

# SPIRIT-LM: Interleaved Spoken and Written Language Model

Tu Anh Nguyen<sup>a,+,†</sup>, Benjamin Muller<sup>a,+,‡</sup>, Bokai Yu<sup>a,+</sup>,  
 Marta R. Costa-jussa<sup>b,+,†</sup>, Maha Elbayad<sup>b,+,‡</sup>, Sravya Popuri<sup>b,+,†</sup>,  
 Paul-Ambroise Duquenne<sup>b,+,†</sup>, Robin Algayres<sup>b,‡</sup>, Ruslan Mavlyutov<sup>b,+,†</sup>, Itai Gat<sup>b,+,†</sup>,  
 Gabriel Synnaeve<sup>c,+,†</sup>, Juan Pino<sup>c,+,†</sup>, Benoît Sagot<sup>c,†</sup>, Emmanuel Dupoux<sup>c,+,‡</sup>

<sup>+</sup> Meta AI, <sup>†</sup> Inria, Paris, <sup>‡</sup> EHESS, ENS-PSL, CNRS, Paris  
 {ntuanh, benjaminmuller, bokai, dpx}@meta.com

## Abstract

We introduce SPIRIT-LM, a foundation multimodal language model that freely mixes text and speech. Our model is based on a pretrained text language model that we extend to the speech modality by continuously training it on text and speech units. Speech and text sequences are concatenated as a single set of tokens, and trained with a word-level *interleaving* method using a small automatically-curated speech-text parallel corpus. SPIRIT-LM comes in two versions: a BASE version that uses speech semantic units and an EXPRESSIVE version that models expressivity using pitch and style units in addition to the semantic units. For both versions, the text is encoded with subword BPE tokens. The resulting model displays both the semantic abilities of text models and the expressive abilities of speech models. Additionally, we demonstrate that SPIRIT-LM is able to learn new tasks in a few-shot fashion across modalities (i.e. ASR, TTS, Speech Classification)<sup>1</sup>.

## 1 Introduction

Prompting Large Language Models (LLMs) has become a standard in Natural Language Processing (NLP) since the release of GPT-3 (Brown et al., 2020). Scaling language models to billions of parameters with massive datasets helps to achieve general-purpose language understanding and generation. Additionally, large-scale language models can solve new tasks by providing the model with a few examples through in-context few-shot learning. Since then, a number of LLMs have been developed (Chowdhery et al., 2022; Hoffmann et al., 2022; Zhang et al.,

2022; Touvron et al., 2023a). Notably, LLaMA (Touvron et al., 2023a) showed that smaller LLMs can achieve very good performance when training longer on more data using optimal-compute scaling laws (Kaplan et al., 2020), making LLMs more accessible for NLP research.

Speech Language Models (SpeechLMs), i.e. language models trained directly on speech, have been introduced (Lakhotia et al., 2021; Algayres et al., 2023; Borsos et al., 2023) and have recently become an active field of research (Wang et al., 2023a; Nguyen et al., 2023b; Hassid et al., 2023; Rubenstein et al., 2023). These models are either trained on speech-only datasets or datasets of specific tasks, e.g. Text-To-Speech (TTS), Automatic Speech Recognition (ASR), or Speech Translation, making the LMs focus on certain modality or tasks potentially losing their generalization capabilities.

Given the increasing quality of text-only LLMs (Brown et al., 2020; Touvron et al., 2023b), one successful approach to generate speech has been to build pipelines that first transcribe input speech with ASR, then generate text using a text-only LLM and finally synthesize the generated text into speech with TTS. However, with such pipelines, modeling and generating expressive speech is constrained out of the language model, leading to poor generation from an expressive point of view.

In this work, we aim to combine the generative abilities and pretrained knowledge of text LLMs with the expressive capacities of speech-language models. We show that LLMs trained on interleaved speech and text can learn speech and text cross-modally and are able to generate language content in either modality. We evaluate the models with comprehension tasks in both speech and text, and extend few-shot prompting to speech-text tasks such as ASR, TTS or Speech Classification. We further extend the semantic speech tokens with expressive tokens that capture the pitch and style

<sup>a,b,c</sup> Equally contributed as co-first, co-second and co-last authors, resp.

<sup>1</sup>Generation samples can be found at: <https://speechbot.github.io/spiritlm>

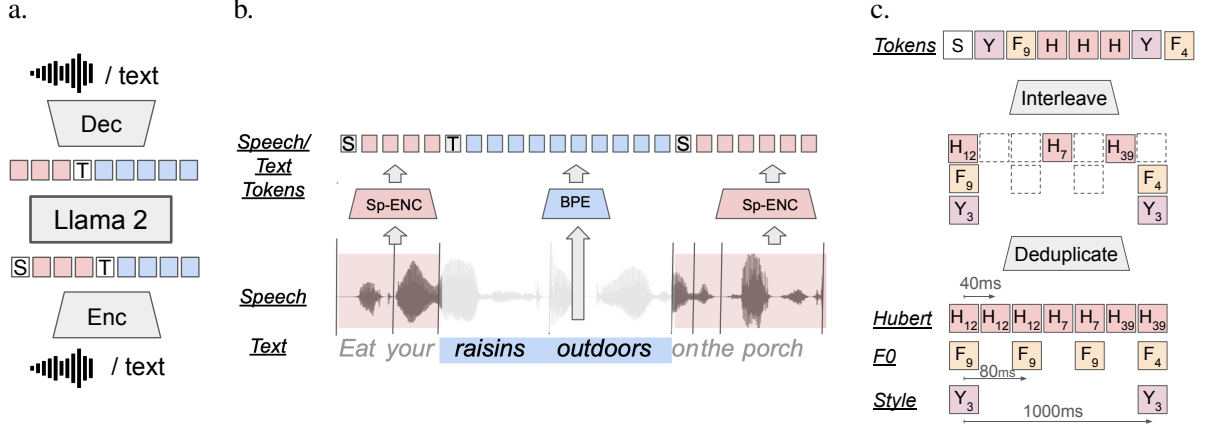


Figure 1: **a. The SPIRIT-LM architecture.** A language model trained with next token prediction; tokens are derived from speech or text with an encoder, and rendered back in their original modality with a decoder. SPIRIT-LM models are trained on a mix of text-only sequences, speech-only sequences, and *interleaved* speech-text sequences. **b. Speech-text interleaving scheme.** Speech is encoded into tokens (pink) using clusterized speech units (Hubert, Pitch, or Style tokens), and text (blue) using BPE. We use special tokens [TEXT] to prefix text and [SPEECH] for speech tokens. During training, a change of modality is randomly triggered at word boundaries in aligned speech-text corpora. Speech tokens are deduplicated and interleaved with text tokens at the modality change boundary. **c. Expressive Speech tokens.** For SPIRIT-LM-EXPRESSIVE, pitch tokens and style tokens are interleaved after deduplication.

of the speech, and evaluate the models with newly introduced sentiment modeling tasks. Our contributions are the following:

- We introduce SPIRIT-LM, a single language model that can generate both speech and text. SPIRIT-LM is based on continuously pre-training LLAMA 2 with *interleaving* speech and text data.
- Similarly to text LLMs, we find that SPIRIT-LM can learn new tasks in the few-shot setting in text, speech and in the cross-modal setting (i.e. speech to text and text to speech)
- To evaluate the expressive abilities of generative models, we introduce the SPEECH-TEXT SENTIMENT PRESERVATION benchmark (noted STSP) that measures how well generative models preserve the sentiment of the prompt within and across modalities for both spoken and written utterances.
- Finally, we propose an expressive version of SPIRIT-LM (SPIRIT-LM-EXPRESSIVE). Using STSP, we show that SPIRIT-LM is the first language model that can preserve the sentiment of text and speech prompts both within and across modalities.

The rest of the paper is structured as follows: We describe relevant related work (Section 2), our methods for model training and evaluation (Section 3), text and speech understanding evaluation results (Section 4), sentiment modeling evaluation (Section 5), an in-depth responsible AI evaluation of SPIRIT-LM with a focus on spoken and written toxicity detection (Section 6), and finally the broader impact of this work (Section 7).

## 2 Related Work

**Textless NLP** Recent progress in Self-Supervised Speech Representation Learning (SSL) (Baeviski et al., 2020; Hsu et al., 2021; Chen et al., 2022; Chung et al., 2021) has made it possible to learn from raw audio speech representations that are good for a variety of downstream tasks (wen Yang et al., 2021). In addition, these methods can be used to derive discrete tokens that operate as a kind of pseudo-text and can be used to learn a language model from raw audio (Lakhotia et al., 2021) which is able to capture both the linguistic content and the prosody (Kharitonov et al., 2022), giving rise to a host of applications: emotion conversion (Kreuk et al., 2022), dialogue generation (Nguyen et al., 2023b), speech classification (Chang et al., 2023). Even though these models are good at capturing expressivity, they

Inference	SPIRIT-LM generations	
	Prompt	Generation
<b>SPIRIT-LM-BASE</b>		
S→S	[SPEECH][Hu34][Hu301][Hu280]..[Hu34] <i>a b c d e</i>	[Hu28][Hu41][Hu123]..[Hu254] ☞ <i>f g h i j k l m n o p q r c s t u v</i>
T→S	[TEXT]The largest country in the world is	[SPEECH][Hu34][Hu20][Hu340]..[Hu489] ☞ <i>Russia. It has about 150 millions inhabitants..</i>
T→S	[TEXT]Here’s a story about a llama that can speak:	[SPEECH][Hu12][Hu41][Hu123]..[Hu254] ☞ <i>This little llama had a friend named doobby..</i>
S→T	[SPEECH][Hu34][Hu71][Hu405]..[Hu34] ☞ <i>Yellowstone national park is an american national park located in</i>	[TEXT] the northwest corner of Wyoming. It is located in the Greater Yellowstone area...
S→T	[SPEECH][Hu34][Hu301][Hu280]..[Hu34] ☞ <i>one two three four five</i>	[TEXT] 6 7 8 9 10...
<b>SPIRIT-LM-EXPRESSIVE</b>		
S→T	[SPEECH][St3][Pi0][Hu34][Hu103]..[Hu22] ☞ <i>Are you really going to do that &lt;Angry&gt;</i>	[TEXT] he said in a voice that was almost a scream i’m afraid
S→T	[SPEECH][St5][Pi5][Hu34][Hu409]..[Hu24] ☞ <i>Are you really going to do that &lt;Disbelief&gt;</i>	[TEXT] she said turning her head quickly and putting out her hand for the glasses
T→S	[TEXT]I am so deeply saddened	[SPEECH][Hu34][St2][Pi9][Hu371][Hu20]..[Hu89] ☞ <i>...this moment is very very hard to me... &lt;Sad&gt;</i>
T→S	[TEXT]Your actions have made me incredibly angry	[SPEECH][Hu37][St1][Pi3][Hu38][Hu111]..[Hu98] ☞ <i>So what you think you could talk about it to me &lt;Angry&gt;</i>

