Probabilistic Language-Aware Speech Recognition

Jen-Tzung Chien Willianto Sulaiman Chung-Hsuan Wang
Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Hsinchu, Taiwan
E-mail: {jtchien, willianto4300.ee07, chwang}@nycu.edu.tw

***Abstract*—Code-switching scheme aims to imitate how speakers alternately switch different languages during speech communication in a multilingual community. However, building an efficient code-switching speech recognition remains a challenging issue, primarily due to the language confusion, which degrades the recognition performance. To address this issue, a novel probabilistic language-aware speech recognition is developed through a parameter-efficient learning where the tunable adapters are configured in a frozen Whisper backbone. In particular, this study incorporates a language-aware calibrator at the prediction stage, introducing the language identity as an additional condition for next token prediction. By merging such a language awareness in conditional generation, the model's ability to distinguish between languages in a code-switching speech is significantly improved. The results on Chinese-English code-switching dataset using SEAME show a 4.9% relative improvement in mixed error rate by using the proposed method with only 5.1% fine-tuned parameters relative to a strong baseline with full fine-tuned parameters.**

I. Introduction

Recent advances in speech recognition have progressed at a remarkable pace, evolving from early models like hidden Markov models combined with statistical n-grams to the sophisticated approaches such as recurrent neural network transducers [1] and long short-term memory networks [2]. Nowadays, state-of-the-art models leverage the transformers [3] to achieve a performance in near-human level. These advancements underscore the growing importance of speech recognition models capable of accurately transcribing speech for a variety of emerging applications, taking empathetic spoken dialogue system [4] as an example. However, most researches on speech recognition have focused on the monolingual models, which are designed to handle the transcription task in a single language. These models primarily aim to capture as much information as possible from the given speech input. Notable examples of recent advances in this area include the conformer [5], which merged the convolution layers into a transformer-based encoder serially, the branchformer [6], which integrated the convolution layers in parallel, and the E-branchformer [7], which further enhanced the branchformer by improving the integration with additional convolution layer. Due to the superior performance of the transformers-based models [8], the current focus of speech recognition research has been shifted from a monolingual to a multilingual setting.

Multilingual setting aims to train a model to handle multiple languages within a single system, opening up an avenue to the generalized approaches to speech recognition. Notable examples were based on XLSR [9] and Whisper [10]. XLSR leveraged Wav2Vec 2.0 [11] as its backbone, fine-tuning the model to be adaptive to a multilingual setting. Whisper, on the other hand, employed a transformer-based encoder-decoder architecture geared with a special prompting format during training, allowing for multitasking and multilingual training to enhance the model capability for speech recognition. Despite these advancements, current multilingual speech recognition models still face significant challenges in accurately transcribing the code-switching speech where multiple languages are spoken within a single utterance [12]. The major challenge in the code-switching setting is the language confusion problem where the model could not distinguish different languages within the input utterance [13], [14].

Previous research has tackled the challenge of language confusion by aligning the predictions for languages with those for tokens [14] or by concatenating the embeddings of speech and language for information fusion in speech recognition and language prediction [13]. Other approaches introduced separate models for individual languages through combining their embeddings [15], [16] or utilized the pretrained multilingual model by modifying the language identity (LID) prompt during the prediction for code-switching setting [17], [18], [19]. Although these methods have improved the code-switching speech recognition, this study argues that language confusion could be further addressed by integrating language condition in an optimization problem for language-aware speech recognition based on variational inference. Based on this argument, this paper presents a new code-switching speech recognition by implementing language awareness in a pre-trained multilingual speech recognition model from a probabilistic perspective during the fine-tuning process based on adapters. Specifically, Whisper is utilized as the backbone, which is augmented with an additional language-aware calibrator and integrated to adjust the next-token prediction depending on language identification. Additionally, parameter-efficient fine-tuning scheme using standard adapter [20] and low-rank adapter (LoRA) [21] are employed within Whisper model to promote efficient adaptation. The experimental results demonstrate the importance of incorporating language awareness in code-switching speech recognition based on the adapters as well as the calibrator.

