Scaling Properties of Speech Language Models

Santiago Cuervo and **Ricard Marxer**
Université de Toulon, Aix Marseille Université, CNRS, LIS. Toulon, France
{santiago.cuervo, ricard.marxer}@lis-lab.fr

Abstract

Speech Language Models (SLMs) aim to learn language from raw audio, without textual resources. Despite significant advances, our current models exhibit weak syntax and semantic abilities. However, if the scaling properties of neural language models hold for the speech modality, these abilities will improve as the amount of compute used for training increases. In this paper, we use models of this scaling behavior to estimate the scale at which our current methods will yield a SLM with the English proficiency of text-based Large Language Models (LLMs). We establish a strong correlation between pre-training loss and downstream syntactic and semantic performance in SLMs and LLMs, which results in predictable scaling of linguistic performance. We show that the linguistic performance of SLMs scales up to three orders of magnitude more slowly than that of text-based LLMs. Additionally, we study the benefits of synthetic data designed to boost semantic understanding and the effects of coarser speech tokenization.

1 Introduction

Inspired by the remarkable ability of preschool children to learn language from raw sensory inputs, Lakhotia et al. (2021) introduced in their seminal paper the *textless NLP* (Natural Language Processing) project. The project aimed to leverage advances in self-supervised speech representation learning for unsupervised unit discovery (Hsu et al., 2021; Chung et al., 2021) and generative neural language models (Brown et al., 2020) to jointly learn the acoustic and linguistic characteristics of a language from audio alone, without access to textual supervision (e.g. lexicon or transcriptions). They formalized this goal in the task of *Generative Spoken Language Modeling* (GSLM), in which a language model is trained on sequences of self-supervised learned speech units.

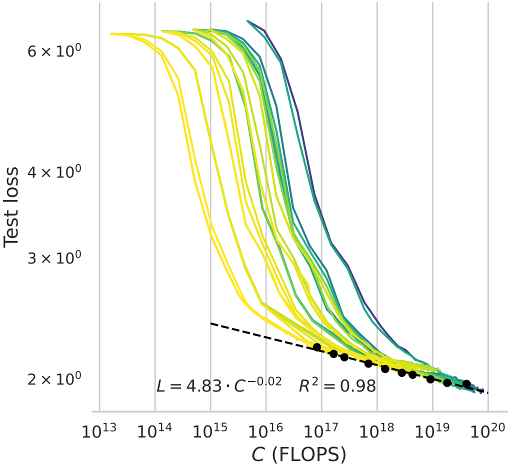

Figure 1: Speech Language Models test loss curves for all our single-epoch runs. Axes are in logarithmic scale. The envelope of minimal loss per FLOP (black dots) follows a power law (dashed line).

Beyond bridging the gap between human and machine language acquisition, the textless NLP project hoped to democratize access to NLP technologies by extending them to the millions of users of languages with little or no textual resources (e.g. due to a lack of standardized orthography). These languages are unlikely to be supported by current technologies, which are heavily dependent on massive volumes of text data. In today's landscape, where NLP-based AI systems are becoming increasingly relevant and pervasive, it is all the more pressing to expand their inclusivity by building speech-based systems that can match the capabilities of their text-based counterparts.

Despite a significant body of research on these Speech-based Language Models (SLMs) (Lakhotia et al., 2021; Kharitonov et al., 2022; Borsos et al., 2023; Hassid et al., 2023), they are still far from matching the syntactic and semantic abilities of text-based systems (Hassid et al., 2023). Therefore, the promise of textless NLP is yet to be realized. However, if the scaling behavior of text-based neu-

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 351–361
November 12-16, 2024 ©2024 Association for Computational Linguistics

ral language models (Brown et al., 2020; Kaplan et al., 2020) holds for the speech modality, we can reasonably expect those abilities to improve as the amount of compute used for training increases.

In this work, we apply recently proposed models of the scaling behavior of neural language models to SLMs, and use them to estimate the scale at which our current methods will match the linguistic performance of Large Language Models (LLMs), generative text-based systems that have achieved remarkably strong performance across a wide range of NLP applications (Brown et al., 2020). The main contributions of this work are:

- We trained over 50 SLMs with different number of parameters and data budgets. We show that the test loss of SLMs follows scaling power laws as those observed in text-based LLMs (Figure 1), and use the methods from Hoffmann et al. (2022) and Muennighoff et al. (2023) to model the scaling behavior of SLMs.

- We establish a strong correlation between the test loss of neural LMs and the downstream metrics commonly used to evaluate their syntactic and semantic abilities. Therefore, the linguistic performance of LMs follows similar scaling laws (Figure 2). We leverage this insight to determine the relative efficiency with scale of SLMs relative to LLMs.

