Interspeech 2024
1-5 September 2024, Kos, Greece

Exploring In-Context Learning of Textless Speech Language Model for Speech Classification Tasks

Kai-Wei Chang[*,1], *Ming-Hao Hsu*[*,2], *Shan-Wen Li*[3], *Hung-yi Lee*[2]

[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]Department of Electrical Engineering, National Taiwan University, Taiwan
[3]Meta AI, USA

kaiwei.chang.tw@gmail.com, qaz159qaz159@gmail.com

Abstract

Ever since the development of GPT-3 in the natural language processing (NLP) field, in-context learning (ICL) has played an essential role in utilizing large language models (LLMs). By presenting the LM utterance-label demonstrations at the input, the LM can accomplish few-shot learning without relying on gradient descent or requiring explicit modification of its parameters. This enables the LM to perform various downstream tasks in a black-box manner. Despite the success of ICL in NLP, little work is exploring the possibility of ICL in speech processing. This study is the first work exploring ICL for speech classification tasks with textless speech LM. We first show that the current speech LM lacks the ICL capability. We then perform warmup training on the speech LM, equipping the LM with demonstration learning capability. This paper explores and proposes the first speech LM capable of performing unseen classification tasks in an ICL manner.

Index Terms: In-context learning, speech language model, prompt tuning, few-shot learning, speech classification

1. Introduction

Large language models (LLMs) [1, 2] have gained significant attention in recent years. With the development of LLMs like GPT-3 [3], researchers have discovered the potential for performing **in-context learning (ICL)** [3, 4]. ICL is a technique that enables LMs to learn to perform new tasks from a small number of demonstrations which are presented at the input of the LM. Formally, we consider a set of data points, denoted as x_i, along with their corresponding labels, denoted as y_i [1]. Additionally, we have a target data point, x_t, for which we want the LM to make an inference. To achieve this, we prepend the demonstrations, consisting of the data points and their labels, to the input sequence as follows: $[x_1, y_1, x_2, y_2, ..., x_n, y_n, x_t]$. By learning the analogy between the data points and their labels, the LM is capable of directly predicting the label of x_t. It is important to note that throughout this learning process, the LM remains fixed, and there is no gradient backward process involved. Instead, the LM relies solely on the input demonstrations to acquire knowledge.

ICL, as an *emergent ability* [9, 10] of the LLM, remains insufficiently optimized due to its disparity between the LM's pre-training task. Regarding this, researchers have attempted to perform *warmup training* [4] in either a supervised [11] or self-supervised [12, 13] manner to enhance the ICL capability. The motivation for building an LM with ICL capability is that, as a new paradigm, ICL offers several advantages. First, ICL simplifies the integration of human knowledge into the LM by providing demonstrations. This process resembles the analogy reasoning process of humans [1]. Second, ICL incorporates demonstrations at the input, eliminating the need for backpropagation and gradient flow to establish connections between data points and their labels. As a result, computational costs are reduced. Third, LLMs are often released as a service in the real-world application [3, 14, 15]. ICL is particularly suitable for LM deployment since it only modifies the input, allowing LLMs to adapt to new tasks defined by the users [14]. Overall, in the NLP field, ICL has emerged as a powerful paradigm for utilizing LLMs. However, despite the recent advancements in large speech LMs, there is a notable lack of research on ICL in the domain of speech processing. Fortunately, the recently developed **textless speech language models (speech LMs)** offer an opportunity for us to explore ICL for speech processing.

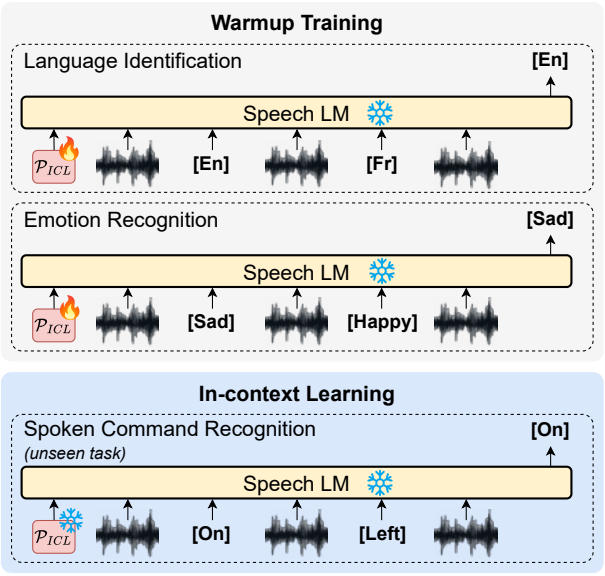

Figure 1: *This figure illustrates the framework of the proposed approach, where warmup training utilizes a variety of training tasks to instill in-context learning abilities in the speech language model. This allows the model to utilize task demonstrations effectively to tackle novel unseen tasks.*

*The first two authors contributed equally.