II. Probabilistic Language-Aware ASR

In general, an automatic speech recognition (ASR) model is built by maximizing the conditional log likelihood $\log p(\boldsymbol{y}|\boldsymbol{x})$ where $\boldsymbol{y}$ denotes the word sequence or the text transcription, and $\boldsymbol{x}$ denotes the speech signal. However, such a model cannot be directly employed in a code-switching setting due to the language confusion when predicting the token identity [13]. Language confusion in an ASR system is basically known as

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

an issue that ASR could not correctly predict the changing languages given its code-switching speech utterance. Previous works have introduced some methods to tackle language confusion for code-switching speech recognition, for example, by aligning the attention maps corresponding to the language identity (LID) special tokens in the Whisper model [17], [18], [19] or treating different languages by introducing separate encoders [15], [16]. These works showed that language needs to be a condition for predicting the next token. Therefore, language-aware speech recognition is formulated to build a model which is trained by maximizing the log conditional likelihood [22]

$$\begin{aligned}\log p(\boldsymbol{y}|\boldsymbol{x}) &= \log \sum_{\boldsymbol{\ell}} p(\boldsymbol{y},\boldsymbol{\ell}|\boldsymbol{x}) = \log \sum_{\boldsymbol{\ell}} \frac{p(\boldsymbol{y},\boldsymbol{\ell}|\boldsymbol{x})q(\boldsymbol{\ell}|\boldsymbol{x})}{q(\boldsymbol{\ell}|\boldsymbol{x})} \\ &= \log \mathbb{E}_{q(\boldsymbol{\ell}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{y},\boldsymbol{\ell}|\boldsymbol{x})}{q(\boldsymbol{\ell}|\boldsymbol{x})}\right] \geq \mathbb{E}_{q(\boldsymbol{\ell}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{y},\boldsymbol{\ell}|\boldsymbol{x})}{q(\boldsymbol{\ell}|\boldsymbol{x})}\right] \\ &= \mathbb{E}_{q_\phi(\boldsymbol{\ell}|\boldsymbol{x})}\left[\log p_{\theta,\psi}(\boldsymbol{y}|\boldsymbol{\ell},\boldsymbol{x})\right] - \mathcal{D}_{\text{kl}}(q_\phi(\boldsymbol{\ell}|\boldsymbol{x})\|p(\boldsymbol{\ell}|\boldsymbol{x})) \triangleq -\mathcal{L}_{\text{vb}}\end{aligned} \tag{1}$$

where the language sequence $\boldsymbol{\ell}$ is seen as the latent variable underlying the observed speech signal $\boldsymbol{x}$, $q_\phi(\boldsymbol{\ell}|\boldsymbol{x})$ is merged as a variational distribution of latent variable $\boldsymbol{\ell}$, $\mathcal{D}_{\text{kl}}$ is the Kullback-Leiblier (KL) divergence, and $\mathcal{L}_{\text{vb}}$ is defined as the variational upper bound of the negative log likelihood $-\log p(\boldsymbol{y}|\boldsymbol{x})$. Maximizing the language-aware log likelihood turns out to minimizing the variational upper bound of negative log likelihood $\mathcal{L}_{\text{vb}}$. It is noted that the code-switching ASR does not only predict the word tokens $\boldsymbol{x}$ but also their language identities $\boldsymbol{\ell}$. Language confusion is caused due to the inability of the model to predict $\boldsymbol{\ell}$ given $\boldsymbol{x}$. Previous works have introduced a language classifier to facilitate ASR model [23], [24] to distinguish different languages [13], [14]. In this study, LID is required and obtained according to the variational probability $q_\phi(\boldsymbol{\ell}|\boldsymbol{x})$. The learning objective for this latent variable model is constructed in a form of Eq. (1), which consists of two terms. The first term is an expectation of language conditional log likelihood $p_{\theta,\psi}(\boldsymbol{y}|\boldsymbol{\ell},\boldsymbol{x})$ with respect to variational probability $q_\phi(\boldsymbol{\ell}|\boldsymbol{x})$ while the second term is an KL divergence between variational posterior and true posterior $p(\boldsymbol{\ell}|\boldsymbol{x})$. In this study, θ denotes the frozen transformer parameter, ψ denotes the adapter parameter and ϕ denotes the language-aware calibrator parameters. The parameters ψ and ϕ are trained by minimizing the variational upper bound $\mathcal{L}_{\text{vb}}$.

III. Language-Aware Code-Switching ASR

This study carries out a language-aware code-switching speech recognition given by a Whisper-based transformer [10] with an encoder-decoder architecture as depicted in Figure 1.

