our architecture. Key to our approach is a training objective that jointly supervises speech recognition, text continuation, and speech synthesis using only paired speech-text pairs, enabling a 'cross-modal' chain-of-thought within a single decoding pass. Our method surpasses existing spoken language models in speaker preservation and semantic coherence. Furthermore, the proposed model improves upon direct initialization in retaining the knowledge of the original LLM as demonstrated through spoken QA datasets. Audio samples and spoken QA dataset can be found on the project website.

1 INTRODUCTION

The goal of natural language processing (NLP) is to develop computational models that can understand and generate human language. By capturing the statistical patterns and structures of text-based natural language, language models can predict and generate coherent and meaningful sequences of words. Combined with the Transformer model architecture (Vaswani et al., 2017), large language models (LLMs) trained on web-scale amounts of text, with proportionate compute and size, have demonstrated remarkable success in NLP tasks (Devlin et al., 2019; Brown et al., 2020; Chowdhery et al., 2022; Zhang et al., 2022b; Scao et al., 2022; Zeng et al., 2023). However, transfering these abilities to *spoken* human language remains a challenging frontier. Spoken dialog systems remain a cascade of separately trained automatic speech recognition (ASR), natural language understanding (NLU) and generation (NLG), and text-to-speech (TTS) systems (Gorin et al., 1997; Jokinen & McTear, 2009), with LLMs now playing the role of a combined NLU and NLG system. However, such cascades introduce latency and additional mechanisms for propagating and rendering non-verbal cues like speaker identity and prosody. Recently, spoken language models (Lakhotia et al., 2021; Kharitonov et al., 2022) and other generative audio models (Dhariwal et al., 2020; Hawthorne et al., 2022; Borsos et al., 2022; Agostinelli et al., 2023) have emerged as a promising avenue for generative speech modeling. These works quantize audio representations (Hsu et al., 2021; Chung et al., 2021; Zeghidour et al., 2022; Défossez et al., 2022) into learned discrete tokens compatible with the same next-token cross-entropy objective as text LLMs, a step that (Nguyen et al., 2022) argued as necessary for generative quality. In this paper, we introduce Spectron, a novel spoken language model that:

- Directly process spectrograms as both input and output. Spectron leverages the audio capabilities of a pre-trained speech encoder through the use of intermediate projection layers.

[*]Equal contribution.
[†]Work Done during internship at Google.

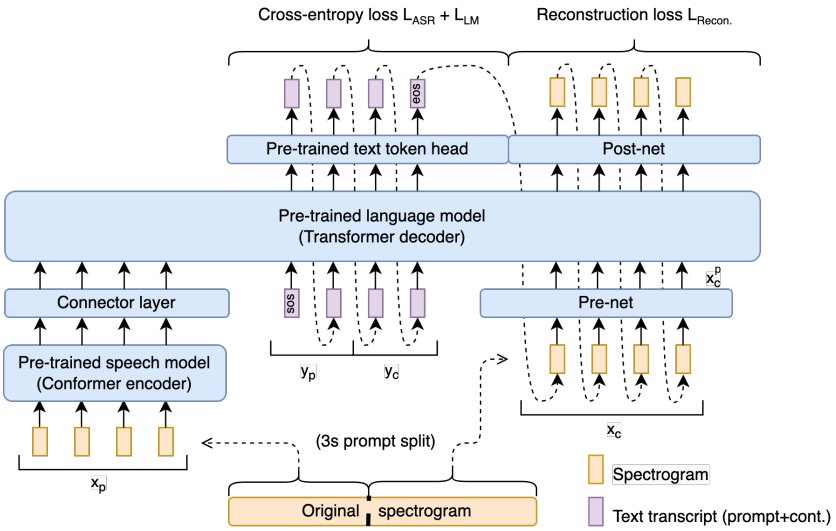

Figure 1: The proposed Spectron model, connects the encoder of a speech recognition model with a pre-trained Transformer decoder language model. At training time, we take speech utterances and split their audio into a *prompt* and its *continuation*. From the prompt speech features, the full (prompt and continuation's) transcript must be reconstructed, as well as the continuation's speech features via newly introduced pre- and post-net speech modules. At inference time, only a prompt is provided; the prompt's transcription, text continuation, and speech continuations are all generated by the model.

- Demonstrably transfer generative ability from a pre-trained LLM, as shown by competitive performance in semantic coherence and spoken question answering over other end-to-end spoken language models.

To quantify this transfer of knowledge, we also introduce two benchmarks for the nascent spoken QA task, which we synthesize from Web Questions and LLaMA generations. Audio samples and LLaMA dataset can be found on the project website.

Our work shows that the inductive biases from a pre-trained speech encoder and a language model decoder enable end-to-end training and state-of-the-art performance without sacrificing representational fidelity. Key to this is a novel end-to-end training objective which implicitly supervises speech recognition, text continuation, and conditional speech synthesis in a joint manner. The language model transcribes and generates text continuations, acting as an 'intermediate scratchpad' (Nye et al., 2021; Wei et al., 2022) to be conditioned on for audio generation. A novel spectrogram regression loss also supervises the model to match the higher-order temporal and feature deltas of the ground truth, based on the idea that the derivatives of the ground truth express rich, longer-range information about the shape of the signal. Our overall scheme is summarized in Figure 1 and described in the rest of this work.

2 Related Work

The dominant approach to spoken language modeling is to use compact discrete speech representations. This allows the application of text-based language models to speech data. Typically, these representations are created by clustering the outputs of a speech encoder using K-means and using the centroids as tokens. The resulting discrete sequences can be easily modeled using Transformer architectures (Vaswani et al., 2017). Below are notable examples of works using this approach; a comparison table is also presented in Appendix A.2.

Generative Spoken Language Modeling (GSLM) (Lakhotia et al., 2021) offers a baseline system that operate on units quantized from pre-trained audio representations, such as HuBERT (Hsu et al., 2021). The quantized units are processed by a Transformer-based model. An additional unit-to-speech

decoder converts the generated units into spectrograms. Spectron's approach is more simple and explicit, where a single model is input with spectograms and outputs raw spectograms.

