Interspeech 2024
1-5 September 2024, Kos, Greece

Online Knowledge Distillation of Decoder-Only Large Language Models for Efficient Speech Recognition

Jeehye Lee, Hyeji Seo**

Kakao Corporation, Sungnam, South Korea
jessie.e@kakaocorp.com, heize.v@kakaocorp.com

Abstract

Large language models (LLMs), which show promising performance in generation tasks, have proven their capabilities to be applied in a wide range of tasks. Although there are several approaches to adapt LLMs as decoder in speech recognition tasks, these can slow down inference speed, which is an important issue for the product-level systems. To address this problem, we introduce online knowledge distillation methods to transfer information from the decoder-only LLMs to a more compact Transformer decoder during the training phase. Implementing our proposed methods on a multilingual low-resource dataset, we achieved a 8.2% relative character error rate (CER) reduction compared to the LLM decoder model with much lower inference cost and a 34.7% relative CER reduction compared to the attention-based encoder-decoder (AED) model. Furthermore, we obtained a 14.9% relative CER reduction along with the same inference cost on a general Korean dataset.

Index Terms: knowledge distillation, speech recognition, large language model

1. Introduction

The recent research in the automatic speech recognition (ASR) field is focused on end-to-end (E2E) frameworks [1, 2, 3, 4, 5, 6] with massive resources. However, acquiring such large-scale speech-text paired datasets is challenging and training massive models requires significant time and cost. To enhance the performance in E2E frameworks, pre-trained models which are trained on large-scale unlabeled speech data with large model sizes are employed [7, 8, 9, 10]. Also, language models (LMs) trained on large amounts of text data are utilized to improve speech recognition performance [11, 12, 13].

The large language models (LLMs), which predict next token when given a sequence of text, have demonstrated notable accomplishments across various natural language tasks such as question answering, knowledge retrieval, text generation and summarization [14, 15]. The capabilities of LLMs are driven by massive training datasets and a vast number of model parameters. Consequently, extensive research has been conducted on leveraging pre-trained LLMs. There are several research aim to extend LLMs to other modalities [16, 17] and methods for aligning embeddings from other modalities with the input space of LLMs have been explored. Furthermore, the works by [18, 19] reveal that audio embeddings align closely with text tokens, performing multilingual speech recognition. Using decoder-only LLMs as a decoder in E2E ASR framework shows promising results in multilingual ASR task [18, 19]. However, high inference costs caused by LLMs are both inevitable and critical. To integrate LLMs into a product-level ASR system, a reasonable inference cost is required.

The knowledge distillation (KD) method, which transfers knowledge from a large teacher model into a small student model, is commonly used for model compression in ASR [20, 21, 22] and has also been applied to the training of small generative language models [23]. In previous studies on ASR tasks, KD approaches have typically focused on using encoders [20, 22, 24] to learn hidden representations of audio features. However, there has been limited exploration into applying KD techniques to decoders, which are responsible for text sequence generation. Therefore, transferring the capabilities of a decoder-only model, which is trained on a large multilingual text dataset, into the decoder of the E2E ASR model, could be effective in improving performance.

However, the efficiency of the vanilla KD methods may be limited when there is a significant disparity between the teacher model and the student model [25]. The teacher model in our experiments is the LLM, which is trained with text embeddings, whereas the student model is a small Transformer decoder with audio embeddings. In order to deal with the large differences in modalities and model size between the two models, we adopt the online KD approach [21, 26] instead of the commonly used KD approach. By jointly training the teacher model and the student model, we can effectively train the student model for the target task of E2E ASR, leveraging the experience of the teacher model trained on various text-related tasks.

In this work, to address high inference costs problem and the disadvantage of the vanilla KD approach, we propose a method to integrate the knowledge of LLMs into a small Transformer decoder through a parameter-efficient fine-tuning process, called Low-Rank Adaptation (LoRA) [27]. Also, the features extracted from the middle of the LLM layers are utilized as distillation targets of the Transformer decoder.

To validate our proposed method, we conducted experiments in various language environments using different foundation models and encoders. After conducting experimental evaluations, we determined that the online collaborative knowledge distillation model outperformed the baseline in both multilingual and Korean environments, maintaining the same inference cost. The selection of the distillation targets led to additional improvement.

