Applied Acoustics 218 (2024) 109883

Contents lists available at ScienceDirect

Applied Acoustics

journal homepage: www.elsevier.com/locate/apacoust

Tri-stage training with language-specific encoder and bilingual acoustic learner for code-switching speech recognition

Xuefei Wang [a,b,1], Yuan Jin [b], Fenglong Xie [b], Yanhua Long [a,c,*]

[a] Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai, China
[b] Xiaohongshu Inc., Shanghai, China
[c] Unisound AI Technology Co., Ltd., Beijing, China

ARTICLE INFO

Keywords:
Code-switching
End-to-end speech recognition
Language-specific encoder
Bilingual acoustic learner

ABSTRACT

Recently, two-pass end-to-end (E2E) automatic speech recognition (ASR) systems with the conformer model followed by a spelling correction backend have demonstrated remarkable progress and exceptional performance in general speech recognition tasks. However, these models may fail when they come to code-switching (CS) speech, where a speaker alternates words of two or more languages within a single sentence or across sentences. In this study, we propose a novel **t**ri-stage **t**raining **two**-pass (TripleT) E2E framework to improve the CS ASR performance by leveraging the individual attributes of each monolingual language. Our framework starts by introducing two symmetric language-specific encoders that are pre-trained using a large monolingual corpus. This improves the high-level acoustic representation of each individual language. Then, a bilingual acoustic learner (BAL) is proposed to combine these language-specific representations and transfer the monolingual acoustic attributes to code-switching properties. Next, these acoustic representations are further utilized to boost the spelling corrector by a context plus acoustic learner with the same structure as BAL. Finally, the whole proposed framework is fine-tuned using the CS corpus to achieve the final CS E2E ASR system. Our experiments are performed on a mixed training dataset consisting of 1000 hours of Mandarin data, 960 hours of English data, and 555.9 hours of Mandarin-English code-switching data. The ASR performances are evaluated on a 23.6 hours CS test set, and results show that our proposed TripleT-E2E framework achieves a 13.4% relative reduction in token error rate compared to a competitive two-pass E2E baseline model.

1. Introduction

In recent years, end-to-end (E2E) automatic speech recognition (ASR) models have gained increasing research attention because of their excellent performance and unified neural network architectures [1–3]. These E2E models have been widely used in a variety of large-scale monolingual [4–6] and multilingual ASR tasks [7,8]. However, it should be noted that ASR performance is significantly degraded when presented with bilingual mixing speech, also known as code-switching (CS) speech. CS refers to the phenomenon of mixing words or phrases from distinct languages by a speaker and it is common in multilingual communities such as Cantonese-English [9], Mandarin-English [10], Frisian-Dutch [11] and Spanish-English [12]. With the increasing frequency of international exchanges, code-switching has become more prevalent, leading to broad research interests in CS ASR in recent speech recognition studies [13–15].

Many recent progresses have been achieved in end-to-end automatic speech recognition [16,17], one of the focuses of these recent works is improving the ASR system performance by adding an additional spelling correction backend module [15] after an E2E ASR decoder in a two-pass manner. Although these models have shown excellent performance in various monolingual speech recognition tasks, building a code-switching ASR system presents a greater challenge. This challenge arises from both the inherent nature of code-switching speech, which involves the mixing of languages, and the severe data sparsity issue specific to code-switching events in both speech and text. Even if the

* Corresponding author at: Shanghai Engineering Research Center of Intelligent Education and Bigdata, Shanghai Normal University, Shanghai, China. Yanhua Long is also with the Key Innovation Group of Digital Humanities Resource and Research, Shanghai Normal University.
E-mail addresses: xuefei_wang@163.com (X. Wang), zhiming@xiaohongshu.com (Y. Jin), fenglongxie@xiaohongshu.com (F. Xie), yanhua@shnu.edu.cn (Y. Long).
[1] This work was done during internship at Xiaohongshu Inc.

https://doi.org/10.1016/j.apacoust.2024.109883
Received 19 July 2023; Received in revised form 22 December 2023; Accepted 19 January 2024
Available online 25 January 2024

individual languages involved are well-resourced, there is a scarcity of available training data for code-switching. Only a limited number of corpora with small amounts of code-switching training data are currently available [18,19]. In contrast to the abundance of training data available for general monolingual ASR tasks, which often consists of thousands of hours of data [20,21], the largest open-resource corpus designed for Mandarin-English code-switching ASR, known as the TALCS corpus [19], still only contains 555.9 hours of training data.

Aiming to alleviate the impact of limited code-switching training data, this study investigates leveraging language-specific acoustic information learned from a large amount of monolingual data to improve both ASR and the corresponding spelling corrector modules in E2E manner for Mandarin-English code-switching speech recognition. Specifically, we propose a new **t**ri-stage **t**raining **two**-pass (TripleT) E2E framework to leverage the individual monolingual language attributes at different training stages. Based on a conformer-based E2E spelling correction architecture, we first introduce two symmetric language-specific encoders, which are well pre-trained on a large monolingual corpus. The purpose is to enhance the high-level acoustic representation for each language independently. Next, a bilingual acoustic learner (BAL) is proposed to combine these language-specific representations and transfer the monolingual acoustic attributes in these representations to a code-switching property. Furthermore, these acoustic representations are leveraged to improve the spelling corrector by a context-plus acoustic learner with the same structure as BAL. The entire proposed framework is then fine-tuned using the CS corpus to achieve the final CS E2E ASR system. All experiments are conducted on the public Mandarin-English code-switching ASR corpus TALCS [19], results show that our proposed TripleT-E2E framework outperforms the Conformer-based two-pass E2E baseline model significantly.

To the best of our knowledge, this is the first study to propose methods to utilize pre-trained language-specific encoders for improving two-pass end-to-end CS ASR system. The rest of this paper is organized as follows. Section 2 presents the review of previous works. In Section 3, we briefly describe the fundamental of the Conformer-based ASR architecture. In Section 4, we introduce the proposed TripleT-E2E framework, including the model structure and tri-stage training strategy. Experimental setup and results are presented in Section 5 and 6. Finally, we conclude the study in Section 7.

2. Review of previous works

Code-switching, or the alternation between languages within a conversation, is a natural linguistic phenomenon common among bilingual and multilingual speakers [22]. Code-switching can be broadly categorized into inter-sentential switching between sentences and intra-sentential switching within a sentence [23]. Notably, in real code-switching ASR scenarios, like the TALCS dataset we used in this study, both inter-sentential and intra-sentential code-switching acoustic events are encompassed. Quantifying code-switching presents considerable difficulty, and there is a continuous effort to establish objective measures for its degree. Various metrics have been proposed in the literature to quantify the code-switching. For example, Hou et al. [24] utilized language entropy and the probability of switching, Bullock et al. [25] explored the numerically dominant language overall, of all verbs, and a subset of system morphemes for measuring intra-sentential code-switching mixing, while Myslín and Levy [26] introduced a normalized IU-position metric. Given these considerations, Code-switching speech introduces various challenges for developing automated ASR systems, including acoustic and language modeling for mixed languages, pronunciation modeling, language identification from speech, etc. To handle these challenges, previous works mainly focus on (i) data sparsity for both acoustic and language modeling; (ii) the co-articulation effects between target modeling units at code-switches.

