Thai Dialect Corpus and Transfer-based Curriculum Learning Investigation for Dialect Automatic Speech Recognition

Artit Suwanbandit[1], *Burin Naowarat*[1], *Orathai Sangpetch*[2], *Ekapol Chuangsuwanich*[1]

[1]Chulalongkorn University, Thailand
[2]CMKL university, Thailand

{6372130221, 6270145221}@student.chula.ac.th, orathai@cmkl.ac.th, ekapol.c@chula.ac.th

Abstract

We release 840 hours of read speech multi-dialect ASR corpora consisting of 700 hours of main Thai dialect, named Thai-central, and 40 hours for each local dialect , named Thai-dialect, with transcripts and their translations to Thai. The dialects, selected to represent different regions of Thailand, are Khummuang, Korat, and Pattani. We also release the baseline dialectal ASR models trained using the curriculum learning approach. We found that the pre-training with the high-resource main dialect and target dialect generally yields the best performance. We believe that the availability of our corpora would contribute to the problem of low-resource Thai dialects. The corpus data will be available on Github[1].

Index Terms: Speech recognition, Dialect ASR, Deep learning, Transfer-learning, curriculum learning

1. Introduction

The development of ASR has made it possible for machines to understand human speech. There are a large number of local dialects in Thailand, mainly influenced by traditions, histories, and ethnicities; such a diverse language landscape has resulted in the decreased accuracy of ASR systems for low-resource local dialects. To address this issue, we have created the Thai-central and Thai-dialect datasets in order to improve ASR systems' performance for local dialects. The Thai-central corpus consists of more than 700 hours of Thai read speech data while the Thai-dialect is a collection of dialectal speech corpora from three different regions in Thailand: Khummuang from the north, Korat from the northeast, and Pattani from the south with each dialectal speech corpus containing more than 40 hours of data. The Thai-dialect corpus is designed so that they are parallel with the Thai-central corpus creating another venue for further research such as speech-to-text and speech-to-speech translation. In this work we also investigate the use of this dataset for dialect and low-resource ASR research.

Deep learning, particularly transformer-based models like Wav2Vec2.0 [1] and HuBERT [2], has become the standard method for ASR since the introduction of the Listen, Attend, and Spell (LAS) model [3]. Recent studies have highlighted the success of deep learning approaches in dialectal Arabic ASR [4]. However, multi-dialect systems have generally been found to perform poorly compared to single-dialect systems. For Japanese dialectal ASR, multi-task learning with dialect identification (DID) and multi-dialect ASR has been proposed, but it was found that incorrect predictions of DID tasks negatively affect ASR performance [5]. Multi-task methods with soft multi-task learning (Soft-MTL) models, such as those presented by Z. Dan, et al. [6] use an additional speech encoder to achieve promising results.

In our work, we observed that Thai dialectal languages often have a high degree of similarity in speech to the main dialect. Chuangsuwanich, et al. [7] demonstrated transfer learning helps the most if the target language is more similar to the source language. Transfer learning has also been shown to be a promising approach for low-resource ASR in several studies [8, 9, 10, 11]. Therefore, we aim to examine the effectiveness of transfer learning in the context of our Thai dialect dataset.

Transfer learning is a simple yet effective approach for improving deep learning performance in low-resource settings. In the field of ASR, [12] investigated the effectiveness of self-supervised pre-training on various English datasets. Additionally, an experiment in [13] explored the impact of supervised and self-supervised pre-training on ASR performance under domain and language mismatch scenarios. The method of utilizing transfer learning in ASR has been referred to as transfer-based learning in [14]. Our work focuses on transfer-based curriculum learning for ASR in Thai-dialect datasets.

Curriculum learning (CL) is an approach for ordering data to optimize training efficiency, inspired by the "starting small" concept. When used correctly, CL can reduce training steps and improve accuracy [15]. According to Gkarakasidis, et al. [14], CL can be categorized into three approaches: metadata-based, adaptive-based, and transfer-based. Their work focused on an adaptive-based approach using loss/metric-based methods. Meanwhile, S. Braun, et al. [16] demonstrated that adding noise over time can improve the robustness of ASR.

In our work, we explore the effectiveness of transfer-based curriculum learning for guiding supervised ASR pre-training and fine-tuning of low-resource dialect ASR using a high-resource main-dialect dataset.