[1]In this paper, we consider ICL following the definition in [4]. That is, the model learns the relationship between the "data point x" and its "label y" by providing them as demonstrations. Recent works, such as VALL-E [5] and AudioBox [6], also claim to perform ICL with large speech models. However, they primarily focus on learning the acoustic condition or speaker identity of the given speech prompt. This topic has been studied in the speech processing literature as "voice cloning [7]" and "style transfer [8]" and is outside the scope of this paper.

10.21437/Interspeech.2024-1932

Textless speech LMs quantize speech representations into *discrete speech tokens* [16] and engage in the next token prediction pre-training task, akin to language models in the NLP field. These textless speech LMs are trained on large speech-only datasets and demonstrate robust generation capabilities without any text supervision. Demonstrating robust capabilities, these speech LMs can generate meaningful discrete tokens conditioned on given speech segments. Notable examples include the Generative Spoken Language Model (GSLM)[17], pGSLM[18], audioLM [19] and others. Although recently there have been researches utilizing text data to help speech language modeling (e.g. TWIST [20], VALL-E [5]), we focus on textless speech LM in this work. The reason is that the textless property offers appealing advantages, including but not limited to: (1) **Language agnostic**: When performing downstream tasks, textless speech LM is not restricted to text prompts with a specific language [5]. (2) **Direct speech modeling**: Textless speech LM directly models the speech signal, which bypasses the need for speech-text paired data that is usually expensive to collect. Furthermore, many languages lack a written or formal textual form. Overall, the textless property offers us a more general and easy-to-scale framework.

We first examined the ICL ability of the current largest open-sourced textless speech LM, GSLM. We observed that GSLM failed to comprehend the provided speech-label pairs for speech classification tasks, indicating the lack of emergent ability to perform ICL. If the LM is capable of performing ICL, it makes predictions by learning the input-label mapping [21], even with random labels [22]. A possible explanation is the LM performs "implicit fine-tuning" during ICL [23]. The preliminary result shows the current speech LM does not possess such emergent ability [2]. As shown in Fig. 1, to build a speech LM with ICL ability, we conduct a simple warmup training with parameter efficient tuning (PET) method, prompt tuning [24, 25]. Warmup training is performed on a set of training tasks to enable the speech LM to understand the demonstrations and make predictions when encountering unseen tasks.

The experimental results indicate that GSLM, when subjected to warmup training, demonstrates the capability to perform ICL not only on seen tasks but also, surprisingly, on unseen tasks. It surpasses the random guessing baseline for all tasks and can outperform the support vector classifier (SVC) in most scenarios. It's worth noting that, in this paper, we aim to show the feasibility of ICL for speech LM, not outperform the current state-of-the-art methods. We believe that as more advanced textless speech LM continues to be developed, the significance of ICL behavior will become increasingly evident, mirroring the observations made in the NLP field (a notable example is the evolution from GPT-2 to GPT-3 [22]). Our contributions are as follows:

- We investigate the in-context learning capability of the existing speech LM and identify its limitations in this regard.
- We introduce the first speech LM that incorporates warmup training, enabling it to perform in-context learning effectively. This is the first speech LM with such capabilities.
- We empirically demonstrate that the speech LM can effectively learn and adapt to unseen tasks through ICL and achieve non-trivial results, surpassing the performance of a random sampling baseline.

[2] In fact, we also examined a more powerful speech LM, TWIST [20], that adopts text data for pre-training. However, we do not observe the ICL capability either.

2. Method

We first investigated the ICL ability of the pre-trained GSLM [17], which is the largest open-sourced textless generative speech LM. In this paper, we focus on speech classification tasks. We provide speech-label pairs as demonstrations at the input and let the model predict the label of a target utterance. Our findings indicate that the current GSLM does not possess ICL ability. As shown in Table 1, "w/o Warmup" is the performance of directly applying ICL on GSLM without warmup training. "Random" is the performance of randomly guessing a label for the speech classification tasks. We find that applying ICL directly to GSLM yields results worse than random guessing.