TWIST (Hassid et al., 2023) Uses the same unit-to-speech and speech-to-unit systems as GSLM, but warm-starts the spoken language model from a text-based language model. They show that this warm-start improves overall performance, and drastically improves the ability to perform StoryCloze tasks. For the textual language model, they use the state of the art and open sourced OPT (Zhang et al., 2022a) and LLama (Touvron et al., 2023) 1.3B and 7B models respectively and show that scaling spoken langauge models also improves performance.

AudioLM (Borsos et al., 2022) utilizes two kinds of quantized representations: w2v-BERT (Chung et al., 2021) as semantic tokens and SoundStream (Zeghidour et al., 2022) as acoustic tokens. SoundStream embeddings undergo discretization using residual vector quantization (RVQ), resulting in a hierarchy of vector quantizers. AudioLM utilizes three transformer models, each corresponding to a different layer of token generation. As Spectron does not involve any quantization, our method naturally preserves the input semantics and accustics characteristics. Further, Spectron offers a single model trained with a unified objective as opposed to the multiple components required by AudioLM.

SpeechGPT (Zhang et al., 2023a) Adapts the LLama 7B to perform speech tasks by using both discrete speech representations and text. They introduce the SpeechInstruct dataset which is used for instruction tuning. SpeechGPT is trained in 3 differnet steps for modality adaptation, cross modal instruction finetuning and chain of modality instruction finetuning (using LoRA (Hu et al., 2021)). The obtained model is capable of generating both speech and text, as well as following instructions in both modalities. Our method, in comparison, is trained using a single reconstruction step and utilizes only publicly available datasets. Using no carefully curated datasets, we demonstrate competitive performance over spoken question answering and superior results for speech continuation, without using any specialized prompts. citet

Spoken language understanding (SLU): A number of recent studies explore the usage of pre-trained language models (LMs) for different SLU tasks. (Gong et al., 2023), (Zhao et al., 2023), and (Liu et al., 2023a) fine-tuned LMs on audio data to perform speech-to-text question answering (QA) tasks. These models were able to answer textual questions about the input audio in a direct manner. (Fathullah et al., 2023) showed that adding an audio encoder to an LM and training using the LoRA algorithm (Hu et al., 2021) enables the LM to perform automatic speech recognition (ASR). (Zhang et al., 2022c) aligned text and audio tokens to perform a large number of SLU tasks. (Peng et al., 2023) showed that LMs can be used to answer questions about spoken language using only text. Spectron, in comparison, produces both textual outputs and spectograms using a single autoregressive decoder. **Multi-modal text-speech training**: (Ao et al., 2021) performed joint training on speech and text data to perform multiple tasks, such as text-to-speech (TTS) and ASR. (Ren et al., 2019) used unsupervised pre-training on text and speech data to perform TTS for low-resource languages. **Textual-guided audio generation**: (Liu et al., 2023b) uses LMs to generate audio scripts and interacting with audio creation APIs. (Huang et al., 2023) augment ChatGPT input/output interface with ASR / TTS capabilities.

3 APPROACH

3.1 ARCHITECTURE

We propose a novel architecture for direct speech continuation. The architecture is initialized with a pre-trained speech encoder denoted as $\mathcal{E}$ and a pre-trained language decoder denoted as LM. The encoder is prompted with a speech utterance as input, which it encodes into continuous linguistic features. These features fed into the decoder as a prefix, and the whole encoder-decoder is optimized to jointly minimize a cross-entropy loss (for speech recognition and transcript continuation) and a novel reconstruction loss (for speech continuation). During inference, one provides a spoken speech prompt, which is encoded and then decoded to give both text and speech continuations.

3.1.1 INPUT PRE-PROCESSING

During training, the proposed model uses supervised speech utterances, which are pairs of speech x and transcripts y for training. The speech input, denoted as x, is a spectrogram that is split into two

segments at position s:

$$x_p = x_{\leq s}, \quad x_c = x_{>s}. \tag{1}$$

The first segment x_p (which we call the *prompt*) is fed into the speech encoder $\mathcal{E}$ to give continuous representations that condition the LM. The second segment x_c (the *continuation*) is used later for a spectrogram reconstruction loss. SpecAugment (Park et al., 2019) is applied for data augmentation. The corresponding transcripts y can be also split at position $\phi(s)$:

$$y_p = y_{\leq \phi(s)}, \quad y_c = y_{>\phi(s)}, \tag{2}$$

where $\phi(s)$ maps the feature index s in x to its text token index in y. Note that $\phi(s)$ is not needed for our training losses.

3.1.2 SPEECH ENCODER

The speech encoder $\mathcal{E}$ is a 600M-parameter Conformer encoder (Gulati et al., 2020) pre-trained on web-scale data (12M hours; Zhang et al., 2023b). It takes the spectrogram of the source speech as input, generating a hidden representation that incorporates both linguistic and acoustic information. The input spectrogram is first subsampled using a convolutional layer and then processed by a series of Conformer blocks. Each Conformer block consists of a feed-forward layer, a self-attention layer, a convolution layer, and a second feed-forward layer. The outputs of the total encoder $\mathcal{E}$ are passed through a layer $\mathcal{P}$ that projects the hidden representations into the embedding dimension of the language model. We denote these final embeddings

$$x_p^{\text{lm}} = \mathcal{P}(\mathcal{E}(x_p)). \tag{3}$$

3.1.3 LANGUAGE MODEL

We use prefix decoder language models with 350M or 1B parameters trained in the manner of PaLM 2 (Google, 2023), which we denote as LM. The LM receives the encoded features of the prompt x_p^{lm} as a prefix. Note that this is the only connection between the speech encoder and the LM decoder; i.e., there is no cross-attention between the encoder and the decoder. This late-stage integration is consistent with work in ASR, which found that joint finetuning of a pre-trained speech encoder and a pre-trained LM decoder into a sequence-to-sequence model can improve performance, even if the integration occurs as a single final layer (Deng et al., 2021); more layers did not improve performance, which they attribute to having sufficiently powerful text representations. During training, the decoder is teacher-forced to predict the text transcription y_p, text continuation y_c, and speech embeddings x_c^p. To convert the speech embeddings to and from spectrograms, we introduce lightweight modules h^{pre} and h^{post}, described in the next section. In all, we get next-step predictions for the concatenation of these text tokens and embeddings:

$$[\hat{y}_p, \hat{y}_c, \hat{x}_c^p] = \text{LM}(x_p^{\text{lm}}, [y_p, y_c, x_c^p]). \tag{4}$$

