pISSN 2005-8063
eISSN 2586-5854
2023. 9. 30.
Vol.15 No.3
pp. 83-88

말소리와 음성과학
Phonetics and Speech Sciences

한국음성학회지

https://doi.org/10.13064/KSSS.2023.15.3.083

Exploring the feasibility of fine-tuning large-scale speech recognition models for domain-specific applications: A case study on Whisper model and KsponSpeech dataset

Jungwon Chang · Hosung Nam*

Department of English Language and Literature, Korea University, Seoul, Korea

Abstract

This study investigates the fine-tuning of large-scale Automatic Speech Recognition (ASR) models, specifically OpenAI's Whisper model, for domain-specific applications using the KsponSpeech dataset. The primary research questions address the effectiveness of targeted lexical item emphasis during fine-tuning, its impact on domain-specific performance, and whether the fine-tuned model can maintain generalization capabilities across different languages and environments. Experiments were conducted using two fine-tuning datasets: Set A, a small subset emphasizing specific lexical items, and Set B, consisting of the entire KsponSpeech dataset. Results showed that fine-tuning with targeted lexical items increased recognition accuracy and improved domain-specific performance, with generalization capabilities maintained when fine-tuned with a smaller dataset. For noisier environments, a trade-off between specificity and generalization capabilities was observed. This study highlights the potential of fine-tuning using minimal domain-specific data to achieve satisfactory results, emphasizing the importance of balancing specialization and generalization for ASR models. Future research could explore different fine-tuning strategies and novel technologies such as prompting to further enhance large-scale ASR models' domain-specific performance.

Keywords: automatic speech recognition, deep learning, Transformers, Whisper

1. Introduction

The advancement of Transformer architecture and Self-Attention (Vaswani et al., 2017) has resulted in a substantial acceleration in the development of end-to-end (E2E) Automatic Speech Recognition (ASR) systems. Moreover, novel architectures such as the Conformer (Gulati et al., 2020) or E-Branchformer (Kim et al., 2023) encoder architecture have surfaced, which has increased the performance of various speech applications including ASR (Guo et al., 2021; Peng et al., 2023a). Another noteworthy approach in ASR via Transformer architecture is self-supervised learning (SSL). A prime example is Wav2Vec 2.0 (Baevski et al., 2020), a model that shares similarities with the prominent BERT model (Devlin et al., 2018) used in natural language processing. By incorporating a Transformer encoder,

* hnam@korea.ac.kr, Corresponding author

Received 15 August 2023; Revised 15 September 2023; Accepted 15 September 2023

Wav2Vec 2.0 and other SSL models can learn speech representations directly from raw audio signals, thereby increasing its efficiency and effectiveness in ASR tasks (Mohamed et al., 2022). This direct training approach contrasts with conventional methods that typically require pre-processing or feature extraction steps. SSL has become another major trend of E2E ASR and has evolved into newer models, such as HuBERT (Hsu et al., 2021) and Wavlm (Chen et al., 2022).

Recently, the increasing trend of utilizing massive models and extensive datasets for E2E speech recognition systems has become a focal point in the industry, led by prominent tech companies. Examples of such models include OpenAI's "Whisper," (Radford et al., 2023) Google's "Universal Speech Model," (Zhang et al., 2023) and Meta AI's "Massively Multilingual Speech Model" (Pratap et al., 2023). These state-of-the-art large-scale models have demonstrated remarkable generalization capabilities and have led to significant advancements in speech recognition technology. However, as these models are designed to perform well across various domains, their effectiveness in specific applications warrants further investigation.

Ever since the Whisper model was made publicly available, various research studies have been conducted. Some attempted speech recognition through additional training for languages not supported by the model (Rouditchenko et al., 2023), while others utilized special tokens to extend the model's capabilities beyond existing speech recognition and translation functions (Peng et al., 2023b). Efforts were made to fine-tune the model to specific speech corpora by various communities, but no systematic research has been conducted to the best of our knowledge. Moreover, there have been no studies attempting to overcome the challenges related to specific lexical items that the Whisper model is unable to recognize, as presented in our research.