The code-switching data sparsity problem is a long-standing issue in the field of CS speech recognition, either for conventional hybrid ASR systems or for end-to-end ASR models. Because in each CS utterance, the speech data is normally dominated by the matrix language, acoustic events of code-switches are extremely sparsity in the CS training corpus. In addition, creating a large-scale code-switching speech recognition corpus with a golden standard manual transcription incurs high costs of both time and money due to the labor-intensive nature of manual transcription, the need for skilled annotators, and the time-consuming process of ensuring accurate and consistent transcriptions across a substantial amount of data. Additionally, the complexity of code-switching scenarios and the diverse linguistic contexts involved contribute to the resource-intensive nature of this task. In recent years, the previous works focused on creating a CS corpus are also limited, such as the SEAME corpus with 30 hours of spontaneous intra-sentential CS speech [27], the OC16-CE80 corpus with 80 hours of training data provided for the Chinese-English mixlingual speech recognition challenge (MixASR CHEN) [28], the Arabic-English cs corpus with 12 hours [29], and the TALCS [19] with around 560 hrs training data that was used in our study, etc. Besides, some other researchers explored new CS data augmentation methods to alleviate the impact of limited code-switches training events, and the augmentation using text-to-speech (TTS) system [30–32], etc. Although significant efforts have been made by the academic and industrial communities in creating CS datasets, the current available CS training data is still insufficient for building a successful code-switching ASR system, especially for the recently proposed end-to-end models that normally require large amounts of training data. Therefore, many recent works focus on techniques to alleviate the data sparsity during acoustic and language modeling.

To alleviate the effect of co-articulation at code-switch points, most previous approaches were proposed based on the conventional hybrid deep neural network-hidden Markov model (DNN-HMM) ASR frameworks. These approaches mainly focused on mix-language phone mapping and phone sharing, for example, (i) combining phone sets from two languages [33,34], (ii) mapping phone sets from two languages [35,36], and (iii) merging similar phone sets from two languages [37–39]. However, with the rapid progress of end-to-end acoustic modeling techniques, the end-to-end ASR systems have been widely used for code-switching ASR [40–42]. These systems simplified the building of the CS ASR system by directly predicting the combined graphemes or characters of different languages from acoustic input. The co-articulation problem in the E2E ASR strategy is significantly diminished since the reliance on pronunciation dictionaries or lexicons is no longer required. However, since the E2E ASR models are very deep neural networks, the training data of code-switches are extremely sparse, how to leverage the pronunciation-related linguistic knowledge between different languages to assist the E2E CS ASR system will be interesting and worthwhile.

The acoustic modeling is always very important for code-switching ASR system building, many previous works have been explored. They mainly focus on how to model the acoustic characteristics of different languages well under the condition of limited CS training data. For example, in [43,44], different ways of phoneme-merging between multiple languages were proposed to find effective mixed-language acoustic modeling target units. In [45–47], authors explored using bi-encoder or multi-encoder-decoder structures. They pre-trained the language-specific encoders using monolingual data in each language. Then the individual language attributes are extracted and combined to obtain the mixture features with bilingual information. In addition, some multi-task learning technologies also have been explored for code-switching ASR. Authors of [48] trained a CTC-Attention model [49] for speech recognition and used a frame-level language identification model to adjust the posteriors. In [50], they proposed a language-related attention mechanism to reduce confusion in multilingual contexts for the E2E code-switching ASR model. Many other related works, such as the recently proposed language-specific acoustic boundary learning [42], token-level language diarization [51], the internal language model estimation for E2E CS ASR [52], etc. All these previous works have been greatly boosted the performance of code-switching speech recognition,

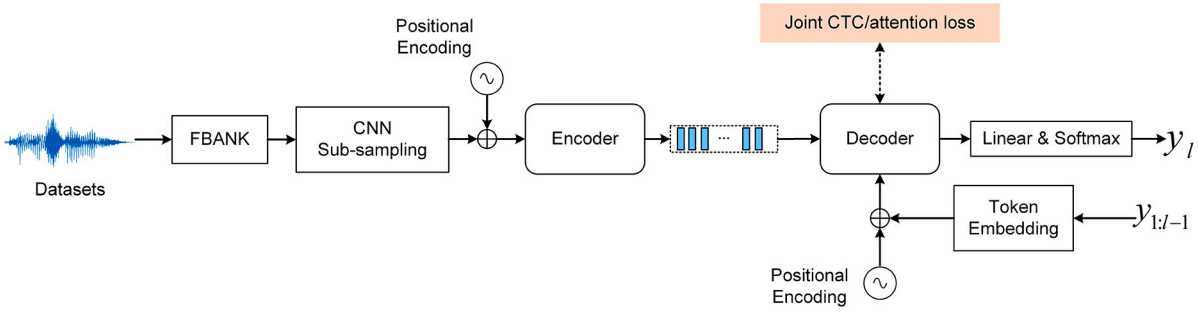

Fig. 1. The architecture of the basic Conformer model.

however, most of them are based on the popular end-to-end ASR model (e.g. attention based Transformer or Conformer [1,53]), or even the traditional hybrid DNN-HMM ASR frameworks [54,55], exploring to improve an E2E ASR model with a spelling correction backend for code-switching ASR is limited.

In this study, we focus on improving Mandarin-English CS ASR performance by taking the E2E ASR model Conformer [53] with a spelling correction backend as our strong baseline. A novel TripleT-E2E framework is proposed, it leverages the attributes of individual languages that learned from large amount of monolingual corpus to improve CS ASR performance. The symmetric language-specific encoders, bilingual acoustic learner (BAL) for enhancing the fundamental ASR module, and the context plus acoustic learner (CAL) for boosting the spelling corrector are investigated.

3. Conformer-based E2E ASR

In this paper, all of our contributions are based on convolution-augmented Transformer (Conformer) E2E ASR model that has been proposed in [53]. The whole architecture of Conformer is shown in Fig. 1. Given the raw audio, we first extract the log Mel-filter bank (FBANK) as the acoustic features, next, these FBANKs are sub-sampled using the CNN Sub-sampling block, which is a crucial step in alleviating computational demands and emphasizing significant information for speech recognition. To provide positional information, positional encoding is incorporated into the features. Then, the positional encoded features are fed into an Encoder-Decoder structure to produce the final speech transcription. Specifically, the Encoder is used to extract high-level acoustic representative embeddings, these acoustic embedding and the positional encoded token embedding (the output token at the previous time step $y_{1:l-1}$) are further transformed by the Decoder block, followed by a linear transformation and a softmax function to produce the final posterior probabilities on a set of target modeling tokens. These tokens are usually the phonemes, characters or sub-word units.

The Encoder and Decoder in Fig. 1 have different structures. In this study, the Encoder is composed of several Conformer blocks with self-attention mechanism, while the Decoder consists of several Transformer blocks with a masked source-target multi-head attention. In order to benefit from the monotonic alignment, the Connectionist Temporal Classification (CTC) loss function and attention-based mechanisms are used to jointly train the model in a multi-objective manner.

Different from the standard Transformer-based E2E ASR system design [56], the Conformer-based E2E ASR replaces the original feed-forward layer in the Transformer encoder block into two half-step feed-forward layers which are inspired by MacaronNet [57]. This replacement enhances the model an improved ability to capture complex patterns and dependencies within the input sequence. For enhancing the local information modeling capability, the Conformer also uses a convolution module that contains a gating mechanism after multi-headed self-attention. By combining self-attention with convolutional techniques, the Conformer encoder is good at capturing both short-range and long-range connections in sequential data. Due to its consistently superior performance across a wide range of ASR tasks, more and more Conformer variants have been explored in recent years [58,59].

