Scaling Properties of Speech Language Models

Anonymous ACL submission

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 the performance of text-based LLMs. Additionally, we study the effects of coarser speech tokenization, and the benefits of synthetic data designed to boost semantic understanding.

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; Devlin et al., 2019) 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.

Despite a significant body of research on these

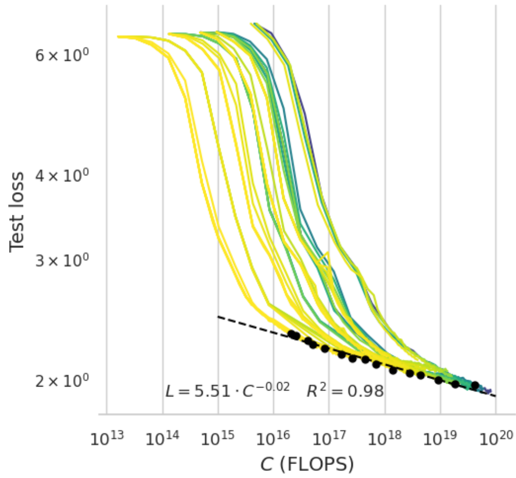

Figure 1: Speech Language Models test loss curves for all our different runs. Axes are in logarithmic scale. The envelope of minimal loss per FLOP (black dots) follows a power law (dashed line).

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 laws of text-based neural language models (Kaplan et al., 2020; Hoffmann et al., 2022) hold for the speech modality, we can 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 scale to 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 pa-

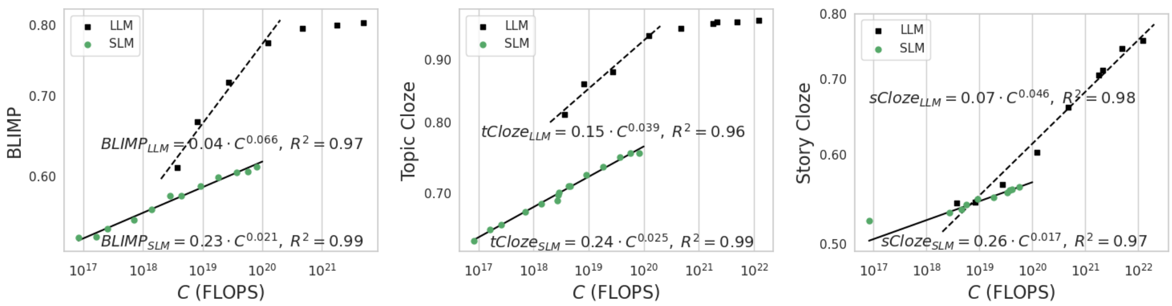

Figure 2: Downstream linguistic performance scaling with compute for LLMs and SLMs. Axes are in logarithmic scale. Syntactic (BLIMP) and semantic (Topic Cloze and Story Cloze) metrics follow a power law before starting to saturate. Linguistic performance scales up to three orders of magnitude more slowly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

rameters and data budgets. We show that the test loss of SLMs follows scaling power laws as those observed in text-based LLMs (Figure 1). We use the method from Hoffmann et al. (2022) 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 estimate the scale at which SLMs will match the linguistic proficiency of LLMs.

- We note that SLMs likely require more context than fits in our models to acquire the semantic understanding measured by our metrics from commonly used speech datasets. 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 semantic downstream performance.

- Based on our previous observation, we studied the use of unigram tokenization to shorten sequences and pack more information in the context window of our models. However, our results suggest that a coarser tokenization is detrimental to SLM performance 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 which 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 $x = (x_1, \ldots, x_T)$, where $x_t \in \mathcal{X}$ for all $1 \leq t \leq T$. An encoder $F : \mathcal{X}^m \rightarrow \mathbb{R}^d$ transforms windows of samples of width m into d dimensional continuous frame representations. Applying F to x yields a sequence of frame representations $z = (z_1, \ldots, z_{T'})$, where usually $T' < T$. Afterwards, a k-means algorithm (MacQueen, 1967) is applied over the encoder outputs to generate a sequence of discrete speech tokens $u = (u_1, \ldots, u_{T'})$, where $u_i \in \{1, \ldots, K\}$ for $1 \leq i \leq T'$, and K is the vocabulary size.

