Adapting OpenAI's Whisper for Speech Recognition on Code-Switch Mandarin-English SEAME and ASRU2019 Datasets

Yuhang Yang*, Yizhou Peng†§, Hao Huang‡, Eng Siong Chng§ and Xionghu Zhong*¶
* Hunan University, China
¶ Corresponding Author. Email: xzhong@hnu.edu.cn
† National University of Singapore, Singapore
‡ School of Computer Science and Engineering, Xinjiang University, China
§ Nanyang Technological University, Singapore

***Abstract*—This paper reports on SOTA results achieved using openAI's Whisper model with adaptation on different adaptation corpus sizes for two established code-switch Mandarin/English corpus - namely SEAME and ASRU2019 corpora.**

Two key experiments were conducted: a) using adaptation data from 1 to 100/200 hours to demonstrate the effectiveness of adaptation, b) examining different language ID setups on Whisper prompt. The Mixed Error Rate results show that the amount of adaptation data may be as low as $1 \sim 10$ hours to achieve saturation in performance gain (SEAME), while the ASRU task continued to show performance with more adaptation data ($>$100 hours). For the language prompt without adaptation, the results show that various prompting strategies produce different outcomes. However, after adaptation, the Whisper model uniformly improves its performance, and the language prompt becomes less critical. We believe that these results can help researchers study the adaptation of Whisper to other code-switch Languages.

I. Introduction

Code-switching is a pervasive linguistic phenomenon in multilingual communities, where speakers alternate between two or more languages within a single speech or conversation. The prevalence of code-switching poses a unique challenge for Automatic Speech Recognition (ASR) systems [1], [2], as it requires the model to have a nuanced understanding of multiple languages simultaneously. Unfortunately, current ASR models often underperform in recognizing code-switching speech due to the scarcity of labeled training data.

In the past few years, the research community has introduced several approaches to tackle the challenges posed by code-switching ASR systems. These approaches can be broadly categorized into three technical aspects: speech, text, and modeling methods. From the speech perspective, strategies have been developed to implement monolingual speech into code-switching ASR systems [3]. Additionally, some researchers propose augmenting pronunciation models to accommodate accents, mispronunciations, and pronunciation variations, thereby addressing the issue of data sparsity [4]. On the text front, multiple techniques have been explored, ranging from augmenting code-switching text from monolingual corpora to build language models (LM) [5]–[8], to employing methods like speech T5 [9], [10], multilingual word-embedding [11]–[14], Internal LM estimation [15]–[17], and LM rescoring [18]. Lastly, from the modeling standpoint, various frameworks have been suggested, such as the Mixture of Experts (MoE) which uses separate encoders and decoders for different languages [19], [20], frame-level Language Identification or Diarization as an auxiliary task [21], [22], and the incorporation of self-supervised models as frontend models for ASR [23].

When superlarge parameter models show their emergent ability to understand and generate language when given a suitable prompt [24], researchers are turning to large-scale foundational models trained on extensive multilingual datasets, e.g., Whisper [25], USM [26] and MMS [27]. These models aim to encapsulate various linguistic rules and contexts, offering more robust performance across different languages and dialects. Using large-scale training data and advanced modeling techniques, these foundational models have the potential to revolutionize the field of ASR, making it more inclusive and accurate for multilingual and code-switching populations.

In this paper, we concentrate on applying varied language labels as prompts during the Whisper model's training and decoding phases. Our investigation centers on the efficacy of these language label prompts in the model's fine-tuning process. Additionally, we propose a prompting approach that considers code-switching as a distinct language. This method derives language embeddings through a weighted combination of the respective language embeddings, such as Mandarin and English, and attempts to enhance model performance under a code-switching scenario.

The paper is organized as follows. Section II is to review recently proposed adaptation methods based on the Whisper model to Code-Switching or specific languages. Section III presents the methods used in our experiments. Section IV briefly summarizes the datasets used and the overall experimental setup. Section V shows our experimental results and section VI shows our ablation study on the training data size. After that, we draw conclusions in Section VII.