2. Thai-central and Thai-dialect corpora

One part of our corpus is collected from standard Thai dialect called Thai-central. The other portion, Thai-dialact, consist of three Thai dialects, Khummuang, Korat, and Pattani. In this section, we describe the recording procedure, corpora information, and the characteristics of the dialects.

2.1. Data collection procedure

2.1.1. Text prompts

Our corpus creation process starts by gathering text prompts which will be used to record spoken utterances from volunteers. In order to improve diversity of the prompts, sentences are obtained from seven different sources. Additional sentences were specifically also created for two special domains of inter-

[1] https://github.com/SLSCU/thai-dialect-corpus

10.21437/Interspeech.2023-1828

est, namely chatting for e-commerce, and daily communication sentences. We call the two domains E-commerce and Survival, respectively. The E-commerce and Survival subsets were generated using templates where nouns, verbs, product names, etc. were replaced. Due to the variety of the templates, the Survival prompts are more challenging than E-commerce. The final breakdown of our text-prompts is shown in Table 1.

For the Thai-central corpus, every sentence from all sources were included. However, the Thai-dialect corpora contain only the E-commerce and Survival sentences, which were translated to their respective dialectal transcriptions. We made sure that the sentences only contain Thai letters and digits, removing equations, parentheses, brackets, and special characters. In addition, numbers, special signs, and times were converted to text beforehand to ensure consistency.

Table 1: *The unique sentences in each sentences source.*

Data source	Data detail	#Sentence
E-commerce	shopping sentences	33063
Survival	daily sentences	6062
BEST2010 [17]	articles, novels, cyclopedias, news	7113
Dek-D	teen webboard	6187
Pantip	general webboard	3499
Wiki	Wikipedia	4147
Others	other sentences	6850
total		66921

2.1.2. Audio recording

The collection process was carried out in between 2020 and 2021. We utilized a crowd-sourcing platform, `https://www.wang.in.th`, to record Thai-central. To minimize human errors, the platform only allows speakers who have passed a prerequisite test to record audio. Moreover, every audio recording is validated by another expert-level crowd worker to ensure data quality. The dialect recordings were done done locally by recruiting participants from target regions. Since our recording were done during the COVID-19 pandemic, most recordings for Thai-dialect were conducted online by sending application links to speakers. All recordings were collected in the wild with minimal environmental control.

2.1.3. Audio Verification

To curate the quality of the recordings, we employed three screening criteria: Voice Activity Detection (VAD), Signal to Noise Ratio (SNR), and signal energy. We utilized our in-house VAD to detect the parts of speech and noise in the audios. Every audio file must meet all of the criteria to be considered acceptable.

2.1.4. Tokenization

Since the Thai writing system does not mark word boundaries, tokenization of the text into words is also required. For the Thai-central corpus, we utilized a maximal-matching method in Pythainlp [18] for word tokenization. However, since there are no established tokenization standards for the dialects, we relied on dialectal specialists to tokenize the sentences. To ensure labeling consistency, we employed maximal matching tokenizers on subsets of the data to flag any potential disagreements with the specialist's tokenization. If there were disagreements between automatic tokenization and the specialist, the sentence would be re-examined. The final tokenization accuracy values are 68.52, 54.81, 72.06 for Khummuang, Korat, and Pattani, respectively.

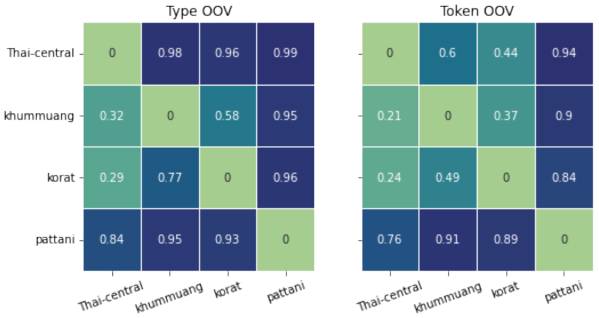

Figure 1: *Matrices showing OOV rate when adopting from the source vocabulary (row) to target transcription (column).*

2.2. Corpus statistics and dialect details

Thai-central is the most comprehensive dialect in our corpus as it includes all seven sentence sources, making it the most generalized. In Thai-dialect, there are three dialects: Khummuang, Korat, and Pattani, which were spoken in the North, Northeastern, and South regions of Thailand, respectively. Thai-central and Thai-dialect information has shown in Table 3. Note that gender statistics were obtained using a pitch-based gender classifier.

