SALM: SPEECH-AUGMENTED LANGUAGE MODEL WITH IN-CONTEXT LEARNING FOR SPEECH RECOGNITION AND TRANSLATION

Zhehuai Chen, He Huang*, Andrei Andrusenko, Oleksii Hrinchuk, Krishna C. Puvvada, Jason Li, Subhankar Ghosh, Jagadeesh Balam, Boris Ginsburg*

NVIDIA, USA

ABSTRACT

We present a novel Speech Augmented Language Model (SALM) with *multitask* and *in-context* learning capabilities. SALM comprises a frozen text LLM, a audio encoder, a modality adapter module, and LoRA layers to accommodate speech input and associated task instructions. The unified SALM not only achieves performance on par with task-specific Conformer baselines for Automatic Speech Recognition (ASR) and Speech Translation (AST), but also exhibits zero-shot in-context learning capabilities, demonstrated through keyword-boosting task for ASR and AST. Moreover, *speech supervised in-context training* is proposed to bridge the gap between LLM training and downstream speech tasks, which further boosts the in-context learning ability of speech-to-text models. Proposed model is open-sourced via NeMo toolkit [1].

Index Terms— LLM, ASR, AST, In-context Learning

1. INTRODUCTION

Large language models (LLMs) have achieved remarkable results on a variety of natural language processing (NLP) benchmarks recently [1, 2]. These models can be trained on massive amounts of unsupervised text data, and learn the knowledge that benefits many downstream text generative tasks. Through instruction tuning, LLMs can be fine-tuned to be more amenable to solving different NLP tasks in general. Additionally, LLMs demonstrate an in-context learning ability, meaning that they can learn from a few examples in the context, even if those examples are unseen in the training data.

These properties of LLMs are attractive to other modalities, including speech. Different interfaces between speech and LLMs have been studied, including text [3–6], quantized audio tokens [7, 8] and continuous audio embeddings [9–13]. Promising results have been shown in speech recognition, translation and synthesis.

In this work, we prompt Megatron LLM[14] using NeMo[15] speech models with different motivations: i) utilize the multitask ability of LLMs to construct a unified model for various speech tasks. ii) augment speech models with the in-context learning (ICL) ability of LLMs. Our main contributions include:

- Propose SALM which performs multitask speech-language modeling in a unified LLM framework. The unified model performs on par with bespoke Conformer baselines in ASR and AST. The speech-LLM solution is open-sourced via NeMo [15].
- Equip speech-to-text models with zero-shot in-context learning ability for the first time, shown by ASR and AST keyword boost.
- Propose *speech supervised in-context training* to further boost ICL ability of speech models.

2. RELATED WORK

The success of LLMs in NLP tasks [1, 2], has motivated growing interest in leveraging them to improve speech modeling. This work focuses on speech-to-text applications. One set of approaches use text as the interface between speech models and LLMs[3–6]. Recently [16] looks into using GPT-2 in the N-best rescoring for contextual ASR. However, some information in the speech modality may be lost due to the hardness in capturing them through text, e.g. speaker information, emotion and accents. In contrast, recent research starts to look at deep integration between speech models and LLMs, e.g. SpeechGPT [7], AudioPaLM [8], LTU [9], etc. [10–13]. Among them, Speech-LLaMA [11, 12] are the most relevant to this work, which share an architecture of prepending continuous audio embeddings to the text embeddings before feeding to a decoder-only LLM.

This work advances the previous works by equipping speech-to-text models with in-context learning (ICL) ability, demonstrated by keyword boosting tasks in ASR and AST. Extending ICL to speech domain is under-explored. Previous works VALL-E[17] and Voicebox[18] focus on the text-to-speech models. Moreover, we will open-source our implementation to accelerate this line of research.

The keyword boosting and contextual speech recognition have been explored in previous speech models. One branch of methods use external keyword LMs and WFSTs to bias the speech model in the inference time [19]. The other branch tries to integrate contextual information into the E2E modeling (CLAS)[20]. This work studies keyword boosting for speech applications with the in-context learning ability of LLMs and compares it to the first branch of methods. The proposed method does not require external context biasing graphs or learning explicit model weights for boosted words.

