Thonburian Whisper: Robust Fine-tuned and Distilled Whisper for Thai

Zaw Htet Aung
Department of Biomedical Engineering
Faculty of Engineering
Mahidol University
Thailand
zawhtetaung.zaw@student.mahidol.ac.th

Thanachot Thavornmongkol
Looloo Technology
Thailand
kimmchi.thanachot@loolootech.com

Atirut Boribalburephan
Looloo Technology
Thailand
atirut.name@loolootech.com

Vittavas Tangsriworakan
Looloo Technology
Thailand
vittavas.tan@student.mahidol.edu

Knot Pipatsriswat
Looloo Technology
Thailand
knot@loolootech.com

Titipat Achakulvisut
Department of Biomedical Engineering
Faculty of Engineering
Mahidol University
Thailand
titipat.ach@mahidol.edu

Abstract

Despite extensive pre-training on a large audio corpus, the Whisper and Distil-Whisper models exhibit considerable challenges in handling Thai speech. This paper presents an approach to enhance pre-trained vanilla Whisper models for Thai automatic speech recognition (ASR). The process involves combining audio datasets, applying audio augmentations during training, and incorporating an audio segmentation strategy. In addition, we show that distilling whisper models can be achieved with less than 1,500 hours of audio while preserving accuracy of student models. The improved models achieve a word error rate (WER) of 11.01%, 6.62%, 5.49%, 11.23%, 7.57% for the small, medium, large, distill-small, and distill-medium Whisper models on Commonvoice 13 dataset. Our models establish as a fine-tuned baseline Whisper ASR for Thai. Furthermore, we demonstrate accuracy of our models with out-of-distribution (OOD) financial datasets while maintaining robustness under environmental noise. The code and pretrained models are available at https://github.com/biodatlab/thonburian-whisper/.

1 Introduction

Automatic Speech Recognition (ASR) converts spoken language into text, which enables several applications such as audio transcription and conversational analysis. Contemporary deep learning-based systems such as Wav2Vec2 (Baevski et al., 2020), Conformer (Gulati et al., 2020), Massively Multilingual Speech (MMS) (Communication et al., 2023), Whisper (Radford et al., 2023), and Seamless M4T (Communication et al., 2023) have demonstrated impressive capabilities in the conversion of spoken languages into text in both English and multilingual audios. However, their performance diminishes when applied to languages with limited audio resources (Bansal et al., 2019). Moreover, adapting these models to accurately transcribe audio in language-specific and specialized domains remains challenging.

Previous efforts to improve Thai ASR models include Wav2Vec2-XLSR (Baevski et al., 2020), Thai Wav2Vec 2.0 (Phatthiyaphaibun et al., 2022) , MMS and Seamless M4T (Communication et al., 2023), which scaled up the Wav2Vec2 architecture to over 1,000 languages. Even though these models perform well in English speech, their performance limitations have been observed in bilingual datasets (Abushariah et al., 2023) and out-of-domain language specific datasets (Jain et al., 2023). This is common in Thai financial audio reports and conference calls, in which most financial terms and company names are dominated by non-native accented English. Inaccurate recognition not only increases the word error rate (WER) but also degrades downstream tasks such as information extraction. End-to-end transformer-based architectures such as OpenAI's Whisper (Radford

et al., 2023) have shown promising results in ASR tasks. Whisper is extensively pretrained on a large multilingual audio corpus of 680,000h, potentially making it a robust and reliable ASR system for Thai speech. This presents an opportunity to combine the strength of transformers and a larger and more diverse datasets to improve the performance of ASR models for Thai.

