Efficient Dialect-Aware Modeling and Conditioning for Low-Resource Taiwanese Hakka Speech Processing

An-Ci Peng[1], Kuan-Tang Huang[1], Tien-Hong Lo[1], Hung-Shin Lee[3], Hsin-Min Wang[2], and Berlin Chen[1]

[1]Dept. Computer Science and Information Engineering, National Taiwan Normal University, Taiwan
[2]Institute of Computer Science, Academia Sinica, Taiwan
[3]United Link Co., Ltd., Taiwan

Abstract

Taiwanese Hakka is a low-resource, endangered language that poses significant challenges for automatic speech recognition (ASR), including high dialectal variability and the presence of two distinct writing systems (Hanzi and Pinyin). Traditional ASR models often encounter difficulties in this context, as they tend to conflate essential linguistic content with dialect-specific variations across both phonological and lexical dimensions. To address these challenges, we propose a unified framework grounded in the Recurrent Neural Network Transducers (RNN-T). Central to our approach is the introduction of dialect-aware modeling strategies designed to disentangle dialectal "style" from linguistic "content", which enhances the model's capacity to learn robust and generalized representations. Additionally, the framework employs parameter-efficient prediction networks to concurrently model ASR (Hanzi and Pinyin). We demonstrate that these tasks create a powerful synergy, wherein the cross-script objective serves as a mutual regularizer to improve the primary ASR tasks. Experiments conducted on the HAT corpus reveal that our model achieves 57.00% and 40.41% relative error rate reduction on Hanzi and Pinyin ASR, respectively. To our knowledge, this is the first systematic investigation into the impact of Hakka dialectal variations on ASR and the first single model capable of jointly addressing these tasks.

Keywords: Low-resource automatic speech recognition, Taiwanese Hakka, Recurrent neural network transducer

1. Introduction

Automatic Speech Recognition (ASR) has undergone significant advancements, transitioning from traditional hybrid models to advanced end-to-end (E2E) architectures. Key innovations in this space include Connectionist Temporal Classification (CTC) (Graves et al., 2006), Attention-based Encoder-Decoder (AED) (Vaswani et al., 2017), and Recurrent Neural Network Transducers (RNN-T) (Graves, 2012).

Among these E2E paradigms, the RNN-T model has garnered significant attention. It offers a robust and flexible framework that achieves an optimal balance between high recognition accuracy and the low-latency streaming capabilities essential for real-time applications (He et al., 2019; Sainath et al., 2020), thereby serving as the foundation of our study. Nonetheless, while the architecture of RNN-T has demonstrated strong performance in widely spoken languages such as English and Mandarin (Wang et al., 2024; Zhang et al., 2022), its application to low-resource and dialectally diverse languages continues to pose substantial challenges.

Taiwanese Hakka, once the second most prominent language in Taiwan (Liao et al., 2023), is now classified as endangered due to the predominance of Mandarin Chinese. The development of robust ASR tools for Taiwanese Hakka are essential for language preservation, cultural archiving, and enhancing media accessibility. However, the construction of such tools encounters two significant challenges. First, the language faces a substantial data scarcity issue: it is not only low-resource but also showcases considerable dialectal variability (Lai, 2001; Shim et al., 2024), complicating the creation of a unified, generalized model from the limited available data. Second, practical applications in educational and linguistic contexts necessitate support for two distinct writing systems: Taiwanese Hakka Hanzi (MOE, 2009, 2010) (Hanzi, 漢字), which utilizes traditional Chinese characters, and the Hakka Romanization System (MOE, 2012, 2024) (Pinyin, 拼音).

Although recent approaches have sought to tackle dialectal variability (Li et al., 2024), many strategies involve adapting specific model components. These approaches typically encompass the training of distinct prediction networks for each dialect, which are subsequently integrated during the post-training phase (Fukuda et al., 2022), or conditioning the decoder on an explicit dialect ID token (Shiota et al., 2022). These methods, however, are often evaluated on relatively high-resource languages and may not be directly applicable to low-resource scenarios where data is insufficient to train robust, specialized components.

Moreover, by concentrating primarily on the prediction network, these approaches largely neglect the fundamental challenge of representation conflation within the shared encoder.

When trained on limited and diverse datasets, a standard encoder is compelled to conflate core linguistic content with dialect-specific variations, affecting both phonological and lexical representations. This conflation poses significant challenges: it necessitates that the RNN-T's prediction network, which serves as its internal language model, learns multiple, redundant mappings for the same linguistic concepts in an inefficient manner.

To address this issue, and as a central contribution of our work, we introduce a dialect integration module consisting of two complementary components[1]. The first part of this module is a dialect-aware modeling strategy. This strategy aims to explicitly condition the model on dialect identity (Punjabi et al., 2021), guiding the encoder to capture dialect-specific acoustic and phonological features. The second part of the dialect integration module is dialect conditioning. This method enables the prediction network to function as a dialect-conditional language model, dynamically adapting its predictions—such as word choice and phonetic-to-character mappings—according to the active dialect, rather than attempting to model all dialects using a single, compromised set of parameters. This targeted conditioning facilitates effective generalization by the model, even in scenarios characterized by limited training data.