In this study, we focus on the fine-tuning of the Whisper model for enhanced recognition of additional vocabulary in a domain-specific context. The research was inspired by the observation that, when decoding the evaluation set of KsponSpeech (Bang et al., 2020), a Korean spontaneous speech dataset, using the Whisper model, consistent failures in recognizing specific nouns were observed. This observation led to a series of research questions that we aim to address throughout this paper:

1. Does fine-tuning the model with only a few utterances with the targeted lexical items enable the recognition of the items?
2. How does the fine-tuning process affect the model's performance on the targeted domain speech compared to its pre-tuned state?
3. Is the fine-tuned model able to maintain the original model's generalization capabilities across different languages and environments?

To address these questions, we conducted a series of experiments by varying the amount of data for fine-tuning process and comparing the performance of the fine-tuned model to its pre-tuned baseline. Through our analysis, we provide a comprehensive understanding of the fine-tuning process, its effectiveness, and the limitations encountered when adapting a large-scale speech recognition model to a specific domain.

2. Experiment

2.1. Whisper Model

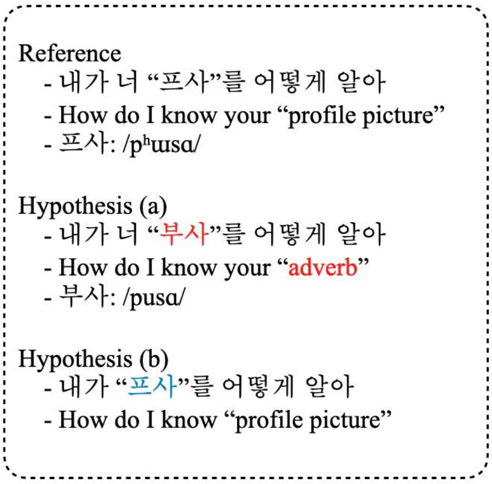
The Whisper model, developed by OpenAI, is a versatile speech model, and the architecture of the model is a standard Transformer based encoder-decoder architecture (Vaswani et al., 2017). The model is trained through multitask learning using 680,000 hours of speech data prepared through weak supervision (Radford et al., 2023). The multitask learning here means training various speech-related tasks at once, which are: speech recognition, speech translation, speech detection, and language identification for 96 languages. The model is divided into tiny, base, small, medium, and large sizes, with the large model being further improved by changes in training techniques and the subsequent release of the "large-v2" model. The performance of multilingual speech recognition and "to English" translation was comparable to the existing state-of-the-art ASR or speech translation models. This paper focuses on the speech recognition feature of the Whisper model, and the large-v2 model was selected for fine-tuning.

2.2. Datasets

2.2.1. Ksponspeech dataset

KsponSpeech is a large-scale spontaneous speech corpus of Korean conversations, which contains 969 hours of general open-domain dialog utterances, spoken by about 2,000 native Korean speakers in a clean environment (Bang et al., 2020). The selection of KsponSpeech in this study was motivated by the belief that the Whisper model may be vulnerable to certain lexical items present in the daily conversations between two speakers, as the adequate recognition of specific vocabulary is a crucial component of the domain-specific fine-tuning process.

However, commonly used evaluation metrics in speech recognition such as character error rate (CER) or word error rate (WER) struggle to accurately assess the recognition of specific lexical items. In Figure 1, the meaning of two speech recognition outputs, Hypothesis (a) and Hypothesis (b), drastically changes depending on whether the lexical item "프사" (profile picture) is correctly decoded. In case of (a), "프사" is incorrectly decoded as "부사" (adverb), which phonetically resembles "프사", thus completely altering the meaning. On the other hand, in case of (b), it correctly recognizes "프사", allowing for a sufficient understanding of the intended meaning through its output. However, when measured with CER or WER, both hypotheses yield identical results. When measuring CER, Hypothesis (a) substitutes the character "프" with "부", while Hypothesis (b) deletes the character "너". Since both recognition results involve only one instance of substitution or deletion respectively, they show an identical CER value of 9.1%. Similarly for WER measurement, Hypothesis (a) substitutes the word "프사" with "부사", and Hypothesis (b) deletes the word "너". As each has only one instance of substitution or deletion respectively, they also show an identical WER value of 20%.



Reference
- 내가 너 “프사”를 어떻게 알아
- How do I know your “profile picture”
- 프사: /pʰɯsɑ/

Hypothesis (a)
- 내가 너 “부사”를 어떻게 알아
- How do I know your “adverb”
- 부사: /pusɑ/

Hypothesis (b)
- 내가 “프사”를 어떻게 알아
- How do I know “profile picture”

CER, character error rate; WER, word error rate.