Table 1: SPIRIT-LM generations with text (T) or speech (S) prompt and elicited to generate text (marked with special token [TEXT]) or speech (marked with special token [SPEECH]). We report the transcribed speech examples under the speech sequence indicated with ☞ and < > (e.g., <Angry>) is appended when the speech is presented with the associated emotion. SPIRIT-LM models are Llama-2 7B models (Touvron et al., 2023a) fine-tuned with text (BPE) and speech tokens where Hubert token (cf. § 3.1) is denoted as [Hu], while [Pi] and [St], used exclusively in SPIRIT-LM-EXPRESSIVE (cf. § 3.2), represent the Pitch token and the Style token, respectively. SPIRIT-LM models enable semantically consistent multimodal generations, few-shot learning for text and speech tasks, cross-modal inference (text to speech and speech to text) and expressive generations. The samples can be found at our demo webpage<sup>1</sup>.

trail text models in capturing semantics when trained with comparable amounts of data (see Nguyen et al., 2020, 2023b). In this work, we use semantic speech tokens extracted from HuBERT (Hsu et al., 2021), possibly combined with pitch and style tokens (as in Kharitonov et al., 2022), and supplement the model training with textual bpe-units.

**Speech and Speech+Text LMs** There has been an increasing number of SpeechLMs since GSLM (Lakhotia et al., 2021). AudioLM (Borsos et al., 2023) utilizes two types of discrete speech tokens: semantic tokens (derived from w2v-BERT, Chung et al., 2021), and acoustic tokens (derived from SoundStream, Zeghidour et al., 2021) to capture semantic and acoustic information from speech respectively. They model speech in a multi-stage fashion (semantic → coarse acoustic → fine-grained acoustic) in order to generate speech in the

same acoustic style as the prompt while being semantically coherent. Vall-E (Wang et al., 2023a) models speech with acoustic tokens (Encodec, Défossez et al., 2022) and perform TTS task by translating phonemes to tokens using an autoregressive LM. Hassid et al. (2023) found that fine-tuning pre-trained TextLMs helps boost the performance of SpeechLMs. SpeechGPT (Zhang et al., 2023a) further fine-tune speechLMs on cross-modal tasks (ASR, TTS) and chain-of-modality Question-Answering (QA) task (Q-speech → Q-text → A-text → A-speech) to perform spoken QA tasks. Similar to SpeechGPT, Spectron (Nachmani et al., 2023) utilizes text as a proxy for spoken QA and speech continuation tasks (speech-prompt → text-prompt → text-continuation → speech-continuation). Unlike previous work, they represent speech using a spectrogram and employ a pre-trained speech encoder (USM, Zhang et al., 2023b) to extract speech features. In the same

	Hours	N Tokens		P Samp.	Epochs
		Speech	Text		
Speech-only	458K	28.2B		33.3%	1.24
Speech+Text	111K	7.0B	1.4B	33.3%	3.81
Text-only			307B	33.3%	0.11

Table 2: **Statistics of training data.** P Samp. is the Sampling Proportion of each subset for a training batch. Epochs is the number of epochs seen for each subset after 100K training steps or equivalently 100B tokens. For Speech+Text datasets, Epochs can be varied for different training tasks as speech & text tokens can be dropped.

spirit, Fathullah et al. (2023) propose replacing the text questions with their speech versions during the fine-tuning of a chat LLAMA 2 model to obtain an end-to-end model able to perform speech question answering, speech translation, and audio summarization tasks. AudioPALM (Rubenstein et al., 2023) and ViOLA (Wang et al., 2023b) both train autoregressive language models on text and speech in a multi-task fashion and focus on Speech Recognition (ASR), Speech Synthesis (TTS) and Speech Translation (AST, S2ST) tasks. Most recently, VoxLM (Maiti et al., 2023) and SUTLM (Chou et al., 2023) jointly trained speech and text LMs on ASR, TTS, and speech/text continuation tasks. Our work is mainly similar to Chou et al. (2023) in the training tasks but with the additional capacity of performing cross-modal generation and expressive speech and text generation. We also study larger models and evaluate their zero-shot and in-context learning capabilities.

### 3 Methods

SPIRIT-LM models are based on continuously pretraining a text-pretrained language model on a combination of text and speech (Figure 1.a). Following Hassid et al., 2023, we continuously pretrain LLAMA 2 (Touvron et al., 2023b) using a collection of text-only datasets, speech-only datasets and aligned speech+text datasets fed to the model with *interleaving*. We evaluate all our models on speech and text comprehension metrics (sWUGGY, sBLIMP, Nguyen et al., 2020; sStoryCloze, tStoryCloze, Hassid et al., 2023; MMLU Hendrycks et al., 2021) and downstream tasks such as ASR, TTS and speech classification.

SPIRIT-LM comes in two versions: SPIRIT-LM-BASE and SPIRIT-LM-EXPRESSIVE. SPIRIT-LM-BASE models speech

Model	#shots	Accuracy $\uparrow$			
		T $\rightarrow$ T	T $\rightarrow$ S	S $\rightarrow$ S	S $\rightarrow$ T
SPIRIT-LM-BASE	0	0.69	0.33	0.33	0.32
SPIRIT-LM-EXPRESSIVE	0	0.68	<b>0.43</b>	<b>0.48</b>	0.33
<i>Few-Shot Prompting</i>					
	3	0.67	0.34	0.42	0.34
SPIRIT-LM-EXPRESSIVE	6	<b>0.70</b>	0.36	0.45	<b>0.37</b>
	9	0.63	0.36	0.46	0.34
Random Predictor		0.33	0.33	0.33	0.33
<i>Cascade Topline</i>					
(ASR) + LLAMA 2 + (TTS)	0	0.64	0.34	0.32	0.36
Prompt Performance	0	0.86		0.96	

Table 3: **Zero-Shot and Few-Shot Performance on the SPEECH-TEXT SENTIMENT PRESERVATION benchmark.** SPIRIT-LM models (trained for 100k steps) are presented with prompts expressing a positive, negative or neutral sentiment. In the speech modality the sentiment is in the audio quality (laughter, cries, etc), and in text it is in the semantic content. The continuation is then elicited across modalities or, as a control, in the same modality, and tested with pre-trained classifiers. The last row (Prompt Performance) presents the performance when we apply the classifier directly on the text or speech prompt.

using HuBERT tokens (Hsu et al., 2021) while SPIRIT-LM-EXPRESSIVE uses the concatenation of HuBERT, pitch and style tokens.

#### 3.1 SPIRIT-LM-BASE

The SPIRIT-LM-BASE model is based on the 7B version of LLAMA 2 trained on Text-only, Speech-only, and aligned Speech+Text datasets.

**Speech Encoder** We use the same HuBERT model as in TWIST (Hassid et al., 2023), which is trained on a mixture of datasets: Multilingual LibriSpeech (Pratap et al., 2020), Vox Populi (Wang et al., 2021), Common Voice (Ardila et al., 2020), Spotify (Clifton et al., 2020), and Fisher (Cieri et al., 2004). The HuBERT model was trained for 4 iterations, with a downsampling factor of 640, resulting in a sample rate of 25hz. For the quantization, we utilized k-means 500 units from TWIST as base units and trained a feed-forward quantizer using data-invariant augmentation technique from Gat et al. (2023). We finally obtained a vocabulary of 501 semantic speech tokens.

**Speech and Text Tokenization** We tokenize text with the default LLAMA’s tokenizer and speech with the HuBERT tokenizer described above. Following previous work, HuBERT tokens are deduplicated for better modeling



Model	Task	WUGGY $\uparrow$		BLIMP $\uparrow$		Topic-StoryCloze $\uparrow$				StoryCloze $\uparrow$				MMLU $\uparrow$
		T	S	T	S	T	S	T $\rightarrow$ S	S $\rightarrow$ T	T	S	T $\rightarrow$ S	S $\rightarrow$ T	T
Previous Work														
GSLM (Lakhotia et al., 2021)		$\emptyset$	64.8	$\emptyset$	54.2	$\emptyset$	66.6	$\emptyset$	$\emptyset$	$\emptyset$	53.3	$\emptyset$	$\emptyset$	$\emptyset$
AudioLM (Borsos et al., 2023)		$\emptyset$	<b>71.5</b>	$\emptyset$	64.7	$\emptyset$	–	$\emptyset$	$\emptyset$	$\emptyset$	–	$\emptyset$	$\emptyset$	$\emptyset$
Voxtlm (Maiti et al., 2023)		<b>80.3</b>	66.1	<b>74.2</b>	57.1	–	–	–	–	–	–	–	–	–
TWIST (Hassid et al., 2023)		$\emptyset$	74.5	$\emptyset$	<b>59.2</b>	$\emptyset$	76.4	$\emptyset$	$\emptyset$	$\emptyset$	55.4	$\emptyset$	$\emptyset$	$\emptyset$
Ours														
SPiRIT-LM-BASE		<b>80.3</b>	69.0	73.3	58.3	<b>98.0</b>	<b>82.9</b>	<b>72.7</b>	<b>88.6</b>	<b>79.4</b>	<b>61.0</b>	<b>59.5</b>	<b>64.6</b>	<b>36.9</b>
SPiRIT-LM-EXPRESSIVE		75.8	65.0	73.6	54.2	97.9	75.4	61.6	73.2	78.9	56.9	54.6	58.8	33.3
Cascade Topline														
(ASR +) LLAMA 2		84.1	79.2	72.8	71.6	98.5	94.76	94.76	94.76	81.9	75.7	75.7	75.7	46.2

Table 4: **Zero- and few-shot comprehension evaluation.** Reporting accuracy based on negative-log-likelihood – normalized by the number of tokens – minimization prediction. MMLU is evaluated in the 5-shots prompting setting. The other tasks are evaluated in the zero-shot setting. T refers to the text modality and S to the Speech modality. We fill with  $\emptyset$  the task and modality that are not supported by the reported system, and with \_ the scores that are not publicly available.

quality. For uni-modal datasets (Text-only and Speech-only), we tokenize the data and prepend them with the corresponding modality token, i.e. "[TEXT]this is a text sentence" or "[SPEECH][Hu262][Hu208][Hu499][Hu105]".