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

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

Neural LMs, parameterized by θ, are neural networks that model the conditional probabilities $P_\theta(w_i | M(w_1, \ldots, w_{i-1}))$, 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 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 of large neural LMs scales with a power-law behavior. Building upon their work, Hoffmann et al. (2022) proposed a parametric function to model the loss of neural LMs (Equation 2) trained for a single epoch:

$$\hat{L}(N, D) = E + \frac{A}{N^\alpha} + \frac{B}{D^\beta} \quad (3)$$

, where N is the number of parameters of the model and D is the number of training tokens. 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 terms captures the fact 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.

Given a set of neural LM training runs yielding a set of (L, N, D) tuples, we can empirically estimate the constants E, A, B, α and β by minimizing the error between the predicted loss and observed loss:

$$\min_{E,A,B,\alpha,\beta} \sum_{\text{Runs } i} G(\hat{L}(N_i, D_i) - L_i) \quad (4)$$

, where G is some error function.

3 Experiments

3.1 Setup

3.1.1 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).

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.

All SLMs are optimized using AdamW (Loshchilov and Hutter, 2019) with weight decay of 0.1, maximum learning rate of 5e-4, cosine learning rate schedule, and a warm-up initial stage of $\max(100, 0.01\, n_{iters})$ steps, where n_{iters} is the number of training steps, and varies for each experiment according to the desired 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 scaling law from Equation 3 we follow Hoffmann et al. (2022) and use the Huber loss (Huber, 1964) with $\delta = 0.03$ as error function.

3.1.2 Evaluation

We use the sBLIMP task (Nguyen et al., 2020) to measure syntactic performance. In sBLIMP, the network is presented with a matched pair of speech segments, grammatical and ungrammatical sentences. The objective is to assign higher probability to the grammatical sentence.

To evaluate semantic understanding we use the spoken StoryCloze 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 benchmark comes in two versions: Story Cloze and Topic Cloze. The difference between them lies in how the negative sample is generated. Spoken Story Cloze uses the same samples as the textual benchmark, which require commonsense reasoning to distin-

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.

guish from the real ending. Topic Cloze measures the ability of the model to stay on topic. In this setup, the negatives are randomly sampled from the whole dataset.

3.1.3 Training data

We use a collection of publicly available speech datasets for training: LibriSpeech (Panayotov et al., 2015), LibriLight (Kahn et al., 2020), SWC (Baumann et al., 2019), Tedlium (Hernandez et al., 2018), People (Galvez et al., 2021), and Vox Populi (Wang et al., 2021b). We hypothesize that the semantic understanding that tasks such as Story Cloze measure is hard to acquire from these 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 consist of too small fragments of audio that lack meaningful causal structure. This led us to propose a new speech dataset: STINYSTORIES, a spoken version of the Tiny Stories dataset (Eldan and Li, 2023), a synthetic text corpus of short stories designed to boost commonsense reasoning in neural LMs. We synthesized STINYSTORIES using the single-speaker TTS system provided by Wang et al. (2021a). 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.

3.2 Results

3.2.1 Gains from sTinyStories

In order to determine if STINYSTORIES meaningfully contributes to the semantic understanding of SLMs, we compare the performance on Topic Cloze and Story Cloze 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. However, the performance gain could be explained by the match between the speakers used to synthesize both STINYSTORIES and Story Cloze, as they were both synthesized using the same single-sepaker TTS system. In order to discard this factor, we synthesize a multi-speaker version of the Story Cloze benchmark using the Bark TTS [1] and repeat the evaluations. The results depicted in Figure 3 show that even with mismatched train and test speakers using STINYSTORIES yields performance gains.

3.2.2 Scaling laws

For each model size, we train multiple SLMs with different data budgets, ranging from 600M to 10B tokens. The resulting learning curves are presented in Figure 1 as a function of compute, and show that the envelope of minimal loss per FLOP follows a power-law.