TABLE I

PROMPT USED IN PROMPTINGWHISPER FOR CS-ASR AND OUR PROPOSED LANGUAGE-FUSION PROMPT FOR BOTH FINETUNING AND DECODING. ⟨|sot|⟩ STANDS FOR ⟨|startoftranscript|⟩ TOKEN AND ⟨|asr|⟩ MEANS THE ⟨|transcribe|⟩ TOKEN IN WHISPER TOKENIZER.

Languages	Default	PromptingWhisper	Language-Fusion
Zh+En	⟨|sot|⟩**⟨|zh|⟩** *or* **⟨|en|⟩**⟨|asr|⟩	⟨|sot|⟩**⟨|zh|⟩⟨|en|⟩** *or* **⟨|en|⟩⟨|zh|⟩**⟨|asr|⟩	⟨|sot|⟩**⟨|en-zh|⟩**⟨|asr|⟩

II. WHISPER APPLICATIONS

The Whisper model [25] is an advanced speech recognition system created by OpenAI that can accurately transcribe audio into text. It is trained on a wide-ranging dataset, enabling it to handle multiple languages and dialects effectively. It is also capable of speech translation and language identification, except for speech recognition.

Whisper stands out for its robust performance in various acoustic settings and its contextual understanding of improved transcription, which makes it distinct for various speech and text research tasks that benefit from the Whisper encoder and decoder separately or simultaneously. Specifically, following the same training pipeline and finetuning the entire Whisper model, performance improvement is obtained for several low-resource language speech recognition [28], [29]. TCPGen also shows its effectiveness of contextual biasing for the Whisper model [30]. With the utilization of the Whisper encoder, deepfake detection can benefit from input features composed of the embeddings extracted from the last layer of the Whisper encoder and traditional MFCC features [31]. With a similar strategy applied, these embeddings can also improve the performance of infant cry classification models compared to the MFCC features [32]. Also, researchers indicate that the embeddings from the Whisper encoder are not only noise robust to ASR task but also contain information that can help classify the noise types, e.g., audio event tagging [33].

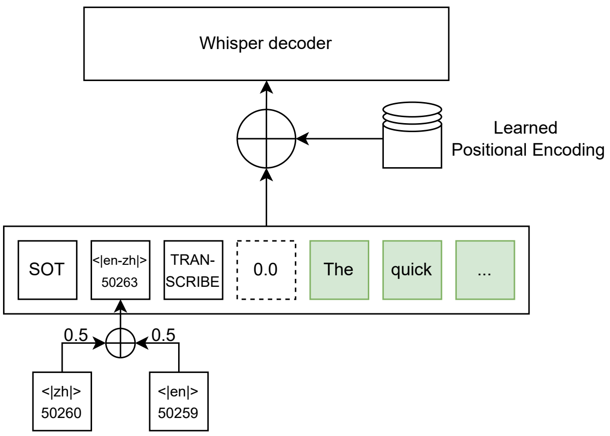

Fig. 1. The proposed Language Fusion method

Researchers usually try to prompt the decoder for various applications when it comes to the Whisper decoder. By replacing the speech recognition transcription label with several Spoken Language Understanding (SLU) labels while keeping the same decoding prompt as the ASR task, Whisper can perform SLU through transfer learning and multitask learning [34]. Also, when customized prompts are fed into the Whisper decoder, improved performance is observed for several zero-shot tasks such as Code-Switching ASR and Speech Translation, which had been previously shown to underperform with the default Whisper prompts. [35].

III. METHODS

A. *PromptingWhisper for CS-ASR*

PromptingWhisper [35] is an innovative approach to speech recognition, focusing on adapting the Whisper model to new and untrained tasks through prompt engineering. This method involves strategically using prompts - specific instructions or inputs - to guide the Whisper model in processing and responding to tasks beyond its original training. PromptingWhisper primarily explores three tasks: audio-visual speech recognition (AVSR), where the model transcribes speech from videos with related visual content; code-switched speech recognition (CS-ASR), which involves recognizing speech that alternates between different languages; and speech translation (ST) for language pairs that the model has not previously encountered.

