Language Awareness in Code-Switching Speech Recognition

Jen-Tzung Chien Bobbi Aditya
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
E-mail: {jtchien, bobbiaditya.ee10}@nycu.edu.tw

***Abstract*—Code-switching speech is observed when two or more languages are spoken in a single utterance. This phenomenon is common in multilingual speech applications. The ability to recognize code-switching speech is essential to develop a bilingual speech recognition system. While the pre-trained multilingual model such as Whisper has shown promising in recognizing monolingual speech from different languages, the previous works in handling code-switching speech are suboptimal due to the language mixing and inaccuracy. This circumstance, existing in attention maps, is tackled through a new strategy, which is fulfilled through a two-stage optimization to ensure Whisper to be aware of language of each word by attending its correct language identity. The experiments on Mandarin-English code-switching speech using SEAME, ASCEND and NTUT show that this method achieves the mixed error rates of 13.8%, 26.7% and 11.8%, surpassing the other methods, respectively, with very few additional adapter parameters.**

I. Introduction

In the multilingual community, it is very common for individuals to use multiple languages in their daily life. One particular scenario that people seamlessly speak and interchange languages in a single utterance, namely the code-switching speech [1], [2], is known as a challenging issue for automatic speech recognition (ASR) in presence of multilingual speakers [3]. In general, code-switching speech can be either inter-sentence or intra-sentence [4], whereas the later is a more common setting where a single sentence contains more than one language. To achieve a desirable performance in code-switching ASR, utilizing the encoder-decoder framework has been a popular choice. In recent years, there has been a rapid advancement in foundation model for multilingual ASR such as XLSR [5], Whisper [6], USM [7], and MMS [8]. Those foundation models have been pre-trained via supervised and unsupervised methods by using a large corpus of audio data and their transcriptions, containing numerous languages. Nevertheless, the effectiveness of these backbone models in handling code-switching inputs remains limited and uncertain, despite their proficiency in capturing inter-sentence speech from multiple languages. In [9], the multilingual language models were employed in code-switching text representation. This study deals with code-switching representation to build multilingual speech and text models for speech recognition.

Recently, the studies on code-switching ASR have been highly attended, and can be categorized into either utilizing a shared encoder for similar representations or adopting the separate encoders for distinct representations. In [10], a shared encoder, integrated with a language identifier, showed promising results. In [11], an emphasis on encouraging the similarity in embeddings within a monolingual language was addressed. In [12], a shared encoder was extended to build individual self attention networks [13], [14] for individual languages. Also, separate encoders [15], [16] were explored for different languages by combining with the transformer-transducer and CTC models [3], [17]. The results on code-switching ASR were improved, but the additional memory and computation were required.

This paper introduces a novel approach to code-switching ASR by leveraging the pre-trained multilingual backbone model where the adapters are implemented. The proposed method focuses on introducing language identification (LID) in attention maps to address the issue of language awareness in the model. This strategy tends to make the model translate the transcribed tokens into their right languages which considerably improve the performance. This study develops such a solution to improve the code-switching performance based on Whisper [18]. A behavioral analysis of the attention maps within transformer decoder is conducted to design a scheme to reflect language-aware behavior in multilingual ASR. Accordingly, a head selection method is introduced to identify the attention heads for LID tokens. An LID task is fulfilled and imposed to those selected heads to ensure a correct LID for each word in a transcribed sequence. Furthermore, a two-stage adapter training process is designed to facilitate a stable and parameter-efficient learning. The experimental results on code-switching Mandarin and English speech recognition over two datasets demonstrate the efficiency and effectiveness of using the proposed language-aware code-switching method.

II. Language Awareness in Whisper

Whisper stands out as a state-of-the-art multilingual ASR model under an encoder-decoder architecture in a transformer. Trained on an extensive dataset containing 680,000 hours of labeled audio data, this model is distinguished from the earlier models like MMS and XLSR [8], [5], which were solely constructed as the pre-trained speech encoders. The significant enhancement in ASR performance was attributed to the inclusion of a pre-trained decoder in using Whisper model, surpassing over the previous benchmark tasks through different speech models. Utilizing the training data $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$, where $\mathbf{x}$ denotes the input speech and $\mathbf{y} = \{\mathbf{y}_n\} = \{y_{nm}\}$ denotes a set of N one-hot vectors in a target transcription with M

979-8-3315-7206-8/25/$31.00

vocabulary words defined in a dictionary, an encoder-decoder model $f_{\theta,\phi}$ can be trained with a frozen Whisper backbone model θ combined by the controllable adapter ϕ [19], [20] as depicted in Figure 1 where sinusoidal and learned positional encodings are merged in encoder and decoder, respectively. This model is constructed by minimizing the classification (CL) loss given by a cross-entropy error function for supervised training for the adapter parameter ϕ

$$\mathcal{L}_{\text{CL}}(\mathbf{x}, \mathbf{y}; \phi) = -\log p(\mathcal{D}|\phi) = -\sum_{n=1}^{N}\sum_{m=1}^{M} y_{nm} \log p_{nm} \quad (1)$$

where $p_{nm} \triangleq p(y_{nm}|f_{\theta,\phi}(\mathbf{y}_1, \ldots, \mathbf{y}_{n-1}, \mathbf{x}))$ denotes the encoder-decoder output for a token $\mathbf{y}_n$ given its history $\{\mathbf{y}_j\}_{j=1}^{n-1}$ with an input speech signal $\mathbf{x}$.

