CLiFT-ASR: A Cross-Lingual Fine-Tuning Framework for Low-Resource Taiwanese Hokkien Speech Recognition

Hung-Yang Sung[1], Chien-Chun Wang[1], Kuan-Tang Huang[1], Tien-Hong Lo[1], Yu-Sheng Tsao[2], Yung-Chang Hsu[2], and Berlin Chen[1]
[1]National Taiwan Normal University, Taiwan
[2]EZAI, Taiwan
{redsheep, 61247033s, 61347002s, teinhonglo, berlin}@ntnu.edu.tw,
{sam, mic}@ez-ai.com.tw

Abstract

Automatic speech recognition (ASR) for low-resource languages such as Taiwanese Hokkien is difficult due to the scarcity of annotated data. However, direct fine-tuning on Han-character transcriptions often fails to capture detailed phonetic and tonal cues, while training only on romanization lacks lexical and syntactic coverage. In addition, prior studies have rarely explored staged strategies that integrate both annotation types. To address this gap, we present CLiFT-ASR, a cross-lingual fine-tuning framework that builds on Mandarin HuBERT models and progressively adapts them to Taiwanese Hokkien. The framework employs a two-stage process in which it first learns acoustic and tonal representations from phonetic Tai-lo annotations and then captures vocabulary and syntax from Han-character transcriptions. This progressive adaptation enables effective alignment between speech sounds and orthographic structures. Experiments on the TAT-MOE corpus demonstrate that CLiFT-ASR achieves a 24.88% relative reduction in character error rate (CER) compared with strong baselines. The results indicate that CLiFT-ASR provides an effective and parameter-efficient solution for Taiwanese Hokkien ASR and that it has potential to benefit other low-resource language scenarios.

Keywords: Automatic speech recognition, low-resource language, Taiwanese Hokkien, cross-lingual transfer, two-stage fine-tuning

1 Introduction

Taiwanese Hokkien is an important dialect in Taiwan with rich cultural and historical significance. However, as Mandarin Chinese dominates education and daily life, the use of Taiwanese Hokkien has been declining, especially among younger generations. A 2020 survey[1] reports that only 7.4% of children regularly use Taiwanese Hokkien. Despite the existence of several speech corpora (Liao et al., 2022; Chou et al., 2023; Lin et al., 2024), the overall amount of annotated data is limited compared to high-resource languages such as Mandarin and English (Zhang et al., 2022; Wang et al., 2021). This data scarcity poses a significant challenge for developing robust Speech Translation (Chen et al., 2023) and automatic speech recognition (ASR) systems.

Existing Taiwanese Hokkien ASR systems face additional challenges due to inconsistent transcription standards. Some systems employ Tai-lo romanization (Chou et al., 2023; Chao et al., 2021), which combines phonetic scripts with tonal markings, making it less intuitive and harder for general users to accept (Khoo, 2019). Other approaches annotate speech with Mandarin characters, but the mapping between Taiwanese Hokkien vocabulary and Mandarin text is often one-to-many or partially aligned, leading to longer and less accurate output sequences (Lin et al., 2024). Using Taiwanese Hokkien Han characters provides a practical alternative that balances readability and phonological detail, improving recognition usability.

To overcome these challenges, we introduce **CLiFT-ASR**[2], a **C**ross-**Li**ngual **F**ine-**T**uning framework for low-resource **A**utomatic **S**peech **R**ecognition that leverages Mandarin HuBERT backbone models and progressively adapts them to Taiwanese Hokkien. The framework follows a two-stage fine-tuning

[1] https://www.stat.gov.tw/News_Content.aspx?Create=1&n=2755&state=1327FD6AD8DCDA52&s=230300&ccms_cs=1&sms=11065/

[2] https://github.com/redsheep913/CLiFT-ASR/

Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 176–183
November 20-22, 2025 ©2025 Association for Computational Linguistics

strategy where it first acquires acoustic-level knowledge from phonetic Tai-lo annotations and then learns language-level structures such as vocabulary and syntax using Taiwanese Hokkien Han characters. Comprehensive experiments on the TAT-MOE corpus demonstrate that CLiFT-ASR achieves a 24.88% relative reduction in character error rate (CER). The framework offers an effective solution for Taiwanese Hokkien ASR and provides guidance for developing ASR systems for other low-resource languages.