**Interleaving Speech and Text** For the aligned Speech+Text datasets, we mix text and speech by interleaving speech and text at the word level (Figure 1.b), making the input look like this "[TEXT]the cat [SPEECH][Hu3][Hu7]..[Hu200][TEXT]the mat"<sup>2</sup>. Our hypothesis is that interleaving training will help the model learn an alignment between speech and text tokens, unlocking better text to speech transfer. The speech and text spans within the sentences are sampled randomly at each training step.

**Speech Decoder** As for speech synthesis from speech tokens, we train a HifiGAN (Kong et al., 2020; Polyak et al., 2021) vocoder on the Espresso dataset. The HifiGAN model is conditioned on HuBERT speech tokens and 1-hot speaker embedding from one of 4 Espresso’s voices.

### 3.2 SPiRIT-LM-EXPRESSIVE

Previous work shows that HuBERT tokens can capture good semantic information from speech but perform badly at expressivity (Nguyen et al., 2023a). Our goal is to have a model that can understand and preserve the emotion in the in-

put speech while being biometric-free. We therefore supplement semantic speech tokens from HuBERT with additional *pitch tokens* and *style tokens* and include them in language model training so that our trained SPiRIT-LM-EXPRESSIVE model can capture and generate more expressive speech.

**Pitch Tokens** Following Polyak et al. (2021) and Kharitonov et al. (2022), we produce pitch tokens using a VQ-VAE (van den Oord et al., 2017) model trained on the F0 of the input speech. Following the implementation of Polyak et al. (2021)<sup>3</sup>, we trained a VQ-VAE model on the Espresso (Nguyen et al., 2023a) dataset with a codebook size of 64 and a downsampling rate of 128, resulting in 12 pitch tokens per second. For training the pitch quantizer, the F0 is extracted using pyannote<sup>4</sup>. However, for the language model training, we extract F0 using FCPE<sup>5</sup>, a fast pitch estimator using Transformer, for inference speed.

**Style Tokens** We extract speechprop features from Duquenne et al. (2023), which capture speech input’s expressive style. The features were pooled with average pooling over input segments of 1 second, making one feature every one second. In order to keep style tokens biometric-free, we further remove speaker information from speechprop features by fine-tuning the features to predict the expressive style on the Espresso dataset which serves as a normalization step to obtain the style features. We finally train a k-means clustering on

<sup>2</sup>with "[Hu3][Hu7]..[Hu200]" being the tokenization of the spoken utterance "sat on"

<sup>3</sup><https://github.com/facebookresearch/speech-resynthesis>

<sup>4</sup><https://github.com/pyannote/pyannote-audio>

<sup>5</sup><https://github.com/CNChTu/FCPE>

the normalized features of Espresso dataset with 100 units.

**Expressive Speech Tokenization** We mix the 3 types of tokens (HuBERT tokens at 25hz, pitch tokens at 12.5hz, style tokens at 1hz) into a single sequence of tokens by sorting the tokens with their corresponding timestamps (Figure 1.c). Similar to SPIRIT-LM-BASE, we deduplicate HuBERT tokens as well as pitch tokens, making the input sequence look like this: "[SPEECH][St10][Pi0][Hu28][Hu22][Pi14][Hu15][Pi32][Hu78][Hu234][Hu468]"

Apart from the speech tokenization, the training details of SPIRIT-LM-EXPRESSIVE are the same as for SPIRIT-LM-BASE.

**Expressive Speech Decoder** We train a Hi-fiGAN model conditioned on HuBERT tokens, pitch tokens, style tokens and 1-hot speaker embedding from Espresso’s voices.

### 3.3 Training Details

Our SPIRIT-LM models are trained on a combination of speech, text and aligned speech+text sequences. We report in Table 2 the amount and sampling proportion of each type of data and list the datasets we use here:

**Text-only datasets** We include a subset of LLaMA (Touvron et al., 2023a) training datasets, where we exclude datasets that are unrelated to speech, like code, totaling 300B text tokens.

**Speech-only datasets** We employ open-sourced large-scale speech datasets, totaling 460K hours of speech or 30B speech tokens.

**Aligned Speech+Text datasets** We use a small subset of speech datasets that came along with text transcriptions. We then collect speech-text alignments at word-level either through the provided dataset or by performing an alignment at the word level using aligner tool from Pratap et al. (2023)<sup>6</sup>. All the alignments are automatically curated, and thus, possible errors in the alignments are admitted. The speech+text datasets comprise of 110K hours of speech or 7B speech tokens (HuBERT) and 1.5B text tokens.

In total, we have 570K hours of speech. As the number of tokens differs a lot in different modali-

ties, we tuned the sampling weights of the datasets so that the model sees each modality (speech, text, speech+text) roughly equal number of times during training.

**Optimization** Following Rubenstein et al. (2023), we extend the embeddings of LLaMa vocabulary with new speech tokens and modality tokens. The new tokens’ embeddings are initialized randomly. We then continue to pre-train the 7B LLAMA 2 model with the constant final learning rate of  $3.0e^{-5}$ , a sequence length of 4k (equivalent to 200 seconds of speech only), and a batch size of 4 per GPU. We trained the model on 64 A100 GPUs, making an efficient batch size of 1M tokens, for 200K steps. Following Xiong et al. (2023) and Rozière et al. (2024), we make a small modification to the RoPE positional encoding by increasing the “base frequency”  $\theta$  of ROPE from 10,000 to 100,000, which has been shown to benefit long-context modeling. Finally, for the speech-text interleaving sampling strategy, we randomly select the word spans so that each text sequence contains 10-30 words and each speech sequence contains 5-15 words, we do this in order to balance the portion of speech tokens and text tokens in the input sequences<sup>7</sup>.

### 3.4 Evaluation

We evaluate SPIRIT-LM checkpoints in a large number of scenarios and use cases. First, to showcase the semantic abilities of our models in speech, we report the transcript of speech generations collected by prompting the model with text or speech sequences. As illustrated in Table 1, SPIRIT-LM is able to generate semantically and expressively consistent speech when prompted with speech tokens or text tokens.

Second, we evaluate our models quantitatively with an extensive collection of benchmarks that require generating text or speech tokens:

**Speech- and Text- only Tasks** We use sWUGGY, sBLIMP, StoryCloze, and speech classification tasks. All these tasks take as input a sequence of speech tokens and measure if the model is able to find the correct sequence among two choices.

<sup>6</sup>[https://pytorch.org/audio/main/tutorials/ctc\\_forced\\_alignment\\_api\\_tutorial.html](https://pytorch.org/audio/main/tutorials/ctc_forced_alignment_api_tutorial.html)

<sup>7</sup>In our initial experiments, we found that changing the length of word spans has little impact on our evaluation metrics, but we do expect a more detailed analysis of this on longer context metrics in further work.

Model	Task	WUGGY $\uparrow$		BLIMP $\uparrow$		Topic-StoryCloze $\uparrow$				StoryCloze $\uparrow$				MMLU $\uparrow$
		T	S	T	S	T	S	T $\rightarrow$ S	S $\rightarrow$ T	T	S	T $\rightarrow$ S	S $\rightarrow$ T	T
SPiRIT-LM <i>variants</i>														
SPiRIT-LM-BASE		<b>80.3</b>	69.0	73.3	58.3	98.0	<b>82.9</b>	<b>72.7</b>	<b>88.6</b>	<b>79.4</b>	<b>61.0</b>	59.5	64.6	36.9
- No Interleaving		74.7	67.1	72.6	57.2	97.7	74.0	57.5	71.9	78.2	60.1	54.2	56.4	32.1
- Randomly-initialize		78.1	<b>69.9</b>	72.9	<b>58.8</b>	97.6	81.8	70.2	88.1	73.7	58.0	58.2	62.5	25.8
- Rope $\theta$ default		78.2	69.5	73.3	57.7	<b>98.2</b>	82.0	72.0	88.3	78.9	60.9	<b>59.8</b>	<b>65.5</b>	34.3
- +ASR+TTS		76.8	68.7	71.7	57.2	97.7	81.6	71.6	86.1	77.4	59.9	58.8	63.5	31.4
<i>Parallel Data Training</i>														
Word-level transcription		74.7	67.1	72.6	57.2	98.0	80.3	57.5	71.9	78.2	60.1	54.2	56.4	32.1
ASR+TTS-only		76.5	69.8	73.3	57.6	97.3	74.9	63.5	71.8	76.3	54.6	53.9	54.0	34.4
<i>Unimodal Models</i>														
Speech Only		67.1	69.5	53.7	58.0	54.8	72.9	52.2	49.4	53.7	54.8	52.6	49.3	27.2
Text Only		72.6	46.8	<b>73.9</b>	52.6	98.2	51.7	47.5	51.7	79.0	50.2	47.3	52.1	<b>40.1</b>

Table 5: **Ablation experiments in Zero- and few-shot comprehension evaluation.** All the models reported are initialized from LLAMA 2 7B (except Randomly-initialize one) and are trained for 100k steps. Reporting accuracy based on negative-log-likelihood – normalized by the number of tokens – minimization prediction. MMLU is evaluated in the 5-shots prompting setting. The other tasks are evaluated in the zero-shot setting. T refers to the text modality and S to the Speech modality. For a full comparison of unnormalized and normalized scoring accuracy, refer to Table 10 in the Appendix.

sWUGGY and sBLIMP are described in detail in [Nguyen et al. \(2020\)](#). Briefly, sWUGGY measures if the model can discriminate between existing spoken words and non-words (e.g., “brick” vs. “blick”). sBLIMP measures if the model can distinguish between a spoken grammatically correct sentence and an ungrammatical spoken variant of the same sentence (e.g., “cats are lazy” vs. “cats is lazy”). Given the beginning of a short spoken story, StoryCloze measures if the model can find the plausible ending among two sentences, which typically requires some high-level semantic understanding and common sense ([Mostafazadeh et al., 2017](#)). We use the spoken version of the original storycloze (S-StoryCloze) and the topic-Storycloze (T-StoryCloze) assembled by [Hassid et al. \(2023\)](#) based on simpler negative samples. All of these tasks have a random baseline performance of 50%. All these tasks are evaluated in the 0-shot prompting setting. We predict the sample with the highest likelihood of the two choices. In addition to speech, these benchmarks are also available in the text modality. We, therefore, measure the text-modeling abilities of SPiRIT-LM on these. In addition, we evaluate SPiRIT-LM on MMLU ([Hendrycks et al., 2021](#)), a popular evaluation benchmark for LLMs in the text modality. Finally, we evaluate SPiRIT-LM on the Intent-Classification task from [Chang et al. \(2023\)](#).