A. Language-Aware Adapters and Calibrator

The baseline transformer encoder θ_e and decoder θ_d are frozen. There are two kinds of adapter introduced in encoder ψ_e and decoder ψ_d [25]. The low-rank adapters (LoRAs) [21] are configured in parallel with self-attention and cross-attention [26] in encoder and decoder while the standard adapter [20], [27], [28] is merged in serial with multi-layer perceptron (MLP). In addition to two adapters in encoder and decoder, an additional tunable parameter is the language calibrator ϕ, allocated in the output of Whisper, which calibrates language variations in a code-switching speech. The prefix as a prompt [29], [30] consisting of the special tokens of "start of transcription", "LID", "transcription" and "no time stamp"

$$\boldsymbol{y}_{\text{prefix}} = [\text{sot}][\text{lid}][\text{trans}][\text{nots}] \tag{2}$$

is formed and augmented with a code-switching sentence to obtain a complete transcription $\boldsymbol{y}$ for ASR. As examined in [17], [18], [19], the model prediction is heavily affected by the prefix or prompt $\boldsymbol{y}_{\text{prefix}}$. This influence via $\boldsymbol{y}_{\text{prefix}}$ is crucial for ASR under multilingual setting, particularly under the context of code-switching speech recognition. In such a setting, the expectation over all possible languages becomes significant because each token prediction carries a probability of switching to another language, thereby complicating the prediction process. Since this study utilizes the transformer-based encoder-decoder framework, the generation of word sequence from the code-switching speech frames is run in an autoregressive way. The expectation in the variational upper bound $\mathcal{L}_{\text{vb}}$ needs to further break down into frame-based calculation in a form of

$$\begin{aligned}\mathcal{L}_{\text{vb}} = -\sum_t \{&\mathbb{E}_{q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})}\left[\log p_{\theta,\psi}(\boldsymbol{y}_t|\boldsymbol{x},\boldsymbol{y}_{<t},\boldsymbol{\ell}_t)\right] \\ &+\mathcal{D}_{\text{kl}}\left(q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})\|p(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})\right)\}.\end{aligned} \tag{3}$$

The learning objective is calculated with respect to language-aware posterior probability $q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})$ for recursive prediction of each token $\boldsymbol{y}_t$ based on the histories of LIDs $\boldsymbol{\ell}_{<t}$ and tokens $\boldsymbol{y}_{<t}$ where KL term acts as a regularization term which regularizes variational distribution $q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})$ to be close to true posterior $p(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})$ at each frame t.

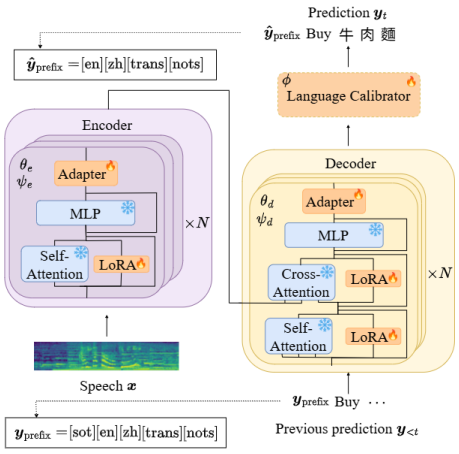

Fig. 1. Overview of a code-switching ASR with adapters $\{\psi_e, \psi_d\}$ and calibrator ϕ, which are configured in a Whisper backbone $\{\theta_e, \theta_d\}$ with N layers. An example of a code-swtiching Chinese-English (zh-en) sentence "buy beef noodle" is shown in presence of a prefix prompt $\boldsymbol{y}_{\text{prefix}}$.

B. Training for Code-Switching ASR

This study utilizes Whisper-small as the backbone for code-switching speech recognition. The learning objective is to minimize $\mathcal{L}_{\text{vb}}$ in Eq. (3) which is composed of two terms. The first term is extended as

$$-\sum_t q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t}) \log p_{\theta,\psi}(\boldsymbol{y}_t|\boldsymbol{x},\boldsymbol{y}_{<t},\boldsymbol{\ell}_t) \triangleq \mathcal{L}_{\text{ce}}(\psi,\phi) \quad (4)$$