To address this limitation and build a speech LM with ICL ability, we propose conducting **warmup training** on the speech LM. For the warmup training, we utilize a set of training tasks denoted as $\mathcal{T}_{train}$ to enhance the speech LM's ICL capability. Specifically, we employ a PET method, prompt tuning [24, 25, 33]. Applying prompt tuning is a design choice, and other tuning methods can also be adopted in the warmup training, as discussed in [4]. However, We conduct prompt tuning for two reasons: (1) Preliminary experiments revealed that fine-tuning the entire model for warmup training is unstable and might lead to inferior ICL performance. (2) Prompt tuning prepends prompt vectors at the input side while keeping the pre-trained speech LM fixed. This preserves its generative capability, which is beneficial for future applications.

In the following sections, we outline our approach to conducting warmup training on the set of training tasks $\mathcal{T}_{train}$ and evaluate the ICL capability of the model on the training tasks $\mathcal{T}_{train}$ (seen tasks) and testing tasks $\mathcal{T}_{test}$ (unseen tasks).

2.1. Warmup Training

Given a speech LM $\mathcal{M}$ that performs next token prediction on discrete speech tokens x:

$$x_{t+1} = \mathcal{M}(x_1, x_2, ..., x_t), \tag{1}$$

where t is the timestep. We first collect a set of training tasks $\mathcal{T}_{train}$ to perform warmup training. Each task T_i in $\mathcal{T}_{train}$ uses its own dataset. To form one training data point for ICL warmup training, we conduct the following procedure:

(1) We randomly sample n utterances and their corresponding labels from a training task as demonstrations. (2) Following GSLM [17], the utterances are first encoded into discrete token sequences $\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n$. In addition, we utilize **random label mapping (RLM)** [34, 35, 22] to map the task's labels (target domain) to the discrete tokens (source domain), resulting in labels for the demonstrations $y_1, y_2, ..., y_n$. (3) Each speech token sequence is then truncated or padded to the same utterance length L, yielding $\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_n$. We found this step critical as it provides a standardized format to the speech LM and simplifies the training. (4) The input data is constructed as

$$X = [\tilde{\boldsymbol{x}}_1, \langle s \rangle, y_1, \langle s \rangle, ..., \tilde{\boldsymbol{x}}_n, \langle s \rangle, y_n, \langle s \rangle, \tilde{\boldsymbol{x}}_t, \langle s \rangle], \tag{2}$$

where "$\langle s \rangle$" is a separation token with trainable embedding, and $\tilde{\boldsymbol{x}}_t$ is the target utterance, which is the data we want the model to predict its label.

During warmup training, we randomly sample the target utterance from the demonstrations, that is, $\tilde{\boldsymbol{x}}_t \in \{\tilde{\boldsymbol{x}}_1, \tilde{\boldsymbol{x}}_2, ..., \tilde{\boldsymbol{x}}_n\}$. The model then learns to compare the target utterance and the demonstrations in order to predict the correct label. We find this step simple but effective since it simplifies the training objective. The model is required to compare the target utterance with each

Table 1: *This table displays the accuracy for ICL with warmup training (w/ Warmup), random guessing (Random), and ICL without warmup training (w/o Warmup) and SVC on seen tasks ($\mathcal{T}_{train}$) and unseen tasks ($\mathcal{T}_{test}$) for the two task groups. The results demonstrate that warmup training enables effective in-context learning, outperforming baselines on diverse speech tasks.*