By having the *same* architecture decode the intermediate text and the spectrograms, we gain two benefits. First, we benefit from the pre-training of the LM in the text domain to continue the prompt in the text domain before synthesizing the speech. Secondly, the predicted text serves as intermediate reasoning, enhancing the quality of the synthesized speech, analogous to improvements in text-based language models when using intermediate scratchpads (Nye et al., 2021) or chain-of-thought (CoT; Wei et al., 2022).

3.1.4 ACOUSTIC PROJECTION LAYERS

To enable the language model decoder to model speech features, we employ a multi-layer perceptron, the pre-net h^{pre} to project the ground truth spectrogram speech continuations x_c to the language model dimension $x_c^p = h^{\text{pre}}(x_c)$. This pre-net h^{pre} compresses the spectrogram input x_c into a lower dimension, creating a bottleneck that aids the decoding process. This bottleneck mechanism prevents the model from repetitively generating the same prediction in the decoding process, as demonstrated in previous work (Shen et al., 2018). To project $\hat{x}_c^p$ from the language model dimension to the spectrogram dimension, the model employs a post-net h^{post}, which is a multi-layer perceptron. This projection is represented by $\hat{x}_c = h^{\text{post}}(\hat{x}_c^p)$.

Both h^{pre} and h^{post} are two-layer multi-layer perceptrons. Additionally, the input text sequence $[y_p, y_c]$ is padded at the beginning with a "start of sequence" (sos) token, while the output sequence is padded with an "end of sequence" (eos) token at the final position.

3.2 TRAINING OBJECTIVE

The training methodology of the proposed approach is depicted in Figure 1. It uses two distinct loss functions: (1) cross-entropy loss, employed for both speech recognition and transcript continuation, and (2) regression loss, employed for speech continuation. During training, all parameters are updated (speech encoder $\mathcal{E}$, projection layer $\mathcal{P}_s$, language model LM, pre-net h^{pre}, and post-net h^{post}).

3.2.1 SPEECH RECOGNITION AND TRANSCRIPT CONTINUATION

The first loss term is a combination of a speech recognition loss $\mathcal{L}_{\text{ASR}}$ and a transcript continuation loss $\mathcal{L}_{\text{LM}}$, which are given by:

$$\mathcal{L}_{\text{ASR}}(y_p, \hat{y}_p) = \text{CE}(y_p, \hat{y}_p), \quad \mathcal{L}_{\text{LM}}(y_c, \hat{y}_c) = \text{CE}(y_c, \hat{y}_c), \tag{5}$$

where CE denotes cross-entropy, which quantifies the dissimilarity between the predicted distribution over $\hat{y}_p, \hat{y}_c$, and the corresponding ground truth distribution over y_p, y_c. This objective increases the likelihood of the text $[y_p, y_c]$ under the conditional distribution modeled by the LM.

3.2.2 SPEECH CONTINUATION

The speech continuation objective is formulated as a regression task, predicting spectrogram frame channels independently given previous spectrogram frame predictions and the ASR and LM context. To promote convergence and improve modeling power, we apply ℓ_1 and ℓ_2 regression losses on the spectrogram (Shen et al., 2020). These losses are applied to the feature-deltas of the spectrogram, and to the time-deltas of the spectrogram up to order K. That is, for a tensor z of dimension $T \times F$ we define:

$$\Delta_k^{\text{time}}(z) = z_{[1:T-k,:]} - z_{[k:T,:]}, \Delta_k^{\text{feat}}(z) = z_{[:,1:F-k]} - z_{[:,k:F]}, \tag{6}$$

$$\mathcal{L}_{1+2}(z, z') = ||z - z'||_1 + ||z - z'||_2^2. \tag{7}$$

For a ground truth spectrogram x_c and the predicted spectrogram $\hat{x}_c$, the speech continuation loss is a combination of three objectives:

$$\mathcal{L}_{\text{s}}(x_c, \hat{x}_c) = \mathcal{L}_{1+2}(x_c, \hat{x}_c), \tag{8}$$

$$\mathcal{L}_{\text{f}}(x_c, \hat{x}_c) = \mathcal{L}_{1+2}(\Delta_1^{\text{feat}}(x_c), \Delta_1^{\text{feat}}(\hat{x}_c)), \tag{9}$$

$$\mathcal{L}_{\text{t}}(x_c, \hat{x}_c) = \sum_{k=1}^{K} \mathcal{L}_{1+2}(\Delta_k^{\text{time}}(x_c), \Delta_k^{\text{time}}(\hat{x}_c)) \tag{10}$$

The overall speech continuation loss is thus given by:

$$\mathcal{L}_{\text{Recon.}}(x_c, \hat{x}_c) = \mathcal{L}_{\text{s}}(x_c, \hat{x}_c) + \mathcal{L}_{\text{f}}(x_c, \hat{x}_c) + \mathcal{L}_{\text{t}}(x_c, \hat{x}_c). \tag{11}$$