For CS-ASR, PromptingWhisper suggests combining two Language Prompts instead of using only one of the two Languages, e.g., Table I shows all the prompts for Mandarin-English CS-ASR. Precisely, following the default prompting rule of Whisper [25], the resulting prompt for speech recognition would be either ⟨|sot|⟩**⟨|zh|⟩**⟨|asr|⟩ or ⟨|sot|⟩**⟨|en|⟩**⟨|asr|⟩ where only one language should be specified, however, PromptingWhisper suggest that given a prompt like ⟨|sot|⟩**⟨|zh|⟩⟨|en|⟩**⟨|asr|⟩ where sequentially inserting two languages would introduce 19% relative performance improvement in average for CS-ASR task among several CS corpora such as SEAME [36] and ASCEND [37]. Likewise, we think ⟨|sot|⟩**⟨|en|⟩⟨|zh|⟩**⟨|asr|⟩which only reverses the order of two languages should also have similar outcome. Therefore, we apply both **⟨|en|⟩⟨|zh|⟩** and **⟨|zh|⟩⟨|en|⟩** language prompts in the following experiments, treating them as PromptingWhisper's suggested prompts.

B. *Language-Fusion Prompt*

Inspired by the **PromptingWhisper**, we introduce a new Language Prompt that fuses the pre-trained English and Chinese Language Embeddings, which is named **Language-Fusion Prompt**, to investigate if this could be beneficial to English-Mandarin CS-ASR task. Typically, as shown in Figure 1, the new embedding is obtained by weighting the pre-trained embeddings corresponding to English and Chinese Language Prompt tokens that are **⟨|en|⟩** (*Token id in Whisper*

Tokenizer is 50259) and ⟨|**zh**|⟩ (*Token id in Whisper Tokenizer is 50260*) with the same weighting factor which is set to 0.5. To keep the same output dimension as the original Whisper Decoder, the ⟨|**ru**|⟩ (*Stands for Russian Language. Token id in Whisper Tokenizer is 50263*) is then replaced with the resulting new Language Prompt which is named as ⟨|**en-zh**|⟩. We selected the Russian language prompt as the replacement because of the substantial acoustic and linguistic differences between Russian and English/Chinese. The decoding prompt is finalized as ⟨|sot|⟩⟨|**en-zh**|⟩⟨|asr|⟩shown in Table I.

TABLE II
OVERALL SPEECH DATA DISTRIBUTION FOR ASR MODEL TRAINING AND TESTING.

Corpus	Subset	Duration(Hrs)
SEAME	Train	93.6
	Dev_{Man}	7.5
	Dev_{Sge}	3.9
ASRU	Train	193.0
	Valid	6.8
	Dev1	20.4
	Dev2	21.3
	Test	20.6

IV. EXPERIMENTS

A. Data

We select two Mandarin-English code-switching ASR data sets to verify the effectiveness of all the methods mentioned in Section III. One is SEAME [36], a conversational Mandarin-English corpus from SouthEast Asia, i.e., Malaysia and Singapore. Another is a Mandarin-English CS data set from China Mainland, released by `Datatang` for a Mandarin-English CS ASR challenge in ASRU2019 [38]. For brevity, we name it ASRU in what follows. Though both data sets are Mandarin-English CS, they are hugely different. Firstly, they are from different areas, which means CS is influenced by different cultural backgrounds. More importantly, SEAME data is conversational speech, while ASRU is read speech, and hence, it is much simpler. Table II reports overall speech data distributions in detail, where Dev_{Man} and Dev_{Sge} are two officially defined test sets for SEAME corpus. Dev_{Man} is dominated by Mandarin and vice versa; the other is dominated by English. Mandarin dominates all ASRU datasets.

B. Model

Due to the limitation of computing resources, all of our experiments are performed with the Whisper-small multilingual model. The encoder is configured with 12 layers, and the decoder consists of 12 layers with 8-head attention. The input feature is 80-dim Mel frequency bins, which are computed on 25-ms windows with a stride of 10 ms. Our models are trained on one A40 GPU with 48GB VRAM. The original batch size is set to 6, and the gradient accumulation is 12, for a total batch size of 72. We use AdamW optimizer with the peak learning rate of $1e^{-5}$, and the warmup lasts 200 steps. The max updating step is set to 30k. Also, mixed-precision training strategy is applied in our experiments. When decoding, we use an average model from the last five epochs, and the decoding beam size is set to 1.