2 Background

2.1 Linguistic Characteristics of Taiwanese Hokkien

Taiwanese Hokkien has a seven-tone system and exhibits tone sandhi, which creates tonal variations that differ from Mandarin (Cheng, 1968). These tonal patterns make automatic speech recognition challenging, as accurate recognition requires modeling both static tones and context-dependent tone changes. Despite these differences, Taiwanese and Mandarin share similar morphological and syntactic structures (Sun, 2006), which allows knowledge transfer from Mandarin-pretrained models. Previous studies show that Mandarin-pretrained ASR models outperform English-pretrained models when recognizing romanized Taiwanese (Tai-lo), indicating that cross-lingual transfer can be effective for end-to-end ASR targeting Taiwanese Han characters (Chou et al., 2023). These observations motivate the design of CLiFT-ASR, which leverages cross-lingual knowledge and adapts it progressively to Taiwanese Hokkien. Note that this work does not focus on modeling tone sandhi phenomena, which is left for future research.

2.2 Orthographic Systems and Their Role in ASR

Taiwanese Hokkien uses two main orthographies: romanization and Han characters. Romanization systems such as Peh-ōe-jī (POJ) and Tai-lo provide systematic phonetic representations (Khoo, 2019). The Ministry of Education has published a recommended set of roughly 700 Han characters for writing Taiwanese, which can be combined with romanization in a mixed-script form known as hàn-lô [3]. For ASR, using Han characters or hàn-lô offers a practical balance between phonetic detail and readability and informs the two-stage fine-tuning strategy employed in CLiFT-ASR.

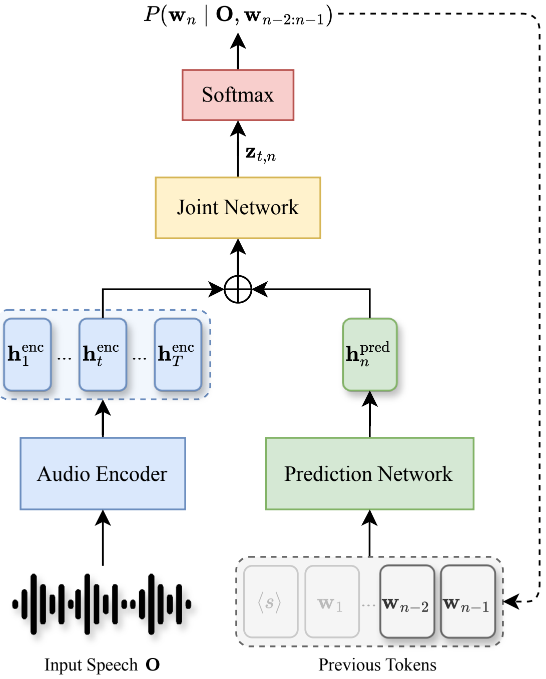

$P(\mathbf{w}_n \mid \mathbf{O}, \mathbf{w}_{n-2:n-1})$
Softmax
$\mathbf{z}_{t,n}$
Joint Network
$\mathbf{h}_1^{\text{enc}}$ ··· $\mathbf{h}_t^{\text{enc}}$ ··· $\mathbf{h}_T^{\text{enc}}$
$\mathbf{h}_n^{\text{pred}}$
Audio Encoder
Prediction Network
$\langle s \rangle$
w_1
$\mathbf{w}_{n-2}$
$\mathbf{w}_{n-1}$
Input Speech $\mathbf{O}$
Previous Tokens

Figure 1: Overview of the proposed CLiFT-ASR. The $\bigoplus$ operator denotes element-wise tensor addition. The dashed arrow indicates that during inference, ground-truth labels are not available, so the model outputs are fed back autoregressively into the prediction network.