Figure 1. Cases with identical CER and WER but different meanings due to lexical item recognition.

Due to the limitations of commonly used evaluation metrics in speech recognition, such as CER and WER, this study devised a new metric called “recognition accuracy for targeted lexical item” to determine whether a specific lexical item has been recognized. When the reference sentence includes a targeted lexical item, we examine the decoding result: if it correctly recognizes the lexical item, we classify it as a correct sentence; if not, it is classified as an incorrect sentence. Therefore, Hypothesis (a) in Figure 1 is treated as an incorrect sentence while Hypothesis (b) is considered correct. The ratio of correctly decoded sentences out of all sentences containing the targeted lexical item was calculated and converted into percentage terms. This metric was devised because correctly recognizing specific lexical item during domain fine-tuning is extremely important for semantic understanding.

The recognition accuracy was considerably low for specific lexical items, irrespective of the model size, as in Table 1. The recognition accuracy generally increases as the size of the model increases. However, we can see all the models are struggling to recognize the lexical items provided at Table 1. Nonetheless, as evidenced by the accuracy figures presented in Table 1, it became apparent that all the models demonstrated vulnerability concerning the specific lexical items. Among these items, some were abbreviations (페북 for Facebook or 페이스북, 프사 for profile picture or 프로필 사진), others were proper nouns referring to companies or organizations (페북 for Facebook, 진에어 for Jin Air), or words that could only be comprehended when the context between conversational partners was known (보증금 and 계절학기).

Table 1. Whisper’s recognition accuracy for targeted lexical items

Lexical item	Size of the Whisper model					
	tiny	base	small	medium	large-v1	large-v2
진에어	0.0	0.0	0.0	12.5	12.5	12.5
페북	0.0	0.0	33.3	0.0	33.3	66.7
컴활	0.0	0.0	0.0	0.0	0.0	16.7
프사	33.3	0.0	0.0	33.3	33.3	33.3
보증금	0.0	0.0	33.3	33.3	33.3	33.3
계절학기	0.0	0.0	0.0	0.0	0.0	0.0

Additionally, the transcriptions of KsponSpeech exhibit frequent code-switching between Korean, English, and numerical values, underscoring the potential importance of incorporating this transcription style in the fine-tuning procedure. The pre-tuned state of the Whisper large-v2 model demonstrated a WER of 29.05% and a CER of 13.95% on the KsponSpeech eval set.

2.2.2. LibriSpeech dataset

The LibriSpeech corpus (Panayotov et al., 2015) is among the most widely utilized ASR corpora, consisting of 1,000 hours of English read speech and transcriptions. Notably, the CER values for the test-clean set and the noise-infused test-other set serve as a standard performance metric for ASR models. In this study, the LibriSpeech corpus was employed to assess the extent to which the fine-tuned model can maintain its inherent generalization ability. The pre-tuned state of the Whisper large-v2 model yielded a CER of 1.77% on the LibriSpeech test-clean set, and a CER of 2.86% on the test-other set.

2.3. Fine-Tuning Setup

2.3.1. Fine-tuning data

Two fine-tuning datasets were prepared. First, a sub-set of sentences extracted from the KsponSpeech corpus training set containing six lexical items with low recognition accuracy, as presented in Table 1. This sub-set of 895 sentences, and less than 2 hours of utterances was designed to investigate whether emphasizing a dataset with specific lexical items could improve their recognition accuracy. This fine-tuning dataset will be noted as “Set A”, and the fine-tuned model will be noted as “Model A”. The second dataset consists of the entirety of the KsponSpeech training set, spanning 969 hours and 620,000 sentences, and was devised to examine whether the recognition accuracy of the selected lexical items would increase even with the minimal proportion of unrecognized lexical items. This fine-tuning dataset will be noted as “Set B”, and the fine-tuned model will be noted as “Model B”.

2.3.2. Hyperparameter setup

For Set A, Whisper model was fine-tuned utilizing the following hyperparameter configurations: A learning rate of 1e-05 was employed, coupled with a training batch size of 16. Adam optimizer (Kingma & Ba, 2014) was selected, incorporating beta 1 (β_1) as 0.9, beta 2 (β_2) as 0.999 and epsilon (ε) set at 1e-8. Furthermore, a linear learning rate scheduler was applied, featuring a warmup period of 50 steps. The overall training process consisted of 300 steps, which would cover the Set A for about 5 epochs.