Text injection is another way for speech models to benefit from text. [21–23] modify the speech models to take both speech and text. [24, 25] scale these up and achieve remarkable success.

3. SALM - SPEECH AUGMENTED LANGUAGE MODEL

This work proposes to conduct supervised speech instruction tuning directly on a text pretrained and instruction fine-tuned LLM. The resultant *SALM* learns to condition on speech prompt, text context and instruction to predict textual outputs for different speech tasks, as shown in Figure 1. The introduced LLM potentially equips speech-to-text models with in-context learning ability.

3.1. Speech and text prompts

We choose Fast Conformer [26] and GPT-style Megatron LLM[14] as the speech and text backbones. Fast Conformer is a carefully

Thanks to Aleksandr Laptev, Somshubra Majumdar, Nithin Koluguri, Paarth Neekhara, Xuesong Yang, Vitaly Lavrukhin, Rafael Valle, Yi Dong, Adi Renduchintala, Sandeep Subramanian, Yang Zhang for discussion.

[1] https://github.com/NVIDIA/NeMo/tree/modular_speechllm

*Equal contribution

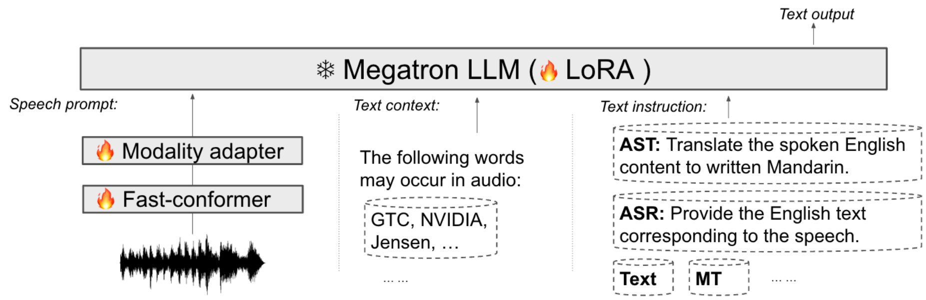

Fig. 1. SALM for Multitask Modeling and In-Context Learning.

redesigned Conformer [27] with a new downsampling schema for better efficiency while preserving state-of-the-art accuracy. We use a 110M pretrained audio encoder from NeMo and a pretrained 2B Megatron LLM with text instruction fine-tuned.

To guide LLM to condition on outputs from the audio encoder, we introduce modality adapter and LoRA layers [28] described below and train these layers through *multitask speech instruction tuning* in Section 3.2. Two Conformer layers with 4X subsampling are used as modality adapter layers in this paper to match the different information rate and modeling space between text and speech. The resultant speech prompt has a frame-shift of 320ms. It is projected to the LLM dimension and prepended to text context and instruction as the input of LLM, shown in Figure 1. Low-rank Adaptation (LoRA) layers with 128 dimensions are added to LLM during the speech instruction tuning. We freeze the LLM and back-propagate the rest.

3.2. Multitask supervised speech instruction tuning

One of the central motivation of combining speech model and LLM in this work is to bring the instruction tuning [29] from NLP to speech multitask learning and provide a unified speech model. We include different speech tasks (ASR, AST and more) with diverse instructions so as to not only promote instruction following but also improve generalization of the aforementioned modality adapter layers on different tasks. This work reuses paired speech and text data from ASR and AST public corpora and randomly prepends task instruction as examples in Figure 1 in the training time.

3.3. In-context learning for speech-to-text tasks

The other main motivation of SALM is to leverage the in-context learning (ICL) ability of LLM in speech tasks. ICL is one of the breakthrough from LLMs, to predict labels for unseen inputs without additional parameter updates. This ability was extended to the speech domain with previous works focusing on the text-to-speech (TTS) application. [17] proposed a neural codec language model that can synthesize speech for unseen speakers without fine-tuning.