In this study, we enhance the existing Whisper ASR models by creating a collection of open Whisper models specifically designed for Thai. We combine multiple audio corpora from various sources for fine-tuning. Our main objective is to build a diverse corpus that captures the range of speech nuances, dialects, and accents in Thai language. To enhance the robustness of our models, temporal and spectral augmentations were introduced during fine-tuning. We experimented with models trained using these enhancements to understand their impact on improving model performance and resilience against varying quality and background noise. Balancing model accuracy with computational efficiency is an important consideration especially for environments with limited resources. Previous works (Gandhi et al., 2023) show that it is possible to compress the Whisper models through knowledge distillation. However, a substantial amount of training data is needed for the distilled models to achieve comparable performance to their counterparts. Our work showed that it is possible to achieve successful model compression for Whisper models using a fraction of training data used in (Gandhi et al., 2023). We show significant reductions in word error rates (WER) in all model sizes compared to vanilla Whisper and other ASR models for Thai. Finally, we demonstrate our model's adaptability on OOD financial data. We release the code and pretrained models which can be used as baselines for Thai Whisper ASR.

1.1 Related Works

Availability of transformer-based multilingual ASR models pretrained on massive datasets marks a milestone in the field of low-moderate resource ASR. Yet, few works have addressed the challenges associated with building a robust ASR for Thai. Naowarat et al. introduced contextualized connectionist temporal classification (CCTC) loss to address spelling inconsistencies in code switching Thai ASR. The contextual prediction capabilities inherent in transformer architectures such as those seen in Whisper models align with the objectives of the CCTC loss. The study focusing on ASR technology for Thai dialects (Suwanbandit et al., 2023) highlighted the importance of understanding tonal variations and employing targeted learning approaches to enhance Thai ASR accuracy. Due to the diverse language landscape of Thailand, models capable of handling dialectical differences are needed. Another advancement is the introduction of fine-tuned Wav2Vec2 models for Thai (Phatthiyaphaibun et al., 2022). Here, they utilized a self-supervised pretrained Wav2Vec2 model and fine-tuned on the Commonvoice dataset. However, the total training data was only 128 hours. A more comprehensive evaluation and pretrained models were needed to understand the model's capabilities. Recent development of transformer-based models such as Whisper (Radford et al., 2023) and Distil-Whisper (Gandhi et al., 2023) have shown to effectively capture robustness in multiple languages. By extending the scope of training to languages with limited resources, such as Thai, we can acquire critical insights into the process of fine-tuning these models. This effort will contribute to the accessibility of Thai ASR within the research community.

2 Materials and Methods

2.1 Datasets

2.1.1 Pretraining datasets

We aim to improve Whisper models to robustly transcribe Thai audios. The first stage is to collect data sets to pretrain the Whisper models. We combine multiple primary data sources for pretraining from publicly available speech and internet audio datasets, including Thai CommonVoice 13 (CMV13) (51.41h) (Ardila et al., 2020), Google Fleurs (8.49h) (Conneau et al., 2023), Gowajee (15h) (Chuangsuwanich et al., 2020) and Thai Elderly Speech (26.56h)[1], and Thai Central Dialect corpus (683.9h) (Naowarat et al., 2021). To make the model generalize to most domains, we scrape audio from various sources on the Internet, first listing 250 generic Thai keywords and exploring their associated queries or topics using Google Trends[2] . We then used the associated queries to search for audios over the Internet and acquired 5,100 uncleaned captioned audios. To clean the captioned audio, the audios are selected if they are (i) pub-

[1] https://github.com/VISAI-DATAWOW/Thai-Elderly-Speech-dataset/releases/tag/v1.0.0