3 Proposed Method

3.1 Model Architecture

The proposed CLiFT-ASR framework builds upon the RNN-Transducer (RNN-T) (Graves, 2012) to align variable-length acoustic sequences with token sequences, as illustrated in Figure 1. Given an input audio signal $\mathbf{O}$ and a sequence of target tokens $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N)$, where N denotes the number of output tokens, CLiFT-ASR estimates a probability distribution over possible tokens at each alignment step. The model consists of three components: an audio encoder, a prediction network, and a joint network. The audio encoder processes T acoustic feature frame vectors $(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T)$ extracted from $\mathbf{O}$ and maps them to a sequence of high-level representations that capture phonetic, tonal, and

[3] https://language.moe.gov.tw/.../D005.pdf

other essential speech information:

$$\mathbf{H}^{\text{enc}} = \text{AudioEncoder}(\mathbf{O}), \tag{1}$$

where $\mathbf{H}^{\text{enc}}$ represents the sequence of encoder hidden states $(\mathbf{h}_1^{\text{enc}}, \ldots, \mathbf{h}_T^{\text{enc}})$. This representation integrates both local and global acoustic patterns, which is crucial for accurately modeling the tonal variations in Taiwanese Hokkien. The prediction network generates a context representation autoregressively, conditioning on the previous two target tokens to form a trigram-style context that models short-term sequential dependencies in the output space:

$$\mathbf{h}_n^{\text{pred}} = \text{PredictionNetwork}(\mathbf{w}_{n-2}, \mathbf{w}_{n-1}). \tag{2}$$

The joint network combines the encoder output at time step t, $\mathbf{h}_t^{\text{enc}}$, which corresponds to the t-th element of $\mathbf{H}^{\text{enc}}$, with the prediction network state to form a joint representation. The conditional distribution over the next token is then obtained by applying a softmax:

$$\mathbf{z}_{t,n} = \text{JointNetwork}(\mathbf{h}_t^{\text{enc}} + \mathbf{h}_n^{\text{pred}}), \tag{3}$$

$$P(\mathbf{w}_n \mid \mathbf{O}, \mathbf{w}_{n-2:n-1}) = \text{Softmax}(\mathbf{z}_{t,n}). \tag{4}$$

This architecture allows CLiFT-ASR to jointly leverage acoustic and linguistic context at each step, which is essential for capturing tonal and phonological patterns in Taiwanese Hokkien.

3.2 Training Strategy

To handle limited annotated Taiwanese Hokkien data, CLiFT-ASR adopts a two-stage fine-tuning framework based on a pre-trained Mandarin HuBERT encoder. In the first stage, the model learns acoustic-level representations from phonetic Tai-lo annotations. Given a training set $\{(\mathbf{O}^{(i)}, \mathbf{W}_{\text{Tai-lo}}^{(i)})\}_{i=1}^{U_{\text{Tai-lo}}}$, the model parameters θ are updated to minimize the negative log-likelihood:

$$\theta' = \arg\min_{\theta} \sum_{i=1}^{U_{\text{Tai-lo}}} -\log P\big(\mathbf{W}_{\text{Tai-lo}}^{(i)} \mid \mathbf{O}^{(i)}; \theta\big). \tag{5}$$

This stage captures fine-grained acoustic and phonetic details, providing a solid foundation for language-level learning. In the second stage, the model is fine-tuned on Taiwanese Hokkien Han character annotations. Starting from the network parameter θ', it is trained on $\{(\mathbf{O}^{(j)}, \mathbf{W}_{\text{Han}}^{(j)})\}_{j=1}^{U_{\text{Han}}}$ to learn vocabulary, syntax, and higher-level linguistic structures:

$$\theta^* = \arg\min_{\theta} \sum_{j=1}^{U_{\text{Han}}} -\log P\big(\mathbf{W}_{\text{Han}}^{(j)} \mid \mathbf{O}^{(j)}; \theta\big). \tag{6}$$

By progressively adapting from phonetic to linguistic representations, CLiFT-ASR effectively leverages cross-lingual knowledge and maximizes the use of limited annotated data, resulting in more accurate and robust Taiwanese Hokkien ASR.

Split	Spk.	Utt.	Hr.
Training	328	86,072	153.33
Development	58	16,357	28.60
Test	54	15,962	26.28
Total	440	118,391	208.21

Table 1: Statistics of the TAT-MOE dataset across training, development, and test splits, including the number of speakers (Spk.), utterances (Utt.), and total duration in hours (Hr.).