- We speculate that SLMs require more context than fits in their context window to acquire from commonly used speech datasets the semantic understanding measured by our metrics. Accordingly, we propose a new speech dataset to boost semantic understanding in SLMs. Specifically, we synthesized a spoken version of the Tiny Stories dataset (Eldan and Li, 2023), and show that its use during pre-training improves downstream semantic performance.

- On the basis of our previous observation, we studied the use of unigram tokenization to shorten sequences and pack more information in the context window of SLMs. However, our results suggest that a coarser tokenization is detrimental to downstream performance.

The training source code, data, and models will be released at https://github.com/tiagoCuervo/slm_scaling.

2 Background

2.1 Generative spoken language modeling

We follow the GSLM framework from Lakhotia et al. (2021). The general GSLM pipeline is composed of three separately trained models: (i) a speech tokenizer, (ii) a language model, and (iii) a vocoder (token-to-waveform) module. In the following, we provide background for the speech tokenizer and LM, as these are the components we use in this work. For details about the vocoder please refer to Lakhotia et al. (2021).

Speech tokenizers transform raw speech waveforms into discrete representations. A speech encoder is used to extract continuous representations that are then transformed into discrete sequences through vector quantization. Formally, let $\mathcal{X} \in \mathbb{R}$ denote the domain of audio samples, a waveform is therefore a sequence of samples $x = (x_1, \dots, x_T)$, where $x_t \in \mathcal{X}$ for all $1 \leq t \leq T$. An encoder $F : \mathcal{X}^m \to \mathbb{R}^d$ transforms windows of samples of width m into d dimensional continuous frame representations. Applying F to x yields a sequence of frame representations $z = (z_1, \dots, z_{T'})$, where usually $T' < T$. Subsequently, a k-means algorithm is applied to the encoder output to generate a sequence of discrete speech tokens $u = (u_1, \dots, u_{T'})$, where $u_i \in \{1, \dots, K\}$ for $1 \leq i \leq T'$, and K is the size of the vocabulary.

Language models aim to learn the joint probability of token sequences $P(w_1, \dots, w_n)$. By the chain rule of probability, the probability of a sequence can be computed as a product of its conditional probabilities:

$$P(w_1, \dots, w_n) = \prod_{i=1}^{n} P(w_i | w_1, \dots, w_{i-1}) \quad (1)$$

Neural LMs, parameterized by θ, are neural networks that model the conditional probabilities $P_\theta(w_i | M(w_1, \dots, w_{i-1}))$, where M is a representation of the previous tokens. The network is optimized to minimize the negative log-likelihood of observed ground truth sequences:

$$L = -\sum_{i=1}^{n} P_\theta(w_i | M(w_1, \dots, w_{i-1})) \quad (2)$$

Nowadays, the network is typically a transformer (Vaswani et al., 2017). LLMs are large transformer

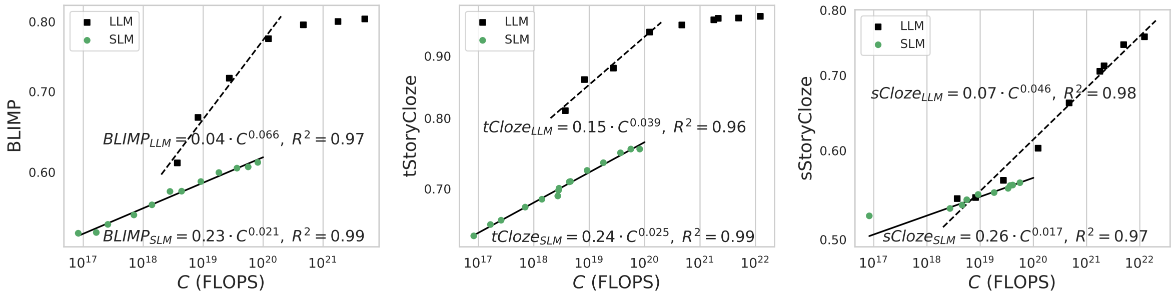

Figure 2: Downstream linguistic performance scaling with compute for LLMs and SLMs. Axes are in logarithmic scale. Syntactic (BLIMP) and semantic (tStoryCloze and sStoryCloze) metrics follow a power law before starting to saturate. Linguistic performance scales up to three orders of magnitude more slowly in SLMs relative to LLMs.

LMs trained on large text corpora (billions of parameters and tokens). SLMs are neural LMs applied to speech tokens u.

2.2 Scaling laws for neural language models

The performance of deep learning models often behaves predictably as a function of model size, dataset size, and compute (Hestness et al., 2017). Kaplan et al. (2020) showed that the loss L (Equation 2) of large neural LMs scales with a power law behavior as a function of these three scale factors:

$$L(C) \propto C^{\gamma}, \quad L(N) \propto N^{\alpha}, \quad L(D) \propto D^{\beta} \quad (3)$$

Where C is the amount of compute (in FLOPS), N is the number of parameters of the model, and D is the number of training tokens.