The basic Conformer block of Encoder in Fig. 1 mainly consists of four modules: (1) the first half-step feed-forward module (FFN1): Processing the input sequence by applying the feed-forward transformation; (2) the multi-head self-attention module (MHSA): Capturing global dependencies by allowing the model to attend to different parts of the input sequence simultaneously; (3) the convolution module (Conv): Incorporating convolutional operations, including a gating mechanism, to enhance local information modeling. and (4) the second half-step feed-forward module (FFN2): Completing the feed-forward transformation, combining the acquired features to generate the ultimate output. Given an input sequence x, the output y of a Conformer block can be mathematically defined as follows:

$$\begin{aligned} x_{\mathrm{FFN}_1} &= x + \frac{1}{2}\mathrm{FFN}(x), \\ x_{\mathrm{MHSA}} &= x_{\mathrm{FFN}_1} + \mathrm{MHSA}(x_{\mathrm{FFN}_1}), \\ x_{\mathrm{Conv}} &= x_{\mathrm{MHSA}} + \mathrm{Conv}(x_{\mathrm{MHSA}}), \\ x_{\mathrm{FFN}_2} &= x_{\mathrm{Conv}} + \frac{1}{2}\mathrm{FFN}(x_{\mathrm{Conv}}), \\ y &= \mathrm{LayerNorm}(x_{\mathrm{FFN}_2}) \end{aligned} \tag{1}$$

The Transformer block structure of the Decoder in Fig. 1 is the same as in standard Transformer-based E2E ASR system design [56]. For further details on the Conformer end-to-end automatic speech recognition, please refer to [53] as your reference.

4. The proposed TripleT-E2E CS ASR framework

In this section, we provide an in-depth exploration of our proposed TripleT-E2E CS ASR framework, specifically designed to enhance the performance of end-to-end code-switching speech recognition. The framework integrates symmetrical language-specific encoders, namely ZH-encoder (for Mandarin) and EN-encoder (for English), pre-trained on monolingual datasets to capture nuanced acoustic information. Additionally, a Bilingual Acoustic Learner (BAL) plays a key role in combining language-specific representations and transferring monolingual attributes to code-switching properties. To boost overall accuracy, the proposed framework also includes a spelling correction(SC) system with a Context plus Acoustic Learner (CAL). The training involves three stages: pre-training, adaptation, and fine-tuning. The whole model architecture is presented in Section 4.1. The tri-stage training strategy and multi-objective learning method are described in detail in Section 4.2 and 4.3, respectively.

4.1. Model architecture

The whole proposed **t**ri-stage **t**raining **two**-pass (TripleT) The E2E framework for the Mandarin-English code-switching speech recognition is illustrated in Fig. 2. In this TripleT-E2E framework, the 'two-pass' means the final ASR inference implements a standard Conformer in

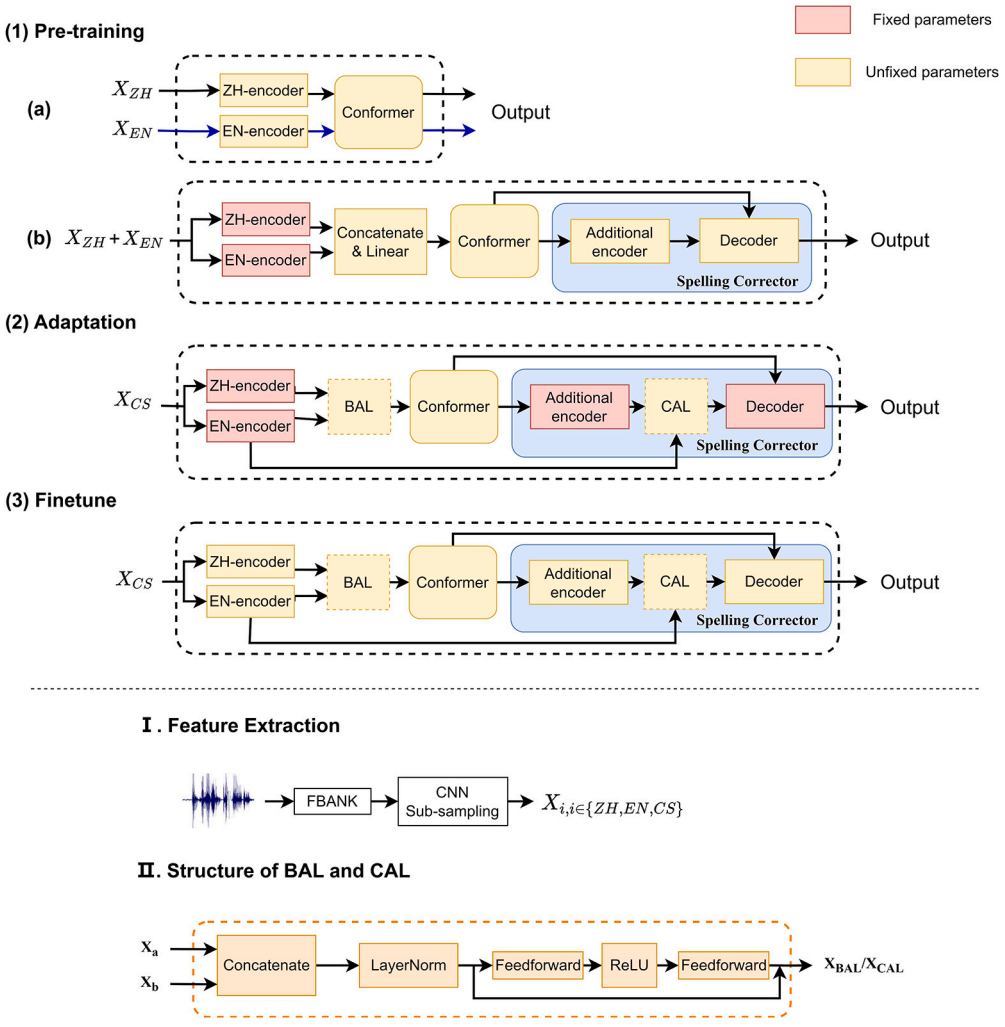

Fig. 2. The framework of proposed tri-stage training two-pass (TripleT) E2E CS ASR.

the first pass and includes spelling correction in the second pass. The three-stage training process comprises the following phases: (1) `Pre-training`: This initial stage focuses on preparing the model through a pre-training process. (2) `Adaptation`: The second stage involves adapting the model to better suit the characteristics of Mandarin-English code-switching speech. (3) `Fine-tuning`: The third and final stage encompasses fine-tuning the model, refining it for optimal performance in the code-switching ASR task. The resultant model from this stage is utilized for ASR inference.

As shown in Fig. 2 I, before the tri-stage model training, we need to extract the FBANK acoustic features and perform CNN sub-sampling to obtain the final model inputs X_i, $i \in \{ZH, EN, CS\}$. In Fig. 2, the model structures of `Pre-training` and the last two stages are very different. Based on the standard Conformer structure that described in Section 3, we additionally add two symmetric language-specific encoders, the ZH-encoder and EN-encoder, one bilingual acoustic learner (BAL), and the spelling corrector backend with a context plus acoustic learner (CAL) module in the TripleT-E2E CS ASR framework. The Conformer module is the Encoder-Decoder structure in Fig. 1. Except for the BAL and CAL modules, all the -encoder and decoder in the proposed TripleT-E2E framework have the same structures as the ones contained in standard Conformer module. Beyond the standard Conformer, the principle details of all our proposed modules are presented below.