In this work, we try to assess the *in-context learning ability in speech understanding tasks*, ASR and AST as examples, and improve upon it. We take the keyword boosting task as the first step towards this direction. Keyword boosting aims at biasing the model to recognize particular words of interest. We define the in-context learning here as: *learning the boosted words from the prompting text context, without back-propagation*. As demonstrated in Figure 2, we provide keywords to the model in the format of optional text context before text instruction. As a contrast, previous keyword boosting works require either learning explicit embeddings and weights for boosted words during training or with external biasing graphs.

3.4. Speech supervised in-context training

Given the differences in both data formats and learning criteria between LLM pretraining and ICL stages, previous NLP research suggests a series of supervised in-context finetuning strategies by constructing in-context training data to enhance ICL capability [30]. With similar motivation, the *speech supervised in-context training* (Speech ICT) is proposed in this work to promote the model to leverage the aforementioned text context in speech understanding.

In the speech instruction tuning stage, we augment the same supervised data by randomly sampling words from the current utterance and other utterances in the dataset, and including them as the optional text context for the utterance as Figure 3. We will later demonstrate in the experiment that this way of in-context training can generalize to unseen words, and corpora in unseen domains.

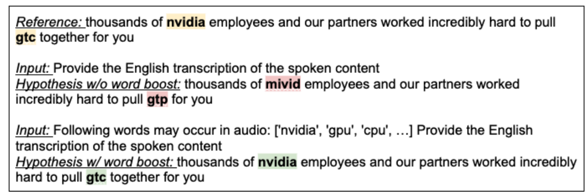

Fig. 2. Example of In-context Learning for Keyword Boosting.

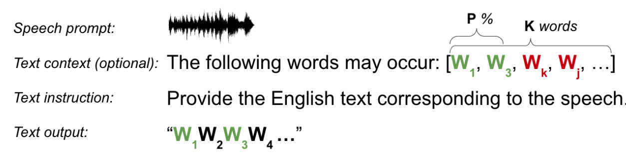

Fig. 3. Demonstration of the Proposed *Speech Supervised In-context Training*. The supervised data is augmented by including the optional text context with a probability of 5%, where K words are sampled with $P\%$ of words from the ground-truth (*positive ratio*).

4. DATA AND EXPERIMENTAL SETUP

Model Details: The whole pipeline is implemented via NeMo toolkit[15]. The audio encoder is initialized from the NGC ASR pretrained Fast Conformer-large[*] or the Conformer self supervised learning (SSL) checkpoint[*], while the modality adapter is randomly initialized. The Megatron LLM [14] we used has 2B parameters, which was trained on 1.1T tokens on a dataset that comprises 70% English, covering web-crawl data, news, conversation, books, and scientific domains, 15% Code from the Stack dataset [31] and 15% non-english text from CommonCrawl [*]. This model was then finetuned on public instruction following datasets like [29].

[*] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/stt_en_fastconformer_transducer_large`

[*] `https://catalog.ngc.nvidia.com/orgs/nvidia/teams/nemo/models/ssl_en_conformer_large`

[*] `https://commoncrawl.org/`

Table 1. SALM Results on ASR and AST tasks

systems	ASR WER LibriSpeech clean	ASR WER LibriSpeech other	AST BLEU MuST-C en-de	AST BLEU MuST-C en-ja
bespoke Fast Conf L+decoder	2.3	5.0	26.0	5.5
+ ASR pretrained encoder	**1.8**	**3.9**	**31.0**	**14.8**
SALM (SSL pretrained)	2.7	6.1	-	-
SALM (ASR pretrained)	2.4	5.3	27.1	15.0
+ nucleus sampling	**2.3**	**4.8**	29.6	16.5
ASR+AST SALM (nucleus s.)	2.6	6.1	**30.7**	**16.8**

Hyper-parameters: We train the model with 64 global batch size, using Adam optimizer with learning rate 1e-4 and weight decay of 1e-3. Cosine annealing with 2000 warm-up steps is applied. Gradients are clipped to 5.0. 8 of A100 GPUs are used for training. We use greedy decoding in the inference by default while nucleus sampling ($t = 0.2$, $p = 0.95$, $k = 50$)[32] is also tested.