Building upon their work, Hoffmann et al. (2022) proposed a parametric function to model the final loss of neural LMs trained for a single epoch as a function of N and D:

$$\hat{L}(N, D) = E + \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} \quad (4)$$

Where the first term is the loss for an ideal LM, and should correspond to the entropy of the distribution of token sequences. The second term captures the approximation error that results from using a neural network with N parameters to approximate the ideal generative process. The final term reflects that the model is not trained to convergence, as a finite number of optimization steps are performed on a sample of size D from the real distribution.

Hoffmann et al. (2022) aimed to solve the problem of optimal allocation of resources given a fixed compute budget C_{avail}. They proposed to approximate the compute needed to train a transformer LM with N parameters on D tokens as $C \approx 6ND$. Then, the problem of optimal allocation of compute for model size and training data is:

$$\min_{N,D} \hat{L}(N, D), \quad \text{s.t.} \quad 6ND = C_{\text{avail}} \quad (5)$$

For which the solution is:

$$\begin{aligned} N_{\text{opt}}(C) &= G\left(\frac{C}{6}\right)^{a} \\ D_{\text{opt}}(C) &= \frac{1}{G}\left(\frac{C}{6}\right)^{b} \end{aligned} \quad (6)$$

With:

$$G = \left(\frac{\alpha A}{\beta B}\right)^{\frac{1}{\alpha+\beta}}, \ a = \frac{\beta}{\alpha+\beta}, \text{ and } b = \frac{\alpha}{\alpha+\beta}$$

Muennighoff et al. (2023) generalized Equation 4 to the case of multi-epoch training by replacing D and N with terms corresponding to the effective data D' and effective model parameters N':

$$\hat{L}(N', D') = E + \frac{A}{N'^{\alpha}} + \frac{B}{D'^{\beta}} \quad (7)$$

Where $D' \leq D$ is the number of effective training tokens, assuming that the value of repeated tokens decays exponentially. Similarly, they note that oversized models offer diminishing returns per parameter, as excess parameters learn the same features and do not add value (in the extreme). They propose an exponential decay model for them, yielding a number of effective parameters $N' \leq N$. They derived the expressions for D' and N' as:

$$\begin{aligned} D' &= U_D + U_D R_D^{*}(1 - e^{\frac{-R_D}{R_D^{*}}}) \\ N' &= U_N + U_N R_N^{*}(1 - e^{\frac{-R_N}{R_N^{*}}}) \end{aligned} \quad (8)$$

Size	Layers	Model dim.	Heads
20M	6	512	8
85M	12	768	12
155M	12	1024	16
309M	24	1024	16
823M	16	2048	32

Table 1: Models description.

Where U_D is the number of unique tokens used, $R_D = \frac{D}{U_D} - 1$ is the number of repetitions (0 for a single epoch), U_N is the number of parameters needed to optimally fit U_D according to Equation 6, $R_N = \frac{N}{U_N} - 1$ is the number of excess parameters, and R_D^* and R_N^* are constants.

The constants E, A, B, α, β, R_D^* and R_N^* can be estimated empirically by fitting Equation 4 or 7 to a set of tuples (N, D, R_N, R_D, L) obtained from training experiments with different budgets.

3 Experimental setup

3.1 Models and training

We adhere to the framework described in Section 2.1. For the speech tokenizer, we use a pre-trained HuBERT model (Hsu et al., 2021) with frame-rate of 25 Hz as the speech encoder F, and a vocabulary size of $K = 500$. This setup reports the best performance among publicly available models (Hassid et al., 2023). For the SLMs we use the Llama architecture (Touvron et al., 2023) with context window of 2050 tokens. Table 1 describes the model sizes used in our experiments. For the LLMs, we use the Pythia suite of pre-trained LLMs (Biderman et al., 2023), ranging in size from 14M to 6.9B parameters (we do not use the largest 12B model), and trained with ∼300B tokens.

All SLMs are optimized using AdamW (Loshchilov and Hutter, 2019) with weight decay of 0.1, maximum learning rate of 5e-4, half-cycle cosine decay learning rate schedule to 5e-5, and a warm-up initial stage of $\max(100, 0.01\, n_{iters})$ steps, where n_{iters} is the number of training steps, which varies for each experiment according to the data budget. We use batch sizes of 64, 128, 256 and 512 for the models with 20M, 85M, 155M and 309M, and 828M parameters, respectively.