To better understand the dialect similarities, we visualized the overlap in vocabulary between each dialect in the corpus using two out-of-vocabulary (OOV) metrics: type OOV and token OOV. Type OOV counts the percentage of non-overlapping unique words, while token OOV counts the frequency of non-overlapping words. As shown in Figure 1, Korat and Khummuang, both being Zhuang-Tai languages, have relatively low type and token OOV values compared to Thai-central. This is due to their shared grammar structure and vocabulary [19]. In contrast, Pattani is the most distinct dialect as it is more closely related to the Malay language from Malaysia, which belongs to the Austronesian language family [20]. In our Pattani corpus, the Thai writing system and alphabets are used instead of Arabic to maintain consistency with the rest of the corpus.

In our corpus, metadata consists of utterance ID, speaker ID, and the transcription. Thai-translated transcriptions were included in each dialect corpus. For the ASR task, we divided each dialect corpus into three sets: Train, Dev, and Test. We ensured that text transcriptions and speaker ID in the Dev, Test, and Train sets did not overlap. Evaluation can be further divided to only evaluate on E-commerce and Survival utterances. For Dev and Test, we set the ratio between Survival and E-commerce to 1:3. The corpus statistics are shown in Table 2.

3. Curriculum Learning for dialect ASR

The main challenge for dialect ASR lies upon the scarcities in their resources. Luckily, dialects are close to others, and one of them usually is a widely spoken language. Previous works leveraged this phenomenon by training multi-dialect ASR systems that learn common acoustic properties from closely-related dialects [5, 6, 11]. However, a multi-dialect ASR has

Table 2: *Statistics of Thai-central and Thai-dialect datasets.*

	Thai-central (Th)			**Khummuang (Kh)**			**Korat (Ko)**			**Pattani (Pa)**		
	Train	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
utterances	335674	5465	5522	23738	806	3003	38624	1080	3852	25039	834	2793
speakers	5422	363	332	277	46	506	442	11	113	761	81	57
duration (hr)	683.9	10.1	10.1	32.4	1.5	6	46.7	1.5	6	40.0	1.25	5
unique transcripts	56702	5087	5132	8029	665	2785	3627	1033	3599	3591	768	2760
type OOV (%)	-	10.72	8.84	-	14.18	31.01	-	13.60	24.22	-	39.96	24.29
token OOV (%)	-	1.17	0.83	-	0.57	0.77	-	1.15	1.13	-	8.06	9.65

Table 3: *Information of each dialect in Thai-central and Thai-dialect corpus*

Dialect	Thai-Central	Khummuang	Korat	Pattani
#Utterance	433,814	60,350	45,307	36,842
#Speaker	6,713	844	571	891
#Sentence	66,921	37,269	8,546	15,570
Hours	862	98	53	70
Male	22.7%	23.6%	19.6%	8.0%
Female	77.3%	76.4%	81.4%	92.0%

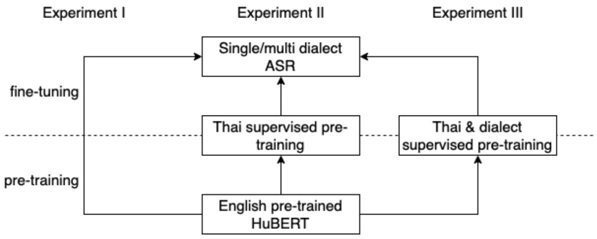
Experiment I
Experiment II
Experiment III
fine-tuning
pre-training
Single/multi dialect ASR
Thai supervised pre-training
Thai & dialect supervised pre-training
English pre-trained HuBERT

Figure 2: *Overview of our experiments.*

to perform well on every dialect, which potentially degrades its effectiveness for smaller dialects. In our work, we explore the use of curriculum learning to study its effect in our dataset.