Typically, Whisper has a special input prompt, denoted as ⟨sot⟩⟨lid⟩⟨trans⟩⟨nots⟩, where ⟨sot⟩ and ⟨nots⟩ denote "start of transcription" and "no time stamps", respectively. Here, ⟨trans⟩ is used to denote the transcription task for ASR. Whisper is feasible to develop different multilingual applications due to the LID token ⟨lid⟩. In [21], the application of Whisper was extended to address the challenge of bilingual code-switching ASR through adding the corresponding LID tokens for two languages. This adaptation led to a notable performance improvement. In order to handle the Mandarin-English code-switching speech recognition, the prompt input for bilingual ASR become ⟨sot⟩⟨lid1⟩⟨lid2⟩⟨trans⟩⟨nots⟩ consisting of two LID tokens for Chinese ⟨zh⟩ and English ⟨en⟩.

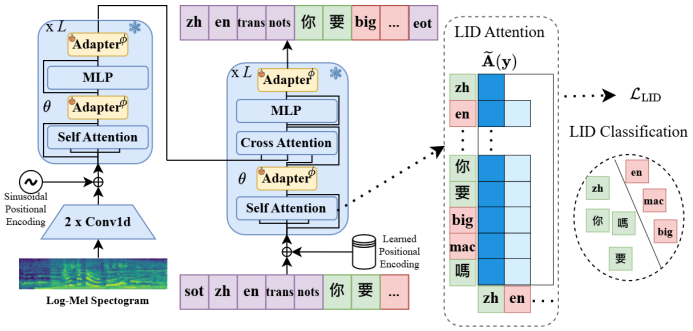

Fig. 1. (left) Adapters ϕ (tunable) are configured in an encoder-decoder framework based Whisper model θ (frozen) with L transformer blocks for code-switching ASR. (right) LID classification is performed in decoder to mitigate language mixing or inaccuracy in Whisper where the self attention map $\tilde{\mathbf{A}}(\mathbf{y})$ from bilingual word sequence $\mathbf{y}$ is shaped to be aware of different languages. An LID loss is imposed and minimized to strengthen language awareness so that each word of a language attends both LID tokens ⟨zh⟩ and ⟨en⟩ for Chinese and English, respectively.

Notably, the settings of LID tokens result in the limitation in Whisper's ability to identify the languages of individual word tokens. The recognition results tend to produce the word tokens without precise language awareness, revealing the lack of using Whisper model to handle the code-switching settings seamlessly. As illustrated in Table I, there are two examples of Whisper transcriptions by using three different settings of LID tokens ⟨lid⟩ where their ground-truth sentences are provided. In the first example, using only Chinese LID

TABLE I

LANGUAGE MIXING OR INACCURACY EXISTS IN CODE-SWITCHING ASR RESULTS USING WHISPER. THREE KINDS OF RESULTS BY USING LID TOKENS ⟨LID⟩ (⟨ZH⟩, ⟨EN⟩ AND ⟨ZH⟩⟨EN⟩) ARE COMPARED. THESE RESULTS TEND TO TRANSLATE SPEECH TO FOLLOW LID TOKENS. THE GROUND-TRUTH EXAMPLES FROM SEAME DATASET "INDONESIANS會比較靠近 (INDONESIANS WILL BE MORE CLOSE)" AND "我住高文THAT SIDE (I LIVE IN GOWEN THAT SIDE)" ARE SHOWN. ORIGINAL WHISPER IS WEAK IN LANGUAGE AWARENESS AS SEEN IN THE RESULTS OF THE THIRD LID SETTING ⟨ZH⟩⟨EN⟩.

