IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 10 July 2024, accepted 28 July 2024, date of publication 2 August 2024, date of current version 12 August 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3437755

RESEARCH ARTICLE

Dynamic Sampling-Based Meta-Learning Using Multilingual Acoustic Data for Under-Resourced Speech Recognition

I-TING HSIEH[1], CHUNG-HSIEN WU[1,2], (Senior Member, IEEE), AND ZHE-HONG ZHAO[2]
[1]Graduate Program of Multimedia Systems and Intelligent Computing, National Cheng Kung University, Tainan 70101, Taiwan
[2]Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan 70101, Taiwan

Corresponding author: Chung-Hsien Wu (chunghsienwu@gmail.com)

This work was supported by the National Science and Technology Council of Taiwan under Contract 111-2221-E-006-150-MY3.

ABSTRACT Under-resourced automatic speech recognition (ASR) has become an active field of research and has experienced significant progress during the past decade. However, the performance of under-resourced ASR trained by existing methods is still far inferior to high-resourced ASR for practical applications. In this paper, speech data from languages that share the most phonemes with the under-resourced language are selected as supplementary resources for meta-training based on the Model-Agnostic Meta-Learning (MAML) strategy. Besides supplementary language selection, this paper proposes a dynamic sampling method instead of the original random sampling method to select support and query sets for each task in MAML to improve meta-training performance. In this study, Taiwanese is selected as the under-resourced language, and the speech corpus of five languages, including Mandarin, English, Japanese, Cantonese, and Thai, are chosen as supplementary training data for acoustic model training. The proposed dynamic sampling approach uses phonemes, pronunciation, and speech recognition models as the basis to determine the proportion of each supplementary language to select helpful utterances for MAML. For evaluation, with the selected utterances from each supplementary language for meta-training, we obtained a Word Error Rate of 20.24% and a Syllable Error Rate of 8.35% for Taiwanese ASR, which were better than the baseline model (26.18% and 13.99%) using only the Taiwanese corpus and other methods.

INDEX TERMS Under-resourced speech recognition, dynamic sampling, model-agnostic meta-learning.

I. INTRODUCTION

In recent years, complex training methods and architectures of automatic speech recognition (ASR) models have been the mainstream of research to improve ASR performance [1], [2], [3], [4], [5], [6]. However, to drive these models, a huge amount of training data is needed to achieve satisfactory performance. OpenAI's Whisper [7] is trained on a corpus of 680,000 hours of data, enabling it to achieve near state-of-the-art performance across multiple languages with rich resources. But not all languages have sufficient training data in general research settings, and collecting large amounts of speech and language data is time-consuming and labor-intensive. Even though Whisper utilizes weak supervision methods to collect data and train models, the improvement in recognition performance for low-resource languages remains limited. Therefore, the use of effective training strategies and data augmentation methods to improve the recognition of low-resourced languages is another topic worth investigating [8], [9].

For under-resourced languages, data augmentation is the main method to increase the duration and diversity of an existing training set without further collecting any new data. Basically, data perturbation is employed to increase the

The associate editor coordinating the review of this manuscript and approving it for publication was Siddhartha Bhattacharyya.

amount of data, such as speed perturbation and volume perturbation. Until recently, SpecAugment [10] was proposed to replace the traditional perturbation method by simply masking the spectrum of speech data. Although data augmentation can improve the recognition ability of the ASR, it does not always work by continuously increasing the data all the time. Besides data augmentation, the field of deep learning often utilizes resource-rich tasks to improve the performance of target tasks. Common methods include fine-tuning [11], transfer learning [12], [13], [14], and multitask learning [15]. However, in the field of ASR, the help from other languages cannot accurately control the contribution of the source domain and the restriction of the architecture causes that the target domain cannot incorporate model-independent knowledge from other domains. The rise of meta learning has the potential to eliminate the above problems and achieve promising results in various fields including ASR field.

In the past, researchers observed that deep learning aims to achieve learning capabilities similar to humans. However, deep learning demands extensive data and computational resources, posing a significant contrast to humans, who can learn from a small amount of data. Meta-learning has been proposed to address this disparity and address the challenges associated with limited data for deep learning. Meta-learning, also known as learning to learn, is a type of few-shot learning approach. This approach differs from traditional fine-tuning methods as it seeks to acquire universal prior knowledge that enables rapid adaptation to unseen target tasks. Meta-learning can acquire prior knowledge through various methods, such as loss-based or gradient-based. Among them, the model-agnostic meta-learning (MAML) proposed by Finn et al. in 2017 [16] is more widely known. MAML is a gradient-based meta-learning method. In order to enhance the generalization of prior knowledge, MAML's training involves multiple tasks. Due to the varying data quantities and training difficulties of each task affecting generalization, past research [13], [17] has often used random sampling to address this issue. And in the training of the MAML, each training iteration randomly samples a fixed number of tasks, and the data for each task is also randomly sampled. The outstanding performance of meta-learning has gradually attracted attention in the field of speech recognition [18], [19], especially MAML applied to under-resourced ASR [20]. The study in [20] verified that the MAML model can be applied to any target language. Reference [21] proposed the method of alphabet unification, using English pronunciation rules based on the Latin alphabet as the standard to align each language in the experiment. Meta-learning was then employed to build automatic speech recognition (ASR) systems to improve the recognition capabilities across different languages. Additionally, [22] applied the fast adaptation characteristics of meta-learning to under-resource accent ASR. In addition to normal speech, researchers have also applied meta-learning to ASR for disordered speech [23].

MAML obtains highly generalizable prior knowledge through random sampling of tasks and data, making it applicable to various tasks. However, solely pursuing generalization by using unrelated models can limit the adaptation speed to the target model. In previous literature, [24] proposed Task Similarity Aware MAML (TSA-MAML). This method obtains group-specific initialization parameters through the relevance between tasks and has been shown to adapt to target tasks faster than MAML. Moreover, when the target task is known in advance, there is no need to consider generalization, and random sampling methods may not be necessary. In our previous research [25], we utilized MAML to train under-resourced ASR with the assumption that the target task is known in advance, and achieved significant improvements. If a new sampling method can take into account the relevance to the target task, the relevance of prior knowledge will be closer to the target task. In previous literature on improving the sampling method for MAML, the study in [26] applied MAML to image classification and conducted sampling based on the granularity of image data labels, using it as an indicator of difficulty. Reference [27] proposed an adaptive sampling method, selecting suitable training data for training based on the current state of the model. Reference [28] proposed a sampling method to prevent catastrophic forgetting and sensitivity to the order of tasks, aiming to avoid overfitting. Reference [29] applied the meta-curriculum learning to neural machine translation. To adapt to a broader range of unseen tasks, the meta-learning process is designed to shift the model's focus from global features to local features as the number of iterations increases. The common goal of the above-mentioned literature is to pursue generalization, employing curriculum learning to progressively train models from simple to hard tasks. However, for a known target task, the emphasis may not be on generalization.

For our approach, we drew inspiration from offline learning and online learning as proposed in [29]. We prepare supplementary tasks related to the target task through background knowledge or design a new sampling method based on feedback from the target task. Although the improved method loses the advantage of being able to quickly adapt to any target model, the target task can learn more relevant features than the original MAML. In this paper, the under-resourced language ASR will be used as the target task, and the languages related to the target language will be selected as the supplementary tasks.

Regarding under-resourced languages, we chose Taiwanese (Hokkien) as the target language. Taiwanese belongs to the category of under-resourced languages, as highlighted in several studies [30], [31]. To deal with the problem of under-resource for Taiwanese ASR, two approaches are proposed in this study and summarized as follows.

- Data supplementation: We selected languages that are related to Taiwanese and share the most phonemes in common as supplementary languages for data augmentation to train the Taiwanese ASR using MAML.