2. Related work

2.1. Multi-task learning

The key idea of multi-task learning (MTL) is to train the encoder with multiple objectives simultaneously to improve robustness [3, 4]. In connectionist temporal classification (CTC)

* : equal contribution

10.21437/Interspeech.2024-163

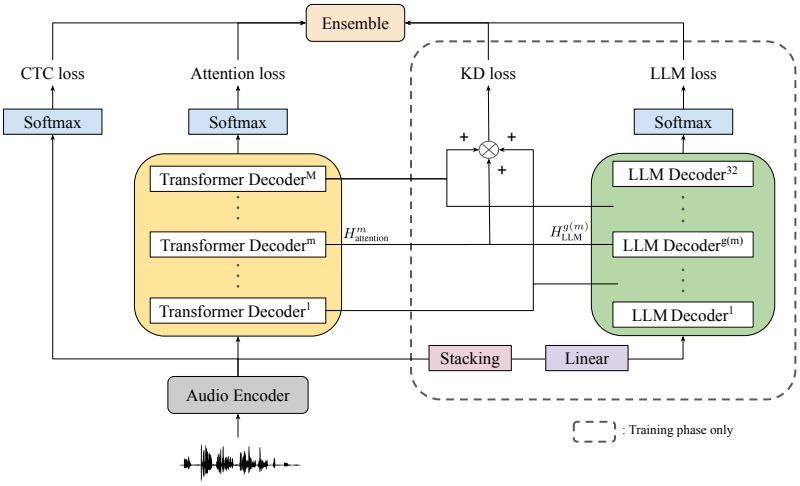

Figure 1: *The proposed model architecture with knowledge distillation methods. The dashed block is used only during the training phase. The audio encoder is shared by all the objectives.*

and attention MTL, CTC is used as an auxiliary task to train an encoder of the attention-based model. The loss function of the attention-based encoder-decoder (AED) model is defined as the weighted sum of the CTC loss and the attention loss as:

$$\mathcal{L}_{\text{AED}} = \alpha\mathcal{L}_{\text{CTC}} + (1-\alpha)\mathcal{L}_{\text{attention}}, \tag{1}$$

with a tunable parameter $\alpha : 0 \leq \alpha \leq 1$. In the LLM decoder model, the attention loss is replaced with the LLM loss [18].

2.2. Knowledge distillation

The KD [20, 22, 24, 28] is a commonly employed transfer learning method that condenses a large teacher model into a smaller student model through the process of training the smaller model to replicate the results of the larger model. The pre-trained large teacher model is frozen during training of a student model in vanilla KD. The teacher model contains several types of knowledge, which can be extracted from the last layer (response-level) [20, 29] and intermediate layers (feature-level) [24, 28]. In the response-level KD (RKD), Kullback-Leibler (KL) divergence is typically used as a distillation loss to align probability distributions:

$$\mathcal{L}_{\text{RKD}} = \tau^2 \text{KL}(\text{softmax}(\frac{p_t}{\tau}), \text{softmax}(\frac{p_s}{\tau})) \tag{2}$$

where τ is the temperature; p_t and p_s represent the logits produced by the teacher model and the student model, respectively.

There are several studies aimed at transferring knowledge from intermediate layers to guide the student network to learn not only the features of the teacher network's final layer but also those of intermediate layers [24, 28]. In the feature-level KD (FKD), the knowledge from the hidden states and attention maps of the Transformer layer can be distilled with mean-squared error (MSE):

$$\mathcal{L}_{\text{FKD}} = \text{MSE}(H_t, H_s W), \tag{3}$$

where H_t and H_s are the hidden states of the teacher and the student networks, respectively; W is the matrix used to map the hidden dimension of the student network to that of the teacher network.

3. Online collaborative knowledge distillation

In this section, we explain our online KD approach, which utilizes multi-task (MT) loss and KD loss collaboratively with the LLM in an end-to-end manner, as shown in Figure 1.

3.1. Collaborative multi-task learning

In conventional MTL, various objective functions are combined to enhance robustness, however we employ MT loss for online KD. The model is trained with a MT loss that incorporates CTC, attention, and LLM losses, while we simultaneously train the small Transformer decoder and the LLM.