4 Experimental Setup

4.1 Dataset

All experiments were conducted on the TAT-MOE subset of the TAT corpus (Liao et al., 2022), a large-scale Taiwanese Hokkien speech resource covering diverse regions of Taiwan. The corpus captures variation in speaker accents and pronunciation, providing a suitable testbed for robust ASR development. Audio recordings were sampled at 16 kHz with 16-bit PCM encoding to ensure consistent acoustic quality. Transcriptions were provided in Hàn-Lô-Tâi-bûn, a mixed orthography combining Han characters and romanized phonetics. Alternative annotations, including Peh-ōe-jī, Tai-lo, and tone-marked Tai-lo, were also available to support different modeling strategies. Table 1 summarizes the number of speakers, utterances, and total duration for the training, development, and test sets. To further evaluate model performance, we included a cleaner test set drawn from the pilot test of the Formosa Speech Recognition Challenge 2020 (FSR-2020) (Liao et al., 2020), referred to as the clean test. The TAT-MOE corpus therefore provides high-quality acoustic data and multiple orthographic representations, making

Model	Parameters (M)	Development		Test		Clean Test	
		CER	Rel.	CER	Rel.	CER	Rel.
Zipformer	65	48.57	-	45.82	-	15.69	-
FSR-2020 Best	-	-	-	-	-	15.62	0.07
Whisper-base	74	27.36	21.21	24.02	21.80	10.05	5.64
HuBERT-base	96	26.16	22.41	24.49	21.33	12.97	2.72
HuBERT-base-cmn	96	24.06	24.51	22.41	23.41	9.08	6.61
CLiFT-ASR	**96**	**22.37**	**26.20**	**20.94**	**24.88**	**8.06**	**7.63**
Whisper-small	244	22.47	26.10	18.68	27.14	7.66	8.03

Table 2: CERs (%) and relative reductions (Rel., %) for Taiwanese Hokkien ASR using various audio encoder initialization strategies.CLiFT-ASR applies a two-stage fine-tuning strategy with the HuBERT-base-cmn encoder. FSR-2020 Best refers to the top-performing model from FSR-2020.

it a valuable benchmark for low-resource Taiwanese Hokkien ASR.

4.2 Data preprocessing

The transcripts in the TAT-MOE dataset were written in Hàn-Lô-Tâi-bûn, a mixed system of Han characters and romanized phonetics. To unify the representation, we first constructed a mapping table using additional corpora to convert romanized segments into the corresponding Han characters. Arabic numerals were also converted into Chinese numerals, and variant or synonymous characters were normalized to a single standardized form. These preprocessing steps reduce inconsistencies and lexical variation in the annotations, thereby improving the stability of training and the accuracy of recognition.

4.3 Model Configuration

All training procedures followed Icefall's official recipes and default settings[4]. To establish a fair baseline and assess the benefit of cross-lingual transfer, we considered two encoder configurations: the baseline Zipformer model and a HuBERT-based Transformer initialized with Mandarin pretrained weights provided by the toolkit. The prediction network adopted Icefall's stateless design for efficient sequence modeling, and the joint network followed the standard implementation for integrating audio encoder and prediction network features into output distributions (Yao et al., 2024; Hsu et al., 2021; Ghodsi et al., 2020). For tokenization, we employed Icefall's byte-level BPE model, which has proven effective for handling large CJK vocabularies in bilingual and multilingual ASR tasks. This configuration enables a direct comparison between a strong baseline and our cross-lingual strategy, ensuring that performance gains are consistent and interpretable.

4.4 Training Details

Speech data were prepared using the Lhotse toolkit (Żelasko et al., 2021). For feature extraction, the Zipformer baseline model used 80-dimensional filter bank (FBank) features, while the HuBERT-based model was fine-tuned directly from raw waveform inputs. In CLiFT-ASR, the first stage was trained for 20 epochs and the second stage for 40 epochs. For comparison, a direct fine-tuning approach without staging was trained for 60 epochs. All models were trained with gradient accumulation over 4 steps to stabilize optimization. To balance computational efficiency and contextual coverage, the maximum audio duration per training sample was limited to 120 seconds. The learning rate was initialized at 0.0005 and scheduled over 40 epochs for smooth convergence. Model embeddings were set to 256 dimensions, and training was initialized from pretrained checkpoints. Optimization was performed with the ScaledAdam optimizer, which applied adaptive learning rates and gradient clipping at 2.0 for stability. A custom learning rate scheduler, Eden, was employed to dynamically adjust the learning rate across both batch and epoch progression (Yao et al., 2024).