Group	Task Type	Task	Dataset	w/ Warmup	Random	w/o Warmup	SVC
Group 1	$\mathcal{T}_{test}$	SCR	Arabic SC [26]	40.9 $\pm$ 1.5	28.0 $\pm$ 1.0	3.9 $\pm$ 0.7	**50.8**
		ER	IEMOCAP [27]	**47.7** $\pm$ 1.8	41.0 $\pm$ 1.4	3.1 $\pm$ 0.4	33.7
	$\mathcal{T}_{train}$	SCR	Google SC v2 [28]	**79.6** $\pm$ 1.1	25.1 $\pm$ 1.6	6.4 $\pm$ 0.8	43.8
		SCR	Lithuanian SC [29]	**80.5** $\pm$ 1.1	25.1 $\pm$ 1.6	8.4 $\pm$ 1.2	37.8
		SCR	Dysarthric Mandarin SC [30]	**57.8** $\pm$ 0.8	24.7 $\pm$ 0.9	9.8 $\pm$ .0.9	12.0
		LID	Voxforge [31]	**32.0** $\pm$ 1.6	23.8 $\pm$ 2.7	1.8 $\pm$ 0.5	29.7
		SD	MUStARD [32]	**64.7** $\pm$ 1.6	54.7 $\pm$ 1.0	1.2 $\pm$ 0.2	60.9
Group 2	$\mathcal{T}_{test}$	SCR	Arabic SC [26]	36.5 $\pm$ 1.2	28.0 $\pm$ 1.0	3.9 $\pm$ 0.7	**50.8**
		SCR	Google SC v2 [28]	**48.0** $\pm$ 0.7	25.1 $\pm$ 1.6	6.4 $\pm$ 0.8	43.8
		SD	MUStARD [32]	**64.1** $\pm$ 1.1	54.7 $\pm$ 1.0	1.2 $\pm$ 0.2	60.9
	$\mathcal{T}_{train}$	SCR	Dysarthric Mandarin SC [30]	**56.0** $\pm$ 1.9	24.7 $\pm$ 0.9	9.8 $\pm$ 0.9	12.0
		SCR	Lithuanian SC [29]	**80.5** $\pm$ 0.9	25.1 $\pm$ 1.6	8.4 $\pm$ 1.2	37.8
		LID	Voxforge [31]	**29.9** $\pm$ 0.5	23.8 $\pm$ 2.7	1.8 $\pm$ 0.5	29.7
		ER	IEMOCAP [27]	**47.6** $\pm$ 1.7	41.0 $\pm$ 1.4	3.1 $\pm$ 0.4	33.7

demonstration and output its corresponding label. The learned behavior benefits ICL in the next stage. We then employ prompt tuning to learn the prompts $\mathcal{P}_{ICL}$ to equip the model with ICL capability. Following [36], a set of prompt vectors $\mathcal{P}_{ICL}$ are prepended at the input of the speech LM. The speech LM then makes the prediction conditioned on the demonstrations X and the prompt vectors $\mathcal{P}_{ICL}$. We then apply cross entropy loss $\mathcal{L}$ on the model prediction and the ground truth label of the target utterance y_t for optimizing the prompts:

$$\mathcal{P}_{ICL} \leftarrow \arg\min_{\mathcal{P}_{ICL}} \mathcal{L}(\mathcal{M}(X; \mathcal{P}_{ICL}), y_t)), \tag{3}$$

where $\mathcal{M}(X; \mathcal{P}_{ICL})$ represents the model prediction conditioned on the input data X and the prompts $\mathcal{P}$, and y_t is the ground truth label corresponding to the target utterance $\tilde{\boldsymbol{x}}_t$.

2.2. In-context Learning

After completing the warmup training, the model becomes ready to perform ICL on the training tasks $\mathcal{T}_{train}$ (seen tasks) and testing tasks $\mathcal{T}_{test}$ (unseen tasks). The process of preparing demonstrations during this stage is similar to the warmup stage as described in the formula (2). The difference is that, in the ICL stage, the target utterance $\tilde{\boldsymbol{x}}_t$ is no longer included in the demonstrations. Instead, its corresponding label $y_t \in \{y_1, y_2, ..., y_n\}$ is included, enabling the model to learn to make predictions based on analogies.

3. Experimental Setup

3.1. Tasks and Datasets

We evaluate speech LM's ICL ability on a diverse set of speech classification tasks with 7 datasets. These tasks include speech command recognition (SCR), emotion recognition (ER), language identification (LID), and sarcasm detection (SD). Also, the datasets involve varying languages, accents, domains, and label distributions, allowing a comprehensive evaluation of the transferability of ICL for speech LM. As shown in Table 1, we select two groups (Group 1 and Group 2) of training tasks $\mathcal{T}_{train}$ and testing tasks $\mathcal{T}_{test}$. These tasks are combined in ways that introduce variety, for instance, tasks with different numbers of classes and different types of tasks. This approach enables us to evaluate our model's performance in ICL and gauge the impact of the training tasks on performance across a range of testing tasks. Our attention is directed explicitly toward the model's capacity to utilize learned ICL knowledge during warmup training when faced with a new task. This capability is essential for many practical uses in a multitude of real-world situations.