### Speech-to-Text and Text-to-Speech Tasks

SPiRIT-LM is trained in both speech and text. For this reason, it has the ability to model tasks that require both text and speech modeling. We evaluate SPiRIT-LM for ASR. We report the Word-Error-Rate (WER) between the generated and the gold transcriptions. For text-to-speech (TTS), we consider our system’s ability to generate the audio corresponding to the inputted text. We measure the performance by transcribing the generated audio with Whisper ([Radford et al., 2023](#)), a state-of-the-art ASR model, and we compare it with the original text with Character-Error-Rate. Both these tasks are evaluated in English with Librispeech clean and other test sets.

### 3.5 Baselines

We compare our results with previously published generative speech systems. All these methods use one or several Transformer ([Vaswani et al., 2017](#)) decoder-only models trained on speech units. They differ in how they are trained (pre-trained from scratch or fine-tuned), the types of speech units they model, and their amount of training data. GSLM ([Lakhotia et al., 2021](#)) is based on speech units (e.g. Hubert) and trained from scratch on speech-unit modeling. TWIST ([Hassid et al., 2023](#)) is a textually pretrained speech model based on Llama-13B ([Touvron et al., 2023a](#)). Au-

Model	Task	LS clean (10 shots)		LS other (10 shots)		IC (30 shots)
		ASR↓	TTS↓	ASR↓	TTS↓	↑
SPiRiT-LM <i>variants</i>						
SPiRiT-LM-BASE		21.9	45.5	29.2	43.8	71.9
+ASR+TTS		<b>6.0</b>	<b>6.7</b>	<b>11.0</b>	<b>7.9</b>	<b>75.8</b>
SPiRiT-LM-EXPRESSIVE		37.9	52.0	50.0	53.6	66.2
<i>Parallel Data Training</i>						
Word-level transcription		113.2	85.2	111.6	75.2	22.6
ASR+TTS only		7.7	8.1	11.9	9.4	7.4
<i>Cascade Topline</i>						
(WHISPER +) LLAMA 2 (+MMS TTS)		3.7	4.0	7.2	4.9	89.6

Table 6: **Few-shot tasks.** We evaluate SPiRiT-LM models for Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) Evaluation on LibriSpeech (LS) and Intent Classification (IC). ASR scores correspond to Word-Error-Rate (% WER) evaluated in the 10-shots setting with a max context length of 1024. TTS scores correspond to the Character-Error-Rate (% CER) in the 10-shots setting with a max context length of 2048. IC scores correspond to accuracy in the 30 shots setting.

dioLM (Borsos et al., 2023) is a cascade system made of a semantic sequence model (using w2v-BERT, Chung et al., 2021) combined with coarse-acoustic and fine-acoustic models (using SoundStream units, Zeghidour et al., 2021). In contrast with SPiRiT-LM, the approach mentioned above only relies on speech units during training, making them speech-only models (i.e. they do not support text understanding nor generation).

We also compare our models to VoxLM (Maiti et al., 2023), a concurrent work on speech and text language modeling. We report the best scores from the original published papers for all the mentioned methods.

As a top-line comparison, we compare our models with cascade models that use LLAMA 2 as a text generative model. For text-to-text (T→T), we only rely on LLAMA 2-7B. For speech-to-speech (S→S), we utilize the cascade model, ASR from WHISPER-MEDIUM (Radford et al., 2023), followed by LLAMA 2, synthesized by MMS-TTS (Pratap et al., 2023).

## 4 Speech and Text Understanding

### 4.1 Lexical, Grammatical and Semantic Knowledge in Text and Speech

We find that SPiRiT-LM-BASE competes with the baselines for WUGGY, BLIMP, and Storycloze in the speech modality while preserving competitive text performance (cf. Table 4). More specifically, SPiRiT-LM-BASE outperforms the baselines by a large margin on StoryCloze, which requires the most advanced speech semantic abilities

compared to the other reported benchmarks.

**Interleaving is critical** We run ablation experiments (cf. Table 5) to understand what leads to this performance by controlling for the training budget and ablating a large number of training parameters. We set the training budget at 100k training steps or 100B tokens.

We compare SPiRiT-LM-BASE to a LLAMA 2 model continuously pretrained with two parallel data training settings. First, the ASR+TTS-only model consists of training with pairs of semantically equivalent sequences of speech and text (e.g. “[TEXT] the cat jumped by the window [TTS][Hu12]..[Hu54]” or “[SPEECH][Hu12]..[Hu54][ASR] the cat jumped by the window”<sup>8</sup>). Second, the Word-level Transcription model consists of training on sequences of pairs of textual and spoken words (e.g. “[TEXT] the [SPEECH][Hu12]..[Hu34] [TEXT] cat [SPEECH][Hu454]..[Hu90]...[TEXT] window [SPEECH][Hu15]..[Hu54]”). Additionally, we compare SPiRiT-LM-BASE to models trained on a single modality (speech or text) and with speech+text but without any interleaving data (cf. No Interleaving in Table 5).

Based on these experiments, we conclude that interleaving training is the primary factor leading to good-quality speech generation. Fine-tuning LLAMA 2 on parallel data leads to lower performance on tasks such as StoryCloze and BLIMP. Notably, fine-tuning the model on speech-only to-

<sup>8</sup>with “[Hu12]..[Hu54]” being the tokenization of the spoken utterance “the cat jumped by the window”



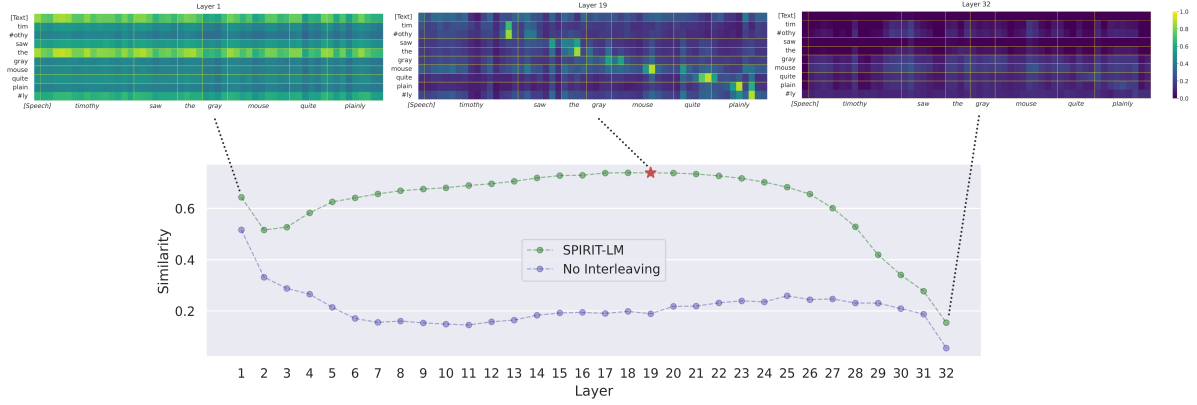


Figure 2: **Alignments of features obtained from Text and Speech Inputs.** **Bottom:** Similarity of speech and text features extracted from different layers of SPIRIT-LM compared with the model training without speech-text interleaving. The similarity is computed as the maximum similarity over speech and text features of the same words and is averaged over a test set. **Top:** Pairwise cosine similarity between text features and speech features of the same sentence extracted from different layers of SPIRIT-LM.

kens leads to a much lower performance (e.g. more than 6 points difference with SPIRIT-LM on spoken StoryCloze). This shows that interleaving training not only helps preserve the text generation abilities of the model but also leads to better speech understanding and generation performance. We measure the importance of the amount of aligned data used for interleaving training in Figure 5. We find that the model’s performance in speech (T-StoryCloze) steadily increases with the amount of aligned data.

As shown in Table 4, SPIRIT-LM-EXPRESSIVE performs lower than SPIRIT-LM-BASE on these tasks, indicating that the expressive speech units lead to moderate lexical, grammatical, and semantic understanding degradation. We explain this with the following intuition. Modeling a given raw speech for SPIRIT-LM-EXPRESSIVE is more costly than for SPIRIT-LM-BASE. Indeed, in contrast with SPIRIT-LM-BASE, SPIRIT-LM-EXPRESSIVE is based on integrating expressive speech units in the sequence during training, in addition to Hubert-tokens. This leads to extending the sequence length in the number of tokens for a fixed raw input speech. This added complexity leads to a degradation of speech modeling performance.

In the text modality, despite being fine-tuned on billions of speech tokens, SPIRIT-LM still performs decently on MMLU (above 33%) and degrades by less than 2 points on WUGGY, BLIMP, and StoryCloze compared to LLAMA 2.

Finally, on these tasks, the cascade approach (ASR with WHISPER followed by LLAMA 2) is above SPIRIT-LM by a large margin.

## 4.2 Cross-Modal Evaluation

SPIRIT-LM can also model sequences that are made of both speech and text tokens.

**Cross-Modal StoryCloze** Based on the text and speech versions of StoryCloze, we build a speech to text (S→T) and text to speech (T→S) Storycloze for which the context is in one modality (e.g. speech) and the hypothesis is in the other modality (e.g. text). As seen in Table 5, we find the performance of SPIRIT-LM-BASE in the text to speech direction (T→S) on par with the speech only performance (S). In contrast, the (S→T) direction is about 5 points above the speech performance (S). This suggests that the model performs better at text generation compared to speech generation even when it is conditioned on a speech sequence.