Symmetric Language-specific Encoders: In our CS ASR task, the symmetric ZH-encoder and EN-encoder are designed for capturing the high-level acoustic representation of the matrix language Mandarin and the embedding of foreign language English, respectively. As depicted in Fig. 2 (a), both encoders are well pre-trained using a large amount of corresponding monolingual data independently. The ZH-encoder is designed to capture the acoustic features of Mandarin, going through a strong pre-training phase with a large dataset of monolingual Mandarin speech. During the pre-training phase, it produces acoustic representations specific to Mandarin. These representations serve as a foundation for effectively capturing the acoustic information in code-switching utterances involving Mandarin. Similarly, the EN-encoder is designed to incorporate English, going through a dedicated pre-training process using a substantial dataset of monolingual English speech. After pre-training, the EN-encoder gains the ability to extract language-specific acoustic information unique to English. This ensures that the encoder is highly effective at capturing the acoustic information of English speech in code-switching scenarios.

Given an utterance in a code-switched language, the extractor will produce language-specific representations, enhancing both the linguistic and acoustic information of both languages. This is particularly advantageous for the embedded foreign language, which may have limited data in the CS training corpus. As a result, the individual language attributes can then be effectively utilized to enhance both the Conformer and spelling corrector of our proposed TripleT-E2E CS ASR framework.

Bilingual Acoustic Learner: The bilingual acoustic learner (BAL) is one component of the final model of our proposed TripleT-E2E framework. As shown in Fig. 2, BAL receives the outputs of two symmetric language-specific encoders and transforms them to be used as input for the Conformer encoder. In Fig. 2, stage (2) and (3), the BAL is specially designed to replace the “Concatenate & Linear” module from

stage (1)(b), this is because during both the "Adaptation" and "Fine-tuning" stages, the model is trained on limited code-switching utterances and generates CS text ground-truth. In this scenario, BAL serves two purposes: effectively combining the acoustic representations from the symmetric language-specific encoders, and transferring the monolingual Mandarin and English acoustic attributes into a code-switching property.

The structure of BAL is illustrated in Fig. 2 II, it consists of two feed-forward layers with a residual connection. In order to make the BAL non-linear, we put a ReLU activation function between the two feed-forward layers. The input embeddings X_a and X_b are first Concatenated and then layer normalized. The output X_{BAL} of the BAL module can be mathematically defined as:

$$\begin{aligned} X_f &= \text{Concatenate}(X_a, X_b), \\ X_F &= \text{LayerNorm}(X_f), \\ X_{BAL} &= X_F + \text{FNN}(\text{ReLU}(\text{FNN}(X_F))) \end{aligned} \tag{2}$$

Spelling Correction with Context plus Acoustic Learner: In addition to the language-specific symmetric encoders and the bilingual acoustic learner, we also incorporate a module called the context plus acoustic learner (CAL) into the standard Conformer based spelling correction model. This inclusion is proposed to further strengthen our proposed TripleT-E2E framework to improve the performance of CS ASR tasks.

The structure of spelling corrector is shown in the module with a blue background in Fig. 2. It contains an additional encoder, a CAL, and a decoder block. The encoder and decoder have the same structures as the ones in Conformer. CAL shares the same structure as our proposed BAL but with different inputs and plays different roles in the whole TripleT-E2E framework. As the name of CAL, it is inserted to enhance the spelling corrector by combining the information between Conformer hypothesis contextual and language-specific acoustic representations. It is worth noting that the language-specific representation CAL input can be only the output of ZH-encode, or EN-encoder, or the additive of both of them. And, the output of CAL in Fig. 2 II is a high-level information representation with combined Conformer hypothesis contextual and language-specific acoustic representations. With this CAL, the final output of spelling corrector can be formalized as:

$$\begin{aligned} &P(y_1, y_2, \ldots, y_d) = \\ &\prod_{d=1}^{D} P(y_d | y_1, y_2, \ldots, y_{d-1}, l_{hyp}, X_{acoustic-asr}, X_{acoustic-sym}) \end{aligned} \tag{3}$$

where the l_{hyp} and $X_{acoustic-asr}$ are the decoder hypothesis and acoustic embedding of encoder output of Conformer module, while the $X_{acoustic-sym}$ is the language-specific acoustic representation produced by the symmetric encoders. In addition, it should be noted that the input of the SC module decoder is the output of the encoder in the conformer, and the input of the SC module additional encoder is the output of the conformer decoder.

4.2. Tri-stage training strategy

Our proposed TripleT-E2E framework for enhancing the CS ASR system is motivated by the substantial difference in data quantity between code-switching and its corresponding monolingual and foreign languages. It is very easy to obtain a large amount of monolingual data with accurate transcriptions, such as the thousands of hours of Mandarin and English ASR training data available from open-source resources like WenetSpeech [20] and Librispeech [60] corpora. However, the largest available corpus we have found for Mandarin-English CS ASR, is the TALCS, only consists of approximately 500 hours of data, with significantly fewer code-switch acoustic events for model training.

Therefore, in order to well utilize the information in a large amount of monolingual training corpus for boosting the CS ASR model training, the following tri-stage training strategy is proposed:

- (1) `Pre-training`. As shown in Fig. 2(1), the Pre-training stage includes two sub-stages, the first one is (a): training two language-specific encoders (ZH/EN-encoder) independently together with the Conformer structure to build two independent monolingual ASR system, one is for Mandarin and the other is for English speech recognition. Because these two systems are trained from a large amount of Mandarin and English training data independently, after this stage, we believe that the ZH/EN-encoder is well-pre-trained and endowed with the ability to produce representative language-specific acoustic embeddings.
  The second pre-training sub-stage (b) is to train the whole Conformer based spelling correction model using the combined Mandarin and English training datasets ($X_{ZH} + X_{EN}$), by fixing the parameters of two pre-trained encoders (ZH/EN-encoder). This sub-stage is to provide a good initialization of Conformer and spelling corrector modules for the next training stage.
- (2) `Adaptation`. In this stage, we froze the parameters of both symmetric language-specific encoders and the spelling corrector that initialized in the above stage (1)(b), only using the limited CS training data (X_{CS}) to adapt the Conformer and the inserted BAL and CAL to endow the model with CS speech recognition ability. As discussed in Section 4.1, the inserted BAL and CAL in this adaptation stage is to leverage the pre-trained monolingual information to enhance the ASR and spelling corrector.
- (3) `Fine-tuning`. In the last training stage, the whole model is fine-tuned using the code-switching training data (X_{CS}) to further improving the final performance of TripleT-E2E CS ASR model, with all the initialized parameters obtained in stage (1) and (2).

After the above tri-stage training, the final model of the proposed TripleT-E2E framework is then used for CS ASR inference. Thanks to the design of symmetric language-specific encoders, BAL and CAL, the tri-stage training makes the final model successfully integrates well pre-trained monolingual information into the CS speech recognition and its backend spelling correction model. Because pre-training on massive monolingual data provides a useful initialization, establishing representative language-specific knowledge to prepare the model for adaptation. And in stage (2), the use of selective adaptation of key components on code-switching data helps transferring the language-specific information to the target CS condition while avoiding over-fitting. The fine-tuning stage updates all the model parameters to a code-switching style and finally improves the Mandarin-English code-switching ASR performance.