The second challenge of supporting two writing systems is addressed by a unified multi-task learning (MTL) framework (Wang et al., 2023, 2021; Hussein et al., 2025). Rather than training separate, specialized models for each task, our unified approach extends a single RNN-T architecture with parameter-efficient prediction networks (Ghodsi et al., 2020), enabling the joint generation of Hanzi and Pinyin without a prohibitive increase in model parameters. These two ASR tasks are complementary: the phonetically-grounded Pinyin task compels the model to maintain acoustically precise representations, while the semantically rich Hanzi task simultaneously requires it to capture high-level linguistic context to resolve ambiguities. This synergy creates a complementary objective that regularizes the shared encoder, forcing it to learn a representation that is both acoustically precise and linguistically context-aware.

Our experiments conducted on the HAT corpus (Liao et al., 2023) validate that our single, unified model achieves significant improvements over all single-task baselines, resulting in 57.00% and 40.41% relative error rate reductions on the Hanzi and Pinyin ASR tasks, respectively. The main contributions of this work are summarized as follows:

1. **Systematic Analysis of Dialectal Variation:** To our knowledge, this is the first systematic study of Hakka dialectal variation's impact on ASR. We propose a novel modeling strategy that explicitly conditions the RNN-T model on dialect identity. This allows the prediction network to function as a dialect-conditional language model, dynamically adapting its output. The result is a more robust and generalized model that provides a path forward for other low-resource, dialect-rich languages.
2. **A Unified Framework for Multi-Orthography Speech Processing:** We present a single, unified framework to jointly perform ASR (in Hanzi and Pinyin) using Taiwanese Hakka as a case study. We demonstrate that this architecture creates a powerful synergy, where the cross-lingual and cross-script objectives mutually regularize the model. This provides a parameter-efficient blueprint for handling similar challenges in other multi-orthography languages.

2. Background

This section provides the foundational context for our research. We begin by exploring the core architecture of the Recurrent Neural Network Transducer (RNN-T), which serves as the basis for our proposed model. Following this, we delve into the linguistic intricacies of Taiwanese Hakka, highlighting its two primary writing systems, Hanzi and Pinyin, as well as the main dialects relevant to this study.

2.1. Neural Transducer (RNN-T)

The neural transducer network, commonly known as the Recurrent Neural Network Transducer (RNN-T) (Graves, 2012), is an end-to-end architecture specifically tailored for sequence-to-sequence tasks. The architecture typically comprises three interrelated components: an Encoder Network, a Prediction Network, and a Joint Network.

The process initiates with the encoder network, $\text{Enc}(\cdot)$. At the current time step t in the input acoustic feature sequence $\mathbf{x}_{1:T}$ (where T denotes the length of the input audio features), the encoder transforms the frame $\mathbf{x}_t$ into a high-level acoustic representation $\mathbf{h}_t^{\text{Enc}}$:

$$\mathbf{h}_t^{\text{Enc}} = \text{Enc}(\mathbf{x}_t). \tag{1}$$

Simultaneously, the prediction network $\text{Pred}(\cdot)$ operates in an autoregressive manner, leveraging the previously predicted token sequence $\hat{\mathbf{y}}_{1:u}$

[1] Code:https://github.com/anci123/hakka_exp

to generate a hidden representation $\mathbf{h}_{u+1}^{\text{Pred}}$ at the $(u+1)$-th output step. This network serves as an internal language model:

$$\mathbf{h}_{u+1}^{\text{Pred}} = \text{Pred}(\hat{\mathbf{y}}_{1:u}). \quad (2)$$

Finally, the joint network, $\text{Joint}(\cdot)$, integrates the outputs $\mathbf{h}_t^{\text{Enc}}$ and $\mathbf{h}_{u+1}^{\text{Pred}}$ from the encoder and the prediction network through a feed-forward layer. This network predicts the posterior probability distribution over the vocabulary:

$$P(\hat{y}_{u+1}|x_t, \hat{\mathbf{y}}_u) = \sigma(\text{Joint}(\mathbf{h}_t^{\text{Enc}}, \mathbf{h}_{u+1}^{\text{Pred}})) \quad (3)$$

While the training objective is defined by maximizing the probability of the correct output sequence **y** given the input sequence **x** across all possible alignments between the input steps t and output steps u, this is computed efficiently using dynamic programming. This joint optimization enables the RNN-T to concurrently learn both robust acoustic and linguistic contexts within a unified differentiable framework.

In recent years, substantial research has focused on enhancing the prediction network within RNN-T architectures (Albesano et al., 2022). A study demonstrates that a stateless variant (RNNT-SLP) exhibits performance comparable to the original recurrent version (Ghodsi et al., 2020). The RNNT-SLP architecture relies solely on the most recent output tokens(the last one or the last two), achieving comparable recognition accuracy while substantially reducing model parameters and architectural complexity. These findings suggest that the prediction network primarily serves the function of alignment rather than language modeling, thereby motivating the development of more compact and parameter-efficient RNN-T designs suitable for low-resource ASR.

2.2. Taiwanese Hakka

The complexity of the Taiwanese Hakka language is rooted in its dialectal variation and its dual-system orthography. Dialect differences manifest in tone realization, phonetics, and lexical usage. Taiwanese Hakka is a language group form of Hakka dialects, such as Sixian (四縣), NanSixian (南四縣), Hailu (海陸), Dapu (大埔), Raoping (饒平), and Zhaoan (詔安). Due to dataset limitations, this study focuses on three dialects, Sixian, Hailu, and NanSixian.