LID	ASR Results
Example 1	*Indonesians*會比較靠近 (ground truth)
⟨zh⟩	印度尼斯會比較靠近
⟨en⟩	Indonesia will be more close
⟨zh⟩⟨en⟩	Indonesia will be more close
Example 2	我住高文*that side* (ground truth)
⟨zh⟩	我住高文在那邊
⟨en⟩	I'll go to Gowen that side
⟨zh⟩⟨en⟩	我住高文deadside

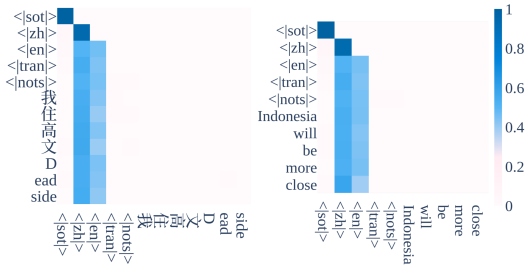

Fig. 2. Examples of showing the self attention maps in decoder corresponding to LID tokens ⟨zh⟩ (left) and ⟨en⟩ (right) where all tokens are attended by both LID tokens. Head $h = 11$ in layer $\ell = 7$ is evaluated.

token ⟨zh⟩, the ASR result comes out the transcription to be all Chinese tokens where the English word "Indonesians" is translated into Chinese tokens 印度尼斯(Yìndù nísī). On the other hand, under the setting of only using English LID token ⟨zh⟩, the model is changed to produce the translation to be all English tokens where the Chinese characters are correctly changed to "will be more close". Even after using two LID tokens ⟨zh⟩⟨en⟩ in the prompt, the model is still unable to detect the switching languages properly, resulting in a wrong transcription which is all English tokens. These results reflect the issue of language mixing or inaccuracy [22], [23]. Whisper model could not understand the correct LID for individual word tokens where the decoder needs to generate. The second example shows another case that using two LID tokens generates a code-switching transcribed sentence but the English tokens "that side" are wrongly recognized as "deadside". To further examine the limitation of Whisper, this study conducts an analysis on the attention map of each head in a decoder layer. An intriguing observation emerges from the analysis of attention patterns, revealing a distinctive pattern of attention values corresponding to two LID tokens. As shown in Figure 2, there is a language mixing in the attention map where each word token equally attends both LID tokens without considering true language of each word. This paper is motivated to address this issue by proposing an

additional LID task in a multi-task learning. In addition to minimizing the CL loss in Eq. (1) for speech recognition, the goal of code-switching ASR is to ensure a correct LID for each word such that the bilingual ASR decoder is allowed to sensitively attend language differences so as to improve the code-switching performance.

III. Language-Aware Code-Switching ASR

First of all, self attention map $\mathbf{A}(\mathbf{y}) \in \mathbb{R}^{N\times N}$ of a head in transformer decoder is calculated from a target sequence $\mathbf{y}$ via

$$\mathbf{A}(\mathbf{y}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\right),\ \text{softmax}(\mathbf{z})_n = \frac{e^{z_n}}{\sum_{j=1}^N e^{z_j}} \quad (2)$$

where softmax function is calculated for each token n in a vector $\mathbf{z}$ from a product matrix due to query or key matrices $\{\mathbf{Q},\mathbf{K}\} \in \mathbb{R}^{N\times d}$, and d is the dimension of individual queries and keys in $\{\mathbf{Q},\mathbf{K}\}$ [24].

A. Language-Aware Attention Heads

It is important to identify the attention heads that characterize language information of a code-switching speech utterance within the model. LID task is implemented by first selecting the attention maps in decoder to reflect LID information and then learning the adapters ϕ to distinguish languages of individual words in attention map $\mathbf{A}(\mathbf{y})$ through LID tokens $\langle\text{zh}\rangle$ and $\langle\text{en}\rangle$. The language-aware attention heads are selected and then adapted to correctly identify the language of each token in a code-switching speech by minimizing the LID loss $\mathcal{L}_{\text{LID}}$. Let $\{\mathbf{A}_{\ell h}\}$ denote the individual attention maps from L layers where each layer has H heads. Notably, the Whisper backbone θ is required to calculate $\mathbf{A}_{\ell h}$. To identify those language-aware attention maps $\mathbf{A}_{\ell h}$, N-dimensional matrix $\mathbf{A} = [A_{nj}]_{N\times N}$ with entry value $A_{ij}(\mathbf{y})$ for self attention [25] in decoder between word tokens i and j is calculated from $\mathbf{y}$. In particular, this paper introduces the language-aware indicator function [26]

$$\mathbb{I}(\mathbf{A}) = \begin{cases} 1, & \text{if} \sum_{n=1}^N \sum_{j\in\Xi} A_{nj}(\mathbf{y}) > \sum_{i=n}^N \sum_{k\in\overline{\Xi}} A_{nk}(\mathbf{y}) \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $\Xi = \{\langle\text{zh}\rangle, \langle\text{en}\rangle\}$ denotes the LID tokens $\langle\text{lid}\rangle$ in a code-switching setting and $\overline{\Xi}$ denotes the remaining $N-2$ tokens which are irrelevant to LID tokens. This indicator detects the basic requirement of a language-aware attention map where the attention values of LID tokens are larger than those of the remaining tokens. The set of language-aware attention maps $\widetilde{\mathbf{A}}$ is then formed in accordance with $\mathbb{I}(\mathbf{A}) = 1$ over all word sequences $\mathbf{y}$ in training data $\mathcal{D}$. In particular, the self attention heads in the decoder layers $\{\mathbf{A}_{\ell h}\}$, selected with top k highest counts through a forward pass over all target sentences $\mathbf{y}$ in $\mathcal{D}$, are determined to setup the set of language-aware attention maps in a way of