which can be calculated as a cross-entropy loss $\mathcal{L}_{\text{ce}}$ between $q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})$ and $p_{\theta,\psi}(\boldsymbol{y}_t|\boldsymbol{x},\boldsymbol{y}_{<t},\boldsymbol{\ell}_t)$. $\boldsymbol{\ell}_t$ is the LID of token $\boldsymbol{y}_t$. In this study, two probabilities q_ϕ and $p_{\theta,\psi}$ are modeled by neural networks for calculating the binomial calibrator parameter ϕ for bilingual languages $\boldsymbol{\ell}_t$ for Chinese and English and multinomial adapter parameters ψ for word tokens $\boldsymbol{y}_t$. Notably, Eq. (4) can be approximated and implemented in two steps which are performed to predict LID $\widehat{\boldsymbol{\ell}}_t$ as well as token $\widehat{\boldsymbol{y}}_t$ before accumulating the learning objective by

$$\widehat{\boldsymbol{\ell}}_t = \arg\max_{\boldsymbol{\ell}_t} q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\widehat{\boldsymbol{\ell}}_{<t}) \quad (5)$$

$$\widehat{\boldsymbol{y}}_t = \arg\max_{\boldsymbol{y}_t} p_{\theta,\psi}(\boldsymbol{y}_t|\boldsymbol{x},\widehat{\boldsymbol{y}}_{<t},\widehat{\boldsymbol{\ell}}_t) \quad (6)$$

$$\mathcal{L}_{\text{ce}}(\psi,\phi) \approx -\sum_t \boldsymbol{y}_t \log \widehat{\boldsymbol{y}}_t \quad (7)$$

where $\boldsymbol{y}_t$ denotes the one-hot vector of ground-truth token at frame t. Here, $\widehat{\boldsymbol{\ell}}_t$ and $\widehat{\boldsymbol{y}}_t$ are obtained by picking up the LID and token with the highest posterior values, respectively. Figure 2 illustrates how the cross-entropy loss $\mathcal{L}_{\text{ce}}(\psi,\phi)$ is calculated based on the prediction of word tokens $\boldsymbol{y}_t$ via the frozen vocabulary head θ and LIDs via the tunable language head ϕ where the latent features $\{\boldsymbol{z}'_t\}$ have been calculated through those adapter parameters ψ.

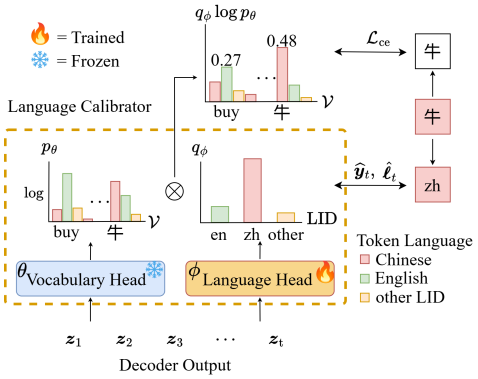

Fig. 2. Illustration of calculating cross-entropy loss $\mathcal{L}_{\text{ce}}$ which is obtained by combining the multinomial outputs from language head q_ϕ and vocabulary head p_θ. Estimated LID $\widehat{\boldsymbol{\ell}}_t$ as ⟨zh⟩ is used to find the token $\widehat{\boldsymbol{y}}_t$ for "cow".

Meanwhile, the KL divergence in Eq. (3) can be minimized by using the following equation

$$\begin{aligned}&\sum_t \mathcal{D}_{\text{kl}}(q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\widehat{\boldsymbol{\ell}}_{<t})\|p(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t}))\\ &= \sum_t q_\phi(\widehat{\boldsymbol{\ell}}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t}) \log \frac{q_\phi(\widehat{\boldsymbol{\ell}}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})}{p(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})} \triangleq \mathcal{L}_{\text{kl}}(\phi).\end{aligned} \quad (8)$$

In the implementation, $p(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})$ is known as a one-hot vector for ground-truth of LID which is seen as the target vector where the calibrator $q_\phi(\boldsymbol{\ell}_t|\boldsymbol{x},\boldsymbol{\ell}_{<t})$ with language head ϕ is learned according to this guidance through minimization of KL divergence. Overall, the optimization problem with respect to language aware adapters and calibrator turns out as

$$\{\psi,\phi\} = \arg\min_{\{\psi,\phi\}} \underbrace{\mathcal{L}_{\text{ce}}(\psi,\phi) + \lambda\mathcal{L}_{\text{kl}}(\phi)}_{\mathcal{L}(\psi,\phi)} \quad (9)$$

where KL loss is seen as a regularization term to guide the output of language calibrator to be close to the ground-truth value and λ is a regularization parameter to tune the relative importance of KL loss compared to cross-entropy loss for token prediction. The adjustment based on λ provides a scheme to mitigate the circumstance when the prediction may misbehave if the language head confidently predicts the wrong language, namely using a smaller λ. The highest overall scores for words and languages are jointly considered.