**ASR & TTS** Similarly to text language models, SPIRIT-LM can be prompted with few-shot examples to perform specific tasks. We illustrate this with ASR and TTS. We show in Table 6 that SPIRIT-LM models reach non-trivial performance in ASR and TTS. We find that few-shot prompting leads to the best performance with 10 shot prompting (cf. Figure 3).<sup>9</sup> Our best SPIRIT-LM-BASE

<sup>9</sup>We note that above 20 shots, we reach the maximum number of tokens that fit in the context for ASR and TTS.

model is at 21.9 Word-Error-Rate in Librispeech clean and 45.5 in Character-Error-Rate in TTS. We observe that when we add parallel ASR and TTS examples during training (cf. +ASR+TTS in Table 6), we can improve the performance from a very large margin. We note that adding ASR and TTS data has a very moderate impact on the rest of the tasks. We report the detailed prompting used for ASR and TTS in the Appendix in Section A.

**Cross-Modal Alignment** To understand better the hidden mechanism that enables SPIRIT-LM to deliver good cross-modal performance while only being trained on *interleaved* data and raw speech and text, we look at the token-level similarity of the model’s features from input sequences of HuBERT tokens and the corresponding BPE tokens. We illustrate this in Figure 2 (bottom), where we compute the maximum similarity over the same words of speech and text features extracted from different layers of SPIRIT-LM. We find that the similarity between spoken and written sequences inside the model increases from layer 2 and layer 20. In comparison, this alignment does not occur when the model is trained without interleaving (cf. Figure 2 bottom). This suggests that interleaving enables the model to map speech sequences with corresponding text sequences. Figure 2 (top) shows the alignments of BPE tokens and HuBERT tokens of the sentence *Timothy saw the gray mouse quite plainly* on layers 1, 19, 32. We see that the middle layers of SPIRIT-LM capture the same semantics information from both input modalities, with high alignments towards the end of each word (last BPE tokens, late HuBERT tokens).

### 4.3 Downstream Speech Classification

Finally, we report in Table 6 the abilities of SPIRIT-LM to perform speech classification task. We experiment with Intent-Classification (IC). We find that the accuracy improves with the number of shots (cf. Figure 3). Our best SPIRIT-LM model reaches up to 79% accuracy (compared to 89% of the topline performance). The detailed prompting used for IC is given in the Appendix A.

**Pretrained Knowledge is Essential for Few-Shot Learning** We report in the Appendix in Figure 6 the task-specific performance of SPIRIT-LM-BASE with regard to the number of training steps compared to a randomly initialized model trained in the same setting. After only 25k training steps, SPIRIT-LM-BASE reaches more

than 75% accuracy on Intent Classification while the randomly initialized model is below 20%. This means that starting from a pretrained LLAMA 2 model is essential for few-shot in-context learning and that our method successfully transfers the pre-trained few-shot learning abilities of the model to the speech modality.

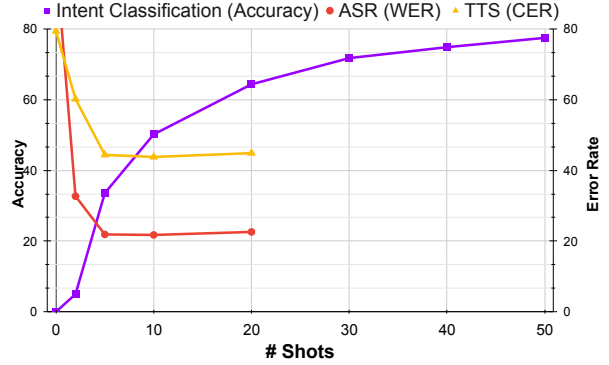


Figure 3: SPIRIT-LM-BASE performance with regard to the number of shots presented to the model context for Intent Classification, ASR and TTS.

## 5 Expressivity Evaluation

One of the core contributions of this work is the expressivity modeling. To measure the expressivity of our model we first evaluate the quality of the introduced pitch and style tokens (§ 5.1). Second, we evaluate our SPIRIT-LM models on the newly introduced SPEECH-TEXT SENTIMENT PRESERVATION benchmark (§ 5.2).

### 5.1 Style and Pitch Tokens Evaluation

We model expressive speech by complementing semantic speech tokens (HuBERT) with Pitch and Style tokens. To evaluate the quality of our tokenization, we use the speech resynthesis task from Nguyen et al. (2023a). It measures how well the resynthesized speech is compared with the original audio in terms of preserved content, expressive style, and pitch.

Table 7 shows the performance of SPIRIT-LM-BASE and SPIRIT-LM-EXPRESSIVE tokenizers compared to Encodec and Hubert-only baselines. We see the SPIRIT-LM-EXPRESSIVE tokenizer can capture good expressive style and pitch from the input speech. Additionally, we observe a very large improvement in Style and Pitch resynthesis when we compare SPIRIT-LM-BASE tokenizer with SPIRIT-LM-EXPRESSIVE.

Model	Metrics	Bitrate	Content	Style	Pitch
		BPS↓	WER↓	EMO↑	FFE↓
Original Audio		-	16.2	65.2	-
Expresso models (Nguyen et al., 2023a)					
Hubert + HifiGAN		550	23.0	22.7	0.30
Hubert + HifiGAN w/ GT Style		550	21.4	61.6	0.27
Encodec (RVQ=1)		500	38.0	41.5	0.09
Encodec (RVQ=8)		4000	19.0	56.7	0.04
SPiRiT-LM Tokenizers					
SPiRiT-LM-BASE		225	23.4	20.4	0.40
SPiRiT-LM-EXPRESSIVE		307	23.2	41.4	0.16

Table 7: **Expressive Speech Resynthesis Evaluation.** Performances of SPIRIT-LM Tokenizers on the Expresso Benchmark (Nguyen et al., 2023a) compared with their systems. The scores are averaged across datasets. For the detailed scores, refer to Table 11 in the Appendix.

## 5.2 The SPEECH-TEXT SENTIMENT PRESERVATION benchmark (STSP)

To evaluate how well our SPIRIT-LM models can understand and generate expressive speech and text, we introduce the SPEECH-TEXT SENTIMENT PRESERVATION benchmark. It is made of a collection of speech and text prompts in the positive, negative or neutral sentiment. Given a spoken or written prompt, the task consists in generating a text or speech sequence of tokens that preserves the sentiment of the prompt.

For instance, in the text-to-X direction (T→T and T→S), given a written sentence bearing sadness, we check if the spoken generated text/utterance is also sad. On the other hand, the direction speech-to-X (S→S and S→T), given a spoken happy-sounding utterance, we check whether the model generates a positively written text or positive utterance.

### 5.2.1 Sentiment-Rich Spoken and Written Prompts

**Speech Prompt** In order to have the read speech of different expressive styles (e.g. *he's done it again* in happy/sad style). We utilize two datasets: 1) *Expressive reading* from EXPRESSO (Nguyen et al., 2023a) consisting of 47 hours of expressive North American English speech where 7 different styles are applied on the same content that does not reflect the emotion being conveyed. We use only the speech from 3 emotions: "happy", "sad" and "default". (we will refer to this dataset as EXPRESSO-READ) 2) EMOV (Adigwe et al.,

The SPEECH-TEXT SENTIMENT PRESERVATION benchmark			
Prompt origin	EXPRESSO-READ	EXPRESSO-ASR	EMOV
Prompt Type	Speech	Text	Speech
#Samples	1020/60/54	1373/479/462	1053/351/351
#Speakers	4	-	3
Classes	Positive(33%) / Negative(33%) / Neutral(33%)		

Table 8: Statistics of the SPEECH-TEXT SENTIMENT PRESERVATION benchmark. (#Samples indicates the number of samples in each train/dev/test split.)

2018), composed of emotional speech from 5 different speakers and 2 languages (North American English and Belgian French). We select only the English speech from 3 speakers when the same content is recorded in three different emotions: "Amused", "Angry" and "Neutral".

**Text Prompt** In order to have expressive text (e.g. *he's such an amazing player* for positive) as prompt, we transcribe<sup>10</sup> *improvised dialog* from EXPRESSO for 4 emotions: "happy", "angry", "sad" and "default" to obtain an aligned Speech-Text dataset. Then we filter the samples if the transcription has less than 10 words (separated by space) or it has one word appearing more than 10 times. We refer to this aligned dataset by EXPRESSO-ASR.

**Sentiment Mapping** To unify different sets of emotional classes, we associate the emotions "happy"/"Amused", "sad"/"Angry" and "default"/"Neutral" to the "positive", "negative" and "neutral" sentiments.

**Data Splits** We split the datasets into train/dev/test subsets for later usage. Table 8 presents a comprehensive statistical overview of the datasets used. For EXPRESSO-READ, we use the original train/dev/test splits; while for the EMOV, we split it randomly into train/dev/test subsets with the ratios of 60/20/20. The EXPRESSO-ASR dataset is also divided into train/dev/set with the ratios of 60/20/20<sup>11</sup>. We use the train and dev subsets to train the sentiment classifiers and the test subset to prompt the SPIRIT-LM models.

<sup>10</sup>The transcription is done by WHISPER-MEDIUM (Radford et al., 2023).

<sup>11</sup>We don't use the original data splits because the amount of data in the dev and test subsets is not enough.

### 5.2.2 Evaluation Metrics

For both tasks, we check if the generated utterance has a sentiment that is consistent with the sentiment of the prompt. We assess the sentiment of the produced utterance using sentiment classifiers and report its accuracy. The accuracy for speech-to-X directions is averaged over EXPRESSO-READ and EMOV.

We obtain text and speech sentiment classifiers by fine-tuning pre-trained text and speech models respectively. For the speech classifier, similar to Nguyen et al. (2023a), we fine-tune the wav2vec2 model<sup>12</sup> on the training sets of EXPRESSO-READ, EXPRESSO-ASR<sup>13</sup> and EMOV. For the text classifier, we fine-tune the 3-classes sentiment classifier from Hartmann et al. (2021) on the transcriptions of the EXPRESSO-ASR training set.