For Set B, all the hyperparameter was identical except the training steps. The overall training steps for Set B was 40,000, which would cover the Set B for about 1 epoch. Both Model A and Model B were fine-tuned to minimize the cross-entropy loss function, and every parameter were updated in the fine-tuning process.

2.3.3. Experimental setup

The process of fine-tuning the model was executed employing the Huggingface’s Transformer library (Wolf et al., 2019), selecting pytorch as the backend engine. Furthermore, the fine-tuning process and assessment of the Whisper model were conducted using a single NVIDIA RTX A6000.

3. Results

3.1. Experiment Results

Model A’s training time was about 50 minutes, taking less than an hour, and the train loss value steadily decreased over 5 epochs until it reached 0.0114. Model B’s training time took approximately 95 hours, with the train loss value converging between 0.23 and 0.26. Although there is a concern that Model A may have overfitted based on the trend of loss values alone, this issue will be addressed in a subsequent section comparing the recognition performance of both models.

3.2. Recognition of Targeted Lexical Items

Table 2 summarizes the recognition accuracy of Model A and Model B for six lexical items which Whisper model could not properly recognize, which were introduced in Section 2.2.1.

Table 2. The recognition accuracy of targeted lexical items

Lexical item	Whisper model			
	large-v2	prompted	Model A	Model B
진에어	12.5	12.5	100	12.5
페북	66.7	100	100	100
컴활	16.7	16.7	100	33.3
프사	33.3	100	100	33.3
보증금	33.3	33.3	100	33.3
계절학기	0.0	100	100	100

For a more in-depth performance comparison, we measured the recognition accuracy of a “prompted” model, which utilized prompts for the pre-tuned model (large-v2), in addition to the pre-tuned model itself, and presented the results in Table 2. Prompt engineering techniques have been gaining significant attention in the field of natural language processing, where a prompt or a set of prompts is provided to a language model to guide its response (Liu et al., 2021). These techniques have also recently attracted interest in the speech domain (Chang et al., 2022).

The Whisper model can employ similar prompting techniques. It can recognize speech input up to 30 seconds in length at a time. For recognizing longer speech segments, the results of the previous 30-second segment are supplied as additional input, which can lead to subtle changes in the recognition output if modified. This supplementary text input acts as a type of prompt, and research is not only exploring the impact of modifying recognition results through prompts but also enabling new functionalities such as audio-visual speech recognition (Peng et al., 2023a). Since the primary focus of this study is not prompt engineering, we provided each lexical item that the Whisper model had difficulty recognizing as prompt when calculating the accuracy of the “prompted” model in Table 2, calculating the recognition accuracy for those specific items.

By providing the targeted lexical item as prompts in the “prompted” column of Table 2, we observed that the model either executed recognition for the specific item (the case of 페북, 프사, 계절학기) or displayed no improvement in recognition accuracy (the case of 진에어, 컴활, 보증금). However, when comparing the overall performance to that of fine-tuned Model A and Model B, prompting appears to be less effective than fine-tuning for targeted item speech recognition. Despite this, the prompts did improve recognition accuracy, suggesting that further research might be warranted in this area.

Set A comprised a subset of sentences containing only the six lexical items listed in Table 2. Consequently, it can be inferred that Model A underwent additional learning for these specific lexical items through Set A. Impressively, it still managed to achieve perfect recognition for all targeted lexical items.

Model B also demonstrated improved recognition accuracy for the six lexical items. However, the extent of the improvement was lesser compared to Model A. This can be attributed to the lower proportion of these lexical items in Set B, leading to a diluted fine-tuning effect for them.

Therefore, it is possible to provide a positive response to research question 1, which asked, “Does fine-tuning the model with only a few utterances with the targeted lexical items enable the recognition of the items?” Based on the example of Model A, the answer is affirmative. In fact, a comparison with Model B suggests that concentrating on fine-tuning utterances containing only the targeted lexical items is more effective for enhancing the recognition rate of the targeted items.