Before being fed into the LLM decoder, the outputs of an audio encoder are stacked to reduce computational load and then projected to align with the dimensions of the LLM decoder. Using LLM as an auxiliary task helps to estimate the desired alignments in long sequences, with information being encompassed within the LLM. The proposed objective of the collaborative multi-task (CMT) model is represented as follows:

$$\mathcal{L}_{\text{CMT}} = \mathcal{L}_{\text{AED}} + \beta\mathcal{L}_{\text{LLM}}, \tag{4}$$

with a tunable parameter β: $0 \leq \beta \leq 1$. In other words, we facilitate the transmission of information from the LLM to the Transformer decoder through online distillation, while simultaneously aligning both networks with the ground truth labels. The output labels of both networks correspond to the tokenizer vocabulary of the LLM.

3.2. Decoder-specific knowledge distillation

To transfer further knowledge from the LLM to the small Transformer decoder, the distillation loss is employed to extract insights obtained from training on a large dataset. In the FKD, we apply an extended version of the loss presented in Equation 3, where the teacher model is the LLM and the student model is the Transformer decoder. The number of intermediate layers M among the Transformer decoder and intermediate layers mapping strategy $g(m)$ need to be defined, which denotes that the m-th layer of the student model learns information from

Table 1: *Language specific CER performance on the FLEURS dataset*

	we	ee	cmn	ssa	sa	sea	cjk
#languages	25	16	12	20	14	11	4
Baseline	9.2	7.0	8.6	12.4	10.3	10.9	15.6
LLM	5.8	4.3	5.7	10.4	6.8	9.1	10.5
CMT	5.7	4.3	5.8	9.4	7.0	8.1	11.9
CKD	4.9	3.8	5.3	9.2	7.3	8.7	11.8
CMT+KD	4.7	3.6	5.0	9.0	6.6	8.8	11.2

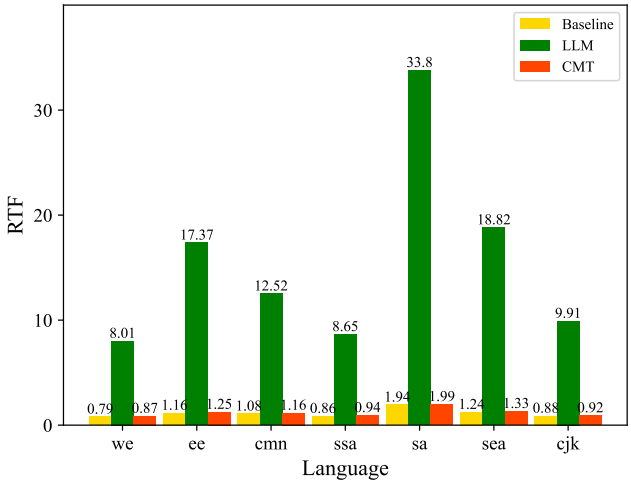

Figure 2: *Comparison of RTF across languages in the FLEURS dataset.*

the $g(m)$-th layer of the teacher model. We have adopted three different mapping strategies to extract only lower local feature, only higher semantic feature, or both uniformly from the LLM. The final FKD loss is defined as:

$$\mathcal{L}_{\text{FKD}} = \sum_{m=1}^{M} \text{MSE}(H_{\text{LLM}}^{g(m)}, H_{\text{attention}}^{m} W), \tag{5}$$

where $H_{\text{LLM}}^{g(m)} \in \mathbb{R}^{l \times d}$ and $H_{\text{attention}}^{m} \in \mathbb{R}^{l \times d'}$ are the hidden state distributions of the intermediate layer in the teacher network and the student network; W is the weight matrix that transforms the hidden dimension of the Transformer decoder d' into the LLM hidden dimension d, in our case $d = 4096$ and $d' = 2048$. The scalar value l refers to the input sequence length.

In the CKD model, we only apply FKD loss instead of LLM loss in Equation 4 as follows:

$$\mathcal{L}_{\text{CKD}} = \mathcal{L}_{\text{AED}} + \gamma \mathcal{L}_{\text{FKD}}, \tag{6}$$

where the weight hyper-parameter γ is set empirically to match the scale of the attention loss. The final loss of the online collaborative knowledge distillation (CMT+KD) model is a weighted sum of the CMT loss and the FKD loss, as follows:

$$\mathcal{L}_{\text{CMT+KD}} = \mathcal{L}_{\text{CMT}} + \gamma \mathcal{L}_{\text{FKD}}. \tag{7}$$

4. Experiments and results

4.1. Experimental setup

In our research, we utilized FLEURS dataset [30], which is comprised of multilingual sentences from 102 languages, with approximately 7-10 hours of training data available for each language. Due to the limited amount of data for each language, this dataset is suitable for illustrating enhancements in multilingual speech recognition performance for low-resource languages through the application of KD in LLMs. Additionally, we employed the Ksponspeech dataset, which has 969 hours of training data, in order to verify the efficacy of our proposed methods in the context of Korean speech recognition.