Speech Recognition: We use the LibriSpeech [33] training set to train SALM, and pick the best checkpoint beasd on the WER on dev sets, which is then evaluated on *test-clean* and *test-other*. Our baseline model uses NGC ASR pretrained Fast Conformer-large encoder and transducer decoder with 114M parameters.

Speech Translation: For speech translation, we use all English audio data available for the Offline Track of IWSLT 2023 [34] paired with pseudo-generated translations to German and Japanese. Our training dataset consists of 2.7M segments which corresponds to 4.8K hours of audio. We used MuST-C v2 tst-COMMON [35] for evaluation. Our baseline model uses NGC ASR pretrained Fast Conformer-large encoder followed by 6-layer Transformer decoder. We used 16384k BPE encodings trained on texts in target language.

Keyword Boosting: For the keyword boosting evaluation we prepared an internal test set based on NVIDIA GTC talks data. The test set is forced aligned and segmented, 8 hours in total. The main feature of such a data set is the presence of a large number of different acronyms, product names, and technical terms, which often have low recognition accuracy for ASR systems. To build the keywords list we selected words and phrases with high occurrences in GTC test set and low recognition accuracy for greedy decoding of baseline transducer model [*]. We include 64 keywords by default and study different numbers. For the evaluation of keywords recognition accuracy we consider precision,P, and recall, R, calculated from keywords according to the alignment of the recognition results with the ground-truths. We also report F-score ($2 * P * R/(P + R)$). The baseline transducer model uses the shallow-fusion approach for the boosting [19]. During beam search decoding, partial hypotheses are rescored according to the context biasing graph. The implementation of the context biasing graph was taken from Icefall toolkit[*] with context score 4. We use modified adaptive expansion search based on [36] with beam width=5, alpha=2, and gamma=8.

5. RESULTS AND ANALYSIS

5.1. Unified model for ASR and AST

Table 1 shows the ASR results on LibriSpeech and AST results on MuST-C. We compare SALM with the bespoke baselines of ASR and AST in the first two rows. ASR baseline uses FastConformer-large encoder and transducer decoder (FC-T). AST baseline uses

[*]Examples: *NVIDIA, GPU, Omniverse, Geforce, NeMo, kubernetes,* etc.

[*]`https://github.com/k2-fsa/icefall/blob/master/icefall`

Table 2. Win and Loss Comparing SALM and Fast Conformer-Transducer, FC-T (errors are shown in red).

	Type	FC-T	SALM
Win	rare word	... a kleptomania like cousin snatcher	... a kleptomaniac like cousin snatcher
	segment	greenhorns flat heads	greenhorns flatheads
Loss	hallucinate	ah lida exclaimed fauchelevent	ah lidah exclaimed shoot up the english transcription ...
	AM	rachel lake rachel lake ...	routen leak routen leak ...
	deletion	six hundred bishops four emperors ... three hundred canonized ...	six hundred [del error] canonized ...

Table 3. ASR Keyword Boosting Results on GTC Talk Test Set.

Systems	boost	WER	F-score (*P/R*)
Fast Conf L-Transducer +	N	16.2	0.36 (0.96/0.22)
ASR pretrained encoder	Y	**15.1**	**0.67 (0.87/0.55)**
SALM	N	17.0	0.35 (0.94/0.21)
	Y	15.8	0.56 (0.74/0.45)
+ nucleus sampling	Y	**14.9**	**0.61 (0.66/0.57)**

transformer decoder instead and one model is trained on each language-pair as found to perform the best. The first row trains from-scratch and the second uses the aforementioned NGC Fast Conformer ASR pretrained encoder.

The SALM model in the 3rd and 4th rows initializes the audio encoder from the aforementioned NGC SSL and ASR checkpoints respectively. The best SALM model in the fifth row with nucleus sampling in the LLM inference outperforms the from-scratch baseline but still behind the stronger baseline in the second row.

For AST, we train one SALM model in the fifth row to support both language pairs and use text instruction as shown by Figure 1 to switch between different pairs. We then further train one SALM model on both AST and ASR data to provide a unified model for both tasks in the last row. The unified model performs better than the two baselines and AST-only SALM. When operating on the ASR task, this model is worse than ASR-only SALM.