V. RESULTS

The experimental results include mainly two parts. First, we show the results before finetuning the Whisper model called the Zero-shot Prompt. Then, by finetuning the Whisper model following different Prompt styles, we show the huge improvements for all Prompt styles among the two Code-Switching datasets.

TABLE III
MEAN ERROR RATES (MERS)(%) WITH DIFFERENT LANGUAGE PROMPT FOR ZERO-SHOT CODE-SWITCHING ASR.

Type	L-Prompt	SEAME		ASRU		
		Dev_{Man}	Dev_{Sge}	Dev1	Dev2	Test
Conformer	N/A	**16.6**	**23.3**	**8.6**	**14.0**	**13.2**
Official	⟨|en|⟩	101.9	83.6	96.0	98.9	105.3
	⟨|zh|⟩	80.8	157.5	27.0	25.3	25.0
	Auto	67.8	84.9	31.1	29.9	29.4
Custom	⟨|en|⟩⟨|zh|⟩	84.0	81.3	98.4	101.1	99.4
	⟨|zh|⟩⟨|en|⟩	98.8	81.2	33.2	32.2	32.3
	⟨|en-zh|⟩	74.0	101.7	33.8	31.9	31.6

A. Zero-Shot Prompts on Whisper-Small

We follow the instructions of PromptingWhisper and apply the suggested Language Prompts and our proposed method to Whisper-small, and the results are shown in Table III. Specifically, **Conformer** follows the configuration from recipe [39] in the ESPnet2 toolkit. **Official** type represents the original Whisper Language Prompt style, where ⟨|en|⟩ and ⟨|zh|⟩ stand for specifying English and Mandarin for the entire test set respectively. *Auto* denotes that we don't manually state the Language Prompt when performing decoding, which means that Whisper would automatically recognize the Language Label for each sentence. **Custom** type includes two combined Language Prompts following PromptingWhisper where ⟨|en|⟩⟨|zh|⟩ stands for English first and Mandarin second in the combined Language Prompt and vice versa. The proposed weighted-sum method is shown as ⟨|en-zh|⟩.

The results show that for all testsets, the Whisper-small model with various Language-Prompts underperforms Conformer models that are trained with corresponding training data. However, different Language-Prompts do significantly affect the performance of Code-Switching speech recognition. Specifically, for the SEAME dataset, ⟨|**en**|⟩ prompt gives better results for Dev_{Sge} while ⟨|**zh**|⟩ shows better performance on Dev_{Man}. The *Auto* prompt shows better performance compared with those with specific language prompts for Dev_{Man} and similar results for Dev_{Sge}. When given customized prompt, ⟨|**en**|⟩⟨|**zh**|⟩ and ⟨|**zh**|⟩⟨|**en**|⟩ prompts show best results on Dev_{Sge}, while Dev_{Man} underperforms the official prompts. Our proposed Language-Fusion prompt ⟨|**en-zh**|⟩ also performs poorly in the zero-shot scenario.