3.2.3 OVERALL LOSS

Using the above notation, our objective is:

$$\mathcal{L}_{\text{total}}(x, y) = \mathcal{L}_{\text{ASR}}(y_p, \hat{y}_p) + \mathcal{L}_{\text{LM}}(y_c, \hat{y}_c) + \mathcal{L}_{\text{Recon.}}(x_c, \hat{x}_c). \tag{12}$$

Since $\mathcal{L}_{\text{ASR}}$ and $\mathcal{L}_{\text{LM}}$ are cross-entropy losses and since $y = [y_p, y_c]$ (Eq.2), the overall speech recognition and transcript continuation loss can be written as:

$$\mathcal{L}_{\text{ASR}}(y_p, \hat{y}_p) + \mathcal{L}_{\text{LM}}(y_c, \hat{y}_c) = \text{CE}(y, \hat{y}) = \mathcal{L}_{\text{CE}}(y, \hat{y}) \tag{13}$$

where $\hat{y}$ is the concatenation of $\hat{y}_p$ and $\hat{y}_c$. This simplifies the overall loss to:

$$\mathcal{L}_{\text{total}}(x, y) = \mathcal{L}_{\text{CE}}(y, \hat{y}) + \lambda_r \mathcal{L}_{\text{Recon.}}(x_c, \hat{x}_c), \tag{14}$$

where λ_r is a weighting coefficient. This simplification eliminates the necessity of the text-speech time alignment $\phi(s)$. Our approach can be seen as jointly optimizing three capabilities:

Speech recognition ($\mathcal{L}_{\text{ASR}}$): The combined model learns to transcribe speech audio into text. As we use pre-trained speech encoder and pre-trained language model, this objective encourages the alignment and integration of each model's functionality.

Transcript continuation ($\mathcal{L}_{\text{LM}}$): This reuses, maintains, and leverages the language model's ability to generate natural text as learned from its training scheme, for example, dialogue for a chat-optimized LM. Depending on the utterance, the decoder may further learn to use paralinguistic cues from the prompt speech to favor certain completions.

Conditional speech synthesis ($\mathcal{L}_{\text{Recon.}}$): We reuse the language model's autoregressive generation ability and direct it toward spectrogram reconstruction. As the teacher-forced transcript is available and the most "accessible" feature, the decoder learns to perform text-to-speech. In this way, the model can synthesize the LM's arbitrary textual continuations at inference time, including words not found in training. Finally, we expect that good spectrogram-level continuations require the preservation of speaker, prosody, and channel effects from the original speech prompt.

3.3 Inference

In inference, the speech prompt x_p is encoded by the speech encoder $\mathcal{E}$ then projected by $\mathcal{P}_s$ to the LM's dimension to give x_p^{lm} (Eq.3). Utilizing x_p^{lm} and the start-of-sentence (sos) token, the language model decodes text in an autoregressive manner: $\hat{y} = \text{LM}([x_p^{\text{lm}}, \text{sos}])$ until eos is emitted, where $\hat{y}$ is a concatenation of the predicted transcript and continuation $[\hat{y}_p, \hat{y}_c]$. Following this, the language model decodes a spectrogram in an autoregressive manner. It predicts the next spectrogram feature estimate $\hat{x}_c(t)$ using prompt features x_p^{lm}, text prediction $\hat{y}$ and past estimated spectrogram features $\hat{x}_c(\leq t-1)$. Past spectrogram estimates $\hat{x}_c(\leq t-1)$ are projected to the language model dimension: $\hat{x}_c^p(\leq t-1) = h^{\text{pre}}(\hat{x}_c(\leq t-1))$. Then, $\hat{x}_c^p(t)$ is predicted at step t: $\hat{x}_c^p(t) = \text{LM}([x_p^{\text{lm}}, \text{sos}, \hat{y}, \hat{x}_c^p(\leq t-1)])$ The decoded output $\hat{x}_c^p(t)$ is then projected to the spectrogram domain using h^{post}: $\hat{x}_c(t) = h^{\text{post}}(\hat{x}_c^p(t))$. Finally, a vocoder converts the predicted spectrogram $\hat{x}_c$ into a waveform signal.

4 Experiments and results

4.1 Data and preprocessing

To empirically evaluate the performance of the proposed approach, we conducted experiments on the Libri-Light dataset (Kahn et al., 2020). Libri-Light is a 60k hour English dataset consisting of unlabelled read speech from LibriVox audiobooks. For our training objective, the dataset was transcribed using a NST (Park et al., 2020) model trained on LibriSpeech (960 hours). We utilized a frozen neural vocoder called WaveFit (Koizumi et al., 2022) to convert the predicted spectrograms into raw audio. The model architecture and hyperparameters as described in Koizumi et al. (2022). Our proposed model was trained using 64 TPUv4 chips (Jouppi et al., 2023), over a duration of 48 hours. A comprehensive table of hyperparameters can be found in the appendix, specifically in Appendix A.1. We consider a predetermined set of 3-second prefixes denoted as $s = 3\text{sec}$. During training utterances with a length of less than 3 seconds are discarded. For Librilight, only 0.04% of utterances are less than 3 seconds. To evaluate our model and the baseline models during testing, we utilize the test-clean test set from LibriSpeech (Panayotov et al., 2015). We employ the first 3 seconds of each utterance in the test set as a prompt to the models, excluding the ground truth transcripts. For semantic and acoustic quality, Spectron was trained with a LM of 350 million parameters, while for the question answering task, Spectron was trained with a LM of 1 billion parameters. The two models are identical except for the LM. In sections 4.2.1 and 4.2.2 a 350M LM was used, in section 4.2.3 a 1B LM was used.