As a result, the probabilistic language-aware speech recognition is carried out by estimating the parameters of the adapters [31] in encoder and decoder ψ as well as the language calibrator ϕ from a collection $\mathcal{D}$ of code-switching speech utterances $\boldsymbol{x}$ and their transcriptions $\boldsymbol{y} = \{\boldsymbol{y}_t\}$. Language calibrator is integrated in ASR pipeline to calibrate the output of ASR due to the influence of languages. Importantly, vocabulary head and language head are used to calculate the multinomial probabilities for the predicted tokens and LIDs, respectively in a code-switching speech. Introducing the language-aware LoRAs in parallel and the adapters in serial on Whisper backbone conducts the parameter-efficient learning where only very few parameters in LoRAs and adapters relative to the whole model dominated by the Whisper backbone parameters are trained. In particular, LoRAs are employed in the attention layers including both self-attention and cross-attention layers [32]. Meanwhile, standard adapters were incorporated after the linear MLP layers or within each residual attention module. Basically, LoRA is employed with the intention to learn lower-level within the attention layer while adapter is employed with the intention to learn higher-level information. Comprehensive details of the entire learning procedure are provided in Algorithm 1. A token-to-language function $\mathcal{M}$ is designed to convert a given token sequence to its language sequence.

IV. Experiments

A. Experimental Settings

This study conducted the experiments by using ESPnet [33], employing the Chinese-English code-switching dataset from Southeast Asia, called the SEAME [34]. Whisper-small was designated as the backbone model, consisting of an encoder and decoder, each with 12 layers (N) and 12 attention heads. The LoRA method with a rank of 10 was used. The adapter consisting of 2 linear layers with a hidden size of 153 was used, The GeLU non-linearity in between layers, and the residual connections were introduced between layers. The language head possessed a similar architecture to the adapter but with

Algorithm 1: Probabilistic Language-Aware Training

Require : paired code-switching data $\mathcal{D} = \{\boldsymbol{x}, \boldsymbol{y}\}$, frozen backbone Whisper $\boldsymbol{\theta}$, calibrator ϕ, adapter ψ, training epochs T, parameter λ, token-to-language map $\mathcal{M}$

initialize adapter and calibrator ψ, ϕ

for $t = 1$ **to** T **do**

  for *each paired sample* $(\boldsymbol{x}, \boldsymbol{y})$ *of a batch in* $\mathcal{D}$ **do**

    calculate feature embedding $\boldsymbol{z} \leftarrow f_{\theta,\psi}(\boldsymbol{x})$

    calculate $p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\ell}) \leftarrow f_{\theta,\psi}(\boldsymbol{z})$

    calculate $q(\boldsymbol{\ell}|\boldsymbol{x}) \leftarrow f_{\phi}(\boldsymbol{z})$

    predict word tokens $\widehat{\boldsymbol{y}}$ by (6)

    calculate loss $\mathcal{L}_{\text{ce}}(\psi, \phi)$ by (7)

    map from token to language $\boldsymbol{\ell} \leftarrow \mathcal{M}(\boldsymbol{y})$

    predict LIDs $\widehat{\boldsymbol{\ell}}$ by (5)

    calculate loss $\mathcal{L}_{\text{kl}}(\phi)$ by (8)

    calculate total loss $\mathcal{L}(\psi, \phi)$ by (9) with λ

    update adapter ψ, calibrator ϕ by $\nabla_{\psi}\mathcal{L}$, $\nabla_{\phi}\mathcal{L}$

a larger hidden size of 192 to capture richer information. The optimization was performed by using the AdamW optimizer with a learning rate of 10^{-3}, as referred from the ESPnet framework. The languages considered in this work were *English*, *Chinese*, and *other* LIDs which were designed as the special tokens. The model underwent fine-tuning (FT) for 15 epochs, and the best 3 model weights obtained during the fine-tuning process were averaged and used as the model for evaluation. The parameter $lambda$ was set to 5 to balance the contribution of each loss function in the experiment.

The proposed model was evaluated with the subset of SEAME called devman, which is mostly Chinese speech, and devsge, which is mostly Singaporean English speech. The metric used was classified into word error rate (WER) (%) for English speech, character error rate (CER) (%) for Chinese speech, mixed error rate (MER) (%) for code-switching speech, and overall MER (%) for all speech including monolingual speech and code-switching speech. Under the proposed architecture, the amount of the resulting additional parameters was only 5.1% relative to that of the whole parameters including those of Whisper backbone and language-aware LoRAs, adapters and calibrator.