$$\widetilde{\mathbf{A}}(\mathbf{y}) = \{\mathbf{A}_{\ell h} | (\ell,h) \in \arg\text{top}\, k_{(\ell',h')} \textstyle\sum_{\mathbf{y}\in\mathcal{D}} \mathbb{I}(\mathbf{A}_{\ell' h'})\}. \quad (4)$$

The hyperparameter k is selected to tune the proportion of the selected $\mathbf{A}_{\ell h}$ from L layers with H heads in each layer.

B. Language-Aware Learning

Next, LID task to deal with language mixing is performed via adapter learning where each token of the resulting language-aware attention maps in $\widetilde{\mathbf{A}}(\mathbf{y})$ is classified to have a correct LID. In case of Mandarin-English code-switching ASR, each word token is classified into Chinese or English. This LID task is carried out by minimizing the LID loss, which is an accumulated cross-entropy between ground-truth LIDs and attention values over those language-aware attention maps $\widetilde{\mathbf{A}}$ along N tokens in word sequence $\mathbf{y}$ from training data $\mathcal{D}$

$$\mathcal{L}_{\text{LID}}(\mathbf{y};\phi_d) = -\sum_{\mathbf{y}\in\mathcal{D}} \sum_{\mathbf{A}\in\widetilde{\mathbf{A}}(\mathbf{y})} \sum_{n=1}^N \sum_{j=1}^N y_{nj}^{\langle\text{lid}\rangle} \log A_{nj} \quad (5)$$

which is the cross-entropy loss for language identification, $y_{nj}^{\langle\text{lid}\rangle}$ is the ground-truth label indicating the correct LID of each token $\mathbf{y}_n$ in a word sequence $\mathbf{y}$, and A_{nj} is the corresponding attention output from the selected LID attention heads $\mathbf{A}_{\ell h}$ in $\widetilde{\mathbf{A}}(\mathbf{y})$. This LID task is seen as a binary Mandarin-English classification. LID loss is minimized to guide the learned adapter ϕ_d in the decoder to properly attend the language changes during word prediction. Through precisely identifying the language of each word, this model is aware of different languages when predicting the word sequence $\mathbf{y}$. The language awareness is strengthened to elevate its code-switching capability in an ASR system.

C. Code-Switching Adapter Learning

In this study, a parameter-efficient learning is performed while the learning of adapters [27] is involved in both encoder ϕ_e and decoder ϕ_d. The backbone transformer θ based on Whisper is frozen. The additional language identification [28] is incorporated through an optimization over a classification task from all training samples $\mathcal{D} = \{\mathbf{x},\mathbf{y}\}$ by solving

$$\{\phi_e,\phi_d\} = \arg\min_{\{\phi_e,\phi_d\}} \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\mathcal{D}} \left[\gamma_c \mathcal{L}_{\text{CL}}(\mathbf{x},\mathbf{y};\phi) + \gamma_l \mathcal{L}_{\text{LID}}(\mathbf{y};\phi_d)\right] \quad (6)$$

where $\phi = \{\phi_e,\phi_d\}$ denotes the adapter parameters and $\{\gamma_c,\gamma_l\}$ denotes the regularization parameters to balance between classification loss and language awareness loss. Adapter ϕ is introduced to enable code-switching capability in using Whisper model without the need for extensive fine-tuning of the entire model. The optimization procedure follows a two-stage approach inspired by [29], [30]. Rather than directly estimating the encoder ϕ_e and decoder ϕ_d in a hybrid model, this paper performs a two-stage learning procedure for a stable learning process. After the parameter initialization, the first stage focuses on estimating the adapter encoder ϕ_e by minimizing $\mathcal{L}_{\text{CL}}$. The Whisper backbone θ is utilized while the adapter decoder ϕ_d is disregarded. Once a converged encoder ϕ_e is estimated, the second stage involves a joint training of ϕ_e and ϕ_d by minimizing both the classification loss $\mathcal{L}_{\text{CL}}$ and the language identification loss $\mathcal{L}_{\text{LID}}$ where the backbone Whisper θ is sufficiently utilized. Language awareness implemented by imposing $\mathcal{L}_{\text{LID}}$ serves as a constraint for adapter learning to

address the code-switching mechanism in speech for bilingual ASR. Algorithm 1 shows a two-stage adapter learning where LID task is merged to enhance language demixing in language-aware code-switching speech recognition.