For the promotion, education and preservation of the language, Taiwanese Hakka utilizes two officially standardized writing systems proposed by the Ministry of Education (MOE) of Taiwan:

Taiwanese Hakka Hanzi (Hanzi), standardized by the MOE between 2009 and 2010 (MOE, 2009, 2010), employs traditional Chinese characters adapted to represent Hakka-specific lexicon and phonetic nuances. The Hanzi form emphasizes the morpho-semantic aspect of the language, encoding Hakka phonological distinctions through character choice and usage conventions.

Dialect	Hanzi	Pinyin
Sixian	佢發譴哩	gi11 fad2 kien31 le24
Nansixian	佢發譴吔	i11 fad2 kien31 e24
Hailu	佢發闊咧	gi55 bod5 ad5 le53

Table 1: Example sentences expressing the same meaning, “He is angry”, across three Taiwanese Hakka dialects. The corresponding Mandarin translation is “他生氣了”. Notably, the Hanzi forms differ in characters such as “譴”, “闊”, and “吔”, reflecting dialect-specific lexical variation. The Pinyin representations also exhibit distinct syllabic realizations, for instance “kien31” versus “ad5”. Moreover, even when the same Hanzi character is shared across dialects, the corresponding pronunciations often diverge, highlighting the phonetic variation inherent to Taiwanese Hakka.

Taiwanese Hakka Romanization System (Pinyin), introduced by the MOE in 2012, uses the Latin alphabet for syllable representation, and the tonemes (tones) are represented using Arabic numerals based on the relative pitch variation of the tones (MOE, 2012, 2024). Pinyin serves as the phonetic transcription standard, closely following the acoustic realization of the spoken language. These two writing forms form the dual targets for our speech recognition tasks, while the inherent differences among the three core dialects (Sixian, NanSixian, and Hailu) serve as the foundation for our dialect-aware modeling experiments. Table 1 shows a writing example of the Hanzi, Pinyin across dialects like Sixian, Nansixian, and Hailu, with the same meaning.

3. Methodology

The overall architecture of our proposed method is illustrated in Figure 1. The system is built upon a multi-task RNN-T framework, comprising three primary components: 1) a shared acoustic encoder; 2) two distinct decoders, each equipped with its dedicated prediction and joint network, tailored for the Hanzi (H) and Pinyin (P) tasks; and 3) a novel dialect-aware module. This dialect-aware module generates a dialect embedding ($\mathbf{e}^d$) that is fused with the shared encoder’s hidden representation ($\mathbf{h}^{\text{Enc}}$). This integration results in a dialect-conditioned representation that is subsequently input into both ASR decoders, allowing them to ad-

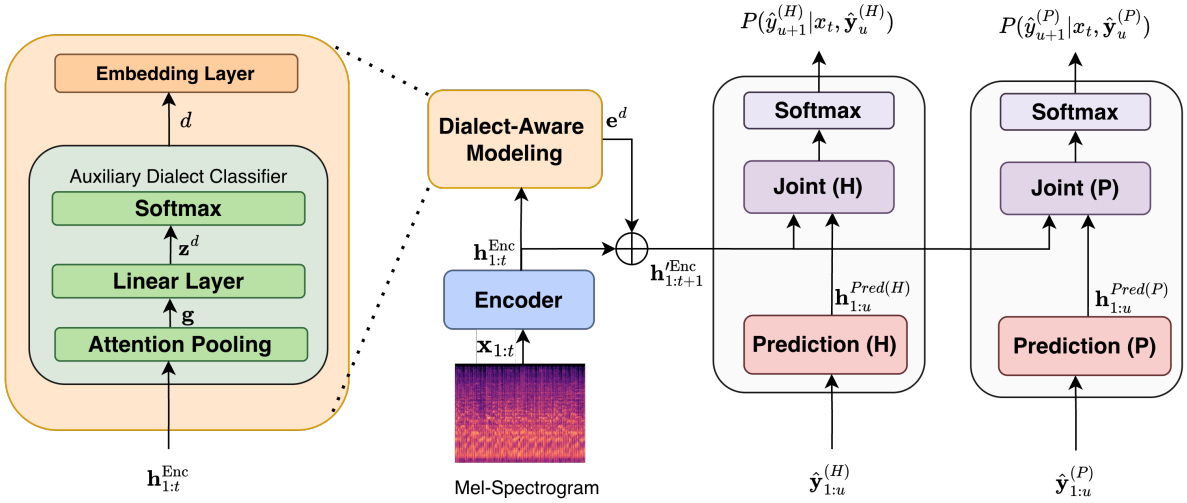


Figure 1: Overall architecture of the proposed method, including an RNN-T architecture with multiple joint networks and prediction networks for Hanzi(H) and Pinyin(P) ASR task, and a dialect-aware module.

just their predictions effectively.

In the following subsections, we will first describe the architecture of the shared-encoder multi-task RNN-T backbone (Section 3.1). We will then detail the components of our dialect-aware module, which includes the Auxiliary Dialect Classifier (ADC) and Dialect Information Integration (DII) (Section 3.2). Finally, we will present several dialect conditioning strategies that adjust the target sequences (Section 3.3).

3.1. Multi-task Learning

We adopt a shared-encoder multi-task RNN-T architecture to simultaneously handle the two required output tasks: Hakka Hanzi recognition and Hakka Pinyin recognition. As illustrated in Figure 1, the architecture consists of a single shared encoder and separate, task-specific prediction networks and joint networks for each output.