We've chosen particular datasets for both training and testing. Our aim for each training dataset group includes generating a balanced dataset to avoid bias. To meet this goal, we sample 10,000 data points, each including 4 demonstrations from the training tasks. The same amount of data points is ensured for every task, providing a balanced dataset as each task gets an equal proportion of demonstrations. If we neglect to do this and simply use the entire dataset, some of the larger datasets would dominate a significant portion of the combined dataset, resulting in a skewed dataset.

3.2. Implementation Detail

We adopt GSLM [17] as our backbone speech LM. Specifically, the GSLM is trained on top of discrete units encoded by HuBERT [37] SSL speech model and K-means clustering algorithm with 100 clusters. In warmup training, we conduct prompt tuning and use a prompt length equal to 5. This approach introduces a small fraction of trainable parameters, specifically less than 0.1% of the total 150 million parameters of the speech LM, simplifying the learning process. As described in Section 2, our initial approach involves enforcing a fixed length for utterances. In the primary experiment in Sec. 4.1, we fix the utterance length L to be 50. This standardization ensures consistent utterance lengths across multiple datasets and simplifies the training. We also investigate the impact of varying utterance lengths in Sec. 4.3. We provide four demonstrations and one target utterance in both warmup training and the ICL stage [3].

Given the limited research of ICL on speech LM, we compare our GSLM with warmup training (**w/ Warmup**) with three baselines: (1) random guessing (**Random**), (2) ICL on GSLM

[3]In our preliminary study, we find four demonstrations as a suitable setup, for it provides a reasonable demonstration number while preventing the "curse of long sequence" problem as discussed in the previous work [36]. How to incorporate more demonstrations and alleviate the long-form problem remains future work.

Table 2: *Guessing Rate Analysis. The results demonstrate that warmup training significantly improves the model's ability to make predictions based on the provided demonstrations in ICL.*

Group	Task Type	Guessing Rate (Avg)	
		w/o Warmup	w/ Warmup
Group 1	$\mathcal{T}_{test}$	13.9	95.1
	$\mathcal{T}_{train}$	21.0	98.4
Group 2	$\mathcal{T}_{test}$	15.0	92.2
	$\mathcal{T}_{train}$	21.7	97.4

without warmup training (**w/o Warmup**), and (3) a support vector classifier (**SVC**). The random guessing method entails making predictions by selecting labels at random from the demonstrations; while SVC is trained using the provided demonstrations to make predictions. We repeat the experiments five times and compute the average accuracy alongside its standard deviation, offering a more fair evaluation of the model's performance.

4. Results

4.1. Main Result

In Table 1, the result shows that ICL with warmup training consistently outperforms both the Random and w/o Warmup baselines and surpasses SVC for most tasks. For Group 1, in unseen tasks $\mathcal{T}_{test}$, ICL with warmup training excels on the Arabic SC dataset with a score of 40.9, notably higher than both Random and w/o Warmup, but lower than SVC's 50.8. In the IEMOCAP dataset, w/ Warmup surpasses all baselines, including SVC. In seen tasks $\mathcal{T}_{train}$, the w/ Warmup method is superior, particularly scoring 79.6 on Google SC v2 and 80.5 on Lithuanian SC, outpacing all baselines. Group 2 follows a similar trend; the w/ Warmup method leads across the board. In test tasks on Google SC v2, it scores 48.0, markedly higher than both Random and w/o Warmup and slightly better than SVC.

Overall, in both Group 1 and Group 2, ICL with Warmup outperforms other baselines. The only exception is the Arabic SC task presented as an unseen task. Although GSLM can perform ICL when such cross-lingual tasks are in $\mathcal{T}_{train}$ (such as Lithuanian SC and Mandarin SC), we hypothesize that for GSLM, it's still a challenge when presenting cross-lingual tasks as unseen tasks $\mathcal{T}_{test}$. These results underline the efficacy of the warmup training for ICL, as it consistently outperforms both baselines across a diverse range of speech tasks. This success opens up new paths for future studies and holds potential for more improvements in this field.