Algorithm 1: Adapter learning for language-aware (LA) code-switching speech recognition

Require: speech-text samples $\mathcal{D} = \{\mathbf{x}, \mathbf{y}\}$, Whisper backbone θ, adapter encoder ϕ_e, adapter decoder ϕ_d, weight γ_c, weight γ_l

fix θ and initialize ϕ_e, ϕ_d

select language-aware attention maps $\widetilde{\mathbf{A}}$

stage 1: learning for encoder ϕ_e

for *each batch* $(\mathbf{x}, \mathbf{y})$ *in* $\mathcal{D}$ **do**

compute $\mathcal{L} = \mathcal{L}_{\text{CL}}(\mathbf{x}, \mathbf{y}; \phi_e)$

compute $\nabla_{\phi_e}\mathcal{L}$ to update ϕ_e

end

stage 2: learning for encoder ϕ_e and decoder ϕ_d

for *each batch* $(\mathbf{x}, \mathbf{y})$ *in* $\mathcal{D}$ **do**

compute $\mathcal{L}_{\text{CL}}(\mathbf{x}, \mathbf{y}; \phi)$ & $\mathcal{L}_{\text{LID}}(\mathbf{y}; \phi_d)$

combine $\mathcal{L} = \gamma_c\mathcal{L}_{\text{CL}}(\mathbf{x}, \mathbf{y}; \phi_e, \phi_d) + \gamma_l\mathcal{L}_{\text{LID}}(\mathbf{y}; \phi_d)$

compute $\nabla_{\phi_e}\mathcal{L}$ & $\nabla_{\phi_d}\mathcal{L}$ to update ϕ_e & ϕ_d

end

IV. Experiments

The experiments were conducted for bilingual code-switching speech recognition under an end-to-end speech processing using ESPnet toolkit [31] where three datasets were used. First, a Mandarin-English code-switching speech corpus in South-East Asia (SEAME) [32] from Singaporean and Malaysian speakers was collected. This dataset consists of 192 hours of Mandarin-English code-switching speech from 156 speakers. Second, ASCEND was collected with the Mandarin-English code-switching conversational speech with China accent consisting of 10.6 hours of clean speech from 23 bilingual speakers [33]. In addition, the NTUT-AB01 [34] dataset, which is a Taiwan-accent Mandarin-English code-switching speech dataset containing 376 minutes of speech utterances, was utilized. Low-resource setting was evaluated.

A. Experimental Settings

This study utilized the Whisper-small model with $H = 12$ heads and $L = 12$ transformer layers in both encoder and decoder. The dimension size of adapter was set to 192. Regularization parameters $\{\gamma_c = 1, \gamma_l = 0.01\}$ were selected. AdamW optimizer [35] with a learning rate of 1×10^{-3} was employed. Each training stage consisted of 15 epochs, and the final model was obtained by averaging the weights over the three best models, based on the validation loss from the best epoch. This paper compared three settings for the proposed language awareness in adapter learning: (a) one-stage adapter (denoted as Adapter-1), (b) one-stage language-aware adapter (denoted as LangAw-1), and (c) two-stage language-aware adapter (denoted as LangAw-2). A recent work, which merged an attention guidance scheme in an adapter learning with two stages (denoted as AttGuid-2) [26], was also included in the comparison. The evaluation was conducted on the devman (DM) set (dominated by Mandarin speech) and devsge (DS) set (dominated by Singaporean English speech) of SEAME. Using SEAME and ASCEND, the evaluation for monolingual utterances (both Mandarin and English) and code-switching utterances was performed. Mixed error rate (MER) measured the performance for code-switching speech utterances and the overall MER (OMER) measured the overall results for code-switching utterances as well as monolingual utterances for Mandarin and English. NTUT-AB01 contained only code-switching speech with the settings as referred in [34]. Only the results of MER were reported. The parameter size and the training time in hours were reported. The computation time was measured by a personal computer with a single GPU using GeForce RTX 3090. The selection of language-aware attention maps in $\widetilde{\mathbf{A}}(\mathbf{y})$ via hyperparameter k was implemented by selecting those maps from heads h in layers ℓ under top 70% of ranked values of $\sum_{\mathbf{y}\in\mathcal{D}} \mathbb{I}(\mathbf{A}_{\ell h})$.

TABLE II
Comparison of MER (%) for code-switching speech, overall MER (OMER) (%) for monolingual and code-switching speech, parameter size and training time in hours for bilingual ASR on devman (DM) and devsge (DS) of SEAME.