### 5.2.3 Evaluation Settings

We tune the generation parameters on the dev sets. In terms of the maximal number of generated tokens, we use 50 for  $T \rightarrow T$  and  $S \rightarrow T$ , 200 for  $T \rightarrow S$ , and 300 for  $S \rightarrow S$ . We use a temperature of 0.8 and nucleus sampling (Holtzman et al., 2020) with a *top\_p* of 0.95 for all the directions. All the SPIRIT-LM models reported have been trained for 100k steps.

**Zero-Shot** We prompt SPIRIT-LM using positive, negative or neutral text/speech input from the test sets of the datasets described in section 5.2.1. Then 1) for  $S \rightarrow S$  and  $T \rightarrow S$ , we classify the generated speech with the speech classifier. 2) for  $T \rightarrow T$  and  $S \rightarrow T$ , we assess the text continuation with the text classifier.

**In-context Few-Shot** We also evaluate SPIRIT-LM in a few-shot setting by constructing a set of few-shot examples (see Appendix B) and feed them as the in-context prompt.

### 5.2.4 Results

We report the results evaluated on the test sets in Table 3. For zero-shot performance, SPIRIT-LM-EXPRESSIVE surpasses SPIRIT-LM-BASE in all directions, with the exception of  $T \rightarrow T$  where they perform comparably. Compared to the cascade baseline, SPIRIT-LM-EXPRESSIVE outperforms it over all the directions except  $S \rightarrow T$ .

<sup>12</sup><https://huggingface.co/facebook/wav2vec2-base>

<sup>13</sup>We use only the speech data

In the case of few-shot results, we observe that few-shot is only beneficial when the output modality is text.

For both zero-shot and few-shot, the sentiment continuation is better preserved within the same modality than across different modalities. Among all the directions,  $S \rightarrow T$  scores the lowest.

The final row of Table 3 also includes an evaluation of performance directly on the input prompt. All prompts receive high scores, suggesting a significant potential for improvement in the preservation of expressivity.

## 6 Responsible Evaluation in Speech and Text

Responsible AI is at the core of our research. This section discusses and evaluates responsibility aspects from SPIRIT-LM. Speech language models (SpeechLMs) have the potential to bring the same benefits as text-based language models (textLMs), such as access to various downstream applications like search, language generation, summarization, translation, chatbots, etc. in the audio modality. This can increase their reach to more use cases and languages, including those that are not written or sparsely written.

Quantifying and working on user safety is a key aspect from generative model development. These models can inadvertently generate content that is harmful, offensive, or inappropriate is essential for generative language models (Deshpande et al., 2023; Touvron et al., 2023a). While safety is a broad concept, we focus on the specific problem of added toxicity in the generation of the SPIRIT-LM. Inspired by conditional language models' previous studies (Seamless et al., 2023a), we define added toxicity as a toxicity increase in the generation compared to the initial source utterance.

### 6.1 Data

We use the HOLISTICBIAS dataset (Smith et al., 2022) and its synthesized speech extension (Seamless et al., 2023a). This dataset has been shown to trigger toxicity for conditional language models (Costa-jussà et al., 2023). The dataset consists of over 472k English sentences. Sentences are created by combining a sentence template, such as "What do you think about *PLURAL NOUN PHRASE*?", a noun (e.g. parents) and a descriptor (e.g., disabled), covering a list of 26 templates and



600 descriptors across 13 demographic axes (e.g., ability, race or gender). We utilize the dataset as the prompt for generating text (T→T) and speech (S→S), respectively.

Task	T→T		S→S	
	ETOX ↓	MuTox ↓	ASR-ETOX ↓	MuTox ↓
SPiRiT-LM-BASE	1.19	2.69	1.06	3.75
(ASR) + LLAMA 2 + (TTS)	1.22	2.63	1.17	2.70

Table 9: **Added Toxicity Detection.** The proportion of sentences with added toxicity divided by the total number of sentences. For the LLAMA 2 baseline, we use a cascaded pipeline made of WHISPER for ASR and MMS for TTS; for SPiRiT-LM-BASE, we use the model trained for 200k steps.

## 6.2 Evaluation Metrics

Similar to Seamless M4T V2 (Seamless et al., 2023b), we use MUTox and ETOX<sup>14</sup> (Costa-jussà et al., 2023) as our toxicity classifiers. For speech, we simply run ASR and evaluate toxicity with ETOX (we refer to this as ASR-ETOX). MUTox can be directly applied on both text and speech generations, without the need for an ASR system.

To compute the added toxicity, we evaluate toxicity at the sentence level, both in the input utterance/prompt and in the generated output. We report the proportion of sentences with added toxicity divided by the total number of sentences. For ETOX and ASR-ETOX, a sentence has added toxicity when there are more toxic words found in the generated content than in the prompt. For MUTox, a sentence has added toxicity when the MUTox scores are more than 0.7 higher in the generated content than in the prompt.

## 6.3 Results

We report results in Table 9. In terms of ETOX, both SPiRiT-LM and (WHISPER) + LLAMA 2 + (MMS-TTS) have comparable results. When evaluated with MUTox, however, SPiRiT-LM shows higher added toxicity especially in S→S. This might come from the fact that there exists more toxic contents in our speech training dataset. We leave the mitigation to future work.

Figure 4 shows the distribution of added toxicity in SPiRiT-LM in terms of the 13 demographic

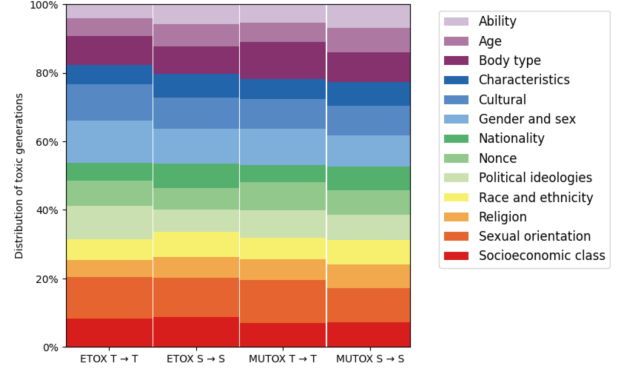


Figure 4: **Toxicity Distribution** Relative Distribution of added toxicity over the 13 demographic axes for T→T and S→S generations. The number of added toxicities are normalized by the number of occurrences in each demographic axis.

axes represented in HOLISTICBIAS and how they vary in modality. We observe that *Gender and sex* and *Sexual orientation* tend to generate more added toxicity than the rest of demographic axes, while *ability* and *nationality* tend to be among the ones that generate the least. There is no big difference in distribution across modalities or metrics.

## 7 Limitations and Broader Impacts

**Harmful applications** SPiRiT-LM also shares the same risks as its generative model predecessors (Touvron et al., 2023a), such as intentionally harmful applications like fake news and spamming as well as unintentionally harmful ones like unfair or biased results, toxic or untrustworthy generations. These risks can be assessed and mitigated using watermarking e.g. (Kirchenbauer et al., 2023) or existing reinforcement learning from human feedback (RLHF) e.g. (Bai et al., 2022). In addition to these traditional text risks, SPiRiT-LM, being a speech model, also extends risks associated with this modality with intentionally harmful applications like impersonating a specific speaker by continuing short speech segments while maintaining speaker identity and prosody. Mitigation measures for this risk include similar ones as with text (speech watermarking (Seamless et al., 2023b) and RLHF). Similarly to text models, unintentionally harm may arise such as the lack of speaker robustness where the model can generate speech continuations inconsistent with the prompt in terms of accent and dialect only for underrepresented groups in the training data. Among the mitigation strategies, we can include: increasing

<sup>14</sup>Freely available at [https://github.com/facebookresearch/seamless\\_communication](https://github.com/facebookresearch/seamless_communication)

the variety of the dataset, compensating for bias in representation of different demographics.

**Future Work** In this paper, we showed how combining style and pitch tokens with semantics tokens and continuously pretraining a text language model delivers very promising multimodal semantic abilities while enabling expressive speech generations. However, several architectural and training improvements could further progress in speech generation.

First, training multimodal models remains a challenge. In this work, we observed that despite training on both speech and text, our SPIRIT-LM models do not perform as well as the initial LLAMA 2 model in text generation. Refining the training procedure could potentially reduce this gap. Second, we restricted our evaluation to English. SPIRIT-LM models were trained on a large amount of non-English data. More investigation is needed to assess the quality and safety of the model in non-English languages. Third, we only experimented with 7B models. Scaling our experiments beyond 7B could lead to much better performance. Finally, the introduced SPIRIT-LM models are foundational models. This means that more work is needed to make them safe and aligned with user expectations. As it is now commonly done with text (Ouyang et al., 2022; Touvron et al., 2023b), fine-tuning a model with instructions and preference data in speech could potentially unlock new experiences such as fully expressive dialog systems.

## 8 Conclusion

We introduced SPIRIT-LM, a speech + text generative language model based on LLAMA 2 that can generate both speech and text in a cross-modal manner. We showed that by alternating speech and text in the input sequence during training, the model is able to generate the content fluidly by changing from one modality to another. We evaluated our models on a collection of speech and text metrics. We plan to make future improvements both in the area of model capability and in transparency and safety.