We studied the effectiveness of curriculum training for Thai dialect ASR using three different approaches, as shown in Figure 2. These three experiments were categorized by how we pre-trained the networks. In experiment I, we directly fine-tuned English HuBERT encoders using dialectal speech. Experiment II and III both comprise two training steps. The first step used the English HuBERT encoder as initialization for training on Thai audios. The second step fine-tuned the weights from the first step using dialect-specific training. We referred to the first step as *pre-training* and the second step as *fine-tuning*. The only difference between experiment II and III is the dataset we used in the pre-training stage. Concretely, we used only Thai-central for experiment II but also included the dialect corpus for experiment III.

3.1. Architecture details

We conducted our experiment using ESPNet toolkit [21] and used transformer-based hybrid CTC/Attention models [22]. The encoder of the model was HuBERT [2], pre-trained on 960 hours of English Librispeech [23]. We used 0.3 as the weight for the CTC prediction head and used the label smoothing of 0.1. In the fine-tuning step, only the encoders were transferred from the pre-trained version, but the decoder was trained from scratch. Since all dialects share the same character set, all fine-tuned models has the same output setup.

3.2. Training details

For each curriculum training step, we trained the models for 200k steps unless there were no improvements in the last 30 epochs. This criteria of 200k steps was applied to both pre-training and fine-tuning steps. We used a maximum training batch size of 562.5 seconds. For decoding, we used a beam size of 10. Since Thai writing does not have the standard use of word boundaries, we used our word tokenizer, described in Section 2.1.4, to re-tokenized predicted transcriptions before computing WER.

3.3. Evaluation

We evaluated the models using the standard Word Error Rate (WER) computed on dialect specific test sets. Since every test set consists of two subsets (Survival and E-commerce), we reported performance for each subset as well. Specifically, we denoted *S-WER* and *E-WER* as the WERs for Survival and E-commerce subsets, respectively. Besides WER, Character Error Rate (CER) are also report to remove any potential tokenization issues. No language model was used in order to directly measure the performance of the acoustic models.

4. Experiments

4.1. Experiment I: Baselines

The results of experiment I, which shows simple fine-tuning baselines from pre-trained English Hubert, are shown in Table 4. **Th**, **Kh**, **Ko**, and **Pa** stand for Thai, Khummuang, Korat, and Pattani, respectively. The fine-tuning can be done on different combination of dialects. As shown in the table, fine-tuning on multi-dialects usually achieve the best WER. Adding Thai-central does not help in the repetitive E-commerce set. On the other hand, it improves the performance on the survival set because of the diversity required in the type of sentences.

Adding more dialects to the training set seems to improve the CER in every dialect. Upon further investigation, we found that adding more languages made the models produce more mispellings which may increase the WER. However, the mispellings are quite reasonable. The misspelling are usually the short-long vowels, tonal characters, dialect confusion, similar consonants, and homophones. This is due to the fact that the dialects share a lot of words with similar roots but slight differences in spelling and pronunciations. CER might be a better metric in these cases, and a multi-dialect system might be the most useful if it is combined with a language model.

We also tried a method proposed in [5] which performed

DID and ASR in a multitask manner. To perform DID, the end of sentence token was replaced with a dialect ID, so that the model learns the discrepancies between dialects. However, the model did not perform well, even though it had the DID accuracy of 96.78%. This is because few incorrect DID predictions dramatically degrade CER and WER values. Our finding also aligns with [5].

Table 4: *Experiment I: English Hubert fine-tuning Baselines*

Dialect	fine-tune	CER	WER	S-WER	E-WER
Kh	Kh	7.19	9.26	15.46	6.94
	Kh+Th	6.65	8.99	12.95	7.51
	Kh+Ko+Pa	5.83	**8.12**	12.98	**6.31**
	Kh+Ko+Pa+Th	**5.82**	8.24	**12.89**	6.49
Ko	Ko	10.84	16.05	38.06	8.73
	Ko+Th	8.55	12.79	**24.18**	8.99
	Kh+Ko+Pa	9.45	14.68	35.42	**7.76**
	Kh+Ko+Pa+Th	**8.09**	**12.50**	24.87	8.38
Pa	Pa	26.25	40.22	63.14	31.78
	Pa+Th	**18.11**	34.03	**49.17**	28.45
	Kh+Ko+Pa	23.04	35.76	58.35	**27.44**
	Kh+Ko+Pa+Th	18.18	**33.37**	49.38	27.47
Multi-task ASR2DID					
Kh	Kh+Ko+Pa+Th	6.61	8.84	13.42	7.13
Ko		9.16	13.73	27.06	9.29
Pa		19.17	37.87	57.96	30.47

4.2. Experiment II: Thai-central pre-training

Table 5 shows how the main dialect, Thai-central, can be useful to other dialects. There are improvements in CER and WER when compared to English pre-training. This finding is congruent with what we found in both single and multi-dialect systems of experiment I. Even though HuBERT has demonstrated excellent performance in English ASR [2], pre-training on Thai-central displays language adaptation, which is the key to dialect ASR. Although Pattani is the distinct dialect, it also improves by 5.9% relatively in WER compared to the English pre-trained model.