4.2. Model Behavior Analysis

Warmup training aims to equip the model with the ability to identify, compare demonstrations, and comprehend the target task. We assess the model's ability to predict labels based on demonstrations by examining its guessing rate during ICL. The guessing rate indicates the probability of the model identifying demonstrated labels. Higher guessing rates demonstrate effective ICL, with the model making predictions derived from the demonstrations. Table 2 compares guessing rates between models with and without warmup training across two groups, including seen and unseen tasks. Models lacking warmup training exhibit low guessing rates in both Group 1 and Group 2. Conversely, warmup training significantly boosts guessing rates to over 90%.

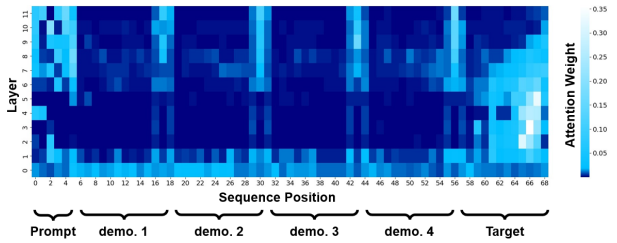

Figure 2: *Model Predicting Attention Weights in ICL. Initially, the model primarily attends to the demonstrations within the first two layers. It then gradually shifts its focus to the target in layers 3 through 7, and finally towards the labels in the demonstrations in the final layers. For the demonstration purpose, we show the scenario where the utterance length L is 10.*

Table 3: *Utterance Length Analysis. Accuracy comparison across different discrete tokens in the utterance is reported.*

# of Tokens		10	30	50	Not Fixed
Group 1	Arabic SC	40.2	**41.7**	40.9	23.7
$\mathcal{T}_{test}$	IEMOCAP	42.2	45.7	**47.7**	44.8

We further examine the model's behavior while predicting the label in ICL. The attention map in the model's attention layers during the execution of ICL is depicted in Figure 2. The figure reveals that the initial two layers mainly concentrate on the demonstrations. The focus then shifts to the target utterance in the middle layers (3rd to 7th) and finally shifts to the demonstrations' labels in the last layers (8th to 12th). Also, the model's continued attention to the prompts $\mathcal{P}_{ICL}$ highlights the utility of warmup training in ICL. From the studies in this section, we can see that the warmup training effectively steers the model to perform ICL for unseen tasks.

4.3. Utterance Length Analysis

As shown in Table 3, the model's performance is influenced by the number of discrete tokens in each utterance. If the length of the utterance is too long, although it might hold sufficient information, the GSLM could struggle with modeling such long sequences as reported in [25]. On the other hand, if utterances are too short, the model may lack the necessary information, leading to random guessing. We expect with more advanced speech LM built, more demonstrations with longer length can be incorporated to boost the performance.

5. Conclusion

This paper presents the first successful application of in-context learning (ICL) to speech LM. We initially investigated the limitations of the current speech LM in performing ICL. With the proposed warmup training, the textless GSLM demonstrates ICL capability on seen tasks and successfully achieves non-trivial results on unseen tasks across diverse datasets. This paper does not aim to achieve competitive performance with ICL but to show its feasibility. We're also aware that the current capacity of the backbone GSLM is restricted. GPT-3 contains 175B parameters, while GSLM is 1000 times smaller, containing 150M parameters, which might cause the limitation of the ICL emergent ability. Future works include investigating ICL on more diverse speech LM and developing more effective warmup strategies for ICL.

6. References

[1] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[4] Q. Dong, L. Li, D. Dai, C. Zheng, Z. Wu, B. Chang, X. Sun, J. Xu, and Z. Sui, "A survey on in-context learning," *arXiv preprint arXiv:2301.00234*, 2022.

[5] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," 2023.

[6] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan *et al.*, "Audiobox: Unified audio generation with natural language prompts," *arXiv preprint arXiv:2312.15821*, 2023.

[7] S. Arik, J. Chen, K. Peng, W. Ping, and Y. Zhou, "Neural voice cloning with a few samples," *Advances in neural information processing systems*, vol. 31, 2018.

[8] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *International conference on machine learning*. PMLR, 2018, pp. 5180–5189.

[9] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler *et al.*, "Emergent abilities of large language models," *Transactions on Machine Learning Research*, 2022.

[10] T. Webb, K. J. Holyoak, and H. Lu, "Emergent analogical reasoning in large language models," *arXiv preprint arXiv:2212.09196*, 2022.