4.3. Multi-objective learning

We propose utilizing a joint loss function to train all modules in the model architecture simultaneously. The combined loss aims to optimize the entire system to maximize gains from each component and enable complementary modeling of various aspects that are important for the code-switching ASR task. The formula is shown below:

$$\mathcal{L} = (1 - \lambda - \mu)\mathcal{L}_{att} + \lambda\mathcal{L}_{ctc} + \mu\mathcal{L}_{ce} \tag{4}$$

with the tuning parameters $\lambda, \mu \in [0, 1]$. $\mathcal{L}_{ctc}$ is the CTC loss of Conformer module that provides initial sequence alignments to ground the model, enabling basic sequence prediction where repetitions are common. $\mathcal{L}_{att}$ is the attention-based Kullback-Leibler divergence (KLD) loss that refines alignments and predictions in the Conformer decoder based on learned relationships between inputs and outputs. The cross-entropy (CE) loss ($\mathcal{L}_{ce}$) in the spelling correction decoder optimizes that component for refined prediction at the word level. In the whole TripleT-E2E framework tri-stage training, except for the (a) step of pre-training

Table 1
Details of monolingual English (LibriSpeech), Mandarin (WenetSpeechM) corpus and the code-switching TALCS datasets.

Corpus	Train		Test	
	#Utt	#Duration(hrs)	#Utt	#Duration(hrs)
LibriSpeech	286808	960	-	-
WenetSpeech-M	1514500	1000	-	-
TALCS	350000	555.9	15000	23.6

Table 2
The composition of Mandarin, English monolingual speech, and Mandarin-English CS speech in TALCS dataset.

		Mandarin	English	Code-Switching
Train	Ratio	53.6%	2.1%	44.2%
	#Utterances	187835	7432	154732
Test	Ratio	53.1%	2.0%	44.8%
	#Utterances	7975	305	6720

stage, two monolingual ASR systems are trained only using $\mathcal{L}_{att}$ and $\mathcal{L}_{ctc}$, all other models are trained with the combined total loss $\mathcal{L}$ to utilize multiple complementary modeling components.

In our experiments, the training dataset is quite large, including the pretrained 960 hrs of LibriSpeech, 1000 hrs WenetSpeech-M and 555.9 hours of Mandarin-English code-switching data, which leads to a quite long model training cycle. Therefore, we did not extensively tune these parameters λ and μ. Instead, we empirically assigned two sets of the λ and μ values on the development set, and then we picked the best one as $\lambda = 0.3$ and $\mu = 0.2$ for all our experiments.

5. Experiments

5.1. Datasets

Two large monolingual corpora and one code-switching corpus are used for our system training and evaluation. Specifically, the 960-hour English LibriSpeech corpus [60] and 1000-hour WenetSpeech-M [20] corpus are taken as our monolingual training data to pre-train the models in the `Pre-training` stage. The TALCS [19] is a dataset containing Mandarin-English code-switching speech data. It includes a set of 555.9 hours for training and a set of 23.6 hours for testing. The detailed description of all training and test sets is shown in Table 1. In the code-switching dataset, the TALCS contains not only CS utterances, but also an amount of monolingual utterances as summarized in Table 2.

5.2. Experimental setup

Features: The input acoustic features used for model training and evaluation consist of 80-dimensional log Mel-filter bank (FBANK) plus one-dimensional pitch features. These features are computed using 25 ms windows with a 10 ms hop size over the raw speech waveform. To normalize the acoustic features, we apply utterance-level cepstral mean and variance normalization (CMVN) on the FBANK features for both training and testing. CMVN helps mitigate the effects of speaker and environment variability, making models more robust to unseen test conditions. During the proposed TripleT-E2E framework model training, the 81-dimensional normalized acoustic features are first passed through a convolutional sub-sampling module that contains two 2D convolutional layers with stride 2. The convolutional layers compress the sequence length in half while retaining important information, making subsequent operations more computationally efficient.

Model configurations: The monolingual Mandarin encoder (ZH-encoder) and English encoder (EN-encoder) in the TripleT-E2E framework consist of 6 Conformer encoder layers each. These layers are responsible for learning language-specific representations for Mandarin and English, respectively. Each Conformer layer includes a 1024-dimensional feed-forward, 256-dimensional self-attention with 4 heads, and convolutional modules.

The additional encoder in the spelling correction module has an identical structure to the ZH-encoder and EN-encoder, it also contains 6 Conformer layers with 1024-dimensional feed-forward and self-attention. It encodes the ASR hypotheses into representations for fusion with acoustic embeddings. The decoder in the spelling corrector has the same structure as the decoder of Conformer model, consisting of 6 transformer [56] decoder layers. The transformer layers have 1024-dimensional feed-forward components and utilize multi-head attention over the outputs of encoder and previous decoder time-steps. They generate the corrected transcription from the combined ASR hypotheses and acoustic representations. All our models are trained using the Adam optimizer [61] with a warmup learning rate schedule [62] over the first 25,000 iterations, where the rate is gradually increased before decaying. After 25,000 iterations, the learning rate follows a cosine annealing schedule.

Modeling units and evaluation metrics: For the Mandarin-English code-switching task, the proposed E2E ASR model utilizes a shared vocabulary consisting of 5,000 byte-pair-encoder (BPE) [63] units and 6,834 Mandarin Chinese characters. BPE units are generated using the SentencePiece library [64] by merging the most frequent character sequences from the training data. They allow open lexical modeling of both languages in a data-driven manner. To evaluate model performance, we use the token error rate (TER) as the ASR performance measure. The token here refers to the unit of Mandarin character and English word, respectively.

6. Results and discussions

6.1. E2E ASR baseline

Table 3 presents different code-switching ASR baseline results of standard Conformer system with spelling correction backend (Conformer+SC). S1 to S3 means training Conformer+SC systems with different training data. S4 is the fine-tuned S3 model using the limited TALCS training set. All systems are trained for the CS ASR task to evaluate the TACLS test set. S1 is trained from the combined 960 hours of English LibriSpeech data and 1000 hours of Mandarin WenetSpeech data. S2 is trained only using the Mandarin-English code-switching TALCS training set. S3 is trained on the pooled monolingual LibriSpeech, WenetSpeech datasets and the code-switched training set.

From Table 3, we see that the error rate numbers for both Mandarin and English parts, as well as the overall test set, are significantly smaller in S2 compared to S1. This suggests that the acoustic and linguistic characteristics of code-switching differ greatly from those of monolingual data, even though the monolingual training sets are much larger than the code-switching training set.

However, when comparing S3 with S2, the error rates are further reduced. This is because the addition of monolingual training data to the code-switching training set provides valuable complementary information, enriching the acoustic and linguistic properties of code-switching. In addition, when we further fine-tune S3 model using CS data, a slight performance improvements are still achieved. Moreover, it is notable that the TERs on the code-switched English part are significantly worse than those on the Mandarin part. This discrepancy suggests that the training data for code-switching, particularly the embedded English words, is far less abundant compared to the Mandarin characters. The CS acoustic events are very sparse even the whole CS training set of TALCS seems relatively large. All these observations from S1 to S4 baseline systems motivate us to propose the symmetric language-specific encoders to enhance the final model of proposed TripleT-E2E CS ASR framework. Given that S4 achieves the best results, it serves as the primary baseline for further comparison with our proposed architectures in the next sections.

Table 3
Different baseline results of standard Conformer system with spelling correction backend (Conformer+SC). For the CS test set performance evaluation, 'All', 'Man' and 'Eng' represent the TER% results on the complete test set, Mandarin characters and English words, respectively.