## References

- Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.
- Robin Algayres, Yossi Adi, Tu Anh Nguyen, Jade Copet, Gabriel Synnaeve, Benoit Sagot, and Emmanuel Dupoux. 2023. [Generative spoken language model based on continuous word-sized audio tokens](#).
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. [Common voice: A massively-multilingual speech corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.
- Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 33, pages 12449–12460, Online.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. 2022. [Training a helpful and harmless assistant with reinforcement learning from human feedback](#).
- Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. 2023. Audioldm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. [Language models are few-shot learners](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.
- Kai-Wei Chang, Yu-Kai Wang, Hua Shen, Iu thing Kang, Wei-Cheng Tseng, Shang-Wen Li, and Hung yi Lee. 2023. [Speechprompt v2: Prompt tuning for speech classification tasks](#). *ArXiv*, abs/2303.00733.
- Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. [Wavlm: Large-scale self-supervised pre-training for full stack speech processing](#). *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.
- 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. *arXiv preprint arXiv:2310.08715*.
- 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 Sankaranarayanan 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. 2022. [Palm: Scaling language modeling with pathways](#).
- Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. [w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training](#). In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.
- Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *LREC*.
- Ann Clifton, Sravana Reddy, Yongze Yu, Aaish Pappu, Rezvaneh Rezapour, Hamed Bonab, Maria Eskevich, Gareth Jones, Jussi Karlgren, Ben Carterette, and Rosie Jones. 2020. [100,000 podcasts: A spoken English document corpus](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5903–5917.
- Marta R Costa-jussà, Eric Smith, Christophe Ropers, Daniel Licht, Jean Maillard, Javier Ferrando, and Carlos Escolano. 2023. Toxicity in multilingual machine translation at scale. In *EMNLP*.
- Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. [Toxicity in chatgpt: Analyzing persona-assigned language models](#).
- Paul-Ambroise Duquenne, Kevin Heffernan, Alexandre Mourachko, Benoît Sagot, and Holger Schwenk. 2023. Sonar expressive: Zero-shot expressive speech-to-speech translation.
- Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*.
- Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangquan, Jay Mahadeokar,

- Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. [Towards general-purpose speech abilities for large language models using unpaired data](#).
- Itai Gat, Felix Kreuk, Tu Anh Nguyen, Ann Lee, Jade Copet, Gabriel Synnaeve, Emmanuel Dupoux, and Yossi Adi. 2023. [Augmentation invariant discrete representation for generative spoken language modeling](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 465–477, Toronto, Canada (in-person and online). Association for Computational Linguistics.
- Jochen Hartmann, Mark Heitmann, Christina Schamp, and Oded Netzer. 2021. The power of brand selfies. *Journal of Marketing Research*.
- Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2023. Textually pre-trained speech language models. *arXiv preprint arXiv:2305.13009*.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *International Conference on Learning Representations*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. 2022. [Training compute-optimal large language models](#).
- 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 Trans. Audio, Speech and Lang. Proc.*, 29:3451–3460.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. [Scaling laws for neural language models](#).
- Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. [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, Dublin, Ireland. Association for Computational Linguistics.
- John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. [A watermark for large language models](#).
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. In *Advances in Neural Information Processing Systems*, volume 33, pages 17022–17033.
- Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu Anh Nguyen, Morgan Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2022. [Text-less speech emotion conversion using discrete & decomposed representations](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11200–11214, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- 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. 2021. [On Generative Spoken Language Modeling from Raw Audio](#). *Transactions of the Association for Computational Linguistics*, 9:1336–1354.
- Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe.



2023. Voxlm: unified decoder-only models for consolidating speech recognition/synthesis and speech/text continuation tasks. *arXiv preprint arXiv:2309.07937*.
- 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. 2023. [Spoken question answering and speech continuation using spectrogram-powered llm](#).
- Tu Anh Nguyen, Wei-Ning Hsu, Antony d’Aiviro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. 2023a. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.
- Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoît Sagot, Abdelrahman Mohamed, and Emmanuel Dupoux. 2023b. [Generative Spoken Dialogue Language Modeling](#). *Transactions of the Association for Computational Linguistics*, 11:250–266.
- 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 Self-Superv. Learn. Speech Audio Process.*, Online.
- Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. [Neural discrete representation learning](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- 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 *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 5206–5210.
- Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhota, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. [Speech resynthesis from discrete disentangled self-supervised representations](#). In *Proc. INTERSPEECH*.
- Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2023. [Scaling speech technology to 1,000+ languages](#).
- 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.
- Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. 2024. [Code llama: Open foundation models for code](#).
- Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsoš, Félix de Chaumont Quitry, Peter Chen,

Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramonovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. [Audiopalm: A large language model that can speak and listen](#).

Seamless, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Iliia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinеш Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023a. [Seamlessm4t: Massively multilingual & multimodal machine translation](#).

Seamless, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi,

Amanda Kallet, Iliia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinеш Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023b. Seamless: Multilingual expressive and streaming speech translation.

Eric Michael Smith, Melissa Hall, Melanie Kambadur, Eleonora Presani, and Adina Williams. 2022. [“I’m sorry to hear that”: Finding new biases in language models with a holistic descriptor dataset](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9211, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams,

- Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.
- Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. [VoxPopuli: A large-scale multi-lingual speech corpus for representation learning, semi-supervised learning and interpretation](#). In *Proceedings of Association for Computational Linguistics*, pages 993–1003.
- Chengyi Wang, Sanyuan Chen, Yu Wu, Zi-Hua Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. 2023a. [Neural codec language models are zero-shot text to speech synthesizers](#). *ArXiv*, abs/2301.02111.
- Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023b. [Viola: Unified codec language models for speech recognition, synthesis, and translation](#).
- Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabza, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2023. [Effective long-context scaling of foundation models](#).
- 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](#).
- Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507.
- Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. [Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities](#).
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona 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. 2022. [Opt: Open pre-trained transformer language models](#).
- 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 Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023b. [Google usm: Scaling automatic speech recognition beyond 100 languages](#).

## A Few-Shot Prompts

### Speech Recognition (ASR)

For ASR, we prompt the model and add special start and end flags. Indeed, we find that without these flags the model tends to hallucinate after transcribing the input sequence.

For SPIRIT-LM, we use the following prompting. We find that 10 examples leads to the best performance. We illustrate the prompting of SPIRIT-LM for ASR with a single few-shot example:

```
[SPEECH] Speech token sequence
[TEXT] <START Transcript> Text transcript <END>
[SPEECH] Speech token sequence
[TEXT]
```

For the models trained with parallel ASR data

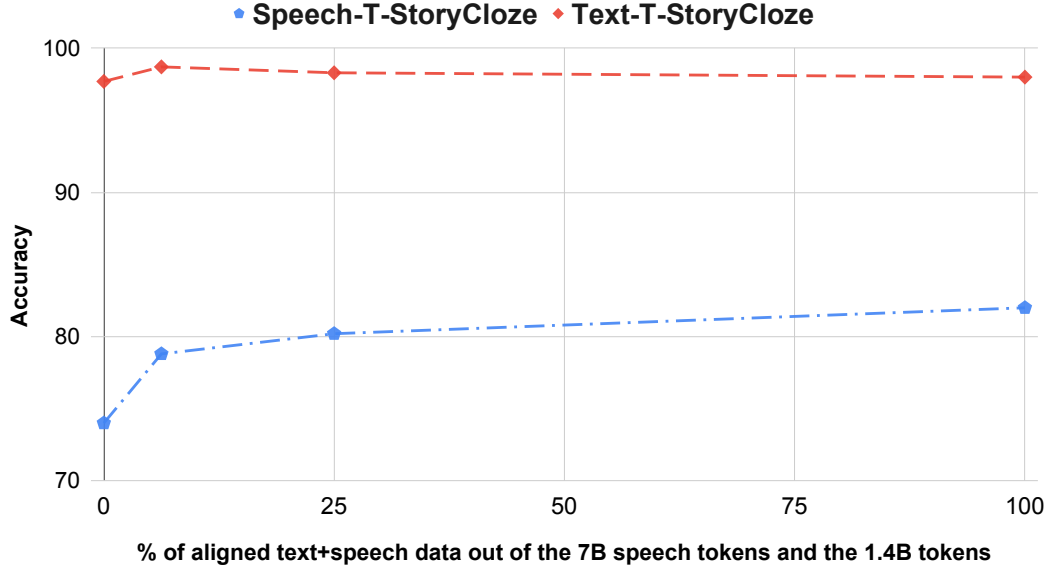


Figure 5: Performance of SPIRIT-LM-BASE on Topic-StoryCloze in speech and text with regard to the sampled amount of aligned speech+text data from 0% to 100% out of the 8.4B tokens aligned tokens. (1.4B text tokens and 7B tokens speech tokens.)

(e.g. SPIRIT-LM-BASE +ASR+TTS), [SPEECH] is replaced with the [ASR] special token to trigger the transcription prediction as seen during training.

### Text-to-Speech (TTS)

We find that prompting SPIRIT-LM with 10-shots leads to the best performance in TTS. We illustrate the prompting with a single example for few-shot learning:

[TEXT] *Input Text* 'stop'  
 [SPEECH] *Speech token sequence* <speech:STOP>  
 [TEXT] *Input Text* 'stop'  
 [SPEECH]

With <speech:STOP>, the spoken utterance “stop” tokenized into speech tokens<sup>15</sup>. For models trained with parallel TTS data (e.g. SPIRIT-LM-BASE +ASR+TTS), the token [SPEECH] is replaced with [TTS].

### Intent Classification

For Intent Classification, we illustrate the prompting used in SPIRIT-LM-BASE with single exam-

ple for few-shot:

[SPEECH] *Speech token sequence* [TEXT]  
 A:activate lights bedroom  
 [SPEECH] *Speech token sequence* [TEXT]  
 A:

For both ASR, TTS and Intent Classification, we postprocess the output of the model using the special tokens and beginning/end of sequence flags in order to extract the predicted text or speech sequence.

## B Construction of Few-Shot examples for Sentiment Continuation

We use S→T as an illustration, the identical process is applied to the remaining modality directions.

1. From the EXPRESSO-READ training set, we select only the speech samples where the waveform length exceeds 200,000, dividing each into two equal parts. The speech in the second segment is then transcribed.<sup>16</sup>
2. We apply the fine-tuned speech classifier and text classifier mentioned in 5.2.2 to the

<sup>15</sup>For SPIRIT-LM-BASE, the spoken word “stop” is tokenized as [Hu481][Hu149][Hu40][Hu48][Hu315][Hu242][Hu428][Hu494][Hu75][Hu497][Hu188][Hu388][Hu109][Hu23][Hu338][Hu23][Hu481]

<sup>16</sup>The transcription is done by WHISPER-MEDIUM (Radford et al., 2023).



speech of the first segment and the transcription of the second segment, respectively. We retain only those pairs where the sentiment of the transcription in the second segment matches that of the speech in the first segment.