4.3. Experiment III: Thai-central and Thai-dialect pre-training

In **Th+Kh** and **Th+Ko** pre-training, single-dialect fine-tuning achieves satisfactory results in the overall CER and WER.

On the other hand, the performance of **Th+Pa** becomes worse compared to just pre-training using **Th**. This suggests training size and distinctiveness of target dialects are factors for pre-training effectiveness. When all languages are used to pre-train (**Th+Kh+Ko+Pa**), both single and multi-dialect systems tends to perform better and are the best overall. The gap in performance between E-commerce and Survival subsets also decreases. The superiority of the multi-dialect system highlights the importance of the similarities between languages used for pre-training and fine-tuning. Our finding also aligns with [12].

4.4. Further study on the effect of the Pattani dialect

Since Pattani has a high OOV rate compared to other dialects, we hypothesize that it might hurt performance of other languages in a multi-dialect system. We therefore compared performances of multi-dialect systems with and without Pattani. Table 6 shows only small differences in both CER and WER compared to other methods. Including distant dialects might not harm the performance of multi-dialect system.

Table 5: *Experiment II&III: different supervised pre-training.*

Dialect	fine-tune	CER	WER	S-WER	E-WER
Th supervised pre-training					
Kh	Kh	6.43	8.35	13.11	6.57
	Kh+Ko+Pa	5.71	7.81	12.15	6.19
Ko	Ko	8.60	12.63	27.71	7.61
	Kh+Ko+Pa	8.57	13.11	29.36	7.70
Pa	Pa	23.19	37.85	60.81	29.40
	Kh+Ko+Pa	**19.13**	**32.62**	**51.68**	25.61
Th+Kh supervised pre-training					
Kh	Kh	6.10	8.13	12.51	6.50
	Kh+Ko+Pa	5.68	7.93	12.64	6.17
Th+Ko supervised pre-training					
Ko	Ko	8.53	12.50	26.61	7.80
	Kh+Ko+Pa	8.43	12.68	27.84	7.63
Th+Pa supervised pre-training					
Pa	Pa	32.40	42.05	63.67	34.09
	Kh+Ko+Pa	21.13	33.81	54.09	26.34
Th+Kh+Ko+Pa supervised pre-training					
Kh	Kh	6.16	8.07	13.02	6.21
	Kh+Ko+Pa	**5.41**	**7.51**	**11.43**	**6.04**
Ko	Ko	8.35	12.39	26.67	**7.63**
	Kh+Ko+Pa	**8.13**	**12.28**	**26.19**	7.64
Pa	Pa	23.86	37.63	60.83	29.08
	Kh+Ko+Pa	19.91	32.66	54.62	**24.57**

Table 6: *Ablation results on using only Th, Ko, and Kh*

Dialect	fine-tune	CER	WER	S-WER	E-WER
Th supervised pre-training					
Kh	Kh+Ko	5.92	8.01	11.95	6.53
Ko		8.64	12.54	28.63	7.82
Th+Kh+Ko supervised pre-training					
Kh	Kh	7.22	8.97	15.09	6.68
	Kh+Ko	**5.49**	**7.53**	**11.68**	**5.98**
Ko	Ko	**8.33**	**12.22**	**26.31**	**7.53**
	Kh+Ko	**8.33**	12.54	27.08	7.69

5. Conclusion

We introduced the Thai-central and Thai-dialect corpus which have 700 hours of Thai and more than 40 hours of Thai dialectal language with parallel transcriptions. We investigated the use transfer-based curriculum learning for creating strong dialectal ASR systems. Using the high-resource main dialect with the target dialect for pre-training and fine-tuning on the target dialect usually gives the most favorable results. Additionally, multi-dialect ASR systems can perform better than single systems in overall performance in Thai-dialectal languages.