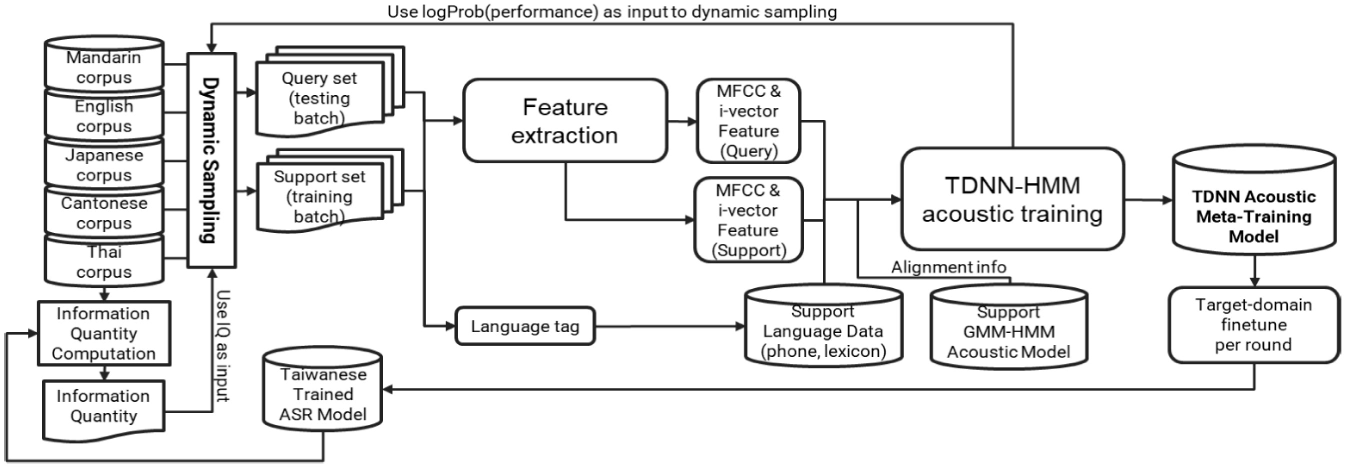

Use logProb(performance) as input to dynamic sampling
Mandarin corpus
English corpus
Japanese corpus
Cantonese corpus
Thai corpus
Dynamic Sampling
Query set (testing batch)
Support set (training batch)
Feature extraction
MFCC & i-vector Feature (Query)
MFCC & i-vector Feature (Support)
TDNN-HMM acoustic training
TDNN Acoustic Meta-Training Model
Alignment info
Information Quantity Computation
Information Quantity
Use IQ as input
Language tag
Support Language Data (phone, lexicon)
Support GMM-HMM Acoustic Model
Target-domain finetune per round
Taiwanese Trained ASR Model

FIGURE 1. System framework of Taiwanese ASR training based on MAML.

- Dynamic sampling: A new sampling method is proposed to replace MAML's original random sampling approach for selecting useful data during Taiwanese ASR training.

II. PROPOSED METHOD

The overall framework for Taiwanese ASR training based on MAML is shown in Figure 1. We will focus on the introduction of the following sections:

- Selection of supplementary languages utilizing the proposed selection method.
- Model construction based on meta-learning principles.
- Dynamic sampling strategy incorporating information quantity (IQ), feedback from the previous round of model training, and data distribution.

A. SUPPLEMENTARY LANGUAGE SELECTION

Due to the insufficiency of the Taiwanese corpus, this study selects available corpus from related languages as supplementary data. The speech corpus of the selected supplementary language could be included into the training data to increase the training volume for training the ASR system. Considering phonemes as the basis for language selection, this paper uses the international phonetic alphabet (IPA), which is common to all languages, for phoneme coverage evaluation. It then converts the IPA of Taiwanese into Tai-Lo Pinyin, the Taiwanese Romanization System.

The languages for selection contain similar pronunciations that share the same phonemes with Taiwanese. If a large portion of the phonemes of the languages for selection are in common with Taiwanese, we can conclude that they are helpful for the training process of Taiwanese ASR. Therefore, we take the overlapping phonemes between the selected supplementary languages and Taiwanese as the main consideration and use the following rule to determine which language should be included. If it satisfies any one of the rule, the language can be considered as a candidate for the supplementary language.

TABLE 1. Statistics of the speech and text corpus and lexicon of each language.

	Speech Corpus	Duration (hour)	Number of Utterances	Number of Words
	Text Corpus			
	Lexicon			
MD	King-ASR-044-sub [32]	75.08	57,718	
	Gigaword [33]		16,941,384	
	MHMC Pre-collected Chinese Lexicon			305,938
EN	Librispeech [34]	100.59	28,539	
	Librispeech [35]		40,418,261	
	Librispeech-lexicon [35]			206,508
JP	Commonvoice [36] JSUT v1.1 [37]	11.41	10,372	
	Wiki backup dumps [38]		250,000	
	重要度順語彙資料庫 [39]			57,374
CA	Commonvoice 5.1 [36]	50.77	41,811	
	Wiki backup dumps [38]		389,136	
	Cifu Lexicon [40]			148,239
TH	Commonvoice 5.1 [36] Gowajee v0.9.1 [41]	18.25	19,444	
	Wiki backup dumps [38]		1,029,680	
	Lexicon-Thai [42]			14,409
	Total	256.1	157,844	732,468

- Languages used in geographically close regions to Taiwan.
- Languages with overlapping phonemes that can compensate for the lack of phonemes in the Taiwanese corpus.

Based on the above rules, we select Mandarin, English, Japanese, Cantonese, and Thai as the supplementary languages. The speech corpus, text corpus, and lexicon of the selected languages that we can obtain are listed in Table 1.

The phonemes of these supplementary languages are compared with those of Taiwanese to cover as many phonemes of Taiwanese as possible in order to achieve a more comprehensive data supplementation. The phoneme coverage for each supplementary language is shown in Table 2. The language notations for each language are denoted as TW for Taiwanese, MD for Mandarin, EN for English, JP for Japanese, CA for Cantonese, and TH for Thai, respectively. Actually, there are still two semi-vowels (ngs and ms) not covered by any of the selected supplementary languages. As the occurrence probabilities of these two semi-vowels are extremely low, we can further collect the speech samples of these two semi-vowels to eliminate the coverage problem without spending too much effort. Besides, since each common phoneme from the various supplementary language has diverse tone value, for the efficiency and rules simplification, the selection of supplementary languages does not consider tone information.

TABLE 2. Common phonemes between Taiwanese and other supplementary languages.

Consonants & Nasal phonemes																	
TW	MD	EN	JP	CA	TH	TW	MD	EN	JP	CA	TH	TW	MD	EN	JP	CA	TH
p	•	•	•	•	•	k	•	•	•	•	•	g		•	•		
ph	•			•	•	kh	•			•	•	j		•	•	•	•
m	•	•	•	•	•	h	•	•	•	•	•	nn	•				
t	•	•	•	•	•	ts	•		•	•	•	ng	•	•	•	•	•
th	•	•		•	•	tsh	•		•	•	•	ng’		•		•	
n	•	•	•	•	•	s	•	•	•	•	•	m’				•	•
l	•	•		•	•	b		•	•		•	nn’				•	•
Vowels						Stop consonant finals						Semivowels					
TW	MD	EN	JP	CA	TH	TW	MD	EN	JP	CA	TH	TW	MD	EN	JP	CA	TH
a	•	•	•	•	•	p’				•	•	ngs					
e	•	•	•	•	•	t’				•	•	ms					
i	•	•	•	•	•	k’				•	•						
o	•	•	•	•	•	h’			•		•						
u	•	•	•	•	•												
er	•	•															

To evaluate whether the common phoneme distribution of speech corpus in each language is similar to that in Taiwanese corpus, this paper uses the Kaldi Speech Recognition Toolkit [43] to conduct phoneme statistics for each language corpus. It then compares the number of occurrences of each phoneme and the number of occurrences of common phonemes with Taiwanese. The overlapping rates of the common phonemes with Taiwanese for each language are: 86.24% for Mandarin, 68.64% for English, 87.99% for Japanese, 86.63% for Cantonese and 71.85% for Thai. This study also lists the amount of data available in each language for each common phoneme. The proportion of each phoneme augmented by the supplementary corpus is shown in Figure 2.

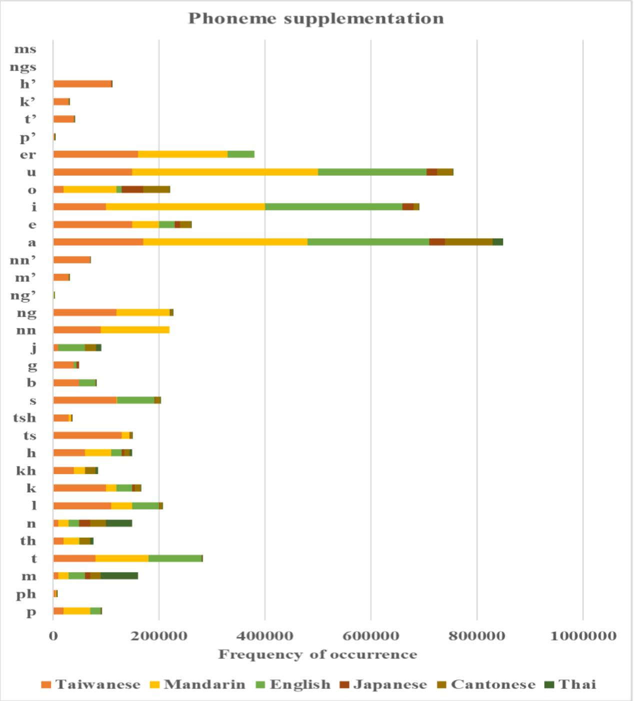


FIGURE 2. The proportion of Taiwanese phonemes supplemented by each language corpus.

B. META-LEARNING FOR TAIWANESE ASR

In this paper, acoustic models are trained using MAML based on the selected languages as supplementary training materials. The training of meta-learning is divided into meta-training and meta-testing. Both are described in the following subsections.