To understand the strengths and weaknesses of LLM based SALM versus the baseline, Table 2 includes the ASR hypothesis comparison. Although SALM suffers from hallucination and long-form deletion problems, it performs better on rare words and proper nouns. We found nucleus sampling can solve some of the former problems and result in better results. Further alleviating these problems will be our future work.

5.2. Zero-shot in-context Learning for keyword boosting

We study the zero-shot in-context learning ability of SALM by taking keyword boosting task as the proxy in Table 3. We took the Fast Conformer-L Transduer (FC-T) initialized with NGC ASR pretrained encoder and trained on LibriSpeech as a strong baseline. Without boosting, the LibriSpeech-trained SALM performs on par (Row 3 v.s. 1). We prompt SALM for keyword boosting in Row 4 with the text context described in Section 3.3. The better result from Row 3 to 4 demonstrates the effectiveness of in-context learning method. Nucleus sampling in the LLM inference can further

Table 4. Improve In-context Learning with Speech ICT. *positive ratio* is defined as the percentage of ground-truth words in augment.

Speech ICT training setup		Eval with 64 words
positive ratio, %	# of keywords	F-score (*P/R*)
n/a	0	0.38 (0.82/0.25)
33%	3	0.52 (0.59/0.47)
33%	64	0.52 (0.62/0.44)
6%	**64**	0.56 (0.74/0.45)
3%	64	0.55 (0.79/0.42)

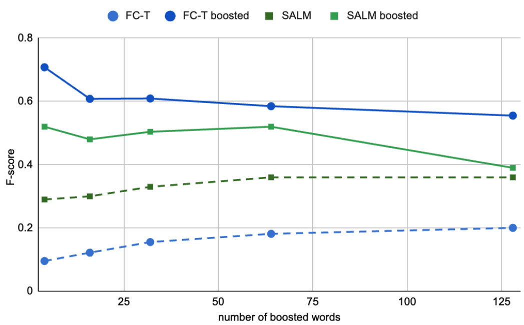

Fig. 4. Scalability of # of Boosted Words Comparing Baseline FC-T and In-context Learning based SALM Keyword Boosting

boost the performance, results in the 5th row. Compared to baseline boosting, this method achieved similar relative boosting gains while not requiring external biasing graphs[19] as in baseline or learning explicit biasing embedding[20].

Table 4 demonstrates the necessity of the proposed Speech ICT. Although the LLM used in SALM has been instruction fine-tuned with text data, SALM in Row 1 without Speech ICT cannot effectively follow the prompt and obtain limited improvement. This shows the challenge of transferring textual knowledge to the speech domain in the current speech and LLM research. Speech ICT provides a route towards solving this problem. Including the augmented in-context training data designed in Section 3.4 significantly improves the performance. Tuning *positive ratio* in the table affects inference precision and recall – the bigger it is the worse precision and better recall. The best 6% is used in the rest.

Figure 4 studies the scalability of the in-context learning based keyword boosting method for SALM. When scaling up the number of boosted words, both baseline boosting method and SALM suffer from worse precision with almost unchanged recall. This behavior is caused by a gain of false accepts associated with an increase in the number of candidate words. We believe this problem in SALM can be alleviated by making LLM better handle the long contexts [1, 2].

We look into the win and loss and different error patterns between SALM and baseline in Table 5. Generally, SALM performs better on shorter words, compound words, and text normalization, while baseline boosting on FC-T performs better on longer words and phrases. Nevertheless, SALM suffers from hallucination and early stopping problems that is seldom seen in the baseline.

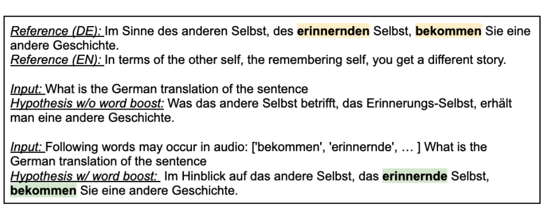

Fig. 5. Example of Using ICL for AST Keyword Boosting.