The shared encoder transforms the input audio feature sequence $\mathbf{x}_{1:T}$ of length T into a hidden representation $\mathbf{h}^{\text{Enc}}_{1:T} \in \mathbb{R}^{T\times D}$, where T denotes the length of the hidden representation and D indicates the dimension of the encoder. This representation is subsequently distributed to two distinct branches, associated with the Hanzi and Pinyin linguistic tasks. The overall ASR loss ($\mathcal{L}_{\text{ASR}}$) is computed by aggregating the individual pruned RNN-T losses (Kuang et al., 2022) from both the Hanzi and Pinyin tasks, thereby enhancing the time and memory efficiency during training.

3.2. Dialect-Aware Modeling

Here we introduce the first component of our dialect integration module: dialect-aware modeling. In the RNN-T architecture, the encoder produces the acoustic representation, significantly impacting the accuracy of the final token sequence prediction. To enhance the robustness of this representation by incorporating explicit dialect knowledge, we investigate two strategies for infusing dialect information into the features generated by the encoder: the first is the adoption of an Auxiliary Dialect Classifier (ADC), and the second is Dialect Information Integration (DII).

3.2.1. ADC: Auxiliary Dialect Classifier

In the ADC strategy, we integrate an auxiliary classification head that is connected to the output of the encoder $\mathbf{h}^{\text{Enc}} \in \mathbb{R}^{T\times D}$. First, this head aggregates the variable-length sequence $\mathbf{h}^{\text{Enc}}$ into a single, fixed-dimensional vector $\mathbf{g} \in \mathbb{R}^{D}$ through the application of attention pooling. Subsequently, this vector $\mathbf{g}$ is forwarded through a linear layer to generate the dialect logits $\mathbf{z}^d$.

During training, the auxiliary loss ($\mathcal{L}_{\text{A}}$) is computed using the Cross-Entropy (CE) loss with the softmax output of the logits, $\sigma(\mathbf{z}^d)$, defined as follows:

$$\mathcal{L}_A = \mathcal{L}_{\text{CE}}(\sigma(\mathbf{z}^d), d), \tag{4}$$

where σ denotes the softmax function, and d corresponds to the ground-truth dialect label.

The final objective $\mathcal{L}_{\text{Final}}$ is formulated as a weighted sum of the primary RNN-T loss ($\mathcal{L}_{\text{ASR}}$) and the auxiliary loss, appropriately balanced by a hyperparameter λ:

$$\mathcal{L}_{\text{Final}} = \mathcal{L}_{\text{ASR}} + \lambda \cdot \mathcal{L}_A. \tag{5}$$

3.2.2. DII: Dialect Information Integration

The DII strategy explicitly conditions the entire sequence transduction process on the known dialect identity. This methodology considers the dialect label as a crucial linguistic feature, necessitating the joiner network to generate output tokens informed by both acoustic features and the specific dialect context.

We first map the discrete dialect ID (e.g., NanSixian, Sixian, Hailu) of an audio segment to a continuous dialect embedding $\mathbf{e}^d \in \mathbb{R}^D$ through an embedding layer. During training, the ground-truth dialect label is used to generate this embedding (i.e., employing teacher-forcing). In contrast, during inference, we leverage the dialect ID predicted by the ADC module to retrieve the corresponding embedding.

Specifically, we employ a prefix conditioning approach whereby the dialect embedding $\mathbf{e}^d$ is concatenated to the acoustic hidden representation $\mathbf{h}^{\text{Enc}}$, thereby yielding the augmented representation $\mathbf{h}'^{\text{Enc}}$:

$$\mathbf{h}'^{\text{Enc}} = [\mathbf{e}^d; \mathbf{h}^{\text{Enc}}], \quad \mathbf{h}'^{\text{Enc}} \in \mathbb{R}^{(T+1)\times(D)}. \tag{6}$$

This augmented sequence $\mathbf{h}'^{\text{Enc}}$ is then fed into the joint networks. By integrating dialect information at the input stage of joint prediction, the model is encouraged to refine its token prediction probabilities according to the respective dialect, thereby enhancing recognition accuracy for that specific dialectal variant.

3.3. Dialect Conditioning

To enhance the model's performance in a multi-dialect scenario, we explore a set of strategies of integrating dialect information by adding the dialect index (denoted as d) directly to the prediction target sequence $\mathbf{y} = (y_1, \ldots, y_U)$ of each decoder, where U is the length of the target sequence. To ensure the dialect index exists within the same vocabulary space for all target sequences (Hanzi and Pinyin), we expand the token table of each decoder by a size of K, where K is the total number of dialect classes. The last K tokens in each expanded token table are reserved to represent the corresponding d for that output target. For instance, if the size of the original Hanzi token table is $V_{\text{Hanzi}} = 500$, and we have $K = 3$ dialects (NanSixian, Sixian, and Hailu), the table size is extended to $V_{\text{Hanzi}} + 3 = 503$. In this extended vocabulary, tokens 500, 501, and 502 explicitly represent the NanSixian, Sixian, and Hailu dialects, respectively. This mechanism is replicated across each ASR task, allowing the decoders to predict not only the linguistic sequence but also the dialectal context simultaneously. These strategies provide an explicit dialect signal that guides the model's predictions based on the source speech's dialect. Concrete examples of these three strategies are illustrated in Table 2.

Strategy	Target Sequence
PSC	佢發閪咧 d
PRSC	d 佢發閪咧
TIC	佢 d 發 d 閪 d 咧 d

Table 2: Illustration of example target sequence in each dialect conditioning strategy. d means the dialect ID.