ID	Training data	Test Set		
		Man	Eng	All
S1	Librispeech[English]+WenetSpeech-M[Mandarin]	19.41	78.38	25.32
S2	TALCS[Code-switching]	7.73	24.55	8.61
S3	Librispeech[English]+WenetSpeech-M[Mandarin]+TALCS[Code-switching]	6.62	24.45	7.97
S4	TALCS[Code-switching]	**6.53**	**24.30**	**7.88**

* The S4 is fine-tuned from the S3 using TALCS data.

Table 4
The results (in TER%) of proposed TripleT-E2E models. 'CAL' represents how the CAL module receives the acoustic embedding from language-specific encoders, and 'ZH, EN and ZH+EN embedding' means the CAL receives the embeddings produced only by ZH-encoder, EN-encoder and the addition of both of them.

ID	Training data	System	CAL	Test Set		
				Man	Eng	All
S4	TALCS[Code-switchinig]	Conformer+SC	-	6.53	24.30	7.88
S5		TripleT-E2E-(3)	w/ ZH+EN embedding	6.59	24.40	7.93
S6	Librispeech[English]+WenetSpeech-M[Mandarin]	TripleT-E2E	w/ ZH embedding	6.17	18.14	6.94
S7	+TALCS[Code-switching]	TripleT-E2E	w/ EN embedding	**6.17**	**17.34**	**6.82**
S8		TripleT-E2E	w/ ZH+EN embedding	6.18	17.77	6.86
S9		TripleT-E2E	w/o acoustic embedding	6.22	18.62	7.08

* The S4 is fine-tuned from the S3 using TALCS data.

6.2. Results of the proposed TripleT-E2E CS ASR

Table 4 presents the performance of different variants of the proposed TripleT-E2E CS ASR framework for improving the Mandarin-English code-switching ASR performance. The Conformer ASR model with spelling correction backend from S4 in Table 3 is taken as our primary baseline for comparison. S5 utilizes the combined large data corpus (Librispeech+ WenetSpeech-M+TALCS) to train the proposed model structure (as shown in Fig. 2 (3)) from scratch with a random initialization, providing an additional comparison. S6 to S9 investigate how the CAL module designed in the spelling corrector affects the adapted and fine-tuned TripleT-E2E final model. Thus, in experiments S6 to S8, the CAL input of the language-specific representation is the output of the ZH encoder, the EN encoder, and the additive of both, respectively. In experiment S9, no language-specific representation is introduced for the CAL input. All the TripleT-E2E models are trained using the Librispeech, WenetSpeech-M and TALCS training data in the tri-stage training manner.

When comparing S3 and S5, it is apparent that in S5, the two symmetric language-specific encoders are not pre-trained, and the entire model is trained from scratch without employing any specific training strategy. The performance is nearly identical to that of S3, showing no significant improvement. Therefore, merely increasing the number of parameters is insufficient for enhancing the performance of code-switching speech recognition. Additionally, contrasting S4 with S5, the fine-tuned model in S4 outperforms the mixed-data training model in S5, so S4 is more suitable to be taken as a strong baseline performance. And, in comparison to the proposed tri-stage training framework (S6-S9 with TripleT-E2E), the system "TripleT-E2E-(3)" with random initialization achieves significantly worse performance, emphasizing the effectiveness of our proposed tri-stage training strategy.

Then, comparing the results of S6 to S8 with S9, it's clear to observe that by introducing the language-specific acoustic embedding does provide useful information to improve the spelling correction backend, especially for the foreign-embedded language English word recognition. Based on the system S9, system S6 to S8 are the ablation experiments to examine the effectiveness of using EN, ZH or both as CAL additional input. Comparing S9 with S6 to S8, we see the WERs on Mandarin words are only slightly reduced by introducing the ZH/EN/ZH+EN embedding as CAL block additional input. In contrast, the WERs on English words are significantly reduced when considering the EN embedding. When comparing S6 with S9, we indeed see a slight improvement in English recognition using ZH Embedding, with the WER reduced from 18.62% to 18.14%. However, when the EN embedding is introduced, the complementary effect of ZH embedding is greatly diminished, explaining why S8 did not outperform S7. And, we believe that the slightly worse performance observed when using both ZH and EN embeddings compared to using only EN embeddings is normal performance variations due to different model initializations across systems. In addition, we see that, the introduction of different types of language-specific embeddings do not have a great impact on Mandarin recognition, with a TER(%) of about 6.17. Comparing S6 to S8, the system S7 achieves the best, it yields a 28.6% relative reduction in TER for English output and 13.4% TER relative reduction for the whole CS test set over the strong baseline S4. It indicates that the English language characteristic has been successfully leveraged to enhance the TripleT-E2E model, and the impact of limited English words or code-switches of CS training data has been greatly alleviated. All these results show that our proposed framework is more effective than the traditional fine-tuning methods and has a stronger generalization to code-switched speech.

6.3. Ablation of tri-stage training method

Fig. 3 presents the results of ablation experiments conducted to evaluate the effectiveness of each stage in the proposed tri-stage training strategy. The pre-training, adaptation, and fine-tuning stages are the same as described in Section 4.2 and illustrated in Fig. 2. All results are derived from the three training stages of system S7, as shown in Table 4, which achieves the best results among the baseline systems.

In Fig. 3, the TER(%) performance for three parts of the code-switching test set is represented by three different color bars. The blue, white, and yellow bars correspond to the token error rate (TER) on the Mandarin part ('Man'), English part ('Eng'), and the entire CS test set ('All'), respectively. It can be observed from Fig. 3 that the TER values on the code-switching test set gradually decrease throughout the tri-stage model training, with the highest TERs in the pre-training stage (a) and the best ASR performance in the final model's fine-tuning stage (c). This indicates that each training stage proposed in Fig. 2 plays a

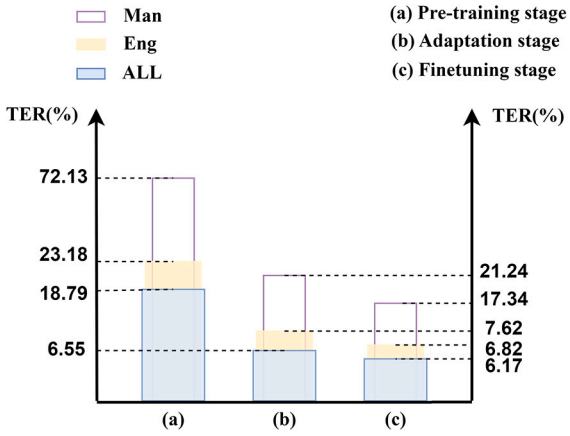

Fig. 3. Ablation study of the proposed tri-stage training strategy.

crucial role in improving the final model's performance and leveraging its potential.

After the tri-stage training, the final model S7 achieves TERs of 6.17%, 17.34%, and 6.82% for the Mandarin part, the English part, and the entire code-switching test set, respectively. Comparing it to the training only on combined monolingual data (stage (a)), adapting the model using code-switching data (stage (b)) provides significant gains, especially in English recognition. This demonstrates the advantages of model adaptation with the proposed BAL and CAL blocks for exploiting language-specific information and learning code-switching characteristics. Further improvements are attained by fine-tuning the model (stage (c)). Compared to conventional fine-tuning using pooled monolingual and code-switching training data in the S4 baseline, the ablation performance presented in Fig. 3 confirms the success of our proposed tri-stage training strategy within the TripleT-E2E CS ASR framework design. We can conclude that incorporating language-specific information and alleviating the acoustic and linguistic mismatch between monolingual and code-switching speech through selective adaptation are both key factors in the success of the final model.