Method	Set	MER	OMER	Param	Time
Original Prompt [21]	DM	38.2	38.2	–	–
	DS	56.4	65.0		
Modified Prompt [21]	DM	33.4	32.7	–	–
	DS	49.6	47.6		
SOTA [31], [21]	DM	–	16.6	47.3M	–
	DS	–	23.3		
AttGuid-2 [26]	DM	13.5	14.2	14.3M (5.6%)	9.91
	DS	18.9	20.8		
Adapter-1	DM	14.4	15.1	14.3M (5.6%)	5.10
	DS	19.7	21.6		
LangAw-1	DM	14.2	14.7	14.3M (5.6%)	5.41
	DS	19.1	21.2		
LangAw-2	DM	12.9	13.8	14.3M (5.6%)	9.51
	DS	18.0	19.9		

B. Experimental Results

Table II shows the results of various code-switching ASR methods in different metrics. The baseline is based on the findings in [21], showing both the original and modified prompts by using the backbone Whisper. State-of-the-art (SOTA) result, following the ESPnet SEAME recipe [31], [21], is included. The results of one-stage adapter training without (Adapter-1) and with (LangAw-1) LID loss are listed, demonstrating the effectiveness of introducing LID loss and showing overall improvement compared to one-stage adapter training without LID loss. Furthermore, introducing a two-stage training (LangAw-2) scheme enhances the results, as indicated by the overall MER. Notably, two-stage adapter with LID loss (LangAw-2) outperforms the methods in [21], [31] across all metrics while requiring significantly fewer trainable parameters. In comparison with the backbone Whisper, the adapters with LID loss

only account for 5.6% of total parameters. Although the MERs based on one-stage adapter learning in Adapter-1 and LangAw-1 are not as low as those of using two-stage adapter learning in AttGuid-2 and LangAw-2, the computation costs using one-stage learning are much smaller than those of using two-stage learning. Applying the same trick of two-stage training, the proposed LangAw-2 obtains lower MER and OMER than the recent work AttGuid-2 in [26] with even smaller amount of training time. This experiment indicates the importance of language identification task in code-switching setting to deal with the issue of language mixing in continuous speech. Among different methods, the lowest overall MER 13.8% is obtained by using LangAw-2. This result is considerably better than 16.6% in [31], [21] and 14.2% in [26].

TABLE III
COMPARISON OF WER (%), CER (%), MER (%) AND OMER (%) USING DIFFERENT METHODS ON VALIDATION AND TEST SETS OF ASCEND.

Method	Set	WER	CER	MER	OMER
Modified Prompt [21]	valid	47.0	33.6	45.6	37.6
	test	48.9	36.0	49.0	43.9
Baseline [33]	valid	–	–	–	25.7
	test				27.0
AttGuid-2 [26]	valid	36.2	20.8	23.3	23.5
	test	37.9	25.9	26.5	27.1
Adapter-1	valid	39.4	22.0	23.6	24.4
	test	38.1	26.7	28.6	27.9
LangAw-1	valid	37.1	20.5	23.2	23.7
	test	37.5	26.0	27.4	27.2
LangAw-2	valid	35.0	20.2	23.0	23.2
	test	37.0	24.6	25.8	26.7

Table III shows the bilingual ASR results on the validation and test sets of ASCEND dataset where the monolingual speech of English and Chinese as well as the code-switching speech via WER, CER, MER and OMER are evaluated by using different methods including modified prompt [21], AttGuid-2 [26], Adapter-1, LangAw-1 and LangAw-2 based on the Whisper model. The baseline system [33], utilizing wav2vec 2.0 [36] as the pre-trained speech encoder, was included in the comparison. The monolingual speech data in ASCEND are limited. From the results, it is found that introducing two-stage methods in AttGuid-2 and LangAw-2 basically stabilizes the learning process and generally works better than one-stage methods in Adapter-1 and LangAw-1. In this comparison, the lowest overall MER 26.7% is achieved by using the proposed LangAw-2.

TABLE IV
COMPARISON OF MER (%) USING DIFFERENT METHODS ON NTUT.

Method	MER	Param	Time
Baseline (XLSR) [34]	28.7	–	–
Original Prompt (Whisper) [21]	21.6	–	–
Modified Prompt (Whisper) [21]	20.7	–	–
AttGuid-2 (Whisper) [26]	12.0	14.3M	0.62
Adapter-1 (Whisper)	12.6	14.3M	0.32
LangAw-1 (Whisper)	12.0	14.3M	0.33
LangAw-2 (Whisper)	11.8	14.3M	0.59

The extended results of code-switching ASR on NTUT dataset in low-resource setting are reported in Table IV. The baseline performance is based on the findings in [34] that the language-dependent adapters were learned for code-switching speech recognition based on the cross-lingual speech representation (XLSR) [5]. For the backbone based on Whisper, introducing LID loss on one-stage adapter training (LangAw-1) does show a significant improvement relative to using the original and modified prompts [21] in terms of MER. However, combining two-stage language-aware adapter (LangAw-2) shows an overall further improvement compared with the one-stage adapter methods (Adapter-1 and LangAw-1) and previous two-stage method (AttGuid-2) [26].