4.2 Baselines

We compare our method against existing spoken language models:

GSLM: We evaluate their best model, the HuBERT-L6 configuration with 200 token units for conditional speech continuation. The model was trained on a filtered subset of LibriLight (Rivière & Dupoux, 2021). **AudioLM:** We utilize the Libri-Light trained model described in their work. The two AudioLM models we compare against differ in the number of SoundStream residual vector quantizer (RVQ) layers they generate. One model generates the top 3 layers (**3-RVQ**), while the other model generates all 12 layers (**12-RVQ**). **TWIST:** We evaluate both the 1.3B version and

7B versions of their models that are based on llama with HuBERT speech representations. Their models where trained on Libri-Light, Spotify podcasts Clifton et al. (2020), People dataset Galvez et al. (2021) and VoxPopuli Wang et al. (2021). **SpeechGPT:** We evaluate their open-sourced model, which is based upon the llama-7B model with HuBERT speech representations. This model is termed SpeechGPT-7B-com, and was trained using all 3 training stages in SpeechGPT. The model was trained using the LibriLight and SpeechInstruct datasets.

4.2.1 SEMANTIC QUALITY

We employ the log-perplexity metric to evaluate the semantic quality of the speech output from the models. We use a state-of-the-art Conformer ASR system (Zhang et al., 2023b) trained on a proprietary English-only dataset to transcribe the speech continuation. Subsequently, we compute the log-perplexity of the predicted transcripts using the GPT-2 Medium model (Radford et al., 2019) from the open-source `transformers` library (Wolf et al., 2020). The results presented in Table 1 demonstrate the performance gains of our method compared to previous approache such as GSLM, where our method achieves an improvement of 170.91 in log-perplexity. Furthermore, when compared to the state-of-the-art AudioLM method, our approach outperforms both the 3-RVQ and 12-RVQ variants, exhibiting enhancements of 12.88 and 14.20 respectively. Moreover, the results in Table 1 reveal that our method exhibits improved performance compared to existing cascade methods.

4.2.2 ACOUSTIC QUALITY

To evaluate acoustic quality, we use the Naturalness Mean Opinion Score (N-MOS) (Nguyen et al., 2023) metric and the average cosine distance between speaker embeddings as measures of speaker similarity. MOS is computed and reported solely for the speech continuations. The average speaker similarity is calculated between the input prompt and its generated continuation. **The Naturalness Mean Opinion Score (N-MOS):** Human evaluators are tasked with assigning a rating on a five-point scale to denote the perceived naturalness of a given speech utterance. Spanning from 1 (indicative of poor quality) to 5 (indicative of excellent quality), N-MOS serves as a valuable tool for discerning the quality of speech naturalness. Test where conducted using 20 randomly sampled utterances from the Librispeech test-clean test set. 30 raters have participated in the tests. The prompt where not available to the raters, and they were requested to evaluate only the naturalness of the continuations. **Avg. speaker similarity:** We compute the speaker similarity between prompt and continuation using the speaker encoder of the PnG NAT TTS model (Morioka et al., 2022). We compute the speaker embeddings of both and measure the cosine similarity between each pair of embeddings. We report the average across the entire test set. As seen in Table 2, our approach performs slightly better than GSLM in terms of N-MOS, with improvements of 0.55. When compared to AudioLM, our approach is comparable to the 3-RVQ version and slightly inferior to the 12-RVQ version, with a decrease of 0.19 in N-MOS. One can see in Table 2 that the results of TWIST are similar to those of GSLM, and Spectron outperforms the 1.3B and 7B versions by 0.4 and 0.65 respectively. SpeechGPT performs slightly inferior to Spectron, which outperforms it by a score of 0.3. Table 3 presents the results for average speaker similarity. Our method demonstrates a significant improvement of 0.31 over the GSLM method. When compared to AudioLM, our method outperforms both the 3-RVQ and 12-RVQ versions, with increases of 0.05 and 0.07 in average speaker similarity, respectively. Moreover, comparing to TWIST 1.3B and 7B, the proposed method improve the average speaker similarity by 0.18 and 0.19, respectively. These results indicate that comparable acoustic quality can be achieved with a simpler approach. Our model is trained end-to-end and utilizes the universal

Table 1: Log-perplexity for completions of LibriSpeech utterances given a 3-second prompt. Lower is better.

Method	**Log-perplexity** (↓)
GSLM	296.99
AudioLM (3-RVQ)	138.96
AudioLM (12-RVQ)	140.28
TWIST (1.3B)	229.53
TWIST (7B)	170.81
SpeechGPT	136.42
SPECTRON (350M)	**126.08**

speech representation of spectrograms. Note that SpeechGPT does not intend to preserve speaker identity and its average speaker similarity is lower.

Table 2: Naturalness Mean-Opinion-Score N-MOS) (Mean $\pm$ SE) metric for completions of LibriSpeech utterances.

Method	**N-MOS** (↑)
GSLM	3.13 ± 0.32
AudioLM (3-RVQ)	3.61 ± 0.29
AudioLM (12-RVQ)	3.87 ± 0.32
TWIST (1.3B)	3.28 ± 0.24
TWIST (7B)	3.03 ± 0.22
SpeechGPT	3.38 ± 0.30
SPECTRON (350M)	3.68 ± 0.29
Ground Truth	4.23 ± 0.33

Table 3: Average Speaker Similarity metric for completions of LibriSpeech utterances.

Method	**Speaker Sim.** (↑)
GSLM	0.11
AudioLM (3-RVQ)	0.37
AudioLM (12-RVQ)	0.35
TWIST (1.3B)	0.24
TWIST (7B)	0.23
SpeechGPT	0.05
SPECTRON (350M)	**0.42**