We analyze the relationship between the upstream test loss and downstream performance metrics for our trained SLMs and the LLMs in the Pythia suite. Figure 4 illustrates the obtained results. Syntactic and semantic downstream 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, as depicted in Figure 2.

We fit the function from Equation 3 to our data using the procedure described in section 2.2. We present the empirically fit scaling law parameters and compare them to the ones obtained for text by Hoffmann et al. (2022) in Table 3.

Equation 3 can be used to determine the optimal N and D to minimize L for a given compute budget C. Hoffmann et al. (2022) obtain $N_{opt} \propto C^a$ and $D_{opt} \propto C^b$, where $a = \frac{\alpha}{\alpha+\beta}$ and $b = \frac{\beta}{\alpha+\beta}$. For both text and speech $a \approx b \approx 0.5$, indicating that

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

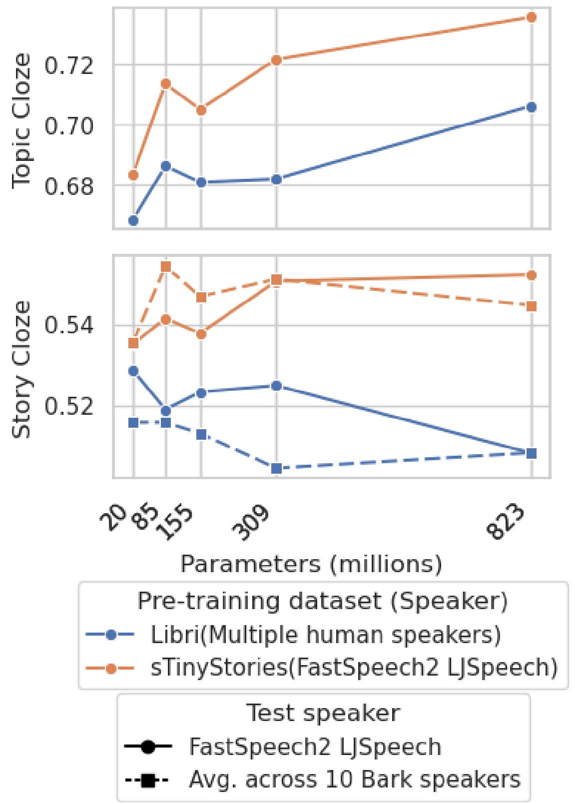

Figure 3: Gains from synthetic data on downstream semantic performance of SLMs. Pre-training on sTinyStories yields consistent improvements on semantic understanding relative to pre-training on audiobooks (LibriSpeech plus LibriLight). Performance gains hold for mismatched train and test speakers.

as compute increases, model size and data should be increased in equal proportions for optimal performance.

3.2.3 Unigram tokenization

As mentioned in section 3.1.3, we believe that the limited context window of SLMs hinders their ability to model the long-range dependencies in language required for causal reasoning. Motivated by this belief, 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 (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 74M to 6.31B tokens. We analyze the scaling behavior of the

	E	A	B	α	β
TEXT	1.69	406.4	410.7	0.34	0.28
SPEECH	1.73	13.92	39.80	0.25	0.24
SPEECH (UNIGRAM)	1.42	3.85	8.90	0.15	0.16

Table 3: Scaling law parameters fit to Equation 3 for different language tokenizations.

upstream and downstream metrics and compare it with SLMs trained on raw HuBERT speech tokens in Figure 5. SpeechLMs trained on unigram compressed speech tokens show better upstream scaling with compute, but worse downstream scaling. Notably, the performance on the StoryCloze benchmark does not seem to scale with compute.

We also fit the function from Equation 3 to our obtained results. The obtained scaling law parameters are presented in Table 3. As before, for a given compute budget, model size and amount of data should scale equally for optimal performance.