[2] https://trends.google.com/trends/

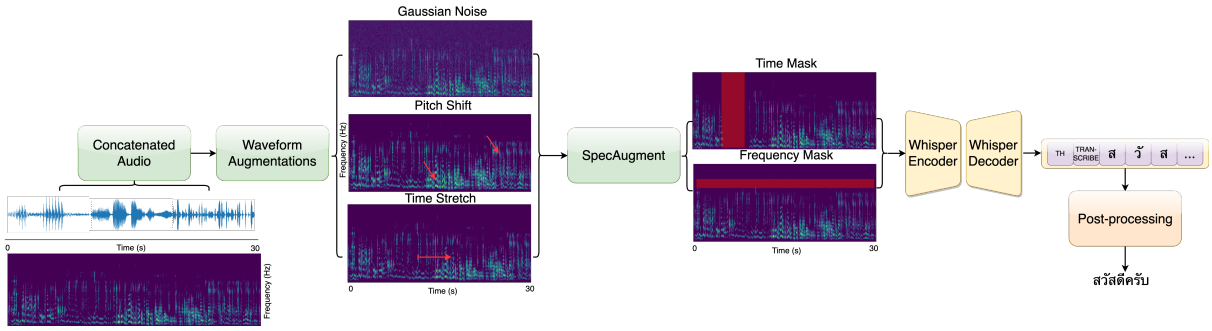

Figure 1: **Schematic of model pretraining**: The Whisper model is fine-tuned on a collection of more than 1.3k hours of Thai audios with additional augmentations including audio clip concatenation, waveform augmentation (Gaussian noise, time stretch, and pitch shift), and spectral augmentations.

licly accessible, (ii) in Thai, (iii) available with Thai subtitles, and (iv) not categorized as music, resulting in a total of 631.89 hours of additional audios. Combining these datasets results in a total of 1,316.76 hours.

2.1.2 OOD Financial audio dataset

We have assembled a specialized dataset tailored for the financial domain to see adaptability and usage of the models in domain-specific use cases. The rationale behind this lies in the presence of distinctive terminology within the financial sector, which is not typically encountered in general-purpose databases. In this effort, we collected around 18 hours of financial reports from earnings call videos, "Oppday"[3]. These records cover quarterly reports from various companies from 2020 to 2021. We used VAD (Team, 2021) to segment each audio file into short snippets ranging approximately from 2 to 4 seconds, resulting in 30,185 audio snippets. We annotate these snippets using the web-based tool 'Audino' (Grover et al., 2021). The audio samples are divided into 28,568 (96.64%, 17.69h) and 1,617 (5.36%, 1.08h) samples of training and testing, respectively. We use the OOD data to evaluate zero-shot generalization and fine-tune our models to see their adaptability compared to other Whisper models.

2.2 Thonburian Whisper Pretraining

The performance of the fine-tuned Whisper for Thai depends on the size of the pre-training and the fine-tuning strategy. Here, we select small, medium, and large (v3) Whisper model sizes for fine-tuning for Thai. We fine-tuned the models using a straightforward approach without augmentation. In addition, we propose a set of augmentation techniques applied during training to make Whisper more robust for Thai audios, which include

- Concatenation of audio clips: The concatenation of short audio to reach the default Whisper input length (30s) improves the efficiency of the sample and training.
- Waveform augmentation: Raw waveform augmentations for collected audios are applied randomly. Techniques include Gaussian noise injection, temporal waveform dilation, and pitch shifting (Jordal et al., 2023).
- SpecAugment: We applied SpecAugment (Park et al., 2019) to mask the features of the spectrogram along the temporal and frequency axes. We set a probability of 0.3 for time masking and apply masking along 10 consecutive time steps. We applied frequency masking across 64 frequency bands with a probability of 0.1.

All models were trained for 10,000 iterations with and without proposed augmentation. We used a batch size of 16, using the deep-speed ZeRO optimizer (Rajbhandari et al., 2020). Pretrained models are evaluated and compared with the vanilla Whisper models, Thai Wav2Vec 2.0, and Seamless-M4T large models.

2.3 Model Distillation

We use the distillation technique proposed by (Gandhi et al., 2023) using layer-based compression of the Whisper decoder layers. Four maximally spaced decoder layers are copied from the teacher model to the student model, while the

[3] https://listed-company-presentation.setgroup.or.th/en

teacher's encoder layers are kept intact (Supplementary Table 4). During distillation, the prediction probabilities of the student model are trained to match those of the teacher model by minimizing the Kullback-Leibler (KL) divergence across the entire spectrum of possible tokens. In contrast to Gandhi et al. which used more than 21k hours of audio, including pseudo-labels, our approach utilizes a more modest distillation dataset of around 1,317 hours. Distillation is carried out in two steps: for the first 10,000 iterations, the optimal alignment between the encoder and decoder layers of the student model is achieved through the guidance of the teacher. Next, the student model is fine-tuned for another 10,000 iterations using the same dataset without relying on KL divergence loss. This approach of dual-step distillation and fine tuning allows the proposed distilled models to preserve accuracy despite utilizing significantly fewer hours of data.