4.2.3 QUESTION ANSWERING

We propose examining whether the models can continue spoken sentences or questions with the appropriate answer. This can can be viewed as spoken generative QA; the correct answer must be produced out of infinite possibilities. Note that except for SpeechGPT, all other methods (including ours) are evaluated in a zero-shot fashion for spoken question answering. Given that the various spoken language models are evaluated with 3-second input contexts, we use TTS to synthesize questions that fit within this duration. The questions are drawn from an existing set WebQuestions and a new test set which we name LLama questions. **WebQuestions** Berant et al. (2013) is an open-ended question answering NLP dataset. The dataset contains open ended questions that are supposed to be answerable by Freebase and are centered around a single named entity. **LLama question** is an open-domain world knowledge QA dataset that we had gathered from the open-source LLama-7B model (Touvron et al., 2023). We had prompted the model to provide questions and short answers regarding various topics. Overall, we had gathered 300 questions in this manner, and had generally verified the answers (the test set will be released). All of the questions are generated using the publicly available Google Cloud TTS service, using the voice en-US-Neural2-C. To compute answer accuracy, we use a Conformer ASR system (Zhang et al., 2023b) to transcribe the answers of the models. If the textual answer is contained in the output of the ASR, we count the answer as being correct. The results presented in Table 4 demonstrate the performance of the proposed model in comparison to other existing models. Specifically, the proposed model exhibits an accuracy of 22.9% on the LLama test set, while SpeechGPT achieves a comparable accuracy of 21.9%. Notably, despite SpeechGPT's utilization of a larger model architecture comprising 7 billion parameters, the proposed method leverages a more modest 1 billion parameter model, yet achieves comparable results. In contrast, TWIST models with 1.3 billion and 7 billion parameters demonstrate notably lower accuracies of 1% and 0.5% respectively. Upon careful examination, it becomes evident that these models predominantly generate completions of input questions rather than providing substantive answers. AudioLM 3-RVQ, AudioLM 12-RVQ and GSLM achieved accuracy of 7%, 6.7% and 4%, respectability, which is likely due to the fact that the underlying Transformer architecture is not pre-trained on a large language model. Similarly, on the Web Questions test set, the proposed model attains an accuracy of 6.1%, while SpeechGPT yields a comparable accuracy of 6.5%. Again, TWIST models with 1.3 billion and 7 billion parameters achieve accuracies of 0.7% and 1.1% respectively, further reinforcing the observed trend of completion-centric behavior rather than direct question answering. Additionally, models such as AudioLM 3-RVQ, AudioLM 12-RVQ, and GSLM exhibit accuracies of 2.3%, 2.3%, and 1.5% respectively, which can likely be attributed to the absence of pre-training on a large-scale language model within the underlying Transformer architecture.

4.2.4 AUDIO SAMPLES

Various samples can be found on the project website. These include generated samples of speech continuation for LibriSpeech dataset as well as spoken question answering Web Questions and LLama Questions test sets.

Table 4: Accuracy (%) on spoken question answering datasets.

Method	**Web Questions** (↑)	**LLama Questions** (↑)	Zero-Shot
GSLM	1.5	4.0	✓
AudioLM (3-RVQ)	2.3	7.0	✓
AudioLM (12-RVQ)	2.3	6.7	✓
TWIST (1.3B)	0.7	1.0	✓
TWIST (7B)	1.1	0.5	✓
SpeechGPT (7B)	6.5	21.9	×
SPECTRON (1B)	6.1	22.9	✓

4.2.5 ABLATION ANALYSIS

To understand the individual impacts of various components within the proposed approach, an ablation study was conducted. We measure the log-perplexity over the test-clean test set of the LibriSpeech dataset (Panayotov et al., 2015). This study involved removing each specific component in isolation. (i) Disabled intermediate loss on text ("-$\mathcal{L}_{\text{CE}}$") (ii) removed spectrogram derivative loss ("-$(\mathcal{L}_{\text{f}} + \mathcal{L}_{\text{t}})$") (iii) removed pre-training of the language model LM, letting it train from scratch (iv) removed pre-training of the speech encoder $\mathcal{E}$ and training it from scratch (v) removed pre-training of both the speech encoder $\mathcal{E}$ and language model LM, training the entire model from scratch. The findings are summarized in Table 5. The results demonstrate that each of the aforementioned components contributes to the overall performance enhancement of the proposed approach. Notably, the ASR & LM cross-entropy loss $\mathcal{L}_{\text{CE}}$ and the spectrogram derivative loss $\mathcal{L}_{\text{f}} + \mathcal{L}_{\text{t}}$ have the most significant impact, leading to a degradation of 661.81 and 588.35 in the log-perplexity score, respectively. Furthermore, the incorporation of the pre-trained speech encoder and pre-trained language model exhibits a discernible decline in performance, resulting in a degradation of 87.17 and 75.63 in the log-perplexity score, respectively. Notably, when both the speech encoder and pre-trained language model are removed, a degradation of 118.31 in the log-perplexity score is observed.

Table 5: Ablation analysis.

Model	**Log-perplexity** (↓)
Proposed SPECTRON	126.08
− $\mathcal{L}_{\text{CE}}$	714.43
− $(\mathcal{L}_{\text{f}} + \mathcal{L}_{\text{t}})$	787.89
− Pre-trained LM	201.71
− Pre-trained speech encoder	213.25
− Pre-trained LM & speech encoder	244.39

5 LIMITATIONS AND FUTURE WORK

The limitation of our work is the high time and space complexity of generating spectrogram frames. Since spectrogram frames are computed with a rate of 12.5 ms, generation of long speech utterances is not possible. We hypothesize that potential solutions include generating multiple spectrogram frames from each hidden representation. Another limitation is that text and spectrogram decoding processes are not parallelizable. This hinders the ability to use Spectron in streaming scenarios and introduces a small latency between audio input and output. We leave the development of a parallelized decoding algorithm for future work. We further recognize that biases in the pre-trained language model may be sustained in our model, we refer to (Google, 2023) for a detailed discussion of ethical considerations for text-based language models.

6 CONCLUSION

We proposed Spectron, a neural direct speech continuation model that can be trained end-to-end and operates in the spectrogram domain. We showed that a pre-trained language model can be given speech recognition and generation capabilities post-hoc, by fine-tuning on continuation tasks using a pre-trained speech encoder and a novel training objective. The result is a model that benefits from the pre-training of both models and outperforms previous spoken language models on various metrics.