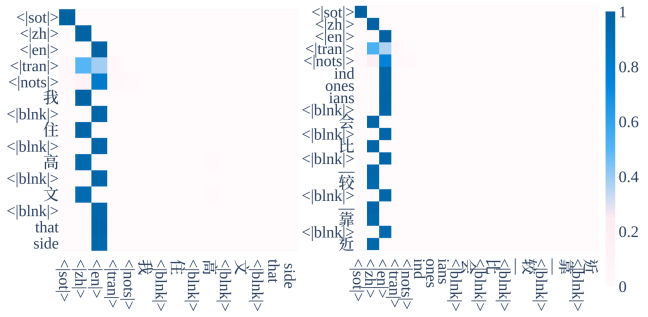

Fig. 3. Examples of showing attention maps corresponding to LID tokens on SEAME after introducing LID loss $\mathcal{L}_{\text{LID}}$. Both examples show that all of word tokens are correctly identified. The attention map of head $h = 11$ in layer $\ell = 8$ is evaluated. ⟨blnk⟩ stands for blank token.

Figure 3 shows the attention maps of two LID tokens obtained from two-stage language-aware adapter. Different from the results in Figure 2, two examples in Figure 3 show that all of word tokens are correctly identified into LID tokens ⟨zh⟩ and ⟨en⟩ by merging with the proposed LID loss. The properties of language demixing and awareness in code-switching ASR can be held. In the experiments, a notable observation is found. A substantial portion of attention heads (110 out of 144 heads) in the model fall under the category of LID token heads where $\mathbb{I}(\mathbf{A}_{\ell h}) = 1$. In the context of transformers, multiple heads in attention mechanisms are designed to capture the diverse relations within a sequence. However, this discovery highlights a significant redundancy within the model where multiple heads capture similar information.

V. CONCLUSIONS

This study has presented the scheme of language awareness and introduced an LID strategy for language demixing in code-switching ASR with two-stage adapter training. By analyzing the attention maps in the transformer decoder, this scheme was discovered as each word token was attended by both LID tokens. A head selection process was applied to identify those language-aware attention maps related to LID tokens along with the process to ensure the correct identification to guide ASR decoder to attend the code-switching in speech signal. The adapter learning of encoder and decoder was performed according to the word classification loss and the language identification loss. Experiments on SEAME, ASCEND and NTUT

datasets demonstrated the efficiency and effectiveness of the proposed method over the previous results while using the reduced trainable parameters. Future studies will be extended by conducting contrastive learning for language discrimination in code-switching ASR. An additional CTC loss will be merged in encoder to enhance the alignment between code-switching speech and bilingual text string.

REFERENCES

[1] S. Poplack, *Syntactic Structure and Social Function of Code-switching*, Centro de Estudios Puertorriqueños, City University of New York, 1978.

[2] Y.-J. Lu, J. Liu, T. Thebaud, L. Moro-Velazquez, A. Rastrow, N. Dehak, and J. Villalba, "CA-SSLR: Condition-aware self-supervised learning representation for generalized speech processing," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50126–50151, 2024.

[3] W. Chen, B. Yan, J. Shi, Y. Peng, S. Maiti, and S. Watanabe, "Improving massively multilingual asr with auxiliary CTC objectives," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[4] K. A. H. Zirker, *Intrasentential vs. Intersentential Code Switching in Early and Late Bilinguals*, Brigham Young University, 2007.

[5] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2021, pp. 2426–2430.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. of International Conference on Machine Learning*, 2023, pp. 28492–28518.

[7] Y. Zhang, W. Han, J. Qin, Y. Wang, A. Bapna, Z. Chen, N. Chen, B. Li, V. Axelrod, G. Wang, et al., "Google USM: Scaling automatic speech recognition beyond 100 languages," *arXiv preprint arXiv:2303.01037*, 2023.

[8] V. Pratap, A. Tjandra, B Shi, P. Tomasello, A. Babu, S. Kundu, A. Elkahky, Z. Ni, A. Vyas, M. Fazel-Zarandi, et al., "Scaling speech technology to 1,000+ languages," *Journal of Machine Learning Research*, vol. 25, no. 97, pp. 1–52, 2024.

[9] Z. X. Yong, R. Zhang, J. Forde, S. Wang, A. Subramonian, H. Lovenia, S. Cahyawijaya, G. Winata, L. Sutawika, J. C. B. Cruz, et al., "Prompting multilingual large language models to generate code-mixed texts: The case of south east asian languages," in *Proc. of Workshop on Computational Approaches to Linguistic Code-Switching*, 2023, pp. 43–63.