Table 5. Comparison of Keyword Boosting with SALM v.s. Baseline Fast Conformer-Transducer (FC-T) w/ boosting (errors in red).

		FC-T	SALM
Win words		nvidia, omniverse, robotic, cybersecurity, ...	gpu, hpc, cudnn, geforce, nvlink, healthcare, ...
	Type	FC-T Hypothesis	SALM Hypothesis
Win	text norm	g t c is the g p u computing developers conference	gtc is the gpu computing developers conference
	boost	we're sophor company	we're a software company
	boost	tim is the virtuality driver	tim is the virtual reality driver
Loss	hallucinate	computer graphics is the driving force of the g p	cyberspace is the driving force of the gpu1 michelangelo sopieness ...
	hallucinate	[del error] ladies and gentlemen	okay ladies and gentlemen clap your hands cla cla ...
	early stop	i am even the composer of the music you are hearing i ai brought to life by vivid deep ...	and i am even the composer of the music you are hearing um are you one cupom

Table 6. SALM based AST Keyword Boosting on MuST-C EN-DE

systems	boost	F-score (*P/R*)
SALM	N	0.20 (0.33/0.15)
	Y	**0.26 (0.25/0.27)**

5.3. In-context learning for dictionary-guided translation

We also conduct initial studies to see whether above keyword-boosting method can be applied to speech translation in Table 6. We select 40 German words from MuST-C EN-DE dev set with high occurrence in references and low occurrence in hypotheses, and boost them through prompting SALM.

Although the overall improvement on F-score is moderate, some successful examples (e.g., Figure 5), show that SALM can correctly pick up the boosted words and the resultant translation is natural. The in-context learning based SALM provides a new route towards dictionary-guided translation task, where users want to guide translation using pre-defined dictionary entries in inference time [37].

6. CONCLUSION

We have described *SALM*, which prompts Megatron LLM[14] using NeMo[15] speech models. We advance recent Speech-LLM works in two dimensions: i) utilize the *multitask* ability of LLMs to construct a unified model for various speech tasks, as demonstrated by performance on par with bespoke ASR and AST baselines. ii) augment speech models with the *in-context learning* (ICL) ability of LLMs. We define and study the ICL of speech-to-text models, and further improve it with *speech supervised in-context training*. We also open-source our implementation to accelerate this line of research. Future plans include solving the demonstrated hallucination, deletion and long context issues in LLM based SALM.

7. REFERENCES

[1] OpenAI, "Gpt-4 technical report," *arXiv*, 2023.

[2] Rohan Anil, Andrew M Dai, Orhan Firat, et al., "Palm 2 technical report," *arXiv:2305.10403*, 2023.

[3] Rongjie Huang, Mingze Li, Dongchao Yang, et al., "Audiogpt: Understanding and generating speech, music, sound, and talking head," *arXiv:2304.12995*, 2023.

[4] Feilong Chen et al., "X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages," *arXiv:2305.04160*, 2023.

[5] Tongzhou Chen, Cyril Allauzen, Yinghui Huang, et al., "Large-scale language model rescoring on long-form data," in *ICASSP*. IEEE, 2023, pp. 1–5.

[6] Rao Ma et al., "N-best t5: Robust ASR error correction using multiple input hypotheses and constrained decoding space," *arXiv:2303.00456*, 2023.

[7] Dong Zhang, Shimin Li, Xin Zhang, et al., "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," *arXiv:2305.11000*, 2023.

[8] Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, et al., "Audiopalm: A large language model that can speak and listen," *arXiv:2306.12925*, 2023.

[9] Yuan Gong, Hongyin Luo, Alexander H Liu, et al., "Listen, think, and understand," *arXiv:2305.10790*, 2023.

[10] Shaoshi Ling, Yuxuan Hu, Shuangbei Qian, et al., "Adapting large language model with speech for fully formatted end-to-end speech recognition," *arXiv:2307.08234*, 2023.