2.4 Zero-shot Performance and Fine-tuning on OOD Financial Dataset

For vanilla Whisper (Radford et al., 2023) models, ThaiWav2Vec 2.0 (Phatthiyaphaibun et al., 2022) and Seamless M4T large (Communication et al., 2023), we evaluated their zero-shot performances on CMV13 test, FLEURS test and Thai Central dev datasets. Furthermore, we test all the models' zero-shot generalization on the OOD dataset and perform fine-tuning of both vanilla and Thonburian Whisper models to see their adaptability in the financial domain.

2.5 Model Robustness Under Environmental Noises

To evaluate the robustness of our pretrained models, we inject environmental disturbances sourced from the ESC-50 dataset (Piczak, 2015) into the FLEURS test set (Conneau et al., 2023). We used 2,000 environmental audio recordings that span 50 semantic categories, each lasting 5 seconds. We selected 40 longest-duration samples and adjusted the amplitude, using the noises according to the signal-to-noise ratio (SNR). Noise samples are duplicated or trimmed depending on the length of the audio to be inserted. This process is repeated across 9 SNR levels, ranging from - 20dB to 20dB with 5dB increments. This results in a corrupted test set that contains 2,000 corrupted audios for each SNR.

2.6 Model Evaluation

We perform naive text post-processing to normalize the output transcript, such as vowel corrections, tone mark orders, and extra white space removal. Evaluation is carried out by calculating the the word error rate (WER), the deletion error rate (DER), the substitution error rate (SER) and the insertion error rate (IER) with Thai word tokenizer, deepcut (Kittinaradorn et al., 2019). IER can be used to indicate the hallucination of the model, i.e., predicting repeated words. Other evaluation includes measurement of the latency in predicting short- and long-form audios (Supplementary Table 4).

3 Results and Discussion

3.1 Model Performance

We evaluate all models on the short-form audios without timestamp prediction on Common Voice 13, FLEURS, and Thai Central development datasets (Table 1). Thonburian Whisper have shown improved performance in all model sizes. They have shown less vulnerability to hallucinations as seen in the lower IERs. The small model gained the highest WER improvement after fine-tuning on the combined Thai dataset where the large Thonburian models obtained the lowest WERs on all our test sets. An interesting observation is that the augmented large model demonstrates a minor decline in performance on Commonvoice 13 and FLUER while slightly outperforming the non-augmented variant in Thai Central development dataset. The augmented models show a higher robustness after a noise corruption with SNR less than -5 dB (Figure 2).

Distilled Thonburian Whisper (S, M) with 1.3k hours of audios have shown comparable performance (less than 1 WER difference on CMV13 and Thai Central Dev) to the original model in all evaluated dataset (Table 1). They have 68.6% and 56.02% less parameters compared to the original S, M models. Hence, the distilled models achieve 1.26x and 3.89x speed up in short-form inference and 1.72x and 2.41x for long-form inference (Table 4). Therefore, the trade-off between accuracy and computational complexity may be justifiable in resource constrained scenarios.