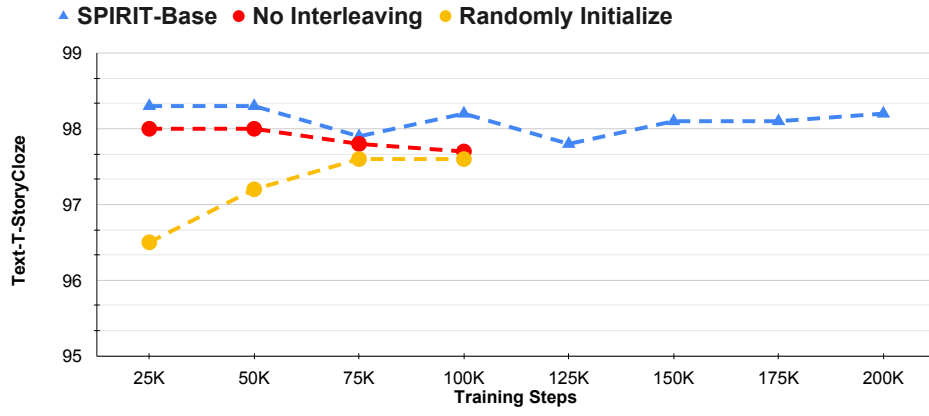
3. At the start of each run, we randomly select 3/6/9 samples from the above subset, ensuring a balanced distribution of samples for each sentiment. These samples are then combined to form the in-context prompt, which is reused for all subsequent iterations.

Model	Task	WUGGY $\uparrow$		BLIMP $\uparrow$		Topic-StoryCloze $\uparrow$				StoryCloze $\uparrow$			
		T	S	T	S	T	S	T $\rightarrow$ S	S $\rightarrow$ T	T	S	T $\rightarrow$ S	S $\rightarrow$ T
Previous Work													
GSLM (Lakhotia et al., 2021)		$\emptyset$	65.4/64.8	$\emptyset$	57.2/54.2	$\emptyset$	56.3/66.6	$\emptyset$	$\emptyset$	$\emptyset$	51.0/53.3	$\emptyset$	$\emptyset$
AudioLM (Borsos et al., 2023)		$\emptyset$	- / 71.5	$\emptyset$	- / <b>64.7</b>	-	-	$\emptyset$	$\emptyset$	$\emptyset$	-	$\emptyset$	$\emptyset$
VoxTlm (Maiti et al., 2023)		- / <b>80.3</b>	- / 66.1	- / <b>74.2</b>	- / 57.1	-	-	-	-	-	-	$\emptyset$	$\emptyset$
TWIST (Hassid et al., 2023)		$\emptyset$	- / <b>74.5</b>	$\emptyset$	- / 59.2	-	- / 76.4	$\emptyset$	$\emptyset$	$\emptyset$	- / 55.4	$\emptyset$	$\emptyset$
SPiRiT-LM variants													
SPiRiT-LM-BASE		<b>95.1/80.3</b>	71.4/69.0	75.7/73.3	<b>63.2/58.3</b>	94.5/98.0	<b>69.2/82.9</b>	<b>66.6/72.7</b>	<b>83.8/88.6</b>	<b>76.6/79.4</b>	<b>56.2/61.0</b>	56.2/59.5	64.3/64.6
+ASR+TTS		94.5/76.8	71.8/68.7	74.3/71.7	62.4/57.2	93.1/97.7	69.1/81.6	66.0/71.6	81.6/86.1	75.3/77.4	55.5/59.9	55.5/58.8	63.5/63.5
Rope $\theta$ default		95.2/78.2	71.7/69.5	75.8/73.3	62.9/57.7	<b>94.5/98.2</b>	69.5/82.0	66.1/72.0	83.5/88.3	76.6/78.9	56.3/60.9	<b>56.4/59.8</b>	<b>64.1/65.5</b>
SPiRiT-LM-EXPRESSIVE		95.2/75.8	66.2/65.0	76.6/73.6	58.7/54.2	94.3/97.9	58.2/75.4	57.7/61.6	81.3/73.2	75.7/78.9	51.8/56.9	52.5/54.6	61.4/58.8
Parallel Data Training													
Word-level transcription		94.7/74.7	71.2/67.1	75.9/72.6	62.8/57.2	94.3/98.0	68.1/80.3	53.9/57.5	67.0/71.9	75.8/78.2	55.0/60.1	51.0/54.2	55.1/56.4
ASR+TTS		94.0/76.5	<b>72.6/69.8</b>	75.7/73.3	62.2/57.6	92.7/97.3	62.7/74.9	56.9/63.5	67.8/71.8	73.6/76.3	50.7/54.6	49.9/53.9	53.5/54.0
Unimodal Ablations													
Speech Only		67.4/67.1	71.8/69.5	54.1/53.7	63.0/58.0	49.7/54.8	62.2/72.9	48.3/52.2	49.0/49.4	48.2/53.7	51.0/54.8	48.1/52.6	49.2/49.3
Text Only		94.5/72.6	53.1/46.8	<b>77.3/73.9</b>	54.6/52.6	<b>94.5/98.2</b>	48.0/51.7	47.3/47.5	51.5/51.7	76.1/79.0	47.0/50.2	47.1/47.3	50.3/52.1
Cascade Topline													
(WHISPER) + LLAMA 2		- / 84.1	- / 79.2	- / 72.8	- / 71.6	- / 98.5	- / 94.76	- / 94.76	- / 94.76	- / 81.9	- / 75.7	- / 75.7	- / 75.7

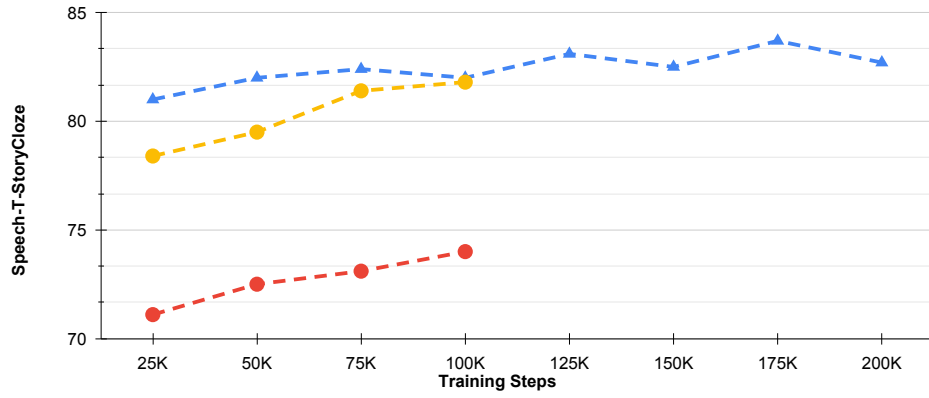
Table 10: Zero-shot Comprehension Evaluation in Speech (S) and Text (T). We report Accuracy / Accuracy-token for all the SPiRiT-LM models. Both metrics are based on selecting the hypothesis (among two choices) with the highest log-likelihood according to the model. The log-likelihood is based on the sum of each token likelihood in the sequence. The Accuracy is computed based on the prediction that maximizes the log-likelihood of the hypothesis. Accuracy-token adds a normalizing step of the log-likelihood by the number of tokens in the hypothesis. The related work performance (except GSLM) comes from the original published papers of each reported system. We recomputed the scores of GSLM on our metrics.

Model	Metrics	Bitrate	Content			Expressive Style			Pitch		
		BPS	Word Error Rate (WER)↓			Classification Accuracy↑			F0 Frame Error (FFE)↓		
			E. Read	LS	Fisher	E. Read	E. Imp.	EmoV	E. Read	E. Imp.	EmoV
Original Audio		-	14.76	3.55	30.26	92.47	75.69	27.46	-	-	-
Expresso models (Nguyen et al., 2023a)											
Hubert + HifiGAN		550	20.64	8.46	39.84	37.02	16.62	14.45	0.31	0.32	0.26
Hubert + HifiGAN cond. on GT Style		550	19.52	8.00	36.67	72.81	62.16	49.71	0.27	0.30	0.25
Codec (RVQ=1)		500	34.36	18.88	60.68	57.76	44.42	22.25	0.08	0.11	0.09
Codec (RVQ=8)		4000	16.85	4.62	35.64	78.65	64.53	26.88	0.04	0.05	0.04
SPiRiT-LM Tokenizers											
SPiRiT-LM-BASE		225	22.90	11.66	35.64	28.25	19.78	13.29	0.41	0.43	0.36
SPiRiT-LM-EXPRESSIVE		307	22.35	10.60	36.58	56.02	47.66	20.52	0.16	0.17	0.16

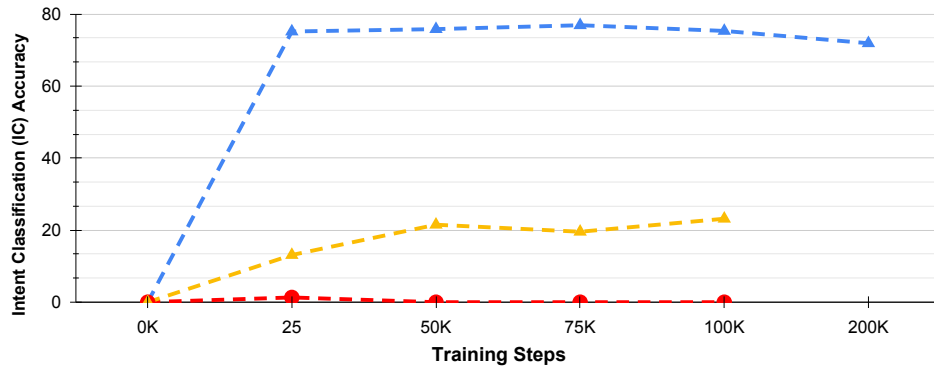
Table 11: **Expressive Speech Resynthesis Evaluation.** Performances of SPiRiT-LM Tokenizers on the Expresso Benchmark (Nguyen et al., 2023a) compared with their Hubert + HifiGAN (with and without conditioning on the Ground Truth Style) and Codec (with 1 and 8 codebooks) systems on various datasets: Expresso Read section (E. Read), Expresso Improvised section (E. Imp), LibriSpeech dev-other (LS, Panayotov et al., 2015), Fisher (Cieri et al., 2004), EmoV (Adigwe et al., 2018). The resynthesis is done with the same input speaker for Expresso subsets and with random Expresso speaker for other datasets. The bitrate is bit-per-second (BPS) computed as  $\log_2(\text{codebook size}) \times n \text{ tokens per second}$ .



(a) Accuracy on Text T-StoryCloze (0-Shot)



(b) Accuracy on Speech T-StoryCloze (0-Shot)



(c) Accuracy on Intent Classification (30-Shot)

Figure 6: Comparing SPIRIT-LM-BASE to a randomly initialized model trained in the same way and to a model trained with no Interleaving data. (i.e. the model is only trained on sequences of raw speech or raw text data without any interleaved aligned data.)