To fit the constants in Equations 4 and 7, we adopt the approaches of Hoffmann et al. (2022) and Muennighoff et al. (2023), utilizing the Huber loss with $\delta = 0.03$ as the error function and L-BFGS as optimizer. Following Muennighoff et al. (2023), we first fit the parameters E, A, B, α, and β using the single-epoch runs, and afterwards fit R_D^* and R_N^* using the multi-epoch runs.

3.2 Evaluation

For upstream performance, we report and use the average loss (Equation 2) on the test set in all cases including the parametric fits. For downstream evaluation we rely on the zero-shot metrics used in the textless NLP literature, which evaluate LMs' linguistic knowledge by comparing likelihoods of positive and negative speech samples. We focus on metrics evaluating syntax and semantic knowledge. In all cases, performance is measured as the binary accuracy with which the model assigns higher likelihood to the positive samples.

Syntax: We use the sBLIMP task from the Zero Resource Speech Challenge (Nguyen et al., 2020). In sBLIMP, the model is presented with minimal pairs of sentences, where one is grammatically correct (positive) and the other is not (negative), targeting specific syntactic contrasts.

Semantics: To evaluate semantic understanding we use the spoken Story Cloze benchmark from Hassid et al. (2023), a spoken version of the StoryCloze textual benchmark (Mostafazadeh et al., 2016), which consists of 4k five-sentence commonsense stories. In StoryCloze, the model receives as input the first four sentences of a story, and has to assign higher probability to the correct final sentence than to an adversarial negative sample.

The spoken Story Cloze benchmark comes in two versions: *sStoryCloze* and *tStoryCloze*. The difference between them lies in how the negative sample is generated. *sStoryCloze* uses the same negative samples as the textual benchmark, which are carefully constructed to evaluate models' ability to grasp causal and temporal commonsense relations. In *tStoryCloze*, the negatives are randomly sampled from the whole dataset, and therefore measures the ability of the model to stay on topic. Since in tStoryCloze the negatives are randomly sampled, they are not specifically designed to violate causal or temporal logic. Instead, they are more likely to be incoherent or irrelevant in a more obvious way, making it an easier task than sStoryCloze.

3.3 Data

3.3.1 Datasets

We use a collection of publicly available English speech datasets for training: LibriSpeech (Panayotov et al., 2015), LibriLight (Kahn et al., 2020),

Dataset	Hours	HuBERT Tokens	Unigram
LibriSpeech	960	67M	38M
LibriLight	53k	3.74B	2.11B
SWC	1k	32M	19M
Tedlium	1.6k	0.11B	67M
People	7k	0.48B	0.29B
Vox Populi	24k	1.64B	1.08B
sTinyStories	72k	4.82B	2.71B
Total	160k	10.89B	6.31B

Table 2: Datasets statistics. The Unigram column corresponds to the dataset of HuBERT tokens compressed through unigram tokenization.

SWC (Baumann et al., 2019), Tedlium (Hernandez et al., 2018), People's Speech (Galvez et al., 2021), and Vox Populi (Wang et al., 2021b); and a novel dataset: sTinyStories, a spoken version of the Tiny Stories dataset (Eldan and Li, 2023) that we synthesized using the single-speaker TTS system provided by Wang et al. (2021a). Tiny Stories is a synthetic text corpus of short stories designed to boost commonsense reasoning in neural LMs. We propose sTinyStories because we hypothesize that the semantic understanding that tasks such as sStoryCloze measure is hard to acquire from commonly used speech datasets. Consider for instance the audiobooks in LibriLight. The data has long-range dependencies spanning multiple pages, whereas our SLMs can ingest roughly a dozen sentences of spoken text in their context window. Other datasets, which were mainly designed to serve as training data for automatic speech recognition systems, consist of too small fragments of audio that lack meaningful causal structure. sTinyStories consists of full stories with causal structure that fit within the context window of our SLMs.

We do not include samples from sTinyStories in our test set, as we intend to use our test loss as measure of the quality with which SLMs model natural language, not synthetic one. For other datasets we use the defined held-out sets for testing. In cases where a held-out set is not defined, we randomly sampled 1% of the data to serve as test set. See Table 2 for dataset sizes.

3.3.2 Data budgets

In order to have a representative set of samples to fit Equations 4 and 7, for each model size, we performed training runs with a ratio of training tokens D to parameters N: $D/N \in \{2, 4, 8, 10, 20, 32, 64, 100\}$. This setup yields single-epoch and multi-epoch runs for the larger models but not for the smaller models (e.g. for the model with 85M parameters the maximum number of training tokens corresponds to 0.99 epochs). To better fit Equation 7, we performed additional experiments so that for each model size there were runs with training epochs in $\{2, 4, 8, 10\}$, with the exception of the 828M parameter model, for which the maximum was 8 epochs.