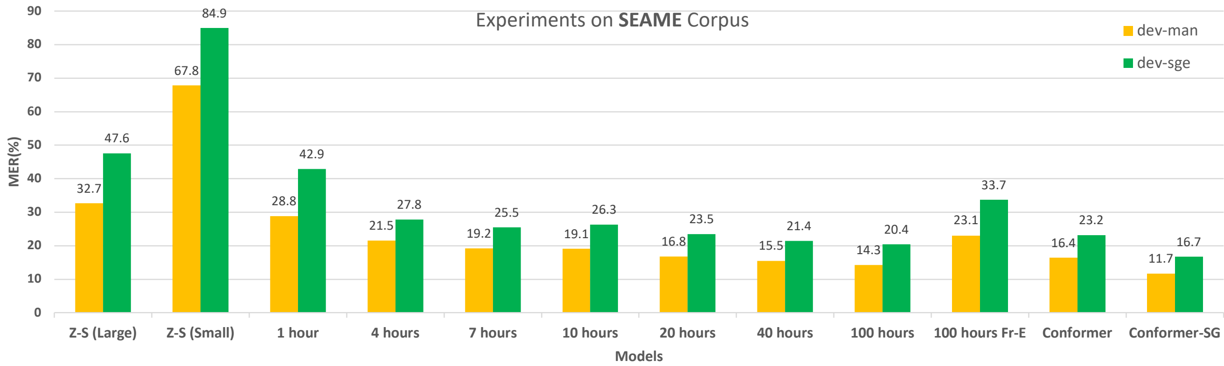

Fig. 2. MER(%) results of SEAME corpus with diverse training data size finetune on Whisper-Small model. **Z-S (Large)** stands for the PromptingWhisper-Large model, and **Z-S (Small)** corresponds to directly decode Whisper-Small model with **Auto** Language Prompt mentioned in Table III. {1,4,7,10,20,40,100} hours indicate the Whisper-Small models that are finetuned with corresponding training data size and ⟨|en|⟩ is fixed during both finetuning and decoding stages. **Fr-E** denotes Freeze-Encoder when fintuning the Whisper-Small model. **Conformer-SG** is our in-house English-Mandarin-Malay trilingual model trained with over 40k hours of Singaporean-accented data.

TABLE IV
MERS(%) WITH DIFFERENT LANGUAGE PROMPT FOR WHISPER-SMALL FINETUNED MODEL. ⟨|ru|⟩ STANDS FOR RUSSIAN LANGUAGE PROMPT.

Type	L-Prompt	SEAME Dev_{Man}	SEAME Dev_{Sge}	ASRU Dev1	ASRU Dev2	ASRU Test
Conformer	N/A	**16.6**	23.3	8.6	14.0	13.2
Official	⟨|en|⟩	**14.3**	**20.4**	**6.3**	10.9	10.3
	⟨|zh|⟩	14.8	20.6	**6.3**	10.8	**10.1**
	⟨|**ru**|⟩	15.5	21.5	6.3	10.8	10.3
Custom	⟨|en|⟩⟨|zh|⟩	15.1	21.1	**6.3**	**10.6**	**10.1**
	⟨|zh|⟩⟨|en|⟩	15.0	21.0	6.5	11.0	10.5
	⟨|en-zh|⟩	15.1	20.9	**6.3**	10.8	**10.1**

When it comes with ASRU dataset, ⟨|**zh**|⟩ and *Auto* prompts show significant performance improvement compared with ⟨|**en**|⟩ **and** ⟨|**en**|⟩⟨|**zh**|⟩ prompts which shows consistency to data composition of ASRU.

B. Finetuning Whisper Model

We finetune the Whisper-Small model given different language prompts and then specify the same language prompt when performing decoding. Table IV shows the MER results of all Whisper models we finetuned on SEAME and ASRU datasets.

The results show that regardless of which Language Prompt we use for finetuning and decoding, all the testsets obtain significant performance improvements and outperform the Conformer models that are trained with corresponding training set.

Specifically, ⟨|**en**|⟩ prompt achieves best performance for both test sets of SEAME corpus and also introduces $2.3 \sim 2.9\%$ absolute MER reduction compared with Conformer model, while ⟨|**en**|⟩⟨|**zh**|⟩ prompt yields optimal results on all Dev and Test sets of ASRU corpus which demonstrates $2.3 \sim 3.4\%$ absolute MER reduction compared with Conformer model. However, We observe that regardless of whether the language prompt is ⟨|**en**|⟩, ⟨|**zh**|⟩, both, or even our proposed fusion prompt, the results are quite similar, with less than a 1% MER gap across all models for each test set. To further explore the effect of using different language prompts, we introduced the Russian Language Prompt ⟨|**ru**|⟩ as a reference in this experiment. The rationale behind using ⟨|**ru**|⟩, a language not present in the training or test data, is to examine whether the choice of an unrelated language prompt would influence the model's performance after finetuning. Our findings are confirmed by the results of ⟨|**ru**|⟩ which demonstrate that regardless of what language prompt is given to the Whisper decoder, once the finetuning process is complete, the performance achieved will be similar.