Table 1: Evaluation Results on Different Datasets

Model	Params	CMV13-Test				Google Fleurs Test				Thai Central Dev			
	(M)	WER	IER	SER	DER	WER	IER	SER	DER	WER	IER	SER	DER
Vanilla (S)	242	38.8	8.6	26.7	3.5	43.0	8.6	30.5	3.9	61.5	5.0	41.3	15.1
Vanilla (M)	764	23.9	4.5	16.8	2.6	30.5	6.5	20.6	3.5	50.6	2.3	32.0	16.3
Vanilla (L)	1,543	12.8	2.1	9.1	1.5	14.7	3.2	9.4	2.0	37.9	1.9	22.6	13.3
Thonburian (S, A)	242	13.1	3.5	8.5	1.1	15.4	3.9	9.5	1.9	8.9	2.6	5.4	0.9
Thonburian (S)	242	11.0	2.2	7.7	1.1	14.1	3.3	8.8	2.0	8.7	2.6	5.1	1.0
Thonburian (M, A)	764	7.4	1.5	5.1	0.8	10.5	2.8	6.2	1.6	6.2	1.7	3.7	0.9
Thonburian (M)	764	6.6	1.0	4.8	0.8	10.2	2.8	5.9	1.5	6.8	2.4	3.7	**0.8**
Thonburian (L, A)	1,543	6.6	1.4	4.5	0.7	9.1	2.3	5.3	1.5	**5.4**	**1.3**	**3.2**	0.9
Thonburian (L)	1,543	**5.5**	**0.8**	**4.0**	**0.7**	**8.7**	**2.0**	**5.2**	**1.5**	6.0	1.8	3.3	0.9
Distilled Thonburian (S)	166	11.2	2.2	7.8	1.2	16.6	4.8	9.8	2.0	8.9	2.6	5.2	1.0
Distilled Thonburian (M)	428	7.6	1.2	5.5	0.9	12.5	3.4	7.3	1.8	6.5	1.6	3.9	1.0
Wav2Vec2 (L)	316	10.3	4.0	5.4	0.9	25.4	9.9	14.0	1.5	26.2	3.5	20.0	2.7
Seamless-M4T (L)	2,360	12.8	1.9	9.3	1.6	20.0	5.1	12.1	2.9	34.2	2.1	23.4	8.7

[1] S,M,L - Small, Medium, Large; A - Augmented

Table 2: Zero-Shot Performance on the OOD Financial Domain Test

Model	WER	IER	SER	DER
Vanilla (S)	72.7	26.6	31.8	14.4
Vanilla (M)	59.7	21.8	24.0	13.9
Vanilla (L)	25.2	3.3	12.4	9.5
Thonburian (S)	32.1	8.2	13.8	10.1
Thonburian (S, A)	33.2	10.9	14.0	8.3
Thonburian (M)	23.6	5.2	10.0	8.4
Thonburian (M, A)	25.4	8.2	10.0	7.2
Thonburian (L)	**18.7**	**2.5**	8.7	7.5
Thonburian (L, A)	19.7	2.6	**8.4**	8.7
Distilled Thonburian (S)	32.4	8.2	14.3	9.9
Distilled Thonburian (M)	27.5	5.4	11.5	10.6
Wav2Vec2 (L)	46.9	10.1	33.1	**3.7**
Seamless-M4T (L)	37.4	7.2	24.7	5.5

Table 3: OOD Fine-Tuning Results on Oppday Test Set

Model	WER	IER	SER	DER
Thonburian (S)	15.3	3.4	10.0	1.9
Thonburian (S, A)	15.2	3.3	9.9	1.9
Thonburian (M)	11.9	2.6	7.7	1.5
Thonburian (M, A)	**11.5**	**2.5**	**7.5**	**1.5**
Distilled Thonburian (S)	17.3	4.2	10.4	2.6
Distilled Thonburian (M)	13.3	3.0	8.6	1.7
Vanilla (S)	21.8	5.5	14.0	2.3
Vanilla (M)	14.7	3.3	9.6	1.7

3.2 OOD in financial domain and fine-tuning capability

Table 2 provides an analysis of how Thai ASR models perform when faced with OOD data without any additional fine tuning. Vanilla whisper models, especially small (72.7% WER) and medium (59.7% WER) ones, exhibit a significant struggle when dealing with audio samples from specific domains such as finance. Frequent code-switching between domain specific terms in English and Thai coupled with non-native accents makes it particularly challenging. In contrast, Thonburian Whisper models show remarkable improvements in performance compared to their vanilla counterparts. Models such as Wav2Vec2 (L) and Seamless-M4T (L) demonstrate higher WERs than Vanilla Whisper (L). In particular, their substitution error rates are much higher. This underscores the varying levels of success in zero-shot generalization across different model architectures. Table 3 shows the results on Oppday test set after fine-tuning on the domain specific data. All Thonburian models perform better than their vanilla Whisper counterparts. Interestingly, even the distilled models can adapt better to OOD data. This suggests that the proposed training