1) META-TRAINING

The purpose of meta-training is to train the model to obtain better initial model parameters for any task. There are usually several different subtasks in the meta-training stage, and each training round will randomly select n tasks as meta-training tasks $T = (t_1, t_2, \ldots, t_n, \ldots, t_N)$. Each meta-training task can build the model with the parameter θ. The data in each meta-training task can be divided into a support set $D^s_{\square} = (D^s_1, D^s_2, \ldots, D^s_n, \ldots, D^s_N)$ and a query set $D^q_{\square} = (D^q_1, D^q_2, \ldots, D^q_n, \ldots, D^q_N)$, which represent the training data set and the test data set of the task, respectively. D^s_n represents the support set in task t_n. The same as the random task selection, the data composition of the support set and query set in each meta-training task is obtained by random sampling. If D^s_n has k data samples, it is called k sample learning. If there are N tasks and k samples in the support set of each task, it is called N way k shot. In meta-training, each meta-training task t_n trains the random initial parameter θ by the support set D^s_n, and uses gradient decent to minimize θ, as shown in (1). η is the learning rate. Since the initial parameters of the target task obtained from each meta-training task focuses on the initial training direction to achieve fast adaptation, each meta-training task only needs to be trained with few times.

$$\hat{\theta}_n = Learn\left(D_n^s;\theta\right) = \theta - \eta \cdot \frac{\partial L\left(D_n^s,\theta\right)}{\partial\theta} \tag{1}$$

After obtaining the parameter $\hat{\theta}$ for each of the n meta-training tasks, MAML uses query set D_n^q to evaluate each meta-training task and obtains the loss value, as shown in (2).

$$L_n^{meta}\left(\theta\right) = L\left(D_n^q, \hat{\theta}_n\right) \tag{2}$$

After the learning process for each meta-training task, the meta-gradient decent is performed to update the model parameters φ based on the loss function $L\left(\varphi\right)$ from each meta-training task. The model remains the same as the model with parameter θ, and is defined as

$$L\left(\varphi\right) = \sum\nolimits_{n=1}^{N} L(D_n^q, \hat{\theta}_n) \tag{3}$$

$$\varphi = \theta - \eta' \cdot \nabla_{\varphi} \sum\nolimits_{n=1}^{N} L\left(D_n^q, \hat{\theta}_n\right) \tag{4}$$

where η' is the meta learning rate. Repeating the entire meta-training process several times to obtain the initial parameters φ that are suitable for fast adaption for the subsequent meta-testing. The update process of initial parameter φ was shown in Figure 3.

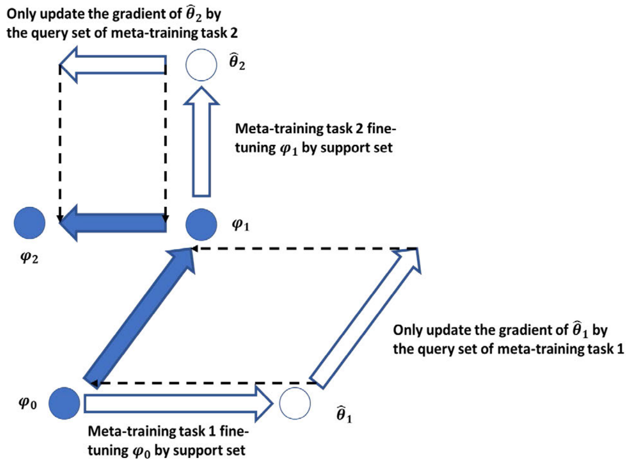

Only update the gradient of $\hat{\theta}_2$ by the query set of meta-training task 2
$\hat{\theta}_2$
Meta-training task 2 fine-tuning φ_1 by support set
φ_1
φ_2
Only update the gradient of $\hat{\theta}_1$ by the query set of meta-training task 1
φ_0
$\hat{\theta}_1$
Meta-training task 1 fine-tuning φ_0 by support set

FIGURE 3. Schematic diagram of the update process of the initial parameter φ.

The above process yields the initial parameters φ. However, this process involves second-order differentiations, making the computation quite extensive. To save time, this paper refers to [44] and adopts the First Order MAML (FOMAML) method, which allows the omission of second-order differentiations without significantly affecting performance. Equation (5) can be rewritten as:

$$\varphi = \theta - \eta' \cdot \nabla_{\hat{\theta}} \sum\nolimits_{n=1}^{N} L\left(D_n^q, \hat{\theta}_n\right) \tag{5}$$

2) META-TESTING

The process of meta-testing is the same as meta-training, but the task extracts the data of the target domain to form the training batches. The model parameters θ after meta-training are used as the initial model and are adapted to parameters suitable for the target domain after the training process, which involves drawing the support and query sets and updating the parameters similar to the case of meta-training.

3) MAML FOR TAIWANESE ASR

As the initial parameters suitable for a new task are obtained, if the target task is single and known, we can adjust the MAML training process to achieve more robust performance. In this paper, there are two adjustment methods proposed for MAML.

In meta-training, through supplementary language selection, we obtain the most helpful meta-training tasks. In each training round, we use all meta-training tasks instead of the randomly selected tasks. Although this tuning method may lose the generalizability of the model, more useful information can be obtained from the meta-training task. In addition, in order to obtain more useful information from meta-training, we also adjust the random sampling method for drawing the support set and query set for each meta-training task. The dynamic sampling method proposed in this paper will replace the random sampling method. The details will be introduced in the next subsection.

Although the meta-testing strategy is similar to the meta-training, the purpose of meta-testing in this study is to further improve the recognition ability of Taiwanese by using a smaller amount of Taiwanese data properly. By modifying the training approach of MAML, in the meta-testing phase, we directly fine-tune all the target domain data (Taiwanese) to form the training dataset based on the initial parameters θ of the model trained by meta-training.

C. DYNAMIC SAMPLING

When using the data from other supplementary domains, it would be helpful to consider the differences between the data from each supplementary domain to improve the efficiency and effectiveness of model training. Hence, this paper introduces a dynamic sampling method for scoring and tagging data to bolster the efficacy of meta-learning and optimize the utilization of data in supplementary domains. In meta-learning, for data selection, random sampling method is generally used to select the supplementary task as well as the data in each supplementary task. The proposed method will replace the random sampling method used in meta-learning. Because the supplementary language is selected based on the phonemes in common with Taiwanese, dynamic sampling only focusses on data selection for each supplementary language. Figure 4 shows the general structure of the dynamic sampling method. The IQ and the training feedback from the current meta-training round (r^{th}) are used to decide which supplementary data should be sampled in the next meta-training round ($(\mathrm{r}+1)^{\mathrm{th}}$). The process is roughly divided into five parts, which are described in the following.

- Determine the IQ of each speech data in the supplementary languages.

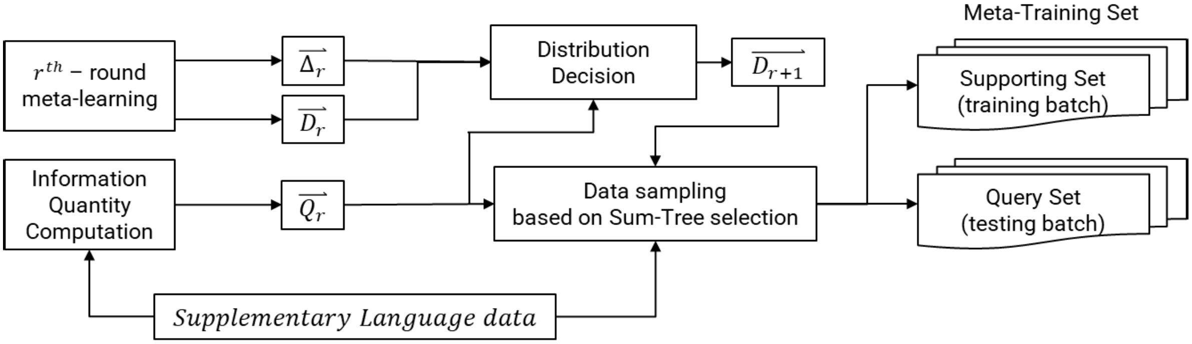

Meta-Training Set
r^{th} – round meta-learning
Information Quantity Computation
$\overrightarrow{\Delta_r}$
$\overrightarrow{D_r}$
$\overrightarrow{Q_r}$
Distribution Decision
$\overrightarrow{D_{r+1}}$
Data sampling based on Sum-Tree selection
Supporting Set (training batch)
Query Set (testing batch)
Supplementary Language data

FIGURE 4. Dynamic sampling framework.