Table 2: *Performance Evaluation of CER and RTF on the Ksponspeech dataset*

Model	CER		RTF	
	eval-clean	eval-other	eval-clean	eval-other
Baseline	8.2	9.3	0.24	0.27
LLM	7.4	8.1	1.69	1.91
CMT	8.0	9.0	0.26	0.28
CKD	7.3	8.0	0.26	0.28
CMT+KD	7.2	8.0	0.26	0.28

The implementation of the model was based on the ESPNet toolkit [31]. We used globally normalized 80-dimensional log Mel-filter bank coefficients as input features. The input features were computed with a window of 25 ms, shifted every 10 ms. To derive rich acoustic representations, the shared audio encoder network was initialized from a wav2vec-BERT 2.0 model (w2v-BERT) for a multilingual setting, which has 635M parameters and has been pre-trained on 4.5M hours of unlabeled speech [10]. In the Korean experiment, the encoder from the Whisper large-v2 model [9] was employed as an audio encoder. This model utilizes a larger amount of Korean training data compared to w2v-BERT and is used to extract more optimized features for Korean. This method allows us to validate the efficacy of our techniques across various foundational models and languages. To reduce the length of acoustic embedding sequence, the 9 consecutive outputs of the encoder were stacked and then projected to 4096-dimensional embeddings, which match the dimensions of the pre-trained decoder-only LLM [18].

The LLaMA-2-7B [14] was used as the LLM and 32000 tokens derived from the LLaMA-2-7B's tokenizer were employed as output labels. For experiments conducted in Korean, the LLaMA-2-7B was replaced with a Korean LLM-7B, which encompasses LLaMA-2-7B tokens and additional Korean tokens. The LLaMA-2-7B and Korean LLM-7B only differ in the training data and the vocabulary size of the tokenizer (i.e. based on the same architecture).

During training, both LLMs were parameter efficiently tuned with LoRA. The Transformer decoder is composed of 6 Transformer layers, each with 8 attention heads and 2048 hidden units. Additional model specifications and training strategies can be found in [18]. The hyper-parameters were set to $M = 6$, $\tau = 1$, $\alpha = 0.3$, $\beta = 0.7$, and $\gamma = 10$, as determined by our validation set. In the inference phase, the LLM was removed and a beam search described in [4] with a beam size of 10 was utilized. The real-time factor (RTF) and character error rate (CER) were used as evaluation metrics and calculated with an NVIDIA A100 GPU.

4.2. Experimental results and discussion

Table 1 compares the CER on the FLEURS dataset using five types of training methods: the baseline (the conventional AED), the LLM decoder model, and our proposed CMT, CKD, CMT+KD models. The proposed CMT model significantly re-

Table 3: *Comparison of CER on FLEURS dataset and Ksponspeech dataset with the layer mapping strategies: uniform distributed selection (uniform), upper-layer focused selection (upper), and lower-layer focused selection (lower)*

Model	FLEURS							Ksponspeech	
	we	ee	cmn	ssa	sa	sea	cjk	eval-clean	eval-other
Baseline	9.2	7.0	8.6	12.4	10.3	10.9	15.6	8.2	9.3
CMT+KD (uniform)	4.7	3.6	5.0	9.0	6.6	8.8	11.2	7.2	8.0
CMT+KD (upper)	4.7	3.6	5.0	9.0	7.0	8.3	10.9	7.1	7.9
CMT+KD (lower)	4.7	3.6	5.0	9.0	7.0	8.1	10.9	7.0	7.9