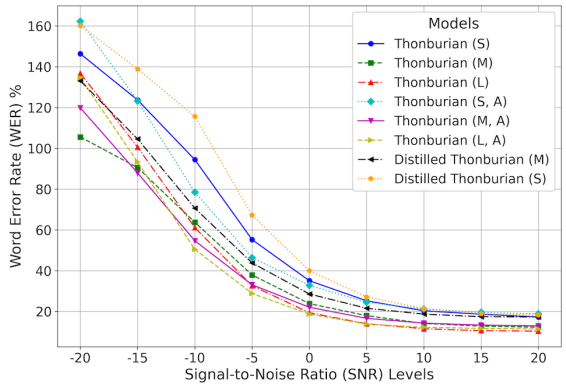

Figure 2: Robustness of the models under environmental noise.

scheme can enhance the adaptability of baseline Whisper models.

3.3 Model robustness under environmental noise

Augmented Thonburian Whisper large consistently outperforms all other models across different SNR levels of environmental noise corruption (Figure 2). From 0 to 20 dB, the non-augmented Thonburian models (S, M, L) show average WERs of 23.3, 16.18 and 13.15 respectively. The augmented models perform similarly under milder noise conditions with average WERs of 23.38, 15.83 and 13.56. As the noise corruptions become more severe (0 to -20 dB), the augmented variants outperform their counterparts. The standard deviations of WER for Thonburian Whisper small, medium and large are 5.10, 3.37, and 2.71 respectively. This suggests their performance is consistent across the SNR levels from 20 to 0, which is a good indicator of the model robustness. The proposed training scheme seems to have significantly improved the performance of the Whisper models under noisy conditions.

4 Conclusion

This study shows that Whisper based automatic speech recognition models can be successfully adapted and enhanced for Thai language. The proposed fine-tuning scheme and a combined corpus of Thai audios have led to substantial improvements in word error rate (WER) over existing baselines and previous works. Furthermore, we show that it is feasible to compress Whisper models through knowledge distillation with a fraction of data used in (Gandhi et al., 2023). This highlights the potential for creating lightweight yet strong ASR solutions for low resource languages. The strong performance of Thonburian Whisper models on the OOD financial dataset showcases their effectiveness and adaptability. This is notable considering the complex terminology, code switching tendencies and accented speech.

Limitations

The suggested audio augmentation methods can help create robust ASR systems designed for noisy environments. However, the differing levels of noise resistance among the models call for a further exploration of optimization tactics that can consistently improve robustness regardless of model sizes. The distilled models, especially the small one, are more adversely affected by noise corruptions. This indicates that there is still room for improvements in the distillation process to enhance robustness. Finally, while this study demonstrated the adaptability to financial domain data, further efforts are necessary to assess how well the models would work in a range of fields and situations for a language as complex and tonally varied as Thai.

Ethics Statement

This research complies with the ACL Ethics Policy[4]. All experiments were conducted using publicly available datasets, namely CMV13-Test, Google Fleurs Test, and Thai Central Dev, which are well-documented and widely accepted in the research community. We ensured that no personal or sensitive information was involved. We recognize the broader impacts of our work in automated speech recognition (ASR). While our models show significant improvements in ASR accuracy, it is essential to apply these advancements responsibly. We advocate for the ethical use of ASR technology to benefit diverse communities and prevent perpetuating biases or inequalities.

Acknowledgements

The authors acknowledge the NSTDA Supercomputer Center(ThaiSC) project #lT200142 for providing computing resources for this work.

[4] https://www.aclweb.org/portal/content/acl-code-ethics

Table 4: Computational Resources Comparison Across Distilled Models

Model	Encoder	Decoder	GPU Memory Usage	Memory Efficiency[1]	Short-Form Speed Up[2]	Long-Form Speed Up[3]
Thonburian (S)	12	12	461MB	-	-	-
Thonburian (M)	24	24	1,420MB	-	-	-
Distilled Thonburian (S)	12	4	317MB	1.45x	1.26x	1.72x
Distilled Thonburian (M)	24	4	816MB	1.74x	3.89x	2.41x

[1] Memory efficiency indicates the relative GPU memory usage effectiveness in FP16.
[2] Short-Form speed up is the time taken to transcribe approximately 6 seconds of audio.
[3] Long-Form speed up refers to the time taken to transcribe approximately 60 seconds of audio.