TABLE I
COMPARISON OF DIFFERENT METHODS FOR CODE-SWITCHING (CS) ASR.

Methods	Test Set	WER	CER	MER	Overall MER	Train Param
Original Prompt FT [19]	Devman	–	–	–	14.3	244M
	Devsge	–	–	–	20.4	(100%)
Modified Prompt FT [19]	Devman	–	–	–	15.0	244M
	Devsge	–	–	–	21.0	(100%)
Language Alignment [14]	Devman	–	–	–	16.7	48.3M
	Devsge	–	–	–	24.0	(100%)
Attention-Guidance [17]	Devman	22.2	14.9	13.5	14.2	14.3M
	Devsge	24.2	**19.4**	19.8	20.8	(5.6%)
Prob Lang Awareness	Devman	**21.7**	**14.6**	**12.9**	**13.6**	12.9M
	Devsge	**22.5**	**19.4**	**18.0**	**19.7**	(5.1%)

B. Experimental Results

Table I shows the overall performance of various methods by using SEAME dataset. Both the results of using "Original Prompt FT" and "Modified Prompt FT" were taken from a strong baseline in [19] where the whole Whisper-small model was fine-tuned with ⟨en⟩ and ⟨zh⟩⟨en⟩ as the LID token, respectively. The result based on the method called language alignment was taken from [14] while the result from the method called attention guidance was obtained in [17]. In the experiments, these related works are compared with the proposed probabilistic language awareness (PLA) in terms of error rates and parameter size. As we can see, the proposed PLA reduces overall MER by absolute 0.6% in Devman and by absolute 1.1% in Devsge in the comparison with state-of-the-art result obtained by using attention guidance (AG) in [17]. The size of tunable parameters using PLA is only 5.1% from the whole model, which is even smaller that 5.6% using AG. This shows the effectiveness of the proposed method for code-switching speech recognition.

C. Experimental Analysis

This study further conducts the experiments on showing the importance of multinomial language head of the calibrator q_{ϕ}. There are three language-head settings for code-switching ASR (CS-ASR). In the first setting, no language head is applied. This setting is evaluated to show the result due to direct fine-tuning without language awareness. The second setting is specified by installing a single linear layer in language head to project the feature dimension to the language dimension. The third setting is arranged with a two-layer language head, which is the case we obtained by PLA given in Table I.

TABLE II
COMPARISON OF CS-ASR WITH VARIOUS LANGUAGE-HEAD SETTINGS.

Head Settings	Test Set	Overall MER	Head Accuracy	Train Param
No Language Head	Devman	14.7	–	12.7M
	Devsge	21.4		(5.0%)
1-Layer Language Head	Devman	13.7	73.4	12.7M
	Devsge	20.0		(5.0%)
2-Layer Language Head	Devman	13.6	74.3	12.9M
	Devsge	19.7		(5.1%)

Table II shows a comparison of overall MER, head accuracy (%) and parameter size under an ablation study using PLA. Relative to the overall MER for the case with no language head, PLA with 2-layer language head reduces the overall MER by absolute 0.9% in Devman and by absolute 1.7% for Devsge, indicating that language head is important for introducing language information for CS-ASR. Furthermore, two-layer language head obtains higher language prediction accuracy by absolute 0.9% and lower overall MER by absolute 0.3% in Devsge when compared with the results obtained by one-layer language head. This finding shows that the performance of language head q_{ϕ} heavily affects the overall performance, hence building an accurate language classifier becomes a necessity for code-switching speech recognition.

V. Conclusions

This study has introduced a probabilistic approach to adapt a multilingual ASR backbone model for code-switching scenarios in presence of Chinese and English by pursuing the property of language awareness during decoding through a language-aware calibrator. By deriving the log-probability of an ASR model with language awareness, a two-layer language head was integrated during the fine-tuning process of the Whisper backbone. The capability of language awareness was built through an KL loss in addition to the standard cross-entropy loss for token prediction. The outputs from this language head were combined with the ASR outputs to generate the prediction that accounted for language context. The experiments conducted on SEAME dataset demonstrate that the proposed method achieves the improved performance with the limited additional parameters compared to the previous methods. The language calibrator embedded in an adapter-based Whisper backbone works for code-switching speech recognition.