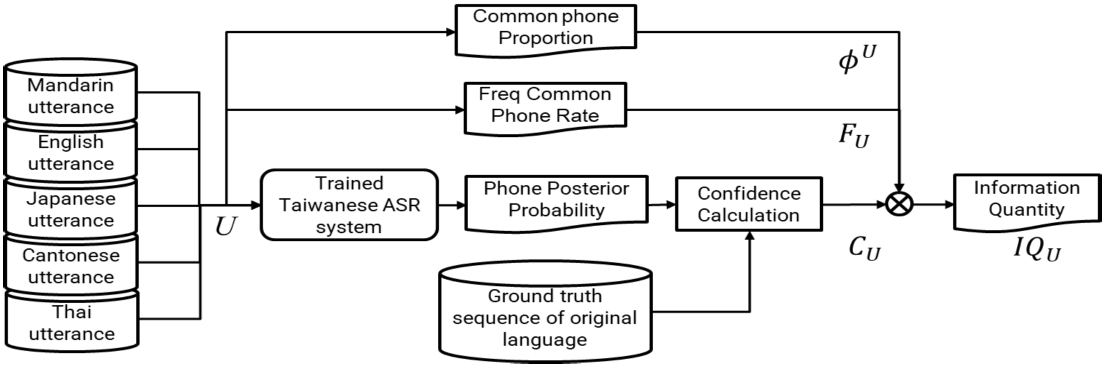


FIGURE 5. Information quantity computation.

- Determine the distribution of data amount to use in each supplementary language for the next meta-training round ((r+1)$^{\text{th}}$).
- Sample the speech data in each supplementary language using the sum-tree algorithm for weighted sampling.

Concerning the computation of IQ and the selection of more informative speech data from supplementary languages, this paper presents the following arguments based on observations of phonemes and language pronunciation:

- When an utterance contains more phonemes in common with Taiwanese, it can provide more information for Taiwanese ASR.
- The utterances containing more phonemes that frequently appear in Taiwanese in each supplementary language can provide further assistance for Taiwanese ASR.

Therefore, to determine the degree of discriminability of a supplementary language in Taiwanese, this paper proposes to use a trained Taiwanese ASR system to recognize the utterances of other supplementary languages and calculate the confidence scores. Afterward, combining prior knowledge about the ratio of common and frequent phonemes in the utterance with posterior information from the feedback of the existing ASR, it is expected to extract more useful training data for model training.

Figure 5 illustrates the calculation of the IQ, with each component detailed in the subsequent sections.

D. INFORMATION QUANTITY

The confidence score of utterance U in each supplementary language is calculated by the frame level Phonetic posteriorgrams (PPG) which are extracted from the Taiwanese ASR (The same as the baseline model described in Table 8.) during decoding utterance U. To understand the ground truth phoneme of each frame of utterance U, the alignment information is obtained from the ASR system for each supplementary language.

For confidence score estimation, cross-entropy and the error rate for each frame are used. The confidence score is shown in (6).

$$C_U = \frac{1}{K}\sum\nolimits_{i=1}^{I} -\log\left(1 - c_{U_i}\right) \tag{6}$$

where C_U represents the confidence value of utterance U in the supplementary language, I denotes the number of frames belonging to the common phonemes in utterance U, K represents the total frame number of utterance U, and c_{U_i} signifies the phoneme correct rate for the i-th frame in utterance U, and the reference phoneme of each frame is provided by the alignment information. If the most probable hypothesis phoneme in the PPG of the frame matches the reference phoneme and is a common phoneme, its hypothetical phoneme probability is considered as the phoneme correctness rate c_{U_i}; else, 0 is assigned to c_{U_i}. In addition, the confidence score does not consider the lexical tone in Taiwanese, so c_{U_i} is obtained by averaging the correct rate of the phonemes disregarding lexical tones.

By using the concept of cross-entropy, where the C_U is high if the value of $(1 - c_{U_i})$ is low, utterances with high correctness have high IQ.

The proportion of common phonemes contained in an utterance affects the amount of training data that can be provided for the language. The proportion of common phonemes ϕ^U is defined as (7).

$$\phi^U = \frac{I}{K} \tag{7}$$

In terms of speech characteristics, the more frequent the phonemes in the corpus of a supplementary language appear in Taiwanese, the more frequent these phonemes could be used. Increasing the amount of the speech data of these phonemes can be beneficial for system training. To determine the frequency of each phoneme used in Taiwanese, ensuring that the distribution of phoneme occurrences closely reflects real-life usage, this paper employs the Taiwanese Across Taiwan corpus [45], a collection of Taiwanese soap operas, for estimating phoneme frequencies. The occurrence statistics for each phoneme in Taiwanese are shown in Figure 6.

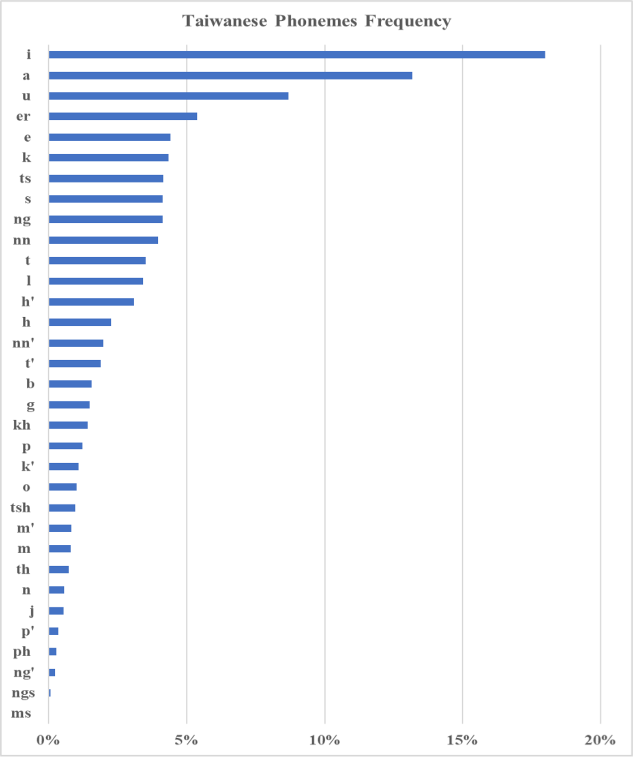


FIGURE 6. Percentage of Taiwanese phonemes used in Taiwanese soap operas.

The frequency distribution of each Taiwanese phoneme is then used to evaluate the usefulness of the phoneme as the training data. Let f_x be the occurrence frequency of Taiwanese phoneme x and f_w be the lowest occurrence frequency among all Taiwanese phonemes. Then, we calculate the ratio of f_x over f_x, followed by taking logarithm to facilitate the calculation of the ratio of each language in the subsequent dynamic sampling and to reduce the disparity in the distribution of individual ratios. For each Taiwanese phoneme x, the log ratio P_x is defined in (8).

$$P_x = \log\left(\frac{f_x}{f_\omega}\right) \tag{8}$$

The ratio for each Taiwanese phoneme is applied to every utterance in the supplementary language. For the frequent phoneme in the supplementary language utterance U, the calculation of the frequency ration at the frame level is shown in (9). If the hypothesis phoneme is the same as the reference phoneme and its phoneme is the common phoneme, then the ratio P of the phoneme for this frame will be calculated, else, 0 is assigned to F_{u_i} for the i-th frame.

$$F_{u_i} = \begin{cases} P_x, \mathit{if}\ x = \mathit{ref.phoneme\ and}\ \in\ \mathit{common\ phoneme} \\ 0, \mathit{else} \end{cases} \tag{9}$$

The percentages of frequent phonemes in the utterance are averaged as shown in (10). K is the total number of frames in utterance U, and I is the number of frames which belong to common phoneme in utterance U.

$$F_U = \frac{1}{K} \times \sum\nolimits_{i=1}^{I} F_{u_i} \tag{10}$$

The confidence score is multiplied with the other two factors as the IQ of utterance U calculated by the speech recognition system, as shown in (11).

$$IQ_U = \phi^U \times F_U \times C_U \tag{11}$$

For a supplementary language l, if the total number of utterances is J, each utterance in language l has its information quantity $\{IQ_1^l, \ldots, IQ_J^l\}$, and the average information quantity q^l of the current training round is the average of the information quantity of all utterances, as shown in (12).

$$q^l = \frac{1}{J} \sum\nolimits_{j=1}^{J} IQ_j^l \tag{12}$$

The acquisition of IQ has been fully described above. Next, we will introduce the training feedback component.

E. FEEDBACK INFORMATION

Let the set of all supplementary languages be $Lan = \{l_1, \ldots, l_5\} = \{MD, EN, JP, CA, TH\}$. Then, for a supplementary language l, the properties of the proportion of the utterances used to determine the next training round are described as follows.