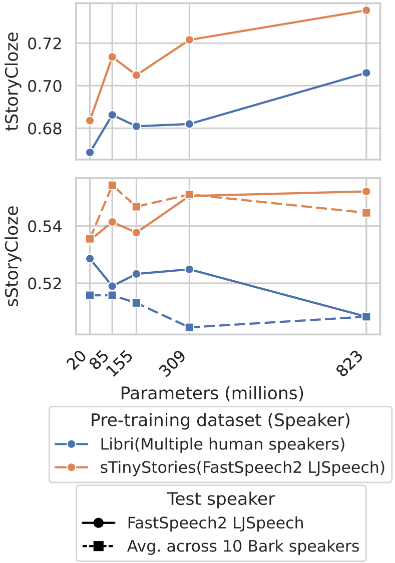

Figure 3: Gains from synthetic data on downstream semantic performance of SLMs. Pre-training on sTinyStories yields consistent improvements on semantic understanding relative to pre-training on audiobooks (LibriSpeech plus LibriLight). Performance gains hold for mismatched train and test speakers.

4 Results

4.1 Gains from sTinyStories

In order to determine if sTinyStories meaningfully contributes to the semantic understanding of SLMs, we compare the performance on tStoryCloze and sStoryCloze of models trained on one epoch of the union of LibriSpeech and LibriLight, against models trained on an equivalent amount of sTinyStories tokens. Figure 3 shows the obtained results. Models trained on sTinyStories consistently outperform those trained on audiobooks across all model scales. A factor that could contribute to the observed performance gain is the match between training and evaluation speakers, as both sTinyStories and Story Cloze were synthesized using the single-sepaker TTS from Wang et al. (2021a). However, we believe this to be unlikely as the speech tokenizer we use likely captures little speaker-specific information (Nguyen et al., 2023). To isolate the potential impact of speaker mismatch between training and evaluation data, we created a multi-speaker version of the sStoryCloze benchmark using Bark TTS [1], and repeat the evaluations. The results, also shown in Figure 3, indicate that even with mismatched train and test speakers training on sTinyStories yields performance gains.

4.2 Benchmarking our setup

To validate our setup, we compared our best performing model with other models in the SLM literature in Table 3. Our model outperformed all other speech-only LMs on the semantic tasks, and performed second best in general, even relative to hybrid speech-text LMs. Notably, our model outperformed models with a larger compute budget. Considering that the models from Hassid et al. (2023) and Nguyen et al. (2024) use similar hyperparameters (same speech tokenizer and the Llama architecture for LMs); the most likely factor to explain the performance difference is the data used. We believe these results further illustrate the benefits from using sTinyStories.

4.3 Scaling laws

We trained multiple SLMs for each model size with different data budgets as described in Section 3.3.2. The resulting learning curves for single-epoch runs are presented in Figure 1 as a function of compute, and show that the envelope of minimal loss per FLOP follows a power law.

4.3.1 Downstream scaling with compute

We analyzed the relationship between the upstream and linguistic downstream performance in SLMs and LLMs. Figure 4 shows the obtained results. Downstream linguistic metrics before saturation are strongly correlated with the upstream test loss in both LLMs and SLMs. Therefore, the envelope of maximum downstream performance per FLOP also follows a power law, i.e. for a downstream performance function Q, $Q \propto C^{\gamma_q}$. The power laws for the different performance metrics are presented in Figure 2 and the exponents in Table 4.

These results allow us to compare the efficiency with scale of LLMs and SLMs. For each metric, we can interpret the ratio between the γ_q exponents of the power laws of LLMs and SLMs as the relative efficiency with scale. For BLIMP, the ratio is $\frac{0.066}{0.021} = 3.14$, indicating that for an increase in compute ΔC yielding a ΔQ in LLM's syntactic performance, SLMs require $10^{3.14}\Delta C$ to get the same ΔQ. Similarly, for tStoryCloze and sStoryCloze the ratios are 1.56 and 2.7, respectively.

4.3.2 Scaling with parameters and tokens

We fitted the functions from Equations 4 and 7 to our data using the procedure described in Section 3.1. We present the empirically fitted scaling law parameters and compare them to the ones obtained for text by Muennighoff et al. (2023) in Table 5. From Equation 6, $N_{opt} \propto C^a$ and $D_{opt} \propto C^b$. For both modalities $a \approx b \approx 0.5$, suggesting that as compute increases, model size and data should be scaled equally for optimal performance. Contrary to text, $R_N^* > R_D^*$, indicating that repeated tokens decay faster than excess parameters (albeit both slower than in text). Therefore, in SLMs, compute allocated to parameters should scale faster than compute allocated for epochs.