References

[1] A. Graves, "Sequence transduction with recurrent neural networks," *arXiv preprint arXiv:1211.3711*, 2012.

[2] H. Soltau, H. Liao, and H. Sak, "Neural speech recognizer: Acoustic-to-word LSTM model for large vocabulary speech recognition," *Proc. of Annual Conference of International Speech Communication Association*, pp. 3707–3711, 2017.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, 2017.

[4] M. Rohmatillah, B. Aditya, L.-J. Yang, B. G. Ngo, W. Sulaiman, and J.-T. Chien, "Promoting mental self-disclosure in a spoken dialogue system," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, vol. 2023, pp. 670–671.

[5] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Proc. of Annual Conference of International Speech Communication Association*, 2020, pp. 5036–5040.

[6] Y. Peng, S. Dalmia, I. Lane, and S. Watanabe, "Branchformer: Parallel MLP-attention architectures to capture local and global context for speech recognition and understanding," in *Proc. of International Conference on Machine Learning*, 2022, pp. 17627–17643.

[7] K. Kim, F. Wu, Y. Peng, J. Pan, P. Sridhar, K. J. Han, and S. Watanabe, "E-branchformer: Branchformer with enhanced merging for speech recognition," in *Proc. of IEEE Spoken Language Technology Workshop*, 2023, pp. 84–91.

[8] J.-T. Chien and S.-E. Li, "Contrastive disentangled learning for memory-augmented transformer," in *Proc. of Annual Conference of International Speech Communication Association*, 2023, pp. 2958–2962.

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

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

[11] 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*, 2020.

[12] Z.-X. Yong, R. Zhang, J. Z. Forde, S. Wang, et al., "Prompting large language models to generate code-mixed texts: The case of south east asian languages," *arXiv preprint arXiv:2303.13592*, 2023.

[13] H. Liu, H. Xu, L. P. Garcia, A. WH 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.

[14] H. Liu, X. Zhang, L. P. Garcia, A. W. H. Khong, E. S. Chng, and S. Watanabe, "Aligning speech to languages to enhance code-switching speech recognition," *arXiv preprint arXiv:2403.05887*, 2024.

[15] P. Chen, F. Yu, Y. Liang, H. Xue, X. Wan, N. Zheng, H. Zhou, and L. Xie, "BA-MoE: Boundary-aware mixture-of-experts adapter for code-switching speech recognition," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop*, 2023, pp. 1–7.

[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] 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.

[18] 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.

[19] 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.

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

[21] E. J. Hu, Y. Shen, P. Wallis, Z. A. Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. of International Conference on Learning Representations*, 2022.

[22] S. Watanabe and J.-T. Chien, *Bayesian Speech and Language Processing*, Cambridge University Press, 2015.

[23] H. Huang, S. Lu, Y. Shan, H. Qu, F. Zhang, W. Guan, Q. Hong, and L. Li, "Dynamic language group-based MoE: Enhancing code-switching speech recognition with hierarchical routing," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2025, pp. 1–5.

[24] C. Y. Kwok, H. Liu, J. Q. Yip, S. Li, and E. S. Chng, "A two-stage LoRA strategy for expanding language capabilities in multilingual ASR models," *IEEE Transactions on Audio, Speech and Language Processing*, pp. 1–16, 2025.

[25] J.-T. Chien and W.-Y. Sun, "Adversarial augmentation and adaptation for speech recognition," in *Proc. of Asia Pacific Signal and Information Processing Association Annual Summit and Conference*, 2024, pp. 1–6.

[26] 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.

[27] 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.

[28] 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.

[29] M. Rohmatillah and J.-T. Chien, "Revise the NLU: A prompting strategy for robust dialogue system," in *Proce. of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024, pp. 10956–10960.

[30] J.-T. Chien, M.-Y. Chen, C.-H. Lee, and J.-H. Xue, "Meta soft prompting and learning," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 5, 2024.

[31] L.-J. Yang and J.-T. Chien, "Continual gated adapter for bilingual codec text-to-speech," in *Proc. of Conference of the Oriental COCOSDA*, 2024.

[32] 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.

[33] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, et al., "ESPnet: End-to-end speech processing toolkit," in *Proc. of Annual Conference of International Speech Communication Association*, 2018, pp. 2207–2211.

[34] D.-C. Lyu, T. P. Tan, E. 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.