Table 4: *Decoding results on English data from FLEURS*

Reference	SEGREGATION AND RECOMBINATION SHUFFLE VARIATION BACK AND FORTH BETWEEN THE TWO POOLS WITH EACH GENERATION
Baseline	**CIGERGATION HEALTH** RECOMBINATION SHUFFLE VARIATION BACK AND **FOR** BETWEEN THE TWO POLES WITH EACH GENERATION
LLM	**SEGREGATION** AND RECOMBINATION SHUFFLE VARIATION BACK AND **FORTH** BETWEEN THE TWO POLES WITH EACH GENERATION
CMT+KD	**SEGREGATION** AND RECOMBINATION SHUFFLE VARIATION BACK AND **FORTH** BETWEEN THE TWO POLES WITH EACH GENERATION
Reference	IT IS THINNER UNDER THE MARIA AND THICKER UNDER THE HIGHLANDS
Baseline	IT IS **SINGER** UNDER THE MORIA AND **SICKER** UNDER THE **HISLANDS**
LLM	IT IS **THINNER** UNDER THE MRIA AND **THICKER** UNDER THE **HIGHLANDS**
CMT+KD	IT IS **THINNER** UNDER THE MARYA AND **THICKER** UNDER THE **HIGHLANDS**

duced the CER by 29.5% relative to the baseline and showed comparable performance with the LLM decoder model, even while excluding the utilization of LLM during the inference phase. This implies that our proposed method efficiently facilitates contextual knowledge about multiple languages acquired from the LLM decoder to the Transformer decoder. Also, it successfully reduces the RTF by approximately 92.2% without any degradation in accuracy as shown in Figure 2.

Further, we investigate the effects of the proposed methods on another language with the different encoder and LLM. Table 2 compares the proposed methods with the baseline and LLM decoder model on the Ksponspeech dataset. The LLM decoder model exhibited a relative CER reduction (CERR) of 9.8% on the eval-clean set and 12.9% on the eval-other set compared to the baseline, but this improvement corresponded with an increase in RTF by a factor of 7. Therefore, we employed an online distillation method to enhance the inference speed while maintaining this performance gain. However, the CMT model only achieved a 2.4% and 3.2% CERR on the eval-clean and eval-other sets, respectively, compared to the baseline. This suggests that when the LLM is already specialized in the language, achieving equivalent performance to the LLM decoder model solely by applying the CMT loss becomes challenging.

Aiming to further improve performance, we applied the FKD loss with the uniformly distributed layer mapping strategy to extract features from both lower and upper layers of the LLM. The CKD model demonstrated enhanced performance across all test cases compared to the baseline and achieved results better than those of the LLM decoder model without any decrease in speed caused by LLM, as shown in Table 1 and Table 2. It achieved an average 31.1% and 12.6% CERR on the FLEURS and the Ksponspeech datasets, respectively, compared to the baseline. It means that FKD of hidden states is effective on both low-resource multilingual and general monolingual environments. Also, the integration of the FKD loss on the CMT loss achieved an average 4.1% additional CERR compared to the CKD model on the FLEURS dataset. As a result, the CMT+KD model showed the lowest CER among all the models, while significantly reducing the inference time when compared to the LLM decoder model. In Korean experiments, the CMT+KD model also exhibited the best performance, showing a relative 12.2% and 14.0% decrease in a CER on eval-clean and eval-other sets compared to the baseline. These results imply that online KD, which performs simultaneous updates on teacher and student models, effectively aligns the audio encoder output to the LLM input embeddings and distills LLM knowledge into the Transformer decoder. Moreover, integrating the distillation loss into the CMT model efficiently minimizes the gap between the LLM and the Transformer decoder.

We also examine three distinct layer mapping strategies on the CMT+KD model: uniform distributed selection (uniform), upper-layer focused selection (upper), and lower-layer focused selection (lower), as demonstrated in Table 3. In the uniform strategy, 6 layers are selected with uniform intervals, where the upper strategy selects the last 6 layers and the lower strategy selects the first 6 layers. The lower strategy shows a further CER reduction than the uniform and the upper strategies, which indicates that LLM exhibits sufficient logical reasoning, multilingual, and recognitive abilities for ASR even in its lower layers. The CMT+KD model with the lower strategy achieved an average 34.7% and 14.9% CERR compared to the baseline on the FLEURS and the Ksponspeech datasets, respectively.

In Table 4, we compare decoding results across distinct models. Several word errors are occurred in the baseline model results, making it difficult to understand the sentence's meaning. Some of these errors are resolved using the linguistic information from the LLM. As shown in the results of the proposed CMT+KD model, critical errors from the baseline model are alleviated with the online KD from the LLM.