4 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 3.1.1). 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. In this work, we focus on the linguistic content of the signal. As reported by (Hassid 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 perhaps most closely related to the work of Hassid et al. (2023). We largely follow their setup in terms of model architecture and evaluation metrics. They showed that linguistic downstream performance of SLMs improves with scale, but did not characterize their scaling behavior. To

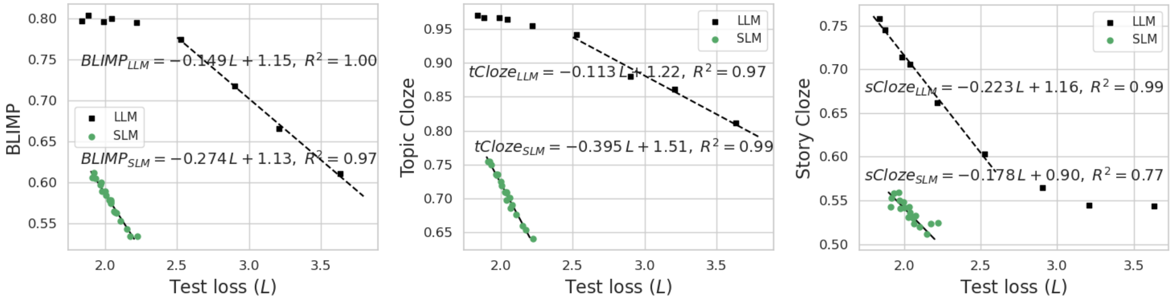

Figure 4: Correlation between downstream linguistic performance and test loss for LLMs and SLMs. Syntactic (BLIMP) and semantic (Topic Cloze and Story Cloze) metrics are strongly linearly correlated with the upstream test loss before saturation.

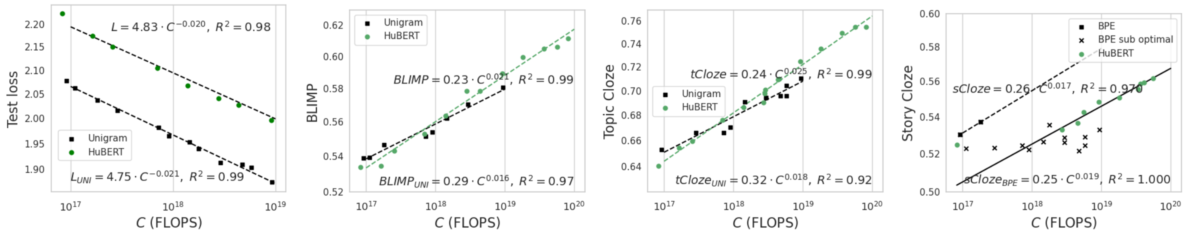

Figure 5: Comparison of the scaling behavior of SLMs trained on raw speech tokens and unigram compressed tokens. Axes are in logarithmic scale. The upstream loss of SLMs trained on unigram tokens scales better with compute, but downstream performance scales worse. Notably, the Story Cloze metric for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

the best of our knowledge, we are the first to characterize the upstream and downstream linguistic performance of SLMs. Furthermore, we compare their scaling behavior with the one of text-based LLMs.

5 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 of achieving neural LMs trained exclusively on speech that match 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. In this regard, recent methods that leverage transfer learning from text-based LLMs (Hassid et al., 2023; Zhang et al., 2023; Nguyen et al., 2024) are likely to be a better choice to achieve highly performant generative speech models. It remains to be seen how knowledge transfer from LLMs performs when the speech data is in a different language than the one the LLM was trained on. If there is no significant cross-lingual knowledge transfer between text and speech modalities, SLMs could still be an attractive choice 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 is a promising research area that could improve their ability to model long range dependencies, and likely their scaling behavior.

6 Conclusions

We have trained a large set of SLMs of different sizes and on different data budgets. Using the collected data from those experiments, we studied the scaling properties of their upstream and downstream performance using recently proposed models of scaling laws for neural LMs. We showed that the pre-training loss and downstream linguistic performance of SLMs and LLMs is highly correlated, and that they both scale predictably according to power-laws. This predictable behavior 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 than those of LLMs. Additionally, we proposed a new speech dataset, sTINYSTORIES, and showed that its use during pre-training improves downstream semantic performance in SLMs. Finally, we explored the use of coarser speech tokenizations as a method to increase the ability of SLMs to model long-range dependencies. However, our results suggest that this is detrimental to downstream performance.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Peter J. Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

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.

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

J. B. MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press.

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.

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 s^2: 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.