Post-Sequence Conditioning (PSC). This strategy appends the dialect ID (d) to the end of the original target sequence. The core idea behind this placement is to provide an explicit signal to the Joint Network, ensuring the dialectal factor is integrated into the loss calculation for the final token prediction. This configuration compels the decoder to first generate the complete linguistic content. The conditioned target sequence processed by this strategy is $\mathbf{y}_{\text{PSC}} = (y_1, \ldots, y_U, d)$.

Pre-Sequence Conditioning (PRSC). In this approach, the dialect ID is prefixed to the beginning of the target sequence. This formulation equips the decoder with essential dialectal context prior to initiating the generation of linguistic tokens, thereby serving as an initial prompt or conditioning cue. The conditioned target sequence is $\mathbf{y}_{\text{PRSC}} = (d, y_1, \ldots, y_U)$.

Token-Interleaved Conditioning (TIC). The proposed strategy interleaves the dialect ID within the output sequence. It incorporates the dialect ID by following the processing of each linguistic token. This dense, token-level integration guarantees that the decoder's prediction for each subsequent token is continuously reinforced and conditioned by the corresponding target dialect. The conditioned target sequence is $\mathbf{y}_{\text{TIC}} = (y_1, d, y_2, d, \ldots, y_U, d)$.

4. Experimental Setup

4.1. Datasets

We evaluated the strategies mentioned in this paper on the Hakka Across Taiwan (HAT) corpus (Liao et al., 2023). The HAT corpus is particularly valuable as it contains speech data spanning three major Taiwanese Hakka dialects: Hailu, Sixian, and NanSixian. To simulate a low-resource scenario and evaluate model robustness across these distinct linguistic varieties, we constructed a dialect-balanced subset for our experiments.

From the corpus, we selected 60 speakers, 20 from each dialect, and partitioned this subset into speaker-disjoint training, development, and test

Split	Duration (hours)	# Utterances
Train	73.91	36,867
Development	3.95	2,330
Test	4.46	2,439
Total	82.32	41,636

Table 3: Statistics of the dialect-balanced subset constructed from the HAT corpus, including total duration and number of utterances in each split.

sets. The training set contains data from 18 speakers per dialect, resulting in a total of 54 speakers. The development and test sets are balanced across dialects, each containing 3 speakers, one from each dialect. This partitioning ensures that all dialects are adequately represented in the training, development, and test data. Detailed statistics of this dataset, including duration and utterance counts for each split, are presented in Table 3.

4.2. Implementation Details

The encoder employs a Zipformer (Yao et al., 2024) architecture. A Zipformer block, which is approximately twice as deep as a Conformer block (Gulati et al., 2020), incorporates self-attention, convolution, and feed-forward modules. Our encoder consists of 6 Zipformer stacking blocks. The number of attention heads for each block is set to $\{4, 4, 4, 8, 4, 4\}$, while the convolution kernel sizes are set to $\{31, 31, 15, 15, 15, 31\}$. The feed-forward dimensions per block are specified as $\{512, 768, 1024, 1536, 1024, 768\}$. For the prediction network, we adopt the RNN-T-SLP architecture (Ghodsi et al., 2020).

Models are optimized using the ScaledAdam optimizer (Yao et al., 2024) with a learning rate set to 0.01. The parameter λ in Eq. 5 is set to 0.5, which is tuned on the development set. All models undergo training for a total of 30 epochs.

4.3. Evaluation Metrics

We evaluate model performance using metrics specifically tailored for each task. For the Hakka ASR tasks (Hanzi and Pinyin), we use character error rate (CER) and syllable error rate (SER). CER quantifies the text fidelity at the character level, defined as: CER $= (S + D + I)/N_{\text{char}}$, where S, D, and I represent the counts of substitutions, deletions, and insertions, respectively; N_{char} is the total number of characters in the reference. The calculation principle for SER mirrors that of CER; however, the unit of measure is based on syllables.

Dialect	**Hanzi** (CER %)	**Pinyin** (SER %)
Sixian	10.61	8.56
NanSixian	8.35	10.60
Hailu	17.27	11.25
Mix	**6.07**	**8.29**

Table 4: Comparison of Hanzi CER and Pinyin SER for single-task models trained on individual dialects. "Mix" denotes training with data combined from all three dialects, resulting in the best overall performance across both tasks.

5. Results and Discussion

Before detailing our main findings, we first validate our foundational choice of using a mixed-dialect training set. Mixing data from multiple dialects allows the model to learn dialect-invariant acoustic and linguistic patterns. As shown in Table 4, training on the "Mix" set (6.07% in CER) significantly outperforms training on any single dialect (e.g., Hailu at 17.27% in CER). This demonstrates the effectiveness of mixed-dialect training in enhancing performance. Therefore, all subsequent experiments in this paper are conducted using this mixed-data configuration.

With this foundational result established, we conduct an analysis across three main aspects of the proposed framework. First, we examine the synergistic effect of multi-task learning, evaluating how jointly modeling multiple linguistic representations (Hanzi and Pinyin) enhances recognition performance across tasks. Next, we analyze the effectiveness of integrating dialect information, demonstrating how dialect-aware and dialect-conditioning strategies contribute to improved modeling and generalization in multi-dialect settings. Finally, we examine the parameter and time efficiency of our models, illustrating that the proposed multi-task and dialect-aware architecture achieves notable accuracy gains with only modest increases in model size and minimal impact on inference speed.