We use Kaldi's chain model which uses the log-probability between the generated phoneme sequence and the correct phoneme sequence as the target function. A higher log probability means that the model performs better, i.e., the model learns better in a particular language, which should lead to higher adaptability. Let the log-rate of a supplementary language l after training be LP_l. For a supplementary language l, the performance score δ^l in the previous training is the final log-rate obtained at the end of the meta-training using the query set to evaluate the model θ, normalized between [0, 1], as shown in (13).

$$\delta^l = norm\left(LP_l\right) \tag{13}$$

We retrieve the dynamic sampling distribution of the current training round as historical information, i.e., the proportion of utterance assigned to each supplementary language in this training round. Then, for a supplementary language l, its data distribution d^l in this training round is a proportion value between [1, 0], as shown in (14).

$$0 \leq d^l \leq 1, d^l \in R \tag{14}$$

Let the time of this training round be r, then the performance score Δ_r, the proportional distribution D_r, and the average information quantity Q_r of the current training round can be expressed as three one-dimensional vectors for 5 supplementary languages respectively, as shown in (15) to (17).

$$\vec{\Delta}_r = \left\{\delta_r^{l_1}, \ldots, \delta_r^{l_5}\right\} \tag{15}$$

$$\vec{D}_r = \left\{d_r^{l_1}, \ldots, d_r^{l_5}\right\} \tag{16}$$

$$\vec{Q}_r = \left\{q_r^{l_1}, \ldots, q_r^{l_5}\right\} \tag{17}$$

Finally, the above information is used to calculate the proportional distribution D_{r+1}, calculated as shown in (18), for each supplementary language in the next training round, and three parameters α, β, γ are set as the weights of the performance score, proportional distribution, and average information quantity, respectively.

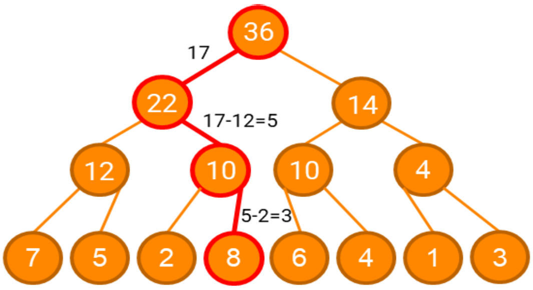
$$\vec{D}_{r+1} = norm\left(\alpha \times \vec{\Delta}_r + \beta \times \vec{D}_r + \gamma \times \vec{Q}_r\right) \tag{18}$$

After obtaining the utterances in each supplementary language through dynamic sampling distribution, this paper further adopts the sum-tree structure, using the IQ of each utterance as a factor, to select the more helpful utterances as the training data.

F. SUM-TREE

Sum-tree is originally used as a data structure for deep Q learning in reinforcement learning. It is used to extract training samples with high temporal difference error (TD error), i.e., poorer performance, and give them more training opportunities. This approach enables the overall training process to prioritize areas that require more learning, thereby expediting model convergence and enhancing its performance.

The structure of the sum-tree is a complete binary tree. This tree is built by storing the priority of each data in each leaf node and adding the priority of each leaf node two by two to its parent node. In this way, the priorities are summed up level by level, and the root of the complete binary tree stores the sum of the priorities of all data.



M=6, segments=[0, 6), [6, 12), [12, 18), [18, 24), [24, 30), [30, 36]
Let the sampling process is executed when v = 17

FIGURE 7. Schematic diagram of the completed sum-tree.

When sampling data using a sum-tree, the sum of all priorities stored in the root of the tree is denoted as R. Assuming that the training round needs to extract M data samples, it is necessary to divide R by M to generate M intervals $[s_1, \ldots, s_m, \ldots, s_M]$, and the size each interval does not exceed the quotient of R divided by M. Then, for selecting the value of v_m in each interval, v_m must follow two rules to traverse from the top to the bottom and left to right of the sum-tree.

- First compare v_m with the value l_e of the left child node of the current node. l_e as parent node is calculated by sum of the values of both child nodes. If $v_m < l_e$, then the value of v_m remains unchanged and the left child tree of the current node is traversed.
- If $v_m > l_e$, subtract the value l_e of the left child node from the value of v_m ($v_m = v_m - l_e$), and traverse to the right child of the current node.

When traversing to the bottom of the tree in this way, the node is the sampling result. For example, the priority distribution of the sum-tree is shown in Figure 7, and the randomly selected value of v is 17 in the 3rd interval ($12 \leq x < 18$, i.e., [12, 18)). In the first branch, v will select left child node according to rule 1 ($17 < 22$). When traversing to the second branch, v will select the right child node according to rule 2 ($17 > 12$, so the right child node is selected to traverse). Finally, v will select the right child node in the third branch using rule 2 ($5 > 2$), so the right child node is selected to traverse and the data with a priority value of 8 will be selected.

III. EXPERIMENTAL RESULTS

A. DATASETS IN TARGET DOMAIN AND HYPERPARAMETER

The data used for training the acoustic models were sampled from the datasets of five languages, consisting of Mandarin, English, Japanese, Cantonese, and Thai. For the training process of the target language (Taiwanese), MHMC balanced speech database [12] and the training set of Taiwanese across Taiwan (TAT-Vol1) [45] were employed. TAT-Vol1 is a Taiwanese speech dataset released by the Formosa Speech Recognition 2020 (FSR2020) challenge. In our experiments, we use the evaluation set from the pilot-test of this challenge as the testing corpus for this paper. The text corpus consisting of 867,260 Taiwanese sentences [46], [47], [48], [49], [50] was used for language model training. The statistics of the Taiwanese speech corpus are shown in Table 3.

TABLE 3. Statistics of taiwanese speech corpus.

Corpus Name	No. of utterances	Duration (hours)
MHMC balanced	96,211	86.3
TAT-Vol1 train	23,104	41.8
TAT-Vol1 evaluation	2,664	4.8
Total	121,979	132.9

Regarding the parameters, in this section, we will cover the remaining parts of Figure 1, including feature extraction, alignment information extraction, acoustic modeling and language modeling. In terms of feature extraction, the raw speech data were sampled at 16 kHz, and the length of each frame was 25ms with 10ms overlap for feature extraction. Then, we extract Mel-scale Frequency Cepstral Coefficients (MFCC) from both the support set and the query set. To mitigate differences arising from recording devices and environments across different audio files, we normalize the MFCC using Cepstral Mean and Variance Normalization (CMVN). Additionally, we adopt the approach outlined in [51], where we extract i-vector and preprocess them with Principal Component Analysis (PCA) and Universal Background Model (UBM). Finally, we combine the MFCC with the i-vector and input them into the acoustic model. Regarding alignment information, it is obtained from the Hidden Markov model with Gaussian mixture emissions (GMM-HMM). The training of the GMM-HMM was carried out sequentially using monophones, triphones, and Linear Discriminant Analysis Maximum Likelihood Linear Transformation (LDA-MLLT). The settings related to the GMM-HMM are referenced in [43] and [52]. In terms of the acoustic model, after obtaining the alignment information, we can proceed with supervised learning of the Hidden Markov model with Time Delay Neural Network (TDNN-HMM) [51]. For the language model, the SRILM tool [53] was used to perform n-gram statistics and probability value conversion. The overall architecture is implemented using Kaldi,[1] with detailed parameters as mentioned above, as shown in Table 4. The related code and the corpus are publicly available.[2]

TABLE 4. Schematic table of hyperparameters and configurations.

Category	Hyperparameters	Values
Feature extraction	Dimension of MFCC	40
	Dimension of i-vector	100
GMM-HMM	Gaussian mixture components	1000
	Number of training iterations (mono/triphone/LDA MLLT)	40/35/35
TDNN-HMM	Dimension of all hidden layer	512
	Number of layers	8
	Number of epochs	4
	Learning Rate	10^{-3} to 10^{-4}
Above is the Baseline model setting		
MAML	meta-training: supplementary language, random sampling meta-testing: target language, baseline settings	
MAML-dynamic sampling	meta-training: supplementary language, dynamic sampling meta-testing: target language, baseline settings	

B. EVALUATION OF UNDER-RESOURCED ASR

In this section, we investigated the effects of parameters and training methods on the performance of the model through the experimental results. The word error rate (WER) and syllable error rate (SER) were adopted as the evaluation criterion for the experiments. The WER was calculated by using words as the unit of calculation, and the text corpus used in this paper was represented by TaiLo Pinyin. Therefore, when each TaiLo Pinyin word was separated into individual characters, it still contained several TaiLo Pinyin syllables, which consisted of more than one character in the text. Therefore, the SER was used here to better represent the recognition accuracy of the syllables in Taiwanese language.

Since there were many variables in system implementation, the most effective combinations of parameters and settings to improve the recognition ability of the model were obtained experimentally. We not only experimented with the number of support sets D_r^s and query sets D_r^q of the tasks extracted in each training round with meta-training using dynamic sampling, but also different combinations of training strategies adopted to examine the improvement in the recognition accuracy of the model. The results for each configuration are shown in Table 5.

To quickly find the optimal parameter configuration, we first conduct experiments to observe which parameter is the most important. For independent experiments based on three parameters, the results showed that the method using only the average information quantity to determine the amount of training data for each supplementary language in the next round at $\gamma = 1$ resulted in a lower WER and SER for each setting. Based on this result, this paper then combined

[1] Kaldi: https://github.com/kaldi-asr/kaldi
[2] https://github.com/qqaazz84415/DY-MAML-TW-ASR

TABLE 5. The WER/SER (%) for the combination of each parameter and data amount.