VI. ABLATION STUDY

In this section, we primarily investigate how the size of the training dataset affects the performance outcomes of finetuning the Whisper model. Additionally, we explore the effects on Whisper's performance when the encoder parameters are frozen during the finetuning process and solely update the decoder parameters.

First, for all the finetuned Whisper models, we fix the language prompts ⟨|**en**|⟩ for SEAME and ⟨|**zh**|⟩ for ASRU2019 experiments. We then randomly subset $(1, 4, 7, 10, 20, 40)$ hours from both SEAME and ASRU training sets and update the Whisper-Small model for 8000 steps to examine the smallest size of data that could obtain a practical CS-ASR system. Also, we try to freeze the entire encoder of the Whisper model while performing finetuning on the whole dataset to determine if finetuning the decoder alone can effectively bridge the gap between multilingual and Code-Switching scenarios. Additionally, this can help to verify whether the Whisper Encoder can overcome the acoustic mismatch that often arises in such diverse linguistic environments.

Figure 2 shows the results on the SEAME corpus. The results show that only 1 hour of training data can produce a model that outperforms PromptingWhisper-Large by around 10% and obtains around 50% MER reduction. When the amount of training data reaches 20 and 40 hours, the model will yield results comparable to or surpass the conformer

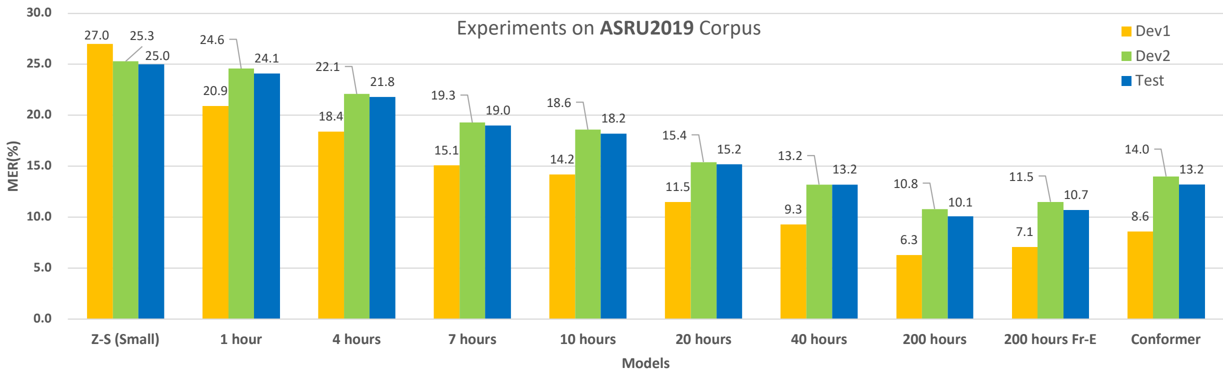

Fig. 3. MER(%) results of ASRU corpus with diverse training data size finetune on Whisper-Small model. **Z-S (Small)** corresponds to directly decode Whisper-Small model with ⟨|zh|⟩ Language Prompt. {1,4,7,10,20,40,200} hours indicate the Whisper-Small models that are finetuned with corresponding training data size and ⟨|zh|⟩ is fixed during both finetuning and decoding stages. **Fr-E** denotes Freeze-Encoder when fintuning the Whisper-Small model.

model trained with the entire training set. However, when the encoder parameters are frozen, the performance drop could be at most 65%, which suggests that the encoder of the Whisper-Small model may exhibit a substantial mismatch with SEAME corpus (Singaporean Accents), potentially leading to degradation in its performance.

In Figure 2, we also present the Conformer-SG, which is a U2++ Conformer streaming ASR model as referenced in [40]. This model, trained on over 40,000 hours of Mandarin-English-Malay speech data, achieves state-of-the-art performance on the SEAME corpus and surpasses the finetuned Whisper-Small model by approximately 18%.