[4] https://github.com/k2-fsa/icefall/

Fine-tuning Strategy	**Frozen**	**Development**	**Test**	**Clean Test**
Direct	None	24.06	22.41	9.08
Two-stage	Audio Encoder	36.82	35.72	26.84
	Prediction Network	25.23	23.91	11.84
	Joint Network	29.58	28.59	18.46
	None	**22.37**	**20.94**	**8.60**

Table 3: CERs (%) on development, test, and clean test sets for different training strategies and parameter freezing configurations. The table compares direct fine-tuning with the proposed two-stage strategy, evaluating the impact of freezing specific components (audio encoder, prediction network, joint network) during the first stage.

4.5 Evaluation Metric

Character error rate (CER) was employed as the primary evaluation metric. CER quantifies the discrepancy between the predicted output and the reference transcription by counting the number of substitutions, deletions, and insertions. It is computed as the ratio of total character errors to the number of characters in the reference:

$$\mathrm{CER} = \frac{S + D + I}{C}, \tag{7}$$

where S, D, and I represent the numbers of substitutions, deletions, and insertions, respectively, and C denotes the total number of characters in the reference. As a character-level measure, CER provides a precise and widely accepted evaluation of recognition accuracy for speech recognition tasks, with lower values indicating better performance. For Taiwanese Hokkien, where annotations include a mix of Han characters and romanized phonetics, CER is particularly suitable because it captures errors across both orthographic forms and effectively reflects the ability of the model to handle tonal and phonological variations.

5 Results and Discussion

5.1 Effects of Language Initialization

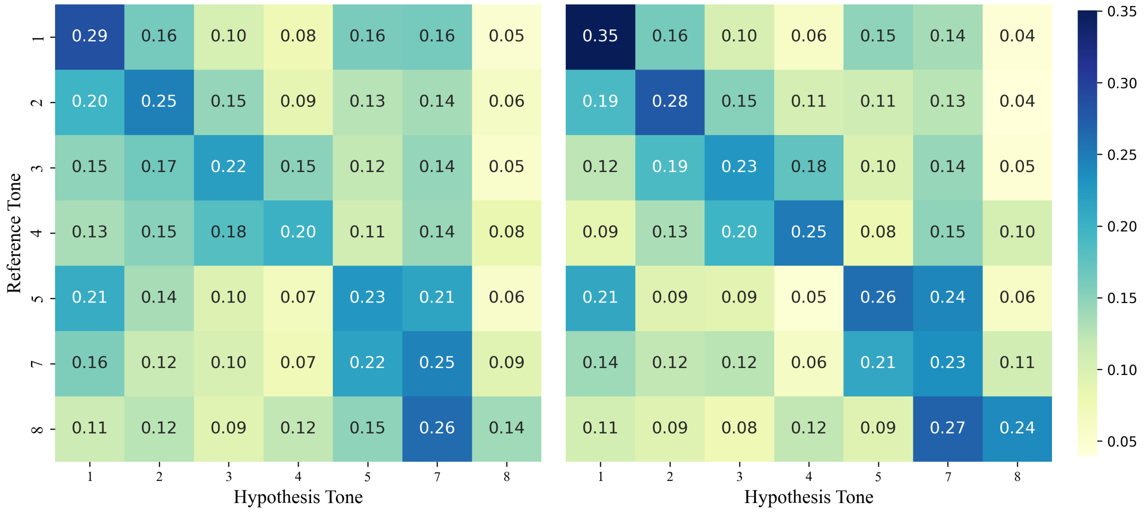
Table 2 summarizes the impact of different encoder initialization strategies on Taiwanese Hokkien ASR performance. The comparison includes Zipformer without pretraining, HuBERT-base pretrained on English, Whisper models with multilingual pretraining, and HuBERT-base-cmn pretrained on Mandarin. CLiFT-ASR, built on the Mandarin-pretrained HuBERT-base-cmn encoder and the proposed two-stage fine-tuning strategy, achieves the strongest overall performance.

Models without language-specific pretraining, such as Zipformer, exhibit the lowest performance, highlighting the difficulty of learning effective acoustic representations from limited Taiwanese data alone. Whisper-base, benefiting from large-scale multilingual pretraining, shows significant improvement and robust generalization across languages. English-pretrained HuBERT-base offers moderate gains, indicating that cross-lingual transfer helps but is constrained by the phonological mismatch between English and Taiwanese Hokkien. Compared with these strong baselines, CLiFT-ASR consistently reduces CER across all evaluation sets, achieving up to 26.2% relative improvement on the development set and 24.88% on the test set. While Whisper-small slightly outperforms CLiFT-ASR on certain splits, it contains more than twice the number of parameters. CLiFT-ASR therefore offers a parameter-efficient solution with substantial gains over competitive baselines, demonstrating the effectiveness of cross-lingual initialization combined with progressive two-stage fine-tuning.