6.4. Model complexity analysis

Just as the Google team proposed enhancing the basic Conformer-based E2E ASR with Spelling Correction (SC) modules in [65], the introduction of SC brings additional model parameters but results in significant performance improvements. Similarly, for our proposed TripleT-E2E ASR framework, we increased model complexity by introducing additional modules over the Conformer+SC baseline. However, as shown in Table 4, the performance improvements are substantial. In comparison to the baseline, our proposed TripleT-E2E ASR framework introduces ZH-encoder, EN-encoder, BAL, and CAL components, encompassing a total of 2 layernorms, 4 feedforward networks, and 12 conformer encoder layers. The addition of these layers naturally increases computational resources and model size during both training and inference. Hence, achieving an optimal tradeoff between the number of model parameters and performance enhancement remains a considerable challenge.

7. Conclusion

In this study, we proposed a new framework called TripleT (tri-stage training two-pass) E2E for improving Mandarin-English code-switching speech recognition. Our framework consists of two inference passes: the first pass uses the Conformer model, while the second pass employs a spelling correction backend. To learn acoustic representations from a large amount of monolingual data, we introduced two symmetric language-specific encoders in the TripleT-E2E CS ASR framework. These language-specific representations are then combined using a bilingual acoustic learner (BAL) to transfer the monolingual acoustic attributes to the code-switching style. Additionally, we utilized pre-trained representations to enhance the spelling corrector through a context plus acoustic learner (CAL) with the same structure as BAL.

We performed experiments on the publicly available TALCS Mandarin-English code-switching ASR corpus. The results demonstrate that our proposed tri-stage training approach, along with the use of language-specific encoders and the BAL and CAL modules, significantly improved the performance of the basic ASR and spelling corrector modules. Compared to the competitive Conformer-based two-pass E2E baseline model, our final TripleT-E2E model achieved a relative reduction of 13.4% in token error rate. Our proposed framework offers a promising direction for future research in this field, highlighting the importance of leveraging language-specific information and utilizing tri-stage training strategies to address the sparse training data problem of code-switching. Our future work will focus on generalizing the proposed framework to other code-switching ASR tasks.

CRediT authorship contribution statement

Xuefei Wang: Conceptualization, Investigation, Methodology, Software, Writing – original draft. **Yuan Jin:** Resources, Validation. **Fenglong Xie:** Formal analysis. **Yanhua Long:** Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The work is supported by the National Natural Science Foundation of China (Grant No. 62071302 and No. 61701306). The authors would like to thank Xiaohongshu Inc. for providing training resources.

References

[1] Chan W, Jaitly N, Le Q, Vinyals O. Listen, attend and spell: a neural network for large vocabulary conversational speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2016. p. 4960–4.
[2] Weninger F, Gaudesi M, Haidar MA, Ferri N. Conformer with dual-mode chunked attention for joint online and offline asr. In: Conference of the international speech communication association (interspeech); 2022. p. 2053–7.
[3] Chiu C-C, Sainath TN, Wu Y, Prabhavalkar R. State-of-the-art speech recognition with sequence-to-sequence models. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2018. p. 4774–8.
[4] Audhkhasi K, Kingsbury B, Ramabhadran B, Saon G. Building competitive direct acoustics-to-word models for English conversational speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2018. p. 4759–63.
[5] Das A, Li J, Zhao R, Gong Y. Advancing connectionist temporal classification with attention modeling. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2018. p. 4769–73.
[6] Rao K, Sak H, Prabhavalkar R. Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In: IEEE automatic speech recognition and understanding workshop (ASRU); 2017. p. 193–9.
[7] Toshniwal S, Sainath TN, Weiss RJ, Li B. Multilingual speech recognition with a single end-to-end model. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2018. p. 4904–8.
[8] Seki H, Watanabe S, Hori T, Le Roux J. An end-to-end language-tracking speech recognizer for mixed-language speech. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2018. p. 4919–23.
[9] Li DC. Cantonese-English code-switching research in Hong Kong: a y2k review. World Engl 2000;19:305–22.