4.4 Unigram tokenization

As mentioned in Section 3.3, we believe that the limited context window of SLMs could cripple their ability to model the long-range dependencies in language required for causal reasoning. Seeking to mitigate this limitation, we apply unigram tokenization to shorten the length of speech token sequences. We use the SentencePiece tokenizer (Kudo and Richardson, 2018) with a vocabulary size of 5000. We choose the vocabulary size on the scale of previous works that have used similar tokenization strategies for speech applications (Chang et al., 2023). The resulting dataset sizes after compression are presented in Table 2.

We train a set of Speech LMs on the compressed datasets, with model sizes up to 309M parameters and data budgets ranging from 740M to 6.31B tokens. We analyze the scaling behavior of the upstream and downstream metrics and compare it with SLMs trained on raw HuBERT speech tokens in Figure 5. SLMs trained on unigram compressed speech tokens show similar upstream scaling with compute, but worse downstream scaling. Notably, the performance on the StoryCloze benchmark does not seem to scale with compute.

We fitted the function from Equation 4 to the results obtained on the compressed dataset. Table 5 presents the resulting scaling law parameters. Similar to the previous findings, for a given compute budget, scaling model size and training data equally

[1] https://github.com/suno-ai/bark

	Parameters	Tokens	BLIMP	tStoryCloze	sStoryCloze
Speech-only language models					
GSLM (Lakhotia et al., 2021)	100M	-	54.2	66.6	53.3
AudioLM (Borsos et al., 2023)	150M	-	<u>**64.7**</u>	-	-
Hassid et al. (2023), Cold-init 1.3B	1.3B	10.8B	56.5	-	-
Nguyen et al. (2024)	7B	100B	58.0	72.9	54.8
Ours (best model)	823M	82B	**61.3**	**78.0**	**56.7**
Speech language models initialized from text language models					
TWIST (Hassid et al., 2023)					
- Warm-init 1.3B	1.3B	10.8B	57.1	70.6	52.4
- Warm-init 7B	7B	36B	59.0	74.1	55.1
- Warm-init 13B	13B	36B	59.2	76.4	55.4
Mutltimodal speech-text language models initialized from text language models					
SpiRit-LM (Nguyen et al., 2024)	7B	100B	58.3	<u>**82.9**</u>	<u>**61.0**</u>
Toplines					
Pythia (Biderman et al., 2023) 6.9B	6.9B	300B	80.0	97.5	76.21
Human (Hassid et al., 2023)	-	-	-	90.2	79.9

Table 3: Models benchmarking. The best model resulting from our experiments obtains the best semantic performance across speech-only models, and the second best overall in all tasks.

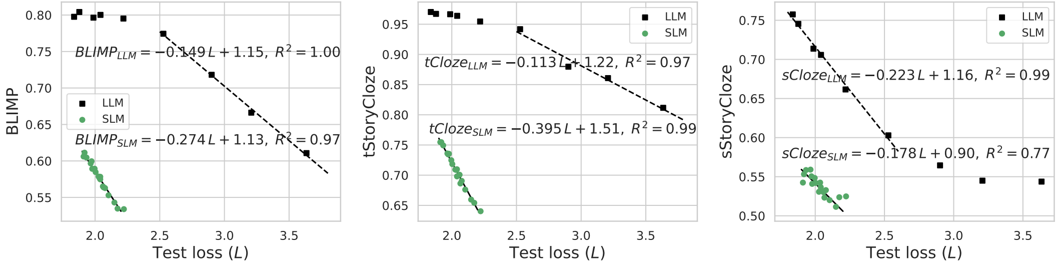

Figure 4: Correlation between downstream linguistic performance and test loss for LLMs and SLMs. Syntactic (BLIMP) and semantic (tStoryCloze and sStoryCloze) metrics are strongly linearly correlated with the upstream test loss before saturation.

Modality	γ_q		
	BLIMP	tCloze	sCloze
Text	0.066	0.039	0.046
Speech	0.021	0.025	0.017

Table 4: γ_q power law coefficients of downstream performance with compute as depicted in Figure 2.

	E	A	B	α	β	R_N^*	R_D^*
Text Muennighoff et al.	1.87	521	1488	0.35	0.35	5.31	15.4
Speech	1.73	13.9	39.8	0.25	0.24	31.0	25.0
Speech (Unigram)	1.42	3.85	8.90	0.15	0.16	-	-

Table 5: Scaling law parameters fit to Equations 4 and 7 for different language tokenizations.

is optimal for performance. Due to the poor downstream results obtained with unigram tokenization and the lack of sufficient compute resources, we did not perform multi-epoch training experiments.