3.3. Character Error Rate (CER) for Evalutation Sets

Table 3 provides the CER of the pre-tuned baseline and fine-tuned models for KsponSpeech and LibriSpeech evaluation sets and the relative CER reduction. The best performance in for each evaluation set has been highlighted in bold. The values in parentheses for Model A and Model B represent the “relative CER reduction”, which is calculated using equation (1).

$$\text{relative CER reduction} = ((X - Y)/X) \times 100 \qquad (1)$$

In equation (1), X refers to the CER of the baseline model, in this case, the pre-tuned Whisper model, and Y indicates the CER value of the fine-tuned model.

Table 3. CER and relative CER reduction of various evaluation sets

Evaluation set	Whisper model		
	large-v2	Model A	Model B
KsponSpeech eval set	13.95	9.44 (32.33)	**9.17** (34.26)
LibriSpeech test-clean	1.77	**1.19** (32.77)	1.33 (24.86)
LibriSpeech test-other	**2.86**	2.87 (−0.35)	3.39 (−18.53)

CER, character error rate.

3.3.1. Results on KsponSpeech

The first row of Table 3 shows that both Model A and Model B achieved over a 4.5 percentage points improvement in CER values for the KsponSpeech evaluation set compared to the pre-tuned baseline, which corresponds to more than a 32 percent relative CER reduction. This suggests that training has been effectively conducted regarding the spontaneous speech pattern and the transcription style of KsponSpeech, which frequently involves code-switching between Korean, English, and numbers. An interesting observation is that with Set A, consisting of only 895 utterances under two hours in duration, the performance gap in CER between Model A and Set B, containing 969 hours, remained within a 0.3 difference.

These findings provide answers to the second research question, which was, “How does the fine-tuning process affect the model’s performance on the targeted domain speech compared to its pre-tuned state?” Fine-tuning enables large-scale ASR models to adapt well to specific domains, and since the large-scale ASR models are already robust across a wide range of speech and acoustic variables, effective fine-tuning can occur even with a limited amount of the domain specific data.

3.3.2. Results on LibriSpeech

In the second and third rows of Table 3, we can observe CER of the models on the LibriSpeech test sets. In the test-clean set, Model A achieved a 32.77% relative CER reduction, and Model B achieved a 24.86% relative CER reduction compared to the pre-tuned models, indicating performance improvements for both cases. Despite being fine-tuned on Korean data, the models’ performance on the test-clean set, an English dataset, has improved. In contrast, for the test-other set, both Model A and Model B exhibited performance degradation. Particularly, Model B showed a significant deterioration in performance with a −18.53% relative CER reduction, indicating an increase in the CER value.

In the case of Model A, the minimal degradation compared to Model B could indicate that the large-scale ASR model has maintained its generalization abilities on various languages and environments. This demonstrates the possibility that, when fine-tuning with a small amount of data, not only can recognition performance be improved for the fine-tuned domain, but the existing generalization ability can also be preserved.

From these findings, we can answer to the third research question, which was, “Is the fine-tuned model able to maintain the original model’s generalization capabilities across different languages and environments?” Model A, fine-tuned with a smaller dataset, maintained generalization capabilities across languages and environments. In contrast, Model B faced performance degradation due to a trade-off between specificity and generalization abilities. These findings offer valuable insights for future ASR model development, emphasizing the importance of balancing specialization and generalization across languages.

4. Conclusion

In this study, we have explored the feasibility of fine-tuning large-scale speech recognition models such as the Whisper model for domain-specific applications, using the KsponSpeech dataset as a case study. Through a series of experiments, we assessed the impact of fine-tuning on recognition accuracy of targeted lexical items and the overall performance in the domain-specific context. Furthermore, we examined whether the fine-tuned model maintained its generalization capabilities across different languages and environments.

Our results indicate that fine-tuning the Whisper model with targeted lexical items can effectively improve the recognition accuracy of those specific items. Additionally, even when fine-tuned with a small subset of data, the performance on domain-specific speech improved significantly, as evidenced by the relative CER reduction on the KsponSpeech dataset. Interestingly, Model A, which was fine-tuned with a smaller dataset, maintained its generalization ability across languages and environments. In contrast, Model B displayed performance degradation in noisy environment, suggesting a potential trade-off between specificity and generalization capabilities.