7 ACKNOWLEDGMENTS

The authors would like to thank Heiga Zen, Neil Zeghidour, Eugene Kharitonov, Tal Schuster, Bryan Richter, Christian Frank, Marco Tagliasacchi, Nadav Bar, and the rest of the Google Research team for helpful discussions and previous work on data preparation. The contribution of Alon Levkovitch is part of a Ph.D. thesis research conducted at Tel-Aviv University.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. MusicLM: Generating music from text. *CoRR*, abs/2301.11325, 2023.

Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, et al. Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing. *arXiv preprint arXiv:2110.07205*, 2021.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1533–1544, 2013.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: a language modeling approach to audio generation. *CoRR*, abs/2209.03143, 2022.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. PaLM: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *ASRU*, pp. 244–250. IEEE, 2021.

Ann Clifton, Aasish Pappu, Sravana Reddy, Yongze Yu, Jussi Karlgren, Ben Carterette, and Rosie Jones. The spotify podcast dataset. *arXiv preprint arXiv:2004.04270*, 2020.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *CoRR*, abs/2210.13438, 2022.

Keqi Deng, Songjun Cao, Yike Zhang, and Long Ma. Improving hybrid CTC/attention end-to-end speech recognition with pretrained acoustic and language models. In *ASRU*, pp. 76–82. IEEE, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *CoRR*, abs/2005.00341, 2020.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, et al. Prompting large language models with speech recognition abilities. *arXiv preprint arXiv:2307.11795*, 2023.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*, 2021.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

Google. PaLM 2 technical report, 2023. `https://ai.google/static/documents/palm2techreport.pdf`.

Allen L. Gorin, Giuseppe Riccardi, and Jeremy H. Wright. How may I help you? *Speech Commun.*, 23(1-2):113–127, 1997.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, pp. 5036–5040. ISCA, 2020.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Défossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. Textually pretrained speech language models. *CoRR*, abs/2305.13009, 2023.

Curtis Hawthorne, Andrew Jaegle, Catalina Cangea, Sebastian Borgeaud, Charlie Nash, Mateusz Malinowski, Sander Dieleman, Oriol Vinyals, Matthew M. Botvinick, Ian Simon, Hannah Sheahan, Neil Zeghidour, Jean-Baptiste Alayrac, João Carreira, and Jesse H. Engel. General-purpose, long-context autoregressive modeling with perceiver AR. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 8535–8558. PMLR, 2022.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jiawei Huang, Jinglin Liu, et al. Audiogpt: Understanding and generating speech, music, sound, and talking head. *arXiv preprint arXiv:2304.12995*, 2023.

Kristiina Jokinen and Michael McTear. Spoken dialogue systems. *Synthesis Lectures on Human Language Technologies*, 2(1):1–151, 2009.

Norman P Jouppi, George Kurian, Sheng Li, Peter Ma, Rahul Nagarajan, Lifeng Nai, Nishant Patil, Suvinay Subramanian, Andy Swing, Brian Towles, et al. TPU v4: An optically reconfigurable supercomputer for machine learning with hardware support for embeddings. *arXiv preprint arXiv:2304.01433*, 2023.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, Tatiana Likhomanenko, Gabriel Synnaeve, Armand Joulin, Abdelrahman Mohamed, and Emmanuel Dupoux. Libri-Light: A benchmark for ASR with limited or no supervision. In *ICASSP*, pp. 7669–7673. IEEE, 2020.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Rivière, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. Text-free prosody-aware generative spoken language modeling. In *ACL (1)*, pp. 8666–8681. Association for Computational Linguistics, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv:1412.6980*, 2014.

Yuma Koizumi, Kohei Yatabe, Heiga Zen, and Michiel Bacchiani. WaveFit: an iterative and non-autoregressive neural vocoder based on fixed-point iteration. In *SLT*, pp. 884–891. IEEE, 2022.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Proc. NeurIPS*, 33:17022–17033, 2020.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, and Emmanuel Dupoux. On generative spoken language modeling from raw audio. *Transactions of the Association for Computational Linguistics*, 9:1336–1354, 2021. doi: 10.1162/tacl_a_00430. URL `https://aclanthology.org/2021.tacl-1.79`.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Yossi Adi, Juan Pino, Jiatao Gu, et al. Textless speech-to-speech translation on real data. *arXiv preprint arXiv:2112.08352*, 2021.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. Music understanding llama: Advancing text-to-music generation with question answering and captioning. *arXiv preprint arXiv:2308.11276*, 2023a.

Xubo Liu, Zhongkai Zhu, Haohe Liu, Yi Yuan, Meng Cui, Qiushi Huang, Jinhua Liang, Yin Cao, Qiuqiang Kong, Mark D Plumbley, et al. Wavjourney: Compositional audio creation with large language models. *arXiv preprint arXiv:2307.14335*, 2023b.

Nobuyuki Morioka, Heiga Zen, Nanxin Chen, Yu Zhang, and Yifan Ding. Residual adapters for few-shot text-to-speech speaker adaptation. *arXiv preprint arXiv:2210.15868*, 2022.

Tu Anh Nguyen, Benoît Sagot, and Emmanuel Dupoux. Are discrete units necessary for spoken language modeling? *IEEE J. Sel. Top. Signal Process.*, 16(6):1415–1423, 2022.

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

Maxwell I. Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models. *CoRR*, abs/2112.00114, 2021.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, pp. 5206–5210. IEEE, 2015.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. SpecAugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*, 2019.

Daniel S Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V Le. Improved noisy student training for automatic speech recognition. *arXiv preprint arXiv:2005.09629*, 2020.

Linkai Peng, Baorian Nuchged, and Yingming Gao. Spoken language intelligence of large language models for language learning. *arXiv preprint arXiv:2308.14536*, 2023.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Yi Ren, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Almost unsupervised text to speech and automatic speech recognition. In *International conference on machine learning*, pp. 5410–5419. PMLR, 2019.