[11] Jian Wu, Yashesh Gaur, Zhuo Chen, et al., "On decoder-only architecture for speech-to-text and large language model integration," *arXiv:2307.03917*, 2023.

[12] Yassir Fathullah, Chunyang Wu, Egor Lakomkin, et al., "Prompting large language models with speech recognition abilities," *arXiv:2307.11795*, 2023.

[13] Mingqiu Wang, Izhak Shafran, Hagen Soltau, et al., "Speech-to-text adapter and speech-to-entity retriever augmented llms for speech understanding," *arXiv:2306.07944*, 2023.

[14] Mohammad Shoeybi et al., "Megatron-lm: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[15] Oleksii Kuchaiev, Jason Li, Huyen Nguyen, et al., "NeMo: a toolkit for building ai applications using neural modules," *arXiv:1909.09577*, 2019.

[16] Guangzhi Sun, Xianrui Zheng, Chao Zhang, and Philip C Woodland, "Can contextual biasing remain effective with whisper and gpt-2?," *arXiv preprint arXiv:2306.01942*, 2023.

[17] Chengyi Wang et al., "Neural codec language models are zero-shot text to speech synthesizers," *arXiv preprint arXiv:2301.02111*, 2023.

[18] Matthew Le et al., "Voicebox: Text-guided multilingual universal speech generation at scale," *arXiv preprint arXiv:2306.15687*, 2023.

[19] Ian Williams, Anjuli Kannan, Petar S Aleksic, et al., "Contextual speech recognition in end-to-end neural network systems using beam search.," in *Interspeech*, 2018.

[20] Golan Pundak, Tara N Sainath, Rohit Prabhavalkar, Anjuli Kannan, and Ding Zhao, "Deep context: end-to-end contextual speech recognition," in *2018 IEEE spoken language technology workshop (SLT)*, 2018.

[21] Samuel Thomas et al., "Integrating text inputs for training and adapting rnn transducer asr models," in *ICASSP*. IEEE, 2022.

[22] Zhehuai Chen, Yu Zhang, Andrew Rosenberg, et al., "Maestro: Matched speech text representations through modality matching," *arXiv:2204.03409*, 2022.

[23] Ziqiang Zhang, Sanyuan Chen, Long Zhou, et al., "Speechlm: Enhanced speech pre-training with unpaired textual data," *arXiv:2209.15329*, 2022.

[24] Yu Zhang, Wei Han, James Qin, et al., "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv:2303.01037*, 2023.

[25] Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, et al., "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv:2308.11596*, 2023.

[26] Dima Rekesh, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Juang, Oleksii Hrinchuk, Ankur Kumar, and Boris Ginsburg, "Fast conformer with linearly scalable attention for efficient speech recognition," *arXiv:2305.05084*, 2023.

[27] Anmol Gulati et al., "Conformer: Convolution-augmented transformer for speech recognition," *arXiv preprint arXiv:2005.08100*, 2020.

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

[29] Jason Wei et al., "Finetuned language models are zero-shot learners," in *ICLR*, 2021.

[30] Qingxiu Dong, Lei Li, Damai Dai, et al., "A survey for in-context learning," *arXiv:2301.00234*, 2022.

[31] Denis Kocetkov et al., "The stack: 3 tb of permissively licensed source code," *Preprint*, 2022.

[32] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi, "The curious case of neural text degeneration," *arXiv preprint arXiv:1904.09751*, 2019.

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

[34] Milind Agarwal et al., "Findings of the IWSLT 2023 Evaluation Campaign," in *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*. 2023, Association for Computational Linguistics.

[35] Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi, "Must-c: A multilingual corpus for end-to-end speech translation," *Computer Speech & Language*, vol. 66, pp. 101155, 2021.

[36] Juntae Kim, Yoonhan Lee, and Eesung Kim, "Accelerating rnn transducer inference via adaptive expansion search," *IEEE Signal Processing Letters*, vol. 27, pp. 2019–2023, 2020.

[37] Tamer Alkhouli, Gabriel Bretschner, and Hermann Ney, "On the alignment problem in multi-head attention-based neural machine translation," *arXiv:1809.03985*, 2018.