References

Ahmad A. M. Abushariah, Hua-Nong Ting, Mumtaz Begum Peer Mustafa, Anis Salwa Mohd Khairuddin, Mohammad A. M. Abushariah, and Tien-Ping Tan. 2023. Bilingual Automatic Speech Recognition: A Review, Taxonomy and Open Challenges. *IEEE Access*, 11:5944–5954. Conference Name: IEEE Access.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. 2019. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 58–68, Minneapolis, Minnesota. Association for Computational Linguistics.

Ekapol Chuangsuwanich, Atiwong Suchato, Korrawe Karunratanakul, Burin Naowarat, Chompakorn CChaichot, Penpicha Sangsa-nga, Thunyathon Anutarases, Nitchakran Chaipojjana, and Yuatyong Chaichana. 2020. Gowajee Corpus. Technical report, Chulalongkorn University, Faculty of Engineering, Computer Engineering Department.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. 2023. SeamlessM4T-Massively Multilingual & Multimodal Machine Translation. ArXiv:2308.11596 [cs].

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. FLEURS: FEW-Shot Learning Evaluation of Universal Representations of Speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805.

Sanchit Gandhi, Patrick von Platen, and Alexander M Rush. 2023. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430*.

Manraj Singh Grover, Pakhi Bamdev, Ratin Kumar Brala, Yaman Kumar, Mika Hama, and Rajiv Ratn Shah. 2021. audino: A Modern Annotation Tool for Audio and Speech. ArXiv:2006.05236 [cs, eess].

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*, pages 5036–5040. ISCA.

Rishabh Jain, Andrei Barcovschi, Mariam Yiwere, Peter Corcoran, and Horia Cucu. 2023. Adaptation of Whisper models to child speech recognition. ArXiv:2307.13008 [cs, eess].

Iver Jordal, Araik Tamazian, Emmanouil Theofanis Chourdakis, Céline Angonin, Tushar Dhyani, askskro, Nikolay Karpov, Omer Sarioglu, Baker-Bunker, kvilouras, Enis Berk Çoban, Florian Mirus, Jeong-Yoon Lee, Kwanghee Choi, MarvinLvn, SolomidHero, and Tanel Alumäe. 2023. iver56/audiomentations: v0.33.0.

Rakpong Kittinaradorn, Titipat Achakulvisut, Korakot Chaovavanich, Kittinan Srithaworn, Pattarawat Chormai, Chanwit Kaewkasi, Tulakan Ruangrong, and Krichkorn Oparad. 2019. DeepCut: A Thai word tokenization library using Deep Neural Network. Version Number: 1.0.

Burin Naowarat, Thananchai Kongthaworn, Korrawe Karunratanakul, Sheng Hui Wu, and Ekapol Chuangsuwanich. 2021. Reducing spelling inconsistencies in code-switching asr using contextualized ctc loss. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6239–6243.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Interspeech 2019*, pages 2613–2617. ArXiv:1904.08779 [cs, eess, stat].

Wannaphong Phatthiyaphaibun, Chompakorn Chaksangchaichot, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. Thai Wav2Vec2.0 with CommonVoice V8. ArXiv:2208.04799 [cs, eess].

Karol J. Piczak. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, pages 1015–1018, New York, NY, USA. Association for Computing Machinery. Event-place: Brisbane, Australia.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust Speech Recognition via Large-Scale Weak Supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518. PMLR. ISSN: 2640-3498.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. ZeRO: memory optimizations toward training trillion parameter models. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC '20, pages 1–16, Atlanta, Georgia. IEEE Press.

Artit Suwanbandit, Jaturong Chitiyaphol, Sutthinan Chuenchom, Kanyarat Kwiecien, Husen Sawal, Ruslan Uthai, Orathai Sangpetch, and Ekapol Chuangsuwanich. 2023. Thai-dialect: Low resource thai dialectal speech to text corpora. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Silero Team. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. Publication Title: GitHub repository.