Data amount / (α, β, γ)	2000 utterances	4000 utterances	6000 utterances
(1, 0, 0)	20.78/8.72	20.84/8.69	21.34/9.01
(0, 1, 0)	20.89/8.68	21.25/8.92	20.98/8.96
(0, 0, 1)	20.75/8.47	20.92/8.64	21.01/8.87
(0.1,0.1,0.8)	20.70/8.46	20.71/8.46	20.83/8.82
(0.1,0.2,0.7)	**20.24/8.35**	20.66/8.48	20.79/8.74
(0.2,0.2,0.6)	20.73/8.41	20.74/8.55	20.86/8.77
(0.2,0.3,0.5)	20.77/8.44	20.83/8.51	20.94/8.85

the ratios of the three parameters (α, β, γ) according to the degree of influence of the above experimental results. The results showed that the model had the lowest WER and SER in each setting for $(\alpha, \beta, \gamma) = (0.1, 0.2, 0.7)$. From these experimental results, it is evident that the average information quantity remains a primary factor influencing the experimental result. However, incorporating a slight reference to the model's feedback can enhance the recognition ability of the model.

After deciding the ratio between the dynamic sampling parameters, the relationship between the number of training data used in each round and the recognition results are presented in Table 6. Table 6 shows the effect of the number of support sets D_r^s and query sets D_r^q on the accuracy of the model for each training task. Besides, the results from random sampling-based MAML are also presented in Table 6. Because of the restriction of the Kaldi system, the training data should consist of at least 1000 utterances for the support set D_r^s and the query set D_r^q, respectively. Due to the possibility of the training data for dynamic sampling-based MAML being fewer than 1000 samples, we omit experiments with 1000 training data for dynamic sampling-based MAML.

TABLE 6. The WER/SER (%) for the different sample method.

Sample methods / Numbers of utterances	Random sampling	Dynamic sampling
1000	20.87/8.89	-/-
2000	20.89/8.86	**20.24/8.35**
4000	21.25/8.90	20.85/8.64
6000	21.36/9.01	20.91/8.83

The experimental results in Table 6 show that the MAML tends to focus on extracting information from a smaller amount of data. The model using 2000 utterances for D_r^s and 2000 utterances for D_r^q for each supplementary language achieved the best performance. There were two probable reasons for the results.

- The goal of meta-training is to obtain the initial parameter for meta-testing task, so the training data volume for each meta-training task does not need to be much. Once each meta-training task has been well-trained, it may be hard to change the training direction for the meta-testing task. For example, when D_r^s and $D_r^q = 6000$, resulting in a total of 60000 data samples from five supplementary languages for each training round, we obtained relatively poor results compared to those with fewer utterances. This proved the importance of the volume of the training data.
- Considering the diversity problem in the supplementary languages, Mandarin had the largest number of utterances, while Japanese had the least. When D_r^s and $D_r^q = 6000$, in each training round, Mandarin had a high probability to select utterances that were different from others. But the total number of Japanese utterances was fewer than the number of utterances to be sampled, so some utterances were duplicated, affecting the performance of the Japanese task due to the low diversity of the training data.

TABLE 7. The WER/SER (%) for the meta-training round.

Meta-training round	WER/SER (%)	Meta-training round	WER/SER (%)
1	20.98/8.84	6	20.90/8.43
2	20.86/8.63	7	20.97/8.57
3	20.78/8.41	8	21.43/9.02
4	**20.24/8.35**	9	21.09/8.97
5	20.72/8.34	10	20.87/8.88

Next, this paper also experimented on the effect of the number of meta-training rounds in regard to recognition accuracy using the settings of the experiments above, where $(\alpha, \beta, \gamma) = (0.1, 0.2, 0.7)$ and both the utterance numbers of the support set and query set were 2000. The experiment was conducted by decoding the test data after each meta-training round and adjusting the model to recognize Taiwanese using the target domain data (i.e., Taiwanese). The recognition result for each round is shown in Table 7.

From the results in Table 7, we can see that the error rate of words and syllables dropped to the lowest when the training rounds are the 4th and the 5th rounds, and then increased unsteadily with the increase of the number of training rounds. This reconfirmed that by simply adding more training rounds may not obtain better performance. Therefore, it is very important to balance the training volume of the data in the target domain and the source domain. Besides, the number of training rounds for the selection of the utterances from the supplementary language should be empirically chosen to achieve the best performance.

In this paper, in order to demonstrate that the dynamic sampling method applied to MAML is effective, we used the parameter settings in Table 4 to construct the baseline model, transfer learning model and the random sampling-based MAML model for comparison. The baseline model was trained from scratch on Taiwanese corpus. Transfer learning was a classical method dealing with under-resource problem. The source domain of the transfer learning was trained with five supplementary languages, and the phoneme set, and the lexicon were the union of the five supplementary languages. After the training process for the source domain, the acoustic model was fine-tuned for the target domain. And the number of utterances in each training round was 2,000. In addition, we compared our experimental results with those of other teams that participated in FSR2020 [54]. In this challenge, the classic example scripts from Kaldi were utilized. Reference [55] proposed the Macaron architecture for the acoustic model. The architecture consists of TDNN-F layers and Time-Restricted Self-Attention (TRSA) layers. TRSA is similar to Transformer but with a greater emphasis on extracting local features. Reference [31] employed perturbation methods such as noise injection and SpecAugment to enhance the recognition ability of ASR. Additionally, we also compared our approach with Whisper. We utilized the Whisper-large-v3 model and adapted it to Taiwanese using AdaLoRA [56]. The experimental results are shown in Table 8.

TABLE 8. Error rate (%) of comparison methods.

Model	**WER**	**SER**
Baseline	26.18	13.99
[54]	-	12.81
[31]	-	13.47
[55]	-	15.30
Whisper	-	9.87
Transfer learning	23.93	11.06
Random sample based MAML [25]	20.89	8.86
Dynamic sample based MAML	**20.24**	**8.35**

The experimental results demonstrated that employing MAML as the training strategy and incorporating additional supplementary language data led to a significant improvement in recognition accuracy, regardless of the sampling method applied. In comparison to the baseline, transfer learning, and random sampling-based MAML, dynamic sampling-based MAML exhibits relative improvement rates of 22%, 15%, and 3% for WER, and 40%, 24%, and 5% for SER, respectively. Compared to other methods [54, 31, 55] and Whisper, the relative improvement rates were 34%, 38%, 45%, and 15%, respectively. These methods incorporate complex attention mechanisms and data augmentation techniques. However, in addressing the challenge of improving low-resource language data, the proposed dynamic sample-based MAML in this paper demonstrates significant advantages.

According to the result of dynamic sampling-based MAML and comparison results, dynamic sampling-based MAML had the best performance as shown in Table 8. Since the difference between this method and the random sampling-based MAML was the sampling method, the following will analyze the adjustment of data amount of the dynamic sampling method for each supplementary language per training round. The dynamic sampling results are given in Table 8, and the corresponding statistical results are shown in Table 9.

TABLE 9. Adjustment in the amount of training data in each language through dynamic sampling.

Meta-training round	MD	EN	JP	CA	TH
1	2000	2000	2000	2000	2000
2	2169	1325	2265	2014	2227
3	2182	1212	2300	1975	2331
4	2178	1200	2299	1972	2351

From the statistics of the training data in Table 9, we could see that the proportion of various supplementary language fluctuated in each training round. The amount of training data in the 4th training round from the greatest to the least were TH, JP, MD, CA, and EN. Among them, the amount of English and Cantonese corpus gradually decreased, especially in English, even though the English corpus had the largest resources. It was not easy to be sampled by the dynamic sampling method because of the few common phonemes. And the amount of Mandarin, Japanese and Thai corpus were increased after the 4th training round, especially in Thai. Unlike English, Thai had the second smallest amount of data among supplementary languages. But it is the supplementary language with the largest common phonemes (27 common phonemes), so it is favored by the dynamic sampling method.

Combined with the experimental results in Table 8 and the statistics in Table 9, we concluded some possible reasons:

- It can be seen from Table 8 that MAML was much better than the baseline model. Since the baseline model did not benefit from any supplementary languages, its performance cannot be compared with other fine-tuned methods such as transfer learning and MAML. Compared with the transfer learning, it is known that too much training in the source domain may make it difficult to transfer the trained knowledge to the target domain.
- In Table 8, the dynamic sampling method slightly outperformed the random sampling method. There are two possible reasons. First, as seen in Table 5, IQ is most beneficial for the dynamic sampling method. However, one of the three parameters that compose IQ, the confidence value C_U, is obtained by calculating the PPG values from the baseline model. The baseline model might provide unstable PPG values due to insufficient training data. Consequently, the quality of the baseline

TABLE 10. Phoneme accuracy (%) through various comparison methods.