5.1. Effect of Multi-Task Learning

Here we analyze the benefits of multi-task learning (MTL) as summarized in Table 5. In comparison to the single-task baseline models, the *MTL*$_{HP}$ architecture demonstrates significant enhancements across both tasks. Specifically, the *MTL*$_{HP}$ model obtains relative improvements of 8.23% in Hanzi CER and 3.61% in Pinyin SER. This indicates a distinct synergistic effect, suggesting that these two ASR tasks are complementary. The phonetically-grounded Pinyin task compels

the model to maintain acoustically precise representations, while the semantically rich Hanzi task simultaneously requires it to capture high-level linguistic context to resolve ambiguities. By training on both, the shared encoder is regularized to learn a more robust representation that benefits both tasks.

5.2. Effectiveness of Integrating Dialect

5.2.1. Dialect-aware Modeling Results

We evaluated the two encoder-side dialect-aware methods, DII and ADC. As shown in Table 5, DII proved to be more effective in reducing the CER to 5.21% and SER to 7.70%. Due to its architectural design as a conditioning module, DII itself cannot predict dialects. Unlike methods that manually integrate dialect information, ADC can predict the dialect of each utterance. While DII achieves a slightly lower error rate (5.21% vs. 5.29%) than ADC, it requires ground-truth IDs during inference. Additionally, the ADC module predicts dialects with near-perfect accuracy (100%) and regularizes hidden states to extract discriminative features, making it essential for real-world ASR when dialect labels are unavailable.

5.2.2. Dialect Conditioning Results

The PSC strategy provided only marginal gains over the baseline (Hanzi CER 5.41%, Pinyin SER 7.76%), indicating that appending a dialect token offers limited benefit. However, the model consistently failed to output the dialect ID during inference, likely due to the RNN-T decoder's lack of an explicit boundary cue to trigger the final token.

In contrast, the PRSC strategy, which places the dialect ID at the beginning of the sequence, severely degraded ASR performance (Hanzi CER 10.54%, Pinyin SER 12.88%) despite achieving high dialect accuracy (88.88%). This suggests that enforcing a fixed initial dialect cue strongly biases the prediction network's state, impeding its ability to generate the initial linguistic tokens.

The TIC strategy stands out as a strongly effective conditioning method, achieving a dramatic performance leap (CER=2.71%, SER=4.98%). This demonstrates a key finding: providing a continuous dialect context directly to the target prediction network is more critical for reducing recognition errors in a multi-dialect scenario. The dense, temporal repetition of the dialect ID throughout the sequence enhances the model stability and robustness for a multi-dialect scenario.

To investigate the model's reliance on TIC, we designed a stress test. During inference, instead of using the model's self-predicted dialect tokens, we force-fed the decoder with dialect IDs of controlled correctness. We varied this correctness from 100% (always the correct ground-truth ID) down to 0% (always an incorrect dialect ID) and measured the impact on ASR performance. As shown in Figure 2, the model trained with TIC achieved baseline-level performance on Pinyin when the provided dialect IDs were entirely incorrect (0%), while Hanzi still outperformed the baseline. Performance gradually improved as the correctness increased, suggesting that accurate dialect cues can effectively guide decoding toward dialect-consistent linguistic forms.

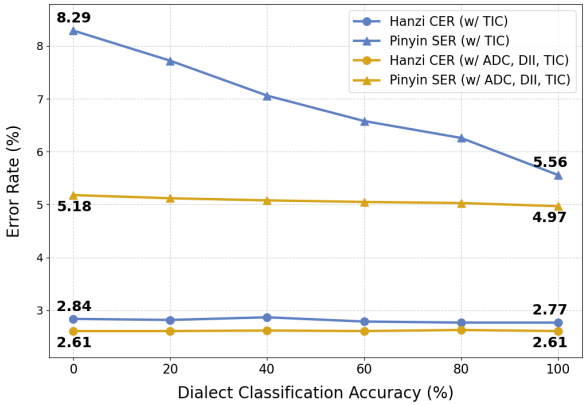


Figure 2: Effect of dialect Acc on ASR performance of our method on the test set. Performance is measured using CER (%) for Hanzi and SER (%) for Pinyin.

The model trained with ADC, DII, and TIC consistently outperformed the TIC-only system across all evaluated correctness levels. This finding indicates that the encoder-side conditioning (via ADC and DII) provides a more robust and comprehensive dialect-aware acoustic representation, rather than being redundant to the decoder-side TIC. By compelling the encoder to capture dialect-specific phonological patterns directly from the audio, these modules create a more stable foundation for the ASR task. This enhancement makes the overall model more resilient, allowing it to maintain superior recognition performance even when the dialect context provided to the decoder is imperfect or noisy.