[10] Lyu D-C, Tan T-P, Chng E-S, Li H. An analysis of a mandarin-English code-switching speech corpus: seame. Age 2010;21:25–8.
[11] Yılmaz E, van den Heuvel H, Van Leeuwen D. Investigating bilingual deep neural networks for automatic recognition of code-switching frisian speech. Proc Comput Sci 2016;81:159–66.
[12] Ardila A. Spanglish: an anglicized Spanish dialect. Hisp J Behav Sci 2005;27:60–81.
[13] Shan C, Weng C, Wang G, Su D. Investigating end-to-end speech recognition for Mandarin-English code-switching. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2019. p. 6056–60.
[14] Zhang S, Yi J, Tian Z, Bai Y. Decoupling pronunciation and language for end-to-end code-switching automatic speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2021. p. 6249–53.
[15] Zhang S, Yi J, Tian Z, Bai Y. End-to-end spelling correction conditioned on acoustic feature for code-switching speech recognition. In: Conference of the international speech communication association (interspeech); 2021. p. 266–70.
[16] Li J, et al. Recent advances in end-to-end automatic speech recognition. APSIPA Trans Signal Inf Process 2022;11(1).
[17] Miao H, Cheng G, Zhang P, Yan Y. Online hybrid ctc/attention end-to-end automatic speech recognition architecture. IEEE/ACM Trans Audio Speech Lang Process 2020;28:1452–65.
[18] Lyu D-C, Tan T-P, Chng ES, Li H. Seame: a mandarin-English code-switching speech corpus in south-East Asia. In: Conference of the international speech communication association (interspeech); 2010. p. 1986–9.
[19] Li C, Deng S, Wang Y, Wang G. Talcs: an open-source Mandarin-English code-switching corpus and a speech recognition baseline. In: Conference of the international speech communication association (interspeech); 2022. p. 1741–5.
[20] Zhang B, Lv H, Guo P, Shao Q. Wenetspeech: a 10000+ hours multi-domain mandarin corpus for speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2022. p. 6182–6.
[21] Chen G, Chai S, Wang G, Du J, Zhang W-Q, Weng C, et al. Gigaspeech: an evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. In: Conference of the international speech communication association (interspeech); 2021. p. 3670–4.
[22] Adel H, Vu NT, Kraus F, Schlippe T, Li H, Schultz T. Recurrent neural network language modeling for code switching conversational speech. In: IEEE international conference on acoustics, speech and signal processing; 2013. p. 8411–5.
[23] Modipa TI, Davel MH. Predicting vowel substitution in code-switched speech. In: Pattern recognition association of South Africa and robotics and mechatronics international conference (PRASA-RobMech); 2015. p. 154–9.
[24] Hou S-Y, Wu Y-L, Chen K-C, Chang T-A, Hsu Y-M, Chuang S-J, et al. Code-switching automatic speech recognition for nursing record documentation: system development and evaluation. JMIR Nursing 2022;5(1):e37562.
[25] Bullock B, Guzmán W, Serigos J, Sharath V, Toribio AJ. Predicting the presence of a matrix language in code-switching. In: The third workshop on computational approaches to linguistic code-switching; 2018. p. 68–75.
[26] Myslín M, Levy R. Code-switching and predictability of meaning in discourse. In: Language; 2015. p. 871–905.
[27] Vu NT, Lyu D-C, Weiner J, Telaar D, Schlippe T, Blaicher F, et al. A first speech recognition system for Mandarin-English code-switch conversational speech. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2012. p. 4889–92.
[28] Wang D, Tang Z, Tang D, Chen Q. Oc16-ce80: a Chinese-English mixlingual database and a speech recognition baseline. In: Conference of the oriental chapter of international committee for coordination and standardization of speech databases and assessment techniques (O-COCOSDA); 2016. p. 84–8.
[29] Hamed I, Denisov P, Li C-Y, Elmahdy M, Abdennadher S, Vu NT. Investigations on speech recognition systems for low-resource dialectal Arabic–English code-switching speech. Comput Speech Lang 2022;72:101278.
[30] Emond J, Ramabhadran B, Roark B, Moreno P, Ma M. Transliteration based approaches to improve code-switched speech recognition performance. In: IEEE spoken language technology workshop (SLT); 2018. p. 448–55.
[31] Shah S, Sitaram S. Using monolingual speech recognition for spoken term detection in code-switched Hindi-English speech. In: International conference on data mining workshops (ICDMW); 2019. p. 1–5.
[32] Long Y, Li Y, Zhang Q, Wei S, Ye H, Yang J. Acoustic data augmentation for Mandarin-English code-switching speech recognition. Appl Acoust 2020;161:107175.
[33] Sivasankaran S, Srivastava BML, Sitaram S, Bali K. Phone merging for code-switched speech recognition. In: Third workshop on computational approaches to linguistic code-switching; 2018.
[34] Manjunath K, Rao KS, Jayagopi DB, Ramasubramanian V. Indian languages asr: a multilingual phone recognition framework with ipa based common phone-set, predicted articulatory features and feature fusion. In: Conference of the international speech communication association (interspeech); 2018. p. 1016–20.
[35] Chen D, Mak BK-W. Multitask learning of deep neural networks for low-resource speech recognition. IEEE/ACM Trans Audio Speech Lang Process 2015;23(7):1172–83.
[36] Hai V, Xiao X, Chng ES, Li H. Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages. IEICE Trans Inf Syst 2014;97(2):285–95.
[37] Yeh CF, Huang CY, Sun LC, Lee LS. An integrated framework for transcribing Mandarin-English code-mixed lectures with improved acoustic and language modeling. In: 7th international symposium on Chinese spoken language processing; 2010. p. 214–9.
[38] Qian Y, Liu J. Mandarin-English bilingual phone modeling and combining mpe based discriminative training for cross-language speech recognition. In: 7th international symposium on Chinese spoken language processing; 2010. p. 103–8.
[39] Wu C-H, Shen H-P, Yang Y-T. Chinese-English phone set construction for code-switching asr using acoustic and dnn-extracted articulatory features. IEEE/ACM Trans Audio Speech Lang Process 2014;22(4):858–62.
[40] Luo N, Jiang D, Zhao S, Gong C, Zou W, Li X. Towards end-to-end code-switching speech recognition. ArXiv preprint. arXiv:1810.13091, 2018.
[41] Zhou X, Yılmaz E, Long Y, Li Y. Multi-encoder-decoder transformer for code-switching speech recognition. In: Conference of the international speech communication association (interspeech); 2020. p. 1042–6.
[42] Fan Z, Dong L, Shen C, Liang Z. Language-specific acoustic boundary learning for Mandarin-English code-switching speech recognition.. In: Conference of the international speech communication association (interspeech); 2023. p. 3322–6.
[43] Li Y, Fung P, Xu P, Liu Y. Asymmetric acoustic modeling of mixed language speech. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2011. p. 5004–7.
[44] Lin H, Deng L, Yu D, Gong Y-f. A study on multilingual acoustic modeling for large vocabulary asr. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2009. p. 4333–6.
[45] Zhang S, Liu Y, Lei M, Ma B. Towards language-universal Mandarin-English speech recognition. In: Conference of the international speech communication association (interspeech); 2019. p. 2170–4.
[46] Lu Y, Huang M, Li H, Guo J. Bi-encoder transformer network for Mandarin-English code-switching speech recognition using mixture of experts. In: Conference of the international speech communication association (interspeech); 2020. p. 4766–70.
[47] Dalmia S, Liu Y, Ronanki S, Kirchhoff K. Transformer-transducers for code-switched speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2021. p. 5859–63.
[48] Li K, Li J, Ye G, Zhao R. Towards code-switching asr for end-to-end ctc models. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2019. p. 6076–80.
[49] Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on machine learning; 2006. p. 369–76.
[50] Zhang S, Yi J, Tian Z, Tao J. Reducing language context confusion for end-to-end code-switching automatic speech recognition. In: Conference of the international speech communication association (interspeech); 2022. p. 3894–8.
[51] Liu H, Xu H, Garcia LP, Khong AW, He Y, Khudanpur S. Reducing language confusion for code-switching speech recognition with token-level language diarization. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2022. p. 1–5.
[52] Peng Y, Liu Y, Zhang J, Xu H, He Y, Huang H, et al. Internal language model estimation based language model fusion for cross-domain code-switching speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE; 2023. p. 1–5.
[53] Gulati A, Qin J, Chiu C-C, Parmar N. Conformer: convolution-augmented transformer for speech recognition. In: Conference of the international speech communication association (interspeech); 2020. p. 5036–40.
[54] Shi X, Feng Q, Xie L. The asru 2019 Mandarin-English code-switching speech recognition challenge: open datasets, tracks, methods and results. ArXiv preprint. arXiv:2007.05916.
[55] Diwan A, Vaideeswaran R, Shah S, Singh A, Raghavan S, Khare S, et al. Multilingual and code-switching asr challenges for low resource Indian languages. In: Conference of the international speech communication association (interspeech); 2021. p. 2446–50.
[56] Vaswani A, Shazeer N, Parmar N, Uszkoreit J. Attention is all you need. Adv Neural Inf Process Syst 2017;30.
[57] Lu Y, Li Z, He D, Sun Z. Understanding and improving transformer from a multi-particle dynamic system point of view. ArXiv preprint. arXiv:1906.02762.
[58] Zeineldeen M, Xu J, Lüscher C, Michel W. Conformer-based hybrid asr system for switchboard dataset. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2022. p. 7437–41.
[59] Deng J, Xie X, Wang T, Cui M. Confidence score based conformer speaker adaptation for speech recognition. ArXiv preprint. arXiv:2206.12045.
[60] Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an asr corpus based on public domain audio books. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2015. p. 5206–10.
[61] Kingma DP, Ba Adam J. A method for stochastic optimization. ArXiv preprint. arXiv:1412.6980.
[62] Gotmare A, Keskar NS, Xiong C, Socher R. A closer look at deep learning heuristics: learning rate restarts, warmup and distillation. ArXiv preprint. arXiv:1810.13243.
[63] Sennrich R, Haddow B, Birch A. Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics, vol. 1. 2016. p. 1715–25.

[64] Kudo T, Richardson J. Sentencepiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 conference on empirical methods in natural language processing; 2018. p. 66–71.

[65] Guo J, Sainath TN, Weiss RJ. A spelling correction model for end-to-end speech recognition. In: IEEE international conference on acoustics, speech and signal processing (ICASSP); 2019. p. 5651–5.