Consonants & Nasal phonemes														
	DS	RS	TF	BL		DS	RS	TF	BL		DS	RS	TF	BL
p	81.22	79.95	77.92	75.88	k	79.89	79.56	79.22	79.19	g	78.89	78.46	77.09	76.45
ph	74.23	73.44	70.12	68.45	kh	79.33	78.86	77.65	76.03	j	80.23	78.23	77.72	73.77
m	81.35	79.55	77.88	74.48	h	80.36	80.03	79.21	77.23	nn	79.49	79.33	79.31	79.01
t	80.36	79.98	79.32	78.57	ts	79.78	79.59	79.23	79.19	ng	79.52	79.48	79.32	79.09
th	79.89	79.67	77.61	74.19	tsh	78.77	77.58	75.73	74.85	ng’	67.88	66.97	65.88	60.58
n	82.55	79.63	79.42	73.98	s	79.66	79.53	79.24	79.13	m’	76.22	75.77	74.63	74.56
l	79.82	79.56	79.26	77.98	b	79.39	78.97	77.63	76.54	nn’	78.33	78.21	77.02	76.99
Vowels					Stop consonant finals					Semivowels				
	DS	RS	TF	BL		DS	RS	TF	BL		DS	RS	TF	BL
a	87.71	84.32	82.95	80.87	p’	70.45	70.22	69.89	69.43	ngs	69.80	69.79	69.77	69.77
e	83.24	79.86	79.28	79.89	t’	78.03	77.86	77.05	76.88	ms	68.05	68.05	68.03	68.89
i	88.93	85.35	83.17	82.39	k’	76.79	76.03	75.66	75.64					
o	83.69	80.66	79.45	75.03	h’	78.02	77.93	77.61	77.59					
u	85.80	82.59	81.63	80.33										
er	84.36	81.87	81.36	80.21										

Average			
DS	**RS**	**TF**	**BL**
79.15	78.08	77.15	75.82

will directly affect the performance of the dynamic sampling method. A well-trained model could potentially widen the performance gap between the dynamic sampling method and the random sampling method. The second reason is the issue of imbalanced supplementary language data. From the statistics in Table 9, we know that Thai was most beneficial to Taiwanese ASR, so Thai had the largest amount of selected data by dynamic sampling. However, from the content of the supplementary language corpus, it is found that the duration of the Thai corpus was the second smallest, which led to that Thai corpus favored by the dynamic sampling method cannot provide enough data with higher IQ for selection. In addition, perhaps the problem of under resource, the performance of dynamic sampling method cannot be further improved with the increase of training rounds. Therefore, it is important for dynamic sampling methods with adequate and balanced supplementary language corpus. But even so, the dynamic sampling method can still improve the performance when the supplementary language corpus is extremely unbalanced (The ratio of duration of the most helpful language, Thai, to the least helpful language, English, is about 1:5), which proved that the dynamic sampling method is helpful for MAML.

In order to verify the benefits of adjusting the amount of training corpus for each supplementary language using the dynamic sampling method for Taiwanese ASR, we conducted a comparison of phoneme accuracy with different methods, including DS (dynamic sampling based MAML), RS (random sampling based MAML), TF (transfer learning), and BL (baseline). The results are presented in Table 10. Recognition accuracy of phonemes was obtained from the PPG, which was calculated from the probability of the phone at the frame level, and alignment information as reference to determine correct phone for each frame. The PPG-based phoneme recognition accuracy was calculated by counting the number of the hypothesis phones which were the same as the reference phones from alignment information at the same frame position.

From the result of Table 10, there was a relatively large gap between the average phoneme accuracy of BL and other three comparison methods. The average phone accuracy of DS was the best in these comparison methods.

This phenomenon not only shows that the accuracy of BL was relatively low due to the amount of training data, but also verifies that it was useful for DS to select favorable training data through dynamic sampling method. The selection of frequently used phonemes is one of the criterions for IQ, observing that the phonemes, such as ‘a’, ‘i’, and ‘u’, shown in Figure 6 are frequently used. The recognition accuracy of the three phones were 3% or above better than the second-best method (RS), and at least 5% better than BL method. According to the amount of training data in each supplementary language from Table 9, Thai was the most favorable supplementary language with the largest amount of the training data after the 4th dynamic sampling adjustment. From Figure 2, it is known that the most contributing phonemes in Thai language were ‘m’ and ‘n’, and the accuracy of these two phonemes was significantly improved compared to other frequently used phonemes due to the increased amount of the training data in Thai. It is also worth mentioning that the two semi-vowels (ngs and ms), which cannot be helped from any supplementary languages, were as accurate as the BL method regardless of the method used.

IV. CONCLUSION

In this study, considering the common phonemes in the high-resourced languages, a total of five supplementary languages are selected as additional training data to increase the training capacity for Taiwanese ASR. MAML is then adopted to learn the training data in each supplementary language. Based on MAML, the initial parameters of the model can obtain information from the supplementary data. In addition, this paper also proposes a dynamic sampling method for MAML instead of the original random sampling method, in which the distribution of supplementary data is scrambled in each training round and the utterances with higher information quantity are extracted to provide more help to model training. After implementing dynamic sampling based on phoneme analysis and the training performance assignment proposed in this paper, further improvement in phoneme accuracy was achieved, with a WER of 20.24% and a SER of 8.35% on the test set. The results show that training the model with more supplementary data can improve the results, and having the model trained with more informative languages and utterances can help improve the recognition of the target under-resourced languages.

The dynamic sampling and IQ determination approach proposed in this paper are based on the knowledge that beneficial to Taiwanese ASR for each utterance and the feedback from the current ASR system. However, the adoption of this method partially relies on rule-based approaches and requires a certain level of knowledge about the target language. To make this method applicable to various languages, future research could explore more objective features as factors influencing dynamic sampling-based MAML.

REFERENCES

[1] L. Deng and X. Li, "Machine learning paradigms for speech recognition: An overview," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 5, pp. 1060–1089, May 2013.

[2] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks," in *Proc. 23rd Int. Conf. Mach. Learn.*, 2006, pp. 369–376.

[3] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2016, pp. 4960–4964.

[4] K. J. Han, J. Pan, V. K. N. Tadala, T. Ma, and D. Povey, "Multistream CNN for robust acoustic modeling," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6873–6877.

[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," 2020, *arXiv:2005.08100*.

[6] S. Karita, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, R. Yamamoto, X. Wang, S. Watanabe, T. Yoshimura, and W. Zhang, "A comparative study on transformer vs RNN in speech applications," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2019, pp. 449–456.

[7] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 28492–28518.

[8] L. Besacier, E. Barnard, A. Karpov, and T. Schultz, "Automatic speech recognition for under-resourced languages: A survey," *Speech Commun.*, vol. 56, pp. 85–100, Jan. 2014.

[9] B. Thai, R. Jimerson, D. Arcoraci, E. Prud'hommeaux, and R. Ptucha, "Synthetic data augmentation for improving low-resource ASR," in *Proc. IEEE Western New York Image Signal Process. Workshop (WNYISPW)*, Oct. 2019, pp. 1–9.

[10] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, Sep. 2019, pp. 2613–2617, doi: 10.21437/interspeech.2019-2680.

[11] D. Yu, L. Deng, and G. E. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in *Proc. NIPS Workshop Deep Learn. Unsupervised Feature Learn.*, Dec. 2010, p. 8.

[12] I.-T. Hsieh, C.-H. Wu, and C.-H. Wang, "Acoustic and textual data augmentation for code-switching speech recognition in under-resourced language," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2020, pp. 302–307.

[13] A. V. Palacios, P. Acharya, A. S. Peidl, M. R. Beck, E. Blanco, A. Mishra, T. Bawa-Khalfe, and S. C. Pakhrin, "SumoPred-PLM: Human SUMOylation and SUMO2/3 sites prediction using pre-trained protein language model," *NAR Genomics Bioinf.*, vol. 6, no. 1, pp. 1–20, Jan. 2024, doi: 10.1093/nargab/lqae011.

[14] S. C. Pakhrin, S. Pokharel, K. F. Aoki-Kinoshita, M. R. Beck, T. K. Dam, D. Caragea, and D. B. Kc, "LMNglyPred: Prediction of human N-linked glycosylation sites using embeddings from a pre-trained protein language model," *Glycobiology*, vol. 33, no. 5, pp. 411–422, Jun. 2023, doi: 10.1093/glycob/cwad033.

[15] S. Parveen and P. Green, "Multitask learning in connectionist robust ASR using recurrent neural networks," in *Proc. Interspeech*, 2003, pp. 1–4.

[16] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, p. 10.

[17] S. C. Pakhrin, S. Pokharel, P. Pratyush, M. Chaudhari, H. D. Ismail, and D. B. Kc, "LMPhosSite: A deep learning-based approach for general protein phosphorylation site prediction using embeddings from the local window sequence and pretrained protein language model," *J. Proteome Res.*, vol. 22, no. 8, pp. 2548–2557, Aug. 2023.

[18] G. I. Winata, S. Cahyawijaya, Z. Lin, Z. Liu, P. Xu, and P. Fung, "Meta-transfer learning for code-switched speech recognition," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 1–22.