5 Related work

Previous works have studied the scaling behavior of neural networks on speech applications. Droppo and Elibol (2021) showed that acoustic models trained with an auto-predictive coding loss follow similar power laws to those observed in neural LMs. Aghajanyan et al. (2023) used the scaling laws from Hoffmann et al. (2022) to model the scaling behavior of the upstream loss of neural LMs on multiple modalities, including speech. They used a speech tokenizer with higher framerate (50 Hz) and vocabulary size ($K = 2000$) than the one we used (Section 3.1). Such fine-grained tokenizers capture a lot of the paralinguistic information in speech (Nguyen et al., 2023). Therefore, their speech tokens can be considered almost a different modality due to the acoustic variance. Furthermore, they do not study the behavior with scale of downstream performance. In this work, we focus on the linguistic content of the signal. As reported by Hassid

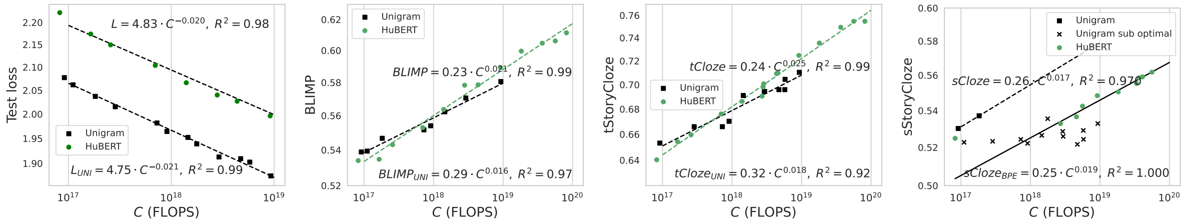

Figure 5: Comparison of the scaling behavior of SLMs trained on raw speech tokens and unigram compressed tokens. Axes are in logarithmic scale. The upstream loss of SLMs trained on unigram tokens scales better with compute, but downstream performance scales worse. Notably, the sStoryCloze metric for SLMs trained on unigram tokens does not seem to improve with increased compute.

et al. (2023), our speech tokenizer performs best on downstream linguistic applications, and is therefore a more suitable choice to study the scaling behavior of the linguistic performance of SLMs.

This paper is most closely related to the work of Hassid et al. (2023). We largely follow their setup in terms of hyperparameters and evaluation metrics. They reported improved linguistic downstream performance with scale in SLMs, but did not characterize their scaling behavior. Our scaling laws allow practitioners to determine the compute needed to attain a specific loss, syntactic and/or semantic performance; and its optimal allocation with respect to parameters and tokens. To the best of our knowledge, we are the first to model the scaling properties of downstream linguistic performance in SLMs, and to study the scaling of the considered downstream metrics on text-based LLMs. This enables a comparison between the two modalities in terms of scaling efficiency.

6 Discussion

Our work showed that the upstream and downstream linguistic performance of our current methods for GSLM scales predictably with compute. This suggests that, with sufficient computational resources, the goal of the textless NLP project—achieving neural LMs trained exclusively on speech, and matching the linguistic proficiency of their text-based counterparts—is achievable. However, the cost of such models could be prohibitive, as we estimate that they will require up to three orders of magnitude more compute than a text-based LLM to achieve equivalent performance. We believe this points to the need for leveraging the rich language representations already learned by text LLMs. This seems to be the current trend in the community, as several recent works have sought to improve SLMs through transfer learning from text-based models (Hassid et al., 2023; Zhang et al., 2023; Nguyen et al., 2024). However, considering one of the grand goals of the textless NLP project—extending the benefits of large-scale language modeling to low-resource or non-written languages—we will have to address the question of how knowledge transfer from text LLMs performs when the speech data is in a different language than the one the text LLM was trained on. If cross-lingual knowledge transfer between text and speech modalities proves to be unfeasible, then purely speech-based SLMs, such as the ones studied here, could still offer a compelling solution for low-resource languages.

We explored the use of synthetic data and coarser tokenization to increase the semantic abilities of SLMs. Our synthetic dataset improved semantic performance, but using a coarser tokenization led to overall degradation of downstream performance. We do not have yet an hypothesis for why coarser tokens degrade performance, as this seems counter-intuitive, and contradicts the findings on other speech applications (Chang et al., 2023). We leave this as an interesting issue to address in future work. Moreover, we believe that working on methods that allow to increase the information density per context-window of SLMs holds promise to improve their scaling behavior.

7 Limitations

Any extrapolation from our models of the scaling behavior of SLMs should be considered optimistic for the following reasons: **1)** Our models for downstream performance ignore the fact that the metrics saturate. As observed in text LLMs, the improvements with scale slow down as performance approaches the saturation value. It is likely that, due to saturation, the compute required to yield a particular performance will be larger than predicted. Moreover, due to the lower density of linguistic information per context window in SLMs relative to LLMs, the saturation values of the metrics may be lower for SLMs. **2)** The LLMs from the Pythia suite that we used in this study are likely overtrained (all models were trained with ∼300B tokens). Optimally trained LLMs (according to Equation 6) should show better performance with scale, and therefore widen the gap with the scaling efficiency of SLMs. **3)** The envelope of minimal loss per FLOP (Figure 1) might show a slight negative curvature at larger scale (Hoffmann et al., 2022), reducing the scaling efficiency.