These findings offer valuable insights for the development of ASR models, emphasizing the importance of balancing specialization and generalization across languages and highlighting the potential of fine-tuning with a minimal amount of domain-specific data to achieve satisfactory results. Future research might expand on these findings, exploring different fine-tuning strategies and incorporating novel technologies such as prompting to further enhance the performance of large-scale speech recognition models in domain-specific settings.

References

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020, December). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Proceedings of the Advances in Neural Information Processing Systems* (pp. 12449-12460). Online Conference.

Bang, J. U., Yun, S., Kim, S. H., Choi, M. Y., Lee, M. K., Kim, Y. J., Kim, D. H., ... Kim, S. H. (2020). KsponSpeech: Korean spontaneous speech corpus for automatic speech recognition. *Applied Sciences*, *10*(19), 6936.

Chang, K. W., Tseng, W. C., Li, S. W., & Lee, H. Y. (2022). SpeechPrompt: An exploration of prompt tuning on generative spoken language model for speech processing tasks. Retrieved from https://arxiv.org/abs/2203.16773

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., ... Wei, F. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, *16*(6), 1505-1518.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Retrieved from https://arxiv.org/abs/1810.04805

Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., Han, W., ... Pang, R. (2020). Conformer: Convolution-augmented transformer for speech recognition. Retrieved from https://arxiv.org/abs/2005.08100

Guo, P., Boyer, F., Chang, X., Hayashi, T., Higuchi, Y., Inaguma, H., Kamo, N., ... Zhang, Y. (2021, June). Recent developments on espnet toolkit boosted by conformer. *Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5874-5878). Toronto, ON.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451-3460.

Kim, K., Wu, F., Peng, Y., Pan, J., Sridhar, P., Han, K. J., & Watanabe, S. (2023, January). E-branchformer: Branchformer with enhanced merging for speech recognition. *Proceedings of the 2022 IEEE Spoken Language Technology Workshop (SLT)* (pp. 84-91). Doha, Qatar.

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. Retrieved from https://arxiv.org/abs/1412.6980

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. Retrieved from https://arxiv.org/abs/2107.13586

Mohamed, A., Lee, H. Y., Borgholt, L., Havtorn, J. D., Edin, J., Igel, C., Kirchhoff, K., ... Watanabe, S. (2022). Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics* *in Signal Processing*, *16*(6), 1179-1210.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). Librispeech: an ASR corpus based on public domain audio books. *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206-5210). South Brisbane, Australia.

Peng, P., Yan, B., Watanabe, S., & Harwath, D. (2023a). Prompting the hidden talent of web-scale speech models for zero-shot task generalization. Retrieved from https://arxiv.org/abs/2305.11095

Peng, Y., Kim, K., Wu, F., Yan, B., Arora, S., Chen, W., Tang, J., ... Watanabe, S. (2023b). A comparative study on E-branchformer vs conformer in speech recognition, translation, and understanding tasks. Retrieved from https://arxiv.org/abs/2305.11073

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A, ... Auli, M. (2023). Scaling speech technology to 1,000+ languages. Retrieved from https://arxiv.org/abs/2305.13516

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492-28518). Honolulu, HI.

Rouditchenko, A., Khurana, S., Thomas, S., Feris, R., Karlinsky, L., Kuehne, H., Harwath, D., ... Glass, J. (2023). Comparison of multilingual self-supervised and weakly-supervised speech pre-training for adaptation to unseen languages. Retrieved from https://arxiv.org/abs/2305.12606

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017, December). Attention is all you need. *Proceedings of the Advances in Neural Information Processing Systems*. Long Beach, CA.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., ... Rush, A. M. (2019). Huggingface's transformers: State-of-the-art natural language processing. Retrieved from https://arxiv.org/abs/1910.03771

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., ... Wu, Y. (2023). Google usm: Scaling automatic speech recognition beyond 100 languages. Retrieved from https://arxiv.org/abs/2303.01037

• **Jungwon Chang**
Ph. D. Student, Dept. of English Language and Literature
Korea University
145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: +82-2-3290-1980
Email: cjw1994cool@korea.ac.kr
Fields of interest: Phonetics, Language Engineering

• **Hosung Nam,** Corresponding author
Professor, Dept. of English Language and Literature
Korea University
145, Anam-ro, Seongbuk-gu, Seoul 02841, Korea
Tel: +82-2-3290-1991
Email: hnam@korea.ac.kr
Fields of interest: Phonetics, Language Engineering