Figure 3 shows the results on the ASRU corpus. The results show a similar conclusion to the one we obtained from experiments with the SEAME corpus. However, the frozen encoder experiment for the ASRU corpus shows much lower performance degradation compared with the SEAME corpus, which suggests the performance gap tends to diminish when dealing with scenarios, such as speech without obvious accents, that closely align with the conditions and characteristics of the Whisper model's training set.

VII. Conclusion

This paper reveals that adapting the Whisper model with code-switching datasets significantly enhances its capability for code-switching speech recognition (CS-ASR), even in contexts with limited resources. Our experiments, which span a variety of linguistic backgrounds, demonstrate that while different prompting strategies yield varied performances prior to adaptation, after adapting the Whisper model with code-switching speech data, these strategies result in similarly enhanced performance, effectively mitigating the complexities inherent in code-switching environments. These adaptations not only bolster the model's overall performance but also align with the broader goal of developing ASR systems that are more inclusive and precise for multilingual users. This research thus marks a crucial step forward in understanding the potential of large foundational models for navigating the intricate dynamics of code-switching in various linguistic scenarios.

VIII. Acknowledgements

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore.

References

[1] G. Ma, W. Wang, Y. Li, Y. Yang, B. Du, and H. Fu, "Lae-st-moe: Boosted language-aware encoder using speech translation auxiliary task for e2e code-switching asr," in *ASRU*, 2023.

[2] Y. Peng, J. Zhang, H. Xu, H. Huang, and E. S. Chng, "Minimum word error training for non-autoregressive transformer-based code-switching asr," in *ICASSP*, 2022.

[3] B. Yan, C. Zhang, M. Yu, *et al.*, "Joint modeling of code-switched and monolingual asr via conditional factorization," in *ICASSP*, 2022.

[4] Y. Long, S. Wei, J. Lian, and Y. Li, "Pronunciation augmentation for mandarin-english code-switching speech recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, 2021.

[5] C.-T. Chang, S.-P. Chuang, and H.-Y. Lee, "Code-switching sentence generation by generative adversarial networks and its application to data augmentation," *arXiv preprint arXiv:1811.02356*, 2018.

[6] Y. Li and P. Fung, "Code-switch language model with inversion constraints for mixed language speech recognition," in *Proceedings of COLING 2012*, 2012.

[7] X. Hu, Q. Zhang, L. Yang, B. Gu, and X. Xu, "Data augmentation for code-switch language modeling by fusing multiple text generation methods.," in *INTERSPEECH*, 2020.

[8] Y. Li and P. Fung, "Code switch language modeling with functional head constraint," in *ICASSP*, 2014.

[9] J. Ao, R. Wang, L. Zhou, *et al.*, "Speecht5: Unified-modal encoder-decoder pre-training for spoken language processing," *arXiv preprint arXiv:2110.07205*, 2021.

[10] B. Yusuf, A. Gandhe, and A. Sokolov, "Usted: Improving asr with a unified speech and text encoder-decoder," in *ICASSP*, 2022.

[11] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[12] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual bert?" *arXiv preprint arXiv:1906.01502*, 2019.

[13] S. Santy, A. Srinivasan, and M. Choudhury, "Bertologicomix: How does code-mixing interact with multilingual bert?" In *Proceedings of Adapt-NLP*, 2021.

[14] G. Lee and H. Li, "Modeling code-switch languages using bilingual parallel corpus," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

[15] A. Kannan, Y. Wu, P. Nguyen, T. N. Sainath, Z. Chen, and R. Prabhavalkar, "An analysis of incorporating an external language model into a sequence-to-sequence model," in *ICASSP*, 2018.

[16] Y. Liu, R. Ma, H. Xu, Y. He, Z. Ma, and W. Zhang, "Internal language model estimation through explicit context vector learning for attention-based encoder-decoder asr," *arXiv preprint arXiv:2201.11627*, 2022.

[17] Y. Peng, Y. Liu, J. Zhang, *et al.*, "Internal language model estimation based language model fusion for cross-domain code-switching speech recognition," *arXiv preprint arXiv:2207.04176*, 2022.