Muennighoff et al. (2023) note that the scaling law coefficients for text LLMs, and consequently the optimal compute allocation, can vary depending on the training datasets used in the scaling study. Commonly used text datasets are significantly larger and more diverse than the academic speech datasets typically used for GSLM, such as those in this study. As a result, these speech datasets represent a more biased sample of the overall distribution of speech data, making scaling laws derived from them less likely to generalize. Therefore, we cannot guarantee that the scaling laws we have developed will be universally applicable to other datasets. However, we do not expect significant deviations that affect the conclusions here presented. Future research could explore validating the predictions from this study on larger and more diverse datasets, such as the recently released Yodas (Li et al., 2023).

8 Conclusions

We have trained a large set of SLMs with different compute budgets and studied the scaling properties of their upstream and downstream performance using recently proposed models of scaling laws for neural LMs. The obtained models allow practitioners to optimally allocate compute to attain a specific loss, syntactic, and/or semantic performance. We showed that the pre-training loss and downstream linguistic performance of SLMs and LLMs is highly correlated, and both scale predictably according to power laws. This allowed us to compare the scaling properties of SLMs and LLMs, from which we established that the linguistic abilities of SLMs scale up to three orders of magnitude more slowly. Additionally, we proposed a new speech dataset, STinyStories, and showed that its use during pre-training improves downstream semantic performance. Finally, we explored the use of coarser speech tokenization as a method to increase the amount of tokens per context window in SLMs, but obtained worse downstream performance.

Acknowledgements

We are grateful to the French National Research Agency for their support through the ANR-20-CE23-0012-01 (MIM) grant, and the Institute of Convergence ILCB, supported by grants from France 2030 (ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX). This work was granted access to the HPC resources of GENCI-IDRIS under the allocation AD011014044.

References

Armen Aghajanyan, Lili Yu, Alexis Conneau, Wei-Ning Hsu, Karen Hambardzumyan, Susan Zhang, Stephen Roller, Naman Goyal, Omer Levy, and Luke Zettlemoyer. 2023. Scaling laws for generative mixed-modal language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Timo Baumann, Arne Köhn, and Felix Hennig. 2019. The spoken wikipedia corpus collection: Harvesting, alignment and an application to hyperlistening. *Lang. Resour. Eval.*, 53(2):303–329.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. 2023. Audiolm: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533.

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.

Xuankai Chang, Brian Yan, Kwanghee Choi, Jeeweon Jung, Yichen Lu, Soumi Maiti, Roshan Sharma, Jiatong Shi, Jinchuan Tian, Shinji Watanabe, Yuya Fujita, Takashi Maekaku, Pengcheng Guo, Yao-Fei Cheng, Pavel Denisov, Kohei Saijo, and Hsiu-Hsuan Wang. 2023. Exploring speech recognition, translation, and understanding with discrete speech units: A comparative study.

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.

J. Droppo and O. Elibol. 2021. Scaling laws for acoustic models. In *Interspeech 2021*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english?

Daniel Galvez, Greg Diamos, Juan Manuel Ciro Torres, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

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. 2023. Textually pretrained speech language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. Tedlium 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, pages 198–208, Cham. Springer International Publishing.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory F. Diamos, Heewoo Jun, Hassan Kianinejad, Md. Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. 2017. Deep learning scaling is predictable, empirically. *CoRR*, abs/1712.00409.

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.

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 Lang.*, 29:3451–3460.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P.E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673.

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. *CoRR*, abs/2001.08361.

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.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. 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.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe 0001. 2023. Yodas: Youtube-oriented dataset for audio and speech. In *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2023, Taipei, Taiwan, December 16-20, 2023*, pages 1–8. IEEE.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and cloze evaluation for deeper understanding of commonsense stories. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 839–849, San Diego, California. Association for Computational Linguistics.

Niklas Muennighoff, Alexander M Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling data-constrained language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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. *CoRR*, abs/2011.11588.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. 2023. Expresso: A Benchmark and Analysis of Discrete Expressive Speech Resynthesis. In *Proc. INTERSPEECH 2023*, pages 4823–4827.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. SpiRit-LM: Interleaved Spoken and Written Language Model.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

Changhan Wang, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Ann Lee, Peng-Jen Chen, Jiatao Gu, and Juan Pino. 2021a. fairseq sˆ2: A scalable and integrable speech synthesis toolkit. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 143–152, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021b. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

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