5.3. Complexity and Efficiency Analysis

Table 6 compares the model complexity and runtime efficiency between the single-task baseline and the proposed multi-task model enhanced with ADC, DII, and TIC strategies. Despite incorporating multiple dialect-aware and multi-task modules, the proposed model increases the total parameter count only moderately, from 71.73 M to 78.71 M. Remarkably, the multi-task model achieves a

Method	Hanzi		Pinyin		Dialect	
Eval.	CER %	Rel. %	SER %	Rel. %	Acc. %	
VAN$_H$	6.07	–	–	–	–	
VAN$_P$	–	–	8.29	–	–	
MTL$_{HP}$	5.57	8.23	7.99	3.61	–	
Dialect-Aware Modeling						
w/ ADC	5.29	12.80	7.85	5.31	**100.00**	
w/ DII	5.21	14.16	7.70	7.16	–	
w/ ADC and DII	5.32	12.35	7.84	5.43	99.96	
Dialect Conditioning						
w/ PSC	5.41	10.87	7.76	6.39	–	
w/ PRSC	10.54	-73.64	12.88	-55.36	88.88	88.80
w/ TIC	2.71	55.35	4.98	39.90	88.76	88.76
w/ ADC , DII and TIC	**2.61**	57.00	**4.97**	40.41	**100.00**	

Table 5: Comparison of CER of Hanzi and SER of Pinyin of the model with each dialect integration strategy. *VAN* (Vanilla) denotes models trained on a single task, while *MTL* denotes joint training on multiple tasks. H and P refer to the Hanzi, Pinyin tasks, respectively. Rel.% indicates relative improvement over the corresponding *VAN* baseline. All strategies listed under "Dialect-Aware Modeling" and "Dialect Conditioning" are built upon the *MTL*$_{HP}$ model.

Model	Params (M)	RTFx
VAN$_H$	71.73	1101.63
MTL$_{HP}$ w/ ADC , DII and TIC	78.71	1309.86

Table 6: Comparison of model parameters (in millions) and RTFx between the single-task baseline and our proposed method. The RTFx represents the mean of 5 independent runs on a single RTX-3090 GPU.

higher RTFx (Inverse Real-Time Factor), indicating that improved acoustic representations can facilitate more efficient decoding. The decoding complexity remains unchanged, as no additional search operations are introduced, and the computational overhead of the auxiliary components is negligible compared to the encoder-decoder backbone. These results demonstrate that our design not only enhances modeling capacity and multi-dialect adaptability but also slightly improves real-time efficiency.

6. Conclusion

This work effectively tackles the inherent challenges of data scarcity and dialectal variability in the development of ASR systems for low-resource languages, with a specific focus on Taiwanese Hakka. To achieve this, we proposed and rigorously evaluated a multi-task RNN-T framework that jointly learns multiple linguistic representations and incorporates dialectal information. By joint training on diverse text representations and mixed-dialect data, the proposed system leverages complementary linguistic and acoustic cues. This approach reinforces the shared encoder representations of the model and enhances performance across all tasks and dialects, thereby enabling improved generalization in multi-dialect scenarios. Furthermore, we demonstrated that the selection of the dialect conditioning strategy is critical. Our proposed Token-Interleaved Conditioning (TIC) strategy, which provides a continuous dialect context, significantly outperformed alternative approaches. The final model, integrating ADC, DII, and TIC, achieved the best overall performance, confirming that explicitly modeling dialect at both the encoder and decoder levels yields the most robust results.

Notably, despite the incorporation of additional modules, the proposed multi-task learning (MTL) framework introduces only a modest increase in parameters while maintaining the same Real-Time Factor (RTF) as the single-task baseline. This observation indicates that the approach achieves significant accuracy improvements without compromising inference efficiency.

In summary, this study emphasizes that the joint modeling of diverse linguistic representations, coupled with the integration of dialectal information, offers a practical and efficient approach for developing robust ASR systems in low-resource, multi-dialect settings.

7. Limitation

Although this study demonstrates the effectiveness of integrating dialect information and multi-task learning for Taiwanese Hakka ASR, certain limitations remain.

First, low-resource languages like Taiwanese Hakka often lack large-scale corpora. Future work could address these challenges by leveraging external information sources, such as fine-tuning on pretrained multilingual speech or language models (e.g., wav2vec (Baevski et al., 2020), Whisper (Radford et al., 2022), or Parakeet-TDT (Sekoyan et al., 2025)), or incorporating large-scale self-supervised representations to enhance both the acoustic and linguistic modeling capabilities under low-resource conditions.

Another limitation of this study is that the Taiwanese Hakka language encompasses more dialects beyond the three examined here—Sixian, Hailu, and NanSixian—such as Dapu, Raoping, and Zhao'an. Due to the lack of publicly available speech resources for these dialects, our experiments could not be extended to cover them. We anticipate that future releases of more comprehensive Taiwanese Hakka datasets will enable broader investigations across additional dialects, allowing for a more complete evaluation of dialectal diversity in Hakka ASR.

8. Bibliographical References

Dario Albesano, Jesús Andrés-Ferrer, Nicola Ferri, and Puming Zhan. 2022. On the prediction network architecture in RNN-T for ASR. In *Proc. Interspeech*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Arxiv preprint arXiv.2006.11477*.

Takashi Fukuda, Samuel Thomas, Masayuki Suzuki, Gakuto Kurata, George Saon, and Brian Kingsbury. 2022. Global RNN transducer models for multi-dialect speech recognition. In *Proc. Interspeech*.

Mohammadreza Ghodsi, Xiaofeng Liu, James Apfel, Rodrigo Cabrera, and Eugene Weinstein. 2020. RNN-Transducer with stateless prediction network. In *Proc. ICASSP*.

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *Arxiv preprint arXiv:1211.3711*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: convolution-augmented transformer for speech recognition. In *Proc. Interspeech*.