6. Acknowledgement

The authors would like to thank the PMU-C grant (Thai Language Automatic Speech Recognition Interface for Community E-Commerce, C10F630122) for the support of this research. We also would like to acknowledge the Apex compute cluster team which provides compute support for this project.

7. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 449–12 460, 2020.

[2] 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.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[4] H. A. Alsayadi, I. Hegazy, Z. T. Fayed, B. Alotaibi, and A. A. Abdelhamid, "Deep investigation of the recent advances in dialectal Arabic speech recognition," *IEEE Access*, 2022.

[5] R. Imaizumi, R. Masumura, S. Shiota, H. Kiya *et al.*, "End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning," *APSIPA Transactions on Signal and Information Processing*, vol. 11, no. 1, 2022.

[6] Z. Dan, Y. Zhao, X. Bi, L. Wu, and Q. Ji, "Multi-task transformer with adaptive cross-entropy loss for multi-dialect speech recognition," *Entropy*, vol. 24, no. 10, 2022. [Online]. Available: https://www.mdpi.com/1099-4300/24/10/1429

[7] E. Chuangsuwanich, Y. Zhang, and J. Glass, "Multilingual data selection for training stacked bottleneck features," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5410–5414.

[8] H. Inaguma, J. Cho, M. K. Baskar, T. Kawahara, and S. Watanabe, "Transfer learning of language-independent end-to-end ASR with language model fusion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6096–6100.

[9] N. Hjortnæs, N. Partanen, M. Rießler, and F. M. Tyers, "The relevance of the source language in transfer learning for ASR," in *Proceedings of the Workshop on Computational Methods for Endangered Languages*, vol. 1, 2021, pp. 63–69.

[10] S. Khare, A. R. Mittal, A. Diwan, S. Sarawagi, P. Jyothi, and S. Bharadwaj, "Low resource ASR: The surprising effectiveness of high resource transliteration," in *Interspeech*, 2021, pp. 1529–1533.

[11] B. Pulugundla, M. K. Baskar, S. Kesiraju, E. Egorova, M. Karafiát, L. Burget, and J. Cernockỳ, "BUT system for low resource Indian language ASR," in *Interspeech*, 2018, pp. 3182–3186.

[12] W.-N. Hsu, A. Sriram, A. Baevski, T. Likhomanenko, Q. Xu, V. Pratap, J. Kahn, A. Lee, R. Collobert, G. Synnaeve, and M. Auli, "Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training," in *Proc. Interspeech 2021*, 2021, pp. 721–725.

[13] A. Misra, D. Hwang, Z. Huo, S. Garg, N. Siddhartha, A. Narayanan, and K. C. Sim, "A comparison of supervised and unsupervised pre-training of end-to-end models," in *Proc. Interspeech 2021*, 2021, pp. 731–735.

[14] G. Karakasidis, "Comparison of new curriculum criteria for end-to-end ASR," Master's thesis, Aalto University. School of Science, 2022. [Online]. Available: http://urn.fi/URN:NBN:fi:aalto-202208285156

[15] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: Association for Computing Machinery, 2009, p. 41–48. [Online]. Available: https://doi.org/10.1145/1553374.1553380

[16] S. Braun, D. Neil, and S.-C. Liu, "A curriculum learning method for improved noise robustness in automatic speech recognition," in *2017 25th European Signal Processing Conference (EUSIPCO)*. IEEE, 2017, pp. 548–552.

[17] T. N. Electronics, "C.T.C.: Benchmark for Enhancing the Standard of Thai Language Processing 2010," 2010.

[18] W. Phatthiyaphaibun, K. Chaovavanich, C. Polpanumas, A. Suriyawongkul, L. Lowphansirikul, and P. Chormai, "Pythainlp: Thai natural language processing in python," *URL: http://doi. org/10.5281/zenodo*, vol. 3519354, 2016.

[19] A. Diller, J. Edmondson, and Y. Luo, *The Tai-Kadai Languages*. Routledge, 2004.

[20] C. Goddard, "Semantic primes and universal grammar in Malay (Bahasa Melayu)," in *Meaning and universal grammar: Theory and empirical findings*. John Benjamins Publishing Company, 2002.

[21] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211.

[22] 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.

[23] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.