[19] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning fast adaptation on cross-accented speech recognition," in *Proc. Interspeech*, Oct. 2020, pp. 1–11.

[20] J.-Y. Hsu, Y.-J. Chen, and H.-Y. Lee, "Meta learning for end-to-end low-resource speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 7844–7848.

[21] R. Zhou, A. Ito, and T. Nose, "Character expressions in meta-learning for extremely low resource language speech recognition," in *Proc. 16th Int. Conf. Mach. Learn. Comput.*, Feb. 2024, pp. 525–529, doi: 10.1145/3651671.3651730.

[22] D. Eledath, A. Baby, and S. Singh, "Robust speech recognition using meta-learning for low-resource accents," in *Proc. Nat. Conf. Commun. (NCC)*, Feb. 2024, pp. 1–6, doi: 10.1109/ncc60321.2024.10485786.

[23] D. Wang, J. Yu, X. Wu, L. Sun, X. Liu, and H. Meng, "Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization," in *Proc. 12th Int. Symp. Chin. Spoken Lang. Process. (ISCSLP)*, Jan. 2021, pp. 1–5, doi: 10.1109/ISCSLP49672.2021.9362068.

[24] P. Zhou, Y. Zou, X.-T. Yuan, J. Feng, C. Xiong, and S. Hoi, "Task similarity aware meta learning: Theory-inspired improvement on maml," in *Uncertainty in Artificial Intelligence*. Breckenridge, CO, USA: PMLR, 2021, pp. 23–33.

[25] I.-T. Hsieh, C.-H. Wu, and Z.-H. Zhao, "Selection of supplementary acoustic data for meta-learning in under-resourced speech recognition," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2022, pp. 409–414, doi: 10.23919/APSIPAASC55919.2022.9979997.

[26] J. Zhang, J. Song, Y. Yao, and L. Gao, "Curriculum-based meta-learning," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 1838–1846.

[27] J. Zhang, J. Song, L. Gao, Y. Liu, and H. T. Shen, "Progressive meta-learning with curriculum," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 5916–5930, Sep. 2022.

[28] T. Wu, X. Li, Y.-F. Li, R. Haffari, G. Qi, Y. Zhu, and G. Xu, "Curriculum-meta learning for order-robust continual relation extraction," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 10363–10369.

[29] R. Zhan, X. Liu, D. F. Wong, and L. S. Chao, "Meta-curriculum learning for domain adaptation in neural machine translation," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 14310–14318.

[30] Y.-F. Yeh, B.-H. Su, Y.-Y. Ou, and J.-F. Wang, "Taiwanese speech recognition based on hybrid deep neural network architecture," in *Proc. 32nd Conf. Comput. Linguistics Speech Process.*, 2020, pp. 102–113.

[31] F.-A. Chao, T.-H. Lo, S.-Y. Weng, S.-H. Chiu, Y.-T. Sung, and B. Chen, "The NTNU Taiwanese ASR system for formosa speech recognition challenge 2020," 2021, *arXiv:2104.04221*.

[32] (2014). *King-ASR-044, Speechocean*. [Online]. Available: https://en.haitianruisheng.com/dataset/c52-5369.htm

[33] C.-R. Huang. (2009). *Tagged Chinese Gigaword Version 2.0*. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2009T14

[34] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*. Australia: IEEE, Apr. 2015, pp. 5206–5210.

[35] V. Panayotov, D. Povey, and S. Khudanpur. (2014). *LibriSpeech Language Models, Vocabulary and G2P Models*. [Online]. Available: https://www.openslr.org/11/ 2024/7/7

[36] (2021). *Mozilla Common voice*. [Online]. Available: https://commonvoice.mozilla.org/zh-TW/datasets 2024/7/7

[37] R. Sonobe, S. Takamichi, and H. Saruwatari, "JSUT corpus: Free large-scale Japanese speech corpus for end-to-end speech synthesis," 2017, *arXiv:1711.00354*.

[38] (2021). *Wikimedia Downloads*. [Online]. Available: https://dumps.wikimedia.org/backup-index.html 2024/7/7

[39] T. Matsushita. (2011). *A Vocabulary Database Sorted in Order of Importance (Vocabularies in order of importance)*. [Online]. Available: http://www17408ui.sakura.ne.jp/tatsum/database.html

[40] R. Lai and G. Winterstein. (2024). *Marseille*. [Online]. Available: https://github.com/gwinterstein/Cifu

[41] E. Chuangsuwanich, A. Suchato, K. Karunratanakul, B. Naowarat, C. Chaichot, and P. Sangsa-nga. (2024). *Gowajee Corpus*. [Online]. Available: https://github.com/ekapolc/gowajee_corpus

[42] W. Phatthiphaiboon. (2017). *Lexicon-Thai*. [Online]. Available: https://github.com/PyThaiNLP/lexicon-thai

[43] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. vesely, "The kaldi speech recognition toolkit," in *Proc. IEEE Workshop Automatic Speech Recognition Understanding*, Aug. 2011, p. 4.

[44] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," 2018, *arXiv:1803.02999*.

[45] (2022). *Taiwanese Across Taiwan Corpus*. [Online]. Available: https://sites.google.com/speech.ntut.edu.tw/fsw/home/tat-corpus?authuser=0

[46] H.-C. Chang, C.-Y. Kao, and H.-C. Lu. (2006). *Taiwanese Language Digital Archive Database*[Online]. Available: https://db.nmtl.gov.tw/site3/plan

[47] (1995). *New Testament Corpus*. [Online]. Available: https://bible.fhl.net/

[48] Y.-Y. Yang and H.-C. Chang. *Collection of Taiwanese Language Corpus and Corpus Statistics of Syllable and Word Frequency of Taiwanese Written Syllables*. [Online]. Available: https://pypi.org/project/hue7jip8/ 2024/7/7

[49] (2024). *iCorpus: Taiwanese Chinese News Corpus*. [Online]. Available: https://iptt.sinica.edu.tw/shares/465

[50] Y.-Y. Yang. (2016). *Taiwanese National School Textbooks*. [Online]. Available: https://github.com/Taiwanese-Corpus/kok4hau7-kho3pun2 20247/7.

[51] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in *Proc. Interspeech*, 2015, p. 5.

[52] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, Sep. 2016, pp. 2751–2755.

[53] A. Stolcke, "SRILM—An extensible language modeling toolkit," in *Proc. Interspeech*, 2002, p. 4.

[54] Y.-F. Liao, C.-Y. Chang, H.-K. Tiun, H.-L. Su; H.-L. Khoo, J. S. Tsay, L.-K. Tan, P. Kang, T.-G. Thiann, U.-G. Iunn, J.-H. Yang, and C.-N. Liang, "Formosa speech recognition challenge 2020 and Taiwanese across Taiwan corpus," in *Proc. 23rd Conf. Oriental COCOSDA Int. Committee Co-ordination Standardisation Speech Databases Assessment Techniques*, Aug. 2020, pp. 65–70.

[55] H.-P. Lin, Y.-J. Zhang, and C.-P. Chen, "Systems for low-resource speech recognition tasks in open automatic speech recognition and formosa speech recognition challenges," in *Proc. Interspeech*, Aug. 2021, pp. 4339–4343.

[56] Q. Zhang, M. Chen, A. Bukharin, N. Karampatziakis, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," in *Proc. 11th Int. Conf. Learn. Represent.*, 2022, pp. 1–20.

I-TING HSIEH received the B.S. and M.S. degrees in electrical engineering from the Southern Taiwan University of Science and Technology, Tainan, Taiwan, in 2017 and 2019, respectively. He is currently pursuing the Ph.D. degree with National Cheng Kung University (NCKU). His research interests include speech signal processing and speech recognition.

CHUNG-HSIEN WU (Senior Member, IEEE) received the B.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1987 and 1991, respectively. Since 1991, he has been with the Department of Computer Science and Information Engineering, NCKU. He was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in summer 2003. He was the Deputy Dean of the College of Electrical Engineering and Computer Science, NCKU, from 2009 to 2015. He became the Chair Professor, in 2017. His research interests include deep learning, affective computing, speech recognition/synthesis, and spoken language processing. He was a APSIPA BoG Member, from 2019 to 2021. He received the 2018 APSIPA Sadaoki Furui Prize Paper Award, in 2018, and the Outstanding Research Award of Ministry of Science and Technology, Taiwan, in 2010 and 2016. He was an Associate Editor of IEEE TRANSACTIONS ON AUDIO, SPEECH AND LANGUAGE PROCESSING, from 2010 to 2014, IEEE TRANSACTIONS ON AFFECTIVE COMPUTING, from 2010 to 2014, *ACM Transactions on Asian and Low-Resource Language Information Processing*, and *APSIPA Transactions on Signal and Information Processing*, from 2014 to 2020.

ZHE-HONG ZHAO received the B.S. and M.S. degrees in computer science and information engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 2019 and 2021, respectively.