[18] G. Liu and L. Cao, "Code-switch speech rescoring with monolingual data," in *ICASSP*, 2021.

[19] Y. Lu, M. Huang, H. Li, J. Guo, and Y. Qian, "Bi-encoder transformer network for mandarin-english code-switching speech recognition using mixture of experts.," in *INTERSPEECH*, 2020.

[20] T. Song, Q. Xu, M. Ge, *et al.*, "Language-specific characteristic assistance for code-switching speech recognition," *arXiv preprint arXiv:2206.14580*, 2022.

[21] C. Shan, C. Weng, G. Wang, *et al.*, "Investigating end-to-end speech recognition for mandarin-english code-switching," in *ICASSP*, 2019.

[22] H. Liu, H. Xu, L. P. Garcia, A. W. Khong, Y. He, and S. Khudanpur, "Reducing language confusion for code-switching speech recognition with token-level language diarization," in *ICASSP*, 2023.

[23] L.-H. Tseng, Y.-K. Fu, H.-J. Chang, and H.-y. Lee, "Mandarin-english code-switching speech recognition with self-supervised speech representation models," *arXiv preprint arXiv:2110.03504*, 2021.

[24] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, 2023.

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

[26] Y. Zhang, W. Han, J. Qin, *et al.*, "Google usm: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[27] V. Pratap, A. Tjandra, B. Shi, *et al.*, "Scaling speech technology to 1,000+ languages," *arXiv preprint arXiv:2305.13516*, 2023.

[28] K. Bhogale, S. Sundaresan, A. Raman, T. Javed, M. M. Khapra, and P. Kumar, "Vistaar: Diverse Benchmarks and Training Sets for Indian Language ASR," in *INTERSPEECH*, 2023.

[29] B. Talafha, A. Waheed, and M. Abdul-Mageed, "N-Shot Benchmarking of Whisper on Diverse Arabic Speech Recognition," in *INTERSPEECH*, 2023.

[30] G. Sun, X. Zheng, C. Zhang, and P. C. Woodland, "Can Contextual Biasing Remain Effective with Whisper and GPT-2?" In *INTERSPEECH*, 2023.

[31] P. Kawa, M. Plata, M. Czuba, P. Szymański, and P. Syga, "Improved DeepFake Detection Using Whisper Features," in *INTERSPEECH*, 2023.

[32] M. Charola, A. Kachhi, and H. A. Patil, "Whisper Encoder features for Infant Cry Classification," in *INTERSPEECH*, 2023.

[33] Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-AT: Noise-Robust Automatic Speech Recognizers are Also Strong General Audio Event Taggers," in *INTERSPEECH*, 2023.

[34] M. Wang, Y. Li, J. Guo, *et al.*, "WhiSLU: End-to-End Spoken Language Understanding with Whisper," in *INTERSPEECH*, 2023.

[35] P. Peng, B. Yan, S. Watanabe, and D. Harwath, "Prompting the hidden talent of web-scale speech models for zero-shot task generalization," *arXiv preprint arXiv:2305.11095*, 2023.

[36] D.-C. Lyu, T.-P. Tan, E.-S. Chng, and H. Li, "Mandarin–english code-switching speech corpus in south-east asia: Seame," *Language Resources and Evaluation*, 2015.

[37] H. Lovenia, S. Cahyawijaya, G. I. Winata, *et al.*, "Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation," *arXiv preprint arXiv:2112.06223*, 2021.

[38] X. Shi, Q. Feng, and L. Xie, "The asru 2019 mandarin-english code-switching speech recognition challenge: Open datasets, tracks, methods and results," *arXiv preprint arXiv:2007.05916*, 2020.

[39] *ESPnet2-SEAME-recipe*, https://github.com/espnet/espnet/tree/master/egs2/seame/asr1, Accessed: 2023-11-27.

[40] D. Wu, B. Zhang, C. Yang, *et al.*, *U2++: Unified two-pass bidirectional end-to-end model for speech recognition*, 2021. arXiv: 2106.05642.