Yanzhang He, Tara N. Sainath, Rohit Prabhavalkar, Ian McGraw, Raziel Alvarez, Ding Zhao, David Rybach, Anjuli Kannan, Yonghui Wu, Ruoming Pang, Qiao Liang, Deepti Bhatia, Yuan Shangguan, Bo Li, Golan Pundak, Khe Chai Sim, Tom Bagby, Shuo-yiin Chang, Kanishka Rao, and Alexander Gruenstein. 2019. Streaming end-to-end speech recognition for mobile devices. In *Proc. ICASSP*.

Amir Hussein, Cihan Xiao, Matthew Wiesner, Dan Povey, Leibny Paola Garcia, and Sanjeev Khudanpur. 2025. HENT-SRT: Hierarchical efficient neural transducer with self-distillation for joint speech recognition and translation. In *Arxiv preprint arXiv:2506.02157*.

Fangjun Kuang, Liyong Guo, Wei Kang, Long Lin, Mingshuang Luo, Zengwei Yao, and Daniel Povey. 2022. Pruned RNN-T for fast, memory-efficient ASR training. In *Proc. Interspeech*.

Huei-ling Lai. 2001. On Hakka BUN: A case of polygrammaticalization. *Language and Linguastic*, pages 137–153.

Qiang Li, Qianyu Mai, Mandou Wang, and Mingjuan Ma. 2024. Chinese dialect speech recognition: a comprehensive survey. *Artificial Intelligence Review*, 57.

Yuan-Fu Liao, Shaw-Hwa Hwang, You-Shuo Chen, Han-Chun Lai, Yao-Hsing Chung, Li-Te Shen, Yen-Chun Huang, Chi-Jung Huang, Hsu Wen Han, Li-Wei Chen, Pei-Chung Su, and Chao-Shih Huang. 2023. Taiwanese Hakka across Taiwan corpus and Formosa speech recognition challenge 2023 - Hakka ASR. In *Proc. O-COCOSDA*.

MOE. 2009. Recommended Taiwanese Hakka Hanji - set 1.

MOE. 2010. Recommended Taiwanese Hakka Hanji - set 2.

MOE. 2012. Taiwanese Hakka Romanization System.

MOE. 2024. Taiwanese Hakka Pinyin Scheme.

Surabhi Punjabi, Harish Arsikere, Zeynab Raeesy, Chander Chandak, Nikhil Bhave, Ankish Bansal, Markus Müller, Sergio Murillo, Ariya Rastrow, Andreas Stolcke, Jasha Droppo, Sri Garimella, Roland Maas, Mat Hans, Athanasios Mouchtaris, and Siegfried Kunzmann. 2021. Joint ASR and Language Identification Using RNN-T: An Efficient Approach to Dynamic Language Switching. In *Proc. ICASSP*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. In *Arxiv preprint arXiv.2212.04356*.

Tara N. Sainath, Yanzhang He, Bo Li, Arun Narayanan, Ruoming Pang, Antoine Bruguier, Shuo-yiin Chang, Wei Li, Raziel Alvarez, Zhifeng Chen, Chung-Cheng Chiu, David Garcia, Alex Gruenstein, Ke Hu, Anjuli Kannan, Qiao Liang, Ian McGraw, Cal Peyser, Rohit Prabhavalkar, Golan Pundak, David Rybach, Yuan Shangguan, Yash Sheth, Trevor Strohman, Mirkó Visontai, Yonghui Wu, Yu Zhang, and Ding Zhao. 2020. A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. In *Proc. ICASSP*.

Monica Sekoyan, Nithin Rao Koluguri, Nune Tadevosyan, Piotr Zelasko, Travis Bartley, Nikolay Karpov, Jagadeesh Balam, and Boris Ginsburg. 2025. Canary-1B-v2 & Parakeet-TDT-0.6B-v3: Efficient and High-Performance Models for Multilingual ASR and AST. In *Arxiv preprint arXiv.2509.14128*.

Ryan Soh-Eun Shim, Kalvin Chang, and David R. Mortensen. 2024. Phonotactic Complexity across Dialects. In *Arxiv preprint arXiv.2402.12998*.

Sayaka Shiota, Ryo Imaizumi, Ryo Masumura, and Hitoshi Kiya. 2022. Dialect-aware semi-supervised learning for end-to-end multi-dialect speech recognition. In *Proc. APSIPA ASC*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proc. NeurIPS*.

Peidong Wang, Tara N. Sainath, and Ron J. Weiss. 2021. Multitask training with text data for end-to-end speech recognition. In *Proc. Interspeech*.

Peidong Wang, Eric Sun, Jian Xue, Yu Wu, Long Zhou, Yashesh Gaur, Shujie Liu, and Jinyu Li. 2023. LAMASSU: A streaming language-agnostic multilingual speech recognition and translation model using neural transducers. In *Proc. Interspeech*.

Xintong Wang, Mingqian Shi, and Ye Wang. 2024. Pitch-Aware RNN-T for Mandarin Chinese mispronunciation detection and diagnosis. In *Arxiv preprint arXiv:2406.04595*.

Zengwei Yao, Liyong Guo, Xiaoyu Yang, Wei Kang, Fangjun Kuang, Yifan Yang, Zengrui Jin, Long Lin, and Daniel Povey. 2024. Zipformer: A faster and better encoder for automatic speech recognition. In *Proc. ICLR*.

Chao Zhang, Bo Li, Zhiyun Lu, Tara N. Sainath, and Shuo-yiin Chang. 2022. Improving the fusion of acoustic and text representations in RNN-T. In *Arxiv preprint arXiv:2201.10240*.