[10] Z. Qiu, Y. Li, X. Li, F. Metze, and W. M. Campbell, "Towards context-aware end-to-end code-switching speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2020, pp. 4776–4780.

[11] Y. Khassanov, H. Xu, V. T. Pham, Z. Zeng, E. S. Chng, C. Ni, and B. Ma, "Constrained output embeddings for end-to-end code-switching speech recognition with only monolingual data," in *Proc. of Annual Conference of International Speech Communication Association*, 2019, pp. 2160–2164.

[12] S. Zhang, J. Yi, Z. Tian, J. Tao, Y. T. Yeung, and L. Deng, "Reducing multilingual context confusion for end-to-end code-switching automatic speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 3894–3898.

[13] J.-T. Chien and C.-W. Wang, "Hierarchical and self-attended sequence autoencoder," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 4975–4986, 2022.

[14] J.-T. Chien and Y.-H. Chen, "Learning continuous-time dynamics with attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1906–1918, 2023.

[15] S. Dalmia, Y. Liu, S. Ronanki, and K. Kirchhoff, "Transformer-transducers for code-switched speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 5859–5863.

[16] T. Song, Q. Xu, M. Ge, L. Wang, H. Shi, Y. Lv, Y. Lin, and J. Dang, "Language-specific characteristic assistance for code-switching speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 3924–3928.

[17] J.-T. Chien and C.-K. Yeh, "Diffusion-based connectionist temporal classification," in *Proc. of IEEE International Workshop on Machine Learning for Signal Processing*, 2025.

[18] Y. Yang, Y. Peng, H. Huang, E. S. Chng, and X. Zhong, "Adapting OpenAI's Whisper for speech recognition on code-switch Mandarin-English SEAME and ASRU2019 datasets," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2024, pp. 1–6.

[19] J.-T. Chien and W.-Y. Sun, "Adversarial augmentation for adapter learning," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–7.

[20] L.-J. Yang and J.-T. Chien, "Continual gated adapter for bilingual codec text-to-speech," in *Proc. of Oriental COCOSDA*, 2024, pp. 1–6.

[21] P. Peng, B. Yan, S. Watanabe, and D. Harwath, "Prompting the hidden talent of web-scale speech models for zero-shot task generalization," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, pp. 396–400.

[22] H. Liu, L. P. Garcia, X. Zhang, A. W. Khong, and S. Khudanpur, "Enhancing code-switching speech recognition with interactive language biases," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10886–10890.

[23] 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 *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[24] J.-T. Chien and Y.-H. Huang, "Latent semantic and disentangled attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 12, pp. 10047–10059, 2024.

[25] J.-T. Chien and Y.-H. Chen, "Continuous-time self-attention in neural differential equation," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2021, pp. 3290–3294.

[26] B. Aditya, M. Rohmatillah, L.-H. Tai, and J.-T. Chien, "Attention-guided adaptation for code-switching speech recognition," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10256–10260.

[27] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP," in *Proc. of International Conference on Machine Learning*, 2019, pp. 2790–2799.

[28] C. Zhang, B. Li, T. Sainath, T. Strohman, S. Mavandadi, S.-Y. Chang, and P. Haghani, "Streaming end-to-end multilingual speech recognition with joint language identification," in *Proc. of Annual Conference of International Speech Communication Association*, 2022, pp. 3223–3227.

[29] M. Rohmatillah and J.-T. Chien, "Hierarchical reinforcement learning with guidance for multi-domain dialogue policy," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 748–761, 2023.

[30] M. Rohmatillah and J. T. Chien, "Corrective guidance and learning for dialogue management," in *Proc. of ACM International Conference on Information & Knowledge Management*, 2021, p. 1548–1557.

[31] 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 *Proc. of Annual Conference of International Speech Communication Association*, 2018, pp. 2207–2211.

[32] D.-C. Lyu, T.-P. Tan, E. S. Chng, and H. Li, "SEAME: A Mandarin-English code-switching speech corpus in south-east Asia," in *Proc. of Annual Conference of International Speech Communication Association*, 2010, pp. 1986–1989.

[33] H. Lovenia, S. Cahyawijaya, G. Winata, P. Xu, Y. Xu, Z. Liu, R. Frieske, T. Yu, W. Dai, E. J. Barezi, Q. Chen, X. Ma, B. Shi, and P. Fung, "ASCEND: A spontaneous Chinese-English dataset for code-switching in multi-turn conversation," in *Proc. of Language Resources and Evaluation Conference*, 2022, pp. 7259–7268.

[34] C.-Y. He and J.-T. Chien, "Learning adapters for code-switching speech recognition," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2023, pp. 344–349.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of International Conference on Learning Representations*, 2019.

[36] 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. 12449–12460, 2020.