Morgane Rivière and Emmanuel Dupoux. Towards unsupervised learning of speech features in the wild. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 156–163. IEEE, 2021.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilic, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, and et al. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100, 2022.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, RJ-Skerrv Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions. In *ICASSP*, pp. 4779–4783. IEEE, 2018.

Jonathan Shen, Ye Jia, Mike Chrzanowski, Yu Zhang, Isaac Elias, Heiga Zen, and Yonghui Wu. Non-attentive tacotron: Robust and controllable neural tts synthesis including unsupervised duration modeling. *arXiv preprint arXiv:2010.04301*, 2020.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pp. 5998–6008, 2017.

Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *ACL 2021-59th Annual Meeting of the Association for Computational Linguistics*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *EMNLP (Demos)*, pp. 38–45. Association for Computational Linguistics, 2020.

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, et al. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*, 2021.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE ACM Trans. Audio Speech Lang. Process.*, 30: 495–507, 2022.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Peng Zhang, Yuxiao Dong, and Jie Tang. GLM-130B: an open bilingual pre-trained model. 2023.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023a.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022a.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona T. Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. OPT: open pre-trained transformer language models. *CoRR*, abs/2205.01068, 2022b.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara N. Sainath, Pedro J. Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. Google USM: Scaling automatic speech recognition beyond 100 languages. *CoRR*, abs/2303.01037, 2023b.

Ziqiang Zhang, Sanyuan Chen, Long Zhou, Yu Wu, Shuo Ren, Shujie Liu, Zhuoyuan Yao, Xun Gong, Lirong Dai, Jinyu Li, et al. Speechlm: Enhanced speech pre-training with unpaired textual data. *arXiv preprint arXiv:2209.15329*, 2022c.

Zihan Zhao, Yiyang Jiang, Heyang Liu, Yanfeng Wang, and Yu Wang. Librisqa: Pioneering free-form and open-ended spoken question answering with a novel dataset and framework. *arXiv preprint arXiv:2308.10390*, 2023.

A APPENDIX

A.1 TABLE OF HYPER-PARAMETERS

Table 6: Model hyper-parameters used in the experiments. ("$\times n$": n layers)

Input & Output	
Sample rate (Hz)	16,000
Mel channels	128
Mel lower band (Hz)	20
Mel upper band (Hz)	8,000
Frame size (ms)	50.0
Frame step (ms)	12.5
SpecAugment	
Freq blocks	2
Time blocks	10
Freq mask max bins	27
Time mask max frames	40
Time block max length ratio	0.05
Speech Encoder	
Conformer dims	1024
Attention heads	8
Conv kernal size	(3, 3)
Conv stride size	(2, 2)
Language Model	
Transformer (dim $\times$ layers)	1024
Dim per head	64
Hidden dims	4096
Num heads	16
Vocab size	256,000
WaveFit vocoder	
Iterations	5
UBlock upsampling factors	[5, 5, 2, 2, 2]
STFT loss resolutions	3
Hann win size, frame shift, FFT size res 1	[160, 32, 512]
Hann win size, frame shift, FFT size res 2	[400, 80, 1024]
Hann win size, frame shift, FFT size res 3	[800, 160, 2048]
Multi-period discriminator	Kong et al. Kong et al. (2020)
Multi-period discriminator loss weight	1.0
Training	
Optimizer	Adam Kingma & Ba (2014)
Learning rate schedule	Vaswani et al. Vaswani et al. (2017)
Learning rate (peak)	3.5×10^{-4}
Warm-up steps	8K
Batch size	128
Continuation loss weight λ_r	0.1
Derivative loss order K	3

A.2 EXTENDED COMPARISON TO PREVIOUS METHODS

Given the variation in pre-trained models across the methods discussed in this paper, we find it fit to conduct a more comprehensive comparison of their performance. This detailed comparison is presented in Table 7. Regarding the Word Error Rate (WER) assessment for ASR systems, the WER scores are sourced from the SUPERB benchmark paper (Yang et al., 2021). WER are reported on the Librispeech test-clean test set. It's important to note that these scores rely solely on the speech encoder type due to limited data availability for all utilized models. For instance, models such as mHuBert (Lee et al., 2021) employed in SpeechGPT and the New Frequency HuBert adapted and trained within TWIST exist solely as tokenization models and lack dedicated ASR model forms. The performance comparison of speech encoders referenced in various methods within this paper is depicted in Table 7. Notably, the performance of the speech encoders on the Librispeech test set is comparable. However,

concerning the language models (LMs) utilized, more variation is evident among the methods. LMs span a spectrum, ranging from larger models such as SpeechGPT and TWIST employing 7B LMs, to intermediate-sized models like Spectron and AudioLM employing approximately 1B LMs, and finally, GSLM utilizing a smaller 200M parameter LM. It is widely acknowledged that LM performance is significantly influenced by model size (Kaplan et al., 2020). Moreover, diverse datasets have been employed across these systems, including Librispeech, Librilight, SpeechInstruct, VoxPopuli (Wang et al., 2021), Common-Voice (Ardila et al., 2019), Spotify (Clifton et al., 2020), and People (Galvez et al., 2021).

Table 7: Comparison of different models mentioned in this paper.

Detail	**Spectron**	**SpeechGPT**	**TWIST**	**GSLM**	**AudioLM**
LM #Params	1B / 350M	7B	1B/7B	∼200M	0.9B
LM Type	Palm2	LLama	OPT/LLama	None	None
Speech Encoder	USM	mHuBert	T-HuBert	HuBert	Wav2Vec
Speech Encoder #Params	600M	317M	317M	317M	600M
WER of Speech Encoder	3.1	2.94	2.94	2.94	3.1
Speech Encoder Dataset	Web-scale Librispeech	VoxPopuli 100k	Librispeech VoxPopuli CommonVoice Spotify Fisher	Librispeech Librilight	Librilight
Training Dataset	Librilight	Librispeech Speech Instruct	Librispeech Spotify People VoxPopuli	Librilight	Librilight
#Training Examples	60k	60k + 38k	150k	60k	60k