5.2 Analysis of Fine-tuning Strategies

Table 3 presents the effects of different fine-tuning strategies and parameter freezing configurations on CLiFT-ASR performance. Compared with direct end-to-end fine-tuning, the proposed two-stage strategy, which first adapts the model on phonetic (romanized) transcriptions and then refines it with Han character targets, consistently improves recognition accuracy across all evaluation sets.

Analyzing parameter freezing during the



(a) Baseline

(b) CLiFT-ASR

Figure 2: Row-normalized substitution confusion matrices for Taiwanese Hokkien tone prediction, comparing the baseline and proposed models. Tone labels are derived from Taibun surface forms without applying sandhi rules.

first stage highlights the contribution of each module. Freezing the audio encoder or joint network restricts the ability of the model to adapt to target phonetics and orthography, leading to notable performance degradation, whereas freezing the prediction network has a milder effect. The lowest CER is achieved when all components are trainable, indicating that full model adaptation within the two-stage fine-tuning strategy enables effective integration of acoustic and linguistic knowledge. These results demonstrate that CLiFT-ASR with a carefully designed multi-stage fine-tuning strategy outperforms direct adaptation and provides a robust solution for low-resource mixed-orthography ASR scenarios.

5.3 Investigation of Tone Confusions

Figure 2 depicts the substitution confusion matrix for tone prediction in Taiwanese Hokkien. The diagonal dominance indicates that most tones are correctly classified, yet tones 5, 7, and 8 exhibit frequent mutual misclassifications. These errors are likely attributed to tone sandhi phenomena, overlapping pitch contours, and speaker-dependent prosodic variations, which complicate accurate tone modeling in ASR. To conduct this analysis, we employed the Taibun tool[5] to convert Taiwanese Han character outputs into Romanized forms with numerical tone labels. By aligning reference and predicted tone sequences, we constructed row-normalized substitution matrices to quantify tone-level confusions.

In the Zipformer baseline, tones 1, 5, and 7 emerge as the most error-prone categories. Tone 1 is correctly recognized only 29% of the time, with 16% of its instances misclassified as tone 2. Tone 5 is frequently misclassified as tone 1 (21%), while tones 7 and 8 show substantial cross-confusions, indicating the limited ability of the baseline model to discriminate between acoustically similar tones. In contrast, the proposed CLiFT-ASR system demonstrates clear improvements across most tonal categories. Tone 1 accuracy increases from 29% to 35%, while tone 4 recognition improves from 20% to 25%. The overall misclassification rate decreases, particularly for tones 5 and 7, where cross-tone errors are substantially reduced. These results highlight the enhanced discriminative capability of the proposed framework. In summary, the tone confusion analysis confirms that CLiFT-ASR effectively reduces inter-tone errors, especially among acoustically similar tone pairs. This improvement can be attributed to the proposed feature design and training strategy, which together provide more robust tonal modeling for Taiwanese Hokkien ASR.

[5] https://github.com/andreihar/taibun/

6 Conclusion

This study presents CLiFT-ASR, a cross-lingual fine-tuning framework designed for low-resource Taiwanese Hokkien ASR. By initializing the audio encoder with Mandarin speech representations and applying an effective two-stage fine-tuning strategy, CLiFT-ASR achieves the best overall performance. The first stage leverages Taiwanese romanization to capture detailed phonetic information, and the second stage adapts to Han character transcriptions to integrate orthographic and syntactic knowledge. This progressive strategy highlights the advantage of aligning acoustic and linguistic representations in stages rather than directly training with limited annotated data. An analysis of tone recognition shows that while general tone recognition is accurate, tones 5, 7, and 8 remain difficult due to tone sandhi, overlapping acoustic patterns, and speaker-specific prosodic variation, all of which complicate precise tone modeling.

7 Future Work

Several directions can be explored to extend the proposed CLiFT-ASR. One promising avenue is targeted data augmentation that balances underrepresented tones. Another is explicit modeling of tone sandhi, which may further reduce tonal confusion. The integration of larger and more diverse pretraining corpora is expected to improve robustness, particularly for conversational speech. Future research may also apply advanced sequence modeling or structured prediction techniques to capture tonal dependencies more effectively. Finally, evaluating multilingual models such as Whisper could provide additional gains through large-scale pre-training and enhanced contextual modeling.