[11] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi, "Metaicl: Learning to learn in context," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 2791–2809.

[12] M. Chen, J. Du, R. Pasunuru, T. Mihaylov, S. Iyer, V. Stoyanov, and Z. Kozareva, "Improving in-context few-shot learning via self-supervised training," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 3558–3573.

[13] Y. Gu, L. Dong, F. Wei, and M. Huang, "Pre-training to learn in context," in *ACL (1)*. Association for Computational Linguistics, 2023, pp. 4849–4870.

[14] T. Sun, Y. Shao, H. Qian, X. Huang, and X. Qiu, "Black-box tuning for language-model-as-a-service," in *International Conference on Machine Learning*. PMLR, 2022, pp. 20 841–20 855.

[15] OpenAI, "Introducing ChatGPT," 2022, https://openai.com/blog/chatgpt.

[16] A. Polyak, Y. Adi, J. Copet, E. Kharitonov, K. Lakhotia, W. Hsu, A. Mohamed, and E. Dupoux, "Speech resynthesis from discrete disentangled self-supervised representations," in *Interspeech*. ISCA, 2021, pp. 3615–3619.

[17] K. Lakhotia *et al.*, "On generative spoken language modeling from raw audio," *Transactions of the Association for Computational Linguistics*, vol. 9, pp. 1336–1354, 2021.

[18] E. Kharitonov *et al.*, "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)*, 2022, pp. 8666–8681.

[19] Z. Borsos *et al.*, "Audiolm: a language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[20] M. Hassid *et al.*, "Textually pretrained speech language models," *arXiv preprint arXiv:2305.13009*, 2023.

[21] J. Wei, J. Wei, Y. Tay, D. Tran, A. Webson, Y. Lu, X. Chen, H. Liu, D. Huang, D. Zhou *et al.*, "Larger language models do in-context learning differently," *arXiv preprint arXiv:2303.03846*, 2023.

[22] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" 2022.

[23] D. Dai, Y. Sun, L. Dong, Y. Hao, S. Ma, Z. Sui, and F. Wei, "Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers," in *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[24] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[25] K.-W. Chang, W.-C. Tseng, S.-W. Li, and H. yi Lee, "An Exploration of Prompt Tuning on Generative Spoken Language Model for Speech Processing Tasks," in *Proc. Interspeech 2022*, 2022, pp. 5005–5009.

[26] L. T. Benamer and O. A. Alkishriwo, "Database for arabic speech commands recognition," in *CEST*, 2020.

[27] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[28] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," 2018.

[29] A. Kolesau and D. Šešok, "Unsupervised pre-training for voice activation," *Applied Sciences*, vol. 10, no. 23, p. 8643, 2020.

[30] Y.-Y. Lin, W.-Z. Zheng, W. C. Chu, J.-Y. Han, Y.-H. Hung, G.-M. Ho, C.-Y. Chang, and Y.-H. Lai, "A speech command control-based recognition system for dysarthric patients based on deep learning technology," *Applied Sciences*, 2021.

[31] K. MacLean, "Voxforge," available: http://www.voxforge.org/home.

[32] S. Castro, D. Hazarika, V. Pérez-Rosas, R. Zimmermann, R. Mihalcea, and S. Poria, "Towards multimodal sarcasm detection (an _obviously_ perfect paper)," in *Proceedings of the 57th Conference of ACL*, 2019, pp. 4619–4629.

[33] K.-W. Chang, M.-H. Chen, Y.-P. Lin, J. N. Hsu, P. K.-M. Huang, C.-y. Huang, S.-W. Li, and H.-y. Lee, "Prompting and adapter tuning for self-supervised encoder-decoder speech model," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.

[34] A. Chen, Y. Yao, P.-Y. Chen, Y. Zhang, and S. Liu, "Understanding and improving visual prompting: A label-mapping perspective," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 133–19 143.

[35] Y.-Y. Tsai, P.-Y. Chen, and T.-Y. Ho, "Transfer learning without knowing: Reprogramming black-box machine learning models with scarce data and limited resources," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9614–9624.

[36] K.-W. Chang, Y.-K. Wang, H. Shen, I. thing Kang, W.-C. Tseng, S.-W. Li, and H. yi Lee, "SpeechPrompt v2: Prompt tuning for speech classification tasks," 2023.

[37] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.