5. Conclusion

We proposed online KD methods using a decoder-only LLM for the AED model to improve speech recognition performance without increasing the inference cost. By jointly training the Transformer decoder model with the LLM decoder model with the online collaborative knowledge distillation process, our proposed model outperformed the baseline with a CERR of 33.9% on multilingual low-resource dataset and 13.1% on Korean dataset. Further, we found that lower-level features of the LLM have enough information for the ASR and it showed slightly better performance of the uniform strategy model with a relative CER of 1.2% and 2.0% on the multilingual low-resource dataset and Korean dataset, respectively.

6. References

[1] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. ICASSP*, 2016, pp. 4960–4964.

[2] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-Based Models for Speech Recognition," in *Proc. NIPS*, vol. 28, 2015, pp. 577–585.

[3] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*, 2017, pp. 4835–4839.

[4] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006, p. 369–376.

[6] J. Kim and J. Lee, "Generalizing RNN-Transducer to Out-Domain Audio via Sparse Self-Attention Layers," in *Proc. Interspeech*, 2022, pp. 4123–4127.

[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. NIPS*, vol. 33, 2020, pp. 12 449–12 460.

[8] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu *et al.*, "WavLM: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, oct 2022.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[10] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman *et al.*, "Seamlessm4t-massively multilingual & multimodal machine translation," *arXiv preprint arXiv:2308.11596*, 2023.

[11] Y. Bai, J. Yi, J. Tao, Z. Tian, and Z. Wen, "Learn spelling from teachers: Transferring knowledge from language models to sequence-to-sequence speech recognition," in *Proc. Interspeech*, 2019, pp. 3795–3799.

[12] K. Choi and H.-M. Park, "Distilling a pretrained language model to a multilingual asr model," in *Proc. Interspeech*, 2022, pp. 2203–2207.

[13] Y. Kubo, S. Karita, and M. Bacchiani, "Knowledge transfer from large-scale pretrained language models to end-to-end speech recognizers," in *Proc. ICASSP*, 2022, pp. 8512–8516.

[14] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[15] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[16] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[17] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.

[18] Y. Fathullah, C. Wu, E. Lakomkin, J. Jia, Y. Shangguan, K. Li, J. Guo, W. Xiong, J. Mahadeokar, O. Kalinli *et al.*, "Prompting large language models with speech recognition abilities," in *Proc. ICASSP*, 2024, pp. 13 351–13 355.

[19] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu *et al.*, "On decoder-only architecture for speech-to-text and large language model integration," in *Proc. ASRU*, 2023, pp. 1–8.

[20] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] Q. Guo, X. Wang, Y. Wu, Z. Yu, D. Liang, X. Hu, and P. Luo, "Online knowledge distillation via collaborative learning," in *Proc. CVPR*, 2020, pp. 11 017–11 026.

[22] S. Tian, K. Deng, Z. Li, L. Ye, G. Cheng, T. Li, and Y. Yan, "Knowledge distillation for ctc-based speech recognition via consistent acoustic representation learning," in *Proc. Interspeech*, 2022, pp. 2633–2637.

[23] Y. Gu, L. Dong, F. Wei, and M. Huang, "MiniLLM: Knowledge distillation of large language models," in *Proc. ICLR*, 2024.

[24] M. Hentschel, Y. Nishikawa, T. Komatsu, and Y. Fujita, "Keep decoding parallel with effective knowledge distillation from language models to end-to-end speech recognisers," in *Proc. ICASSP*, 2024, pp. 10 876–10 880.

[25] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *Proc. AAAI*, vol. 34, no. 04, 2020, pp. 5191–5198.

[26] D. Chen, J.-P. Mei, C. Wang, Y. Feng, and C. Chen, "Online knowledge distillation with diverse peers," in *Proc. AAAI*, vol. 34, no. 04, 2020, pp. 3430–3437.

[27] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, 2022.

[28] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu, "Tinybert: Distilling bert for natural language understanding," in *Proc. EMNLP*, 2020, pp. 4163–4174.

[29] H. Futami, H. Inaguma, S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, "Distilling the Knowledge of BERT for Sequence-to-Sequence ASR," in *Proc. Interspeech*, 2020, pp. 3635–3639.

[30] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *Proc. SLT*, 2023, pp. 798–805.

[31] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba *et al.*, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.