8 Limitations

Although CLiFT-ASR achieves competitive improvements, the current design relies on a stateless RNN-Transducer framework. The stateless prediction network constrains the ability to model long-range dependencies, which may reduce accuracy in recognizing tonal patterns and complex tone sandhi. Compared with recent large-scale pretrained models, the architecture also has limited capacity to exploit fully contextualized acoustic representations. These limitations suggest that adopting more expressive architectures with stronger context modeling could further advance Taiwanese Hokkien ASR.

References

Fu-An Chao, Tien-Hong Lo, Shi-Yan Weng, Shih-Hsuan Chiu, Yao-Ting Sung, and Berlin Chen. 2021. The NTNU Taiwanese ASR system for formosa speech recognition challenge 2020. In *Proc. IJCLCLP*.

Peng-Jen Chen, Kevin Tran, Yilin Yang, Jingfei Du, Justine Kao, Yu-An Chung, Paden Tomasello, Paul-Ambroise Duquenne, Holger Schwenk, Hongyu Gong, Hirofumi Inaguma, Sravya Popuri, Changhan Wang, Juan Pino, Wei-Ning Hsu, and Ann Lee. 2023. Speech-to-speech translation for a real-world unwritten language. In *Findings of ACL*.

Robert L. Cheng. 1968. Tone sandhi in Taiwanese. *Linguistics*, 6(41):19–42.

Yi-Hui Chou, Kalvin Chang, Meng-Ju Wu, Winston Ou, Alice Wen-Hsin Bi, Carol Yang, Bryan Y. Chen, Rong-Wei Pai, Po-Yen Yeh, Jo-Peng Chiang, Iu-Tshiann Phoann, Winnie Chang, Chenxuan Cui, Noel Chen, and Jiatong Shi. 2023. Evaluating self-supervised speech models on a Taiwanese Hokkien corpus. In *Proc. IEEE ASRU*.

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 *Proc. ICML*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Hui-lu Khoo. 2019. The dynamics of Southern Min in Taiwan: From Southern Min dialects to “Taigi” . In Chris Shei, editor, *The Routledge Handbook of Chinese Discourse Analysis*, pages 596–610. Routledge.

Yuan-Fu Liao, Chia-Yu Chang, Hak-Khiam Tiun, Huang-Lan Su, Hui-Lu Khoo, Jane S. Tsay, Le-Kun Tan, Peter Kang, Tsun-guan Thiann, Un-Gian Iunn, Jyh-Her Yang, and Chih-Neng Liang. 2020. Formosa speech recognition challenge 2020 and Taiwanese across Taiwan corpus. In *Proc. O-COCOSDA*.

Yuan-Fu Liao, Jane S. Tsay, Peter Kang, Hui-Lu Khoo, Le-Kun Tan, Li-Chen Chang, Un-Gian Iunn, Huang-Lan Su, Tsun-Guan Thiann, Hak-Khiam Tiun, and Su-Lian Liao. 2022. Taiwanese across Taiwan corpus and its applications. In *Proc. O-COCOSDA*.

Jiayan Lin, Shenghui Lu, Hukai Huang, Wenhao Guan, Binbin Xu, Hui Bu, Qingyang Hong, and Lin Li. 2024. MinSpeech: A corpus of Southern Min dialect for automatic speech recognition. In *Proc. Interspeech*.

Chaofen Sun. 2006. *Chinese: A linguistic introduction*. Cambridge University Press.

Xiong Wang, Zhuoyuan Yao, Xian Shi, and Lei Xie. 2021. Cascade RNN-transducer: Syllable based streaming on-device Mandarin speech recognition with a syllable-to-character converter. In *Proc. IEEE SLT*.

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*.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, Di Wu, and Zhendong Peng. 2022. WenetSpeech: A 10000+ hours multi-domain Mandarin corpus for speech recognition. In *Proc. ICASSP*.

Piotr Żelasko, Daniel Povey, Jan Yenda Trmal, and Sanjeev Khudanpur. 2021. Lhotse: A speech data representation library for the modern deep learning ecosystem. In *Proc. NeurIPS*.