Synthetic Taiwanese Minnan–Mandarin Code-Switching Speech from Parallel Text for Low-Resource Speech Translation

Pang-Cheng Wu[1], *Chia-Hua Wu*[2], *Tai-Shih Chi*[1], *Hsin-Min Wang*[2]

[1] Institute of Electrical and Computer Engineering, National Yang Ming Chiao Tung University, Taiwan

[2] Institute of Information Science, Academia Sinica, Taiwan

andywu.ee13@nycu.edu.tw, maxwu@iis.sinica.edu.tw, tschi@nycu.edu.tw, whm@iis.sinica.edu.tw

Abstract

This study investigates the construction of synthetic code-switching (CS) speech for low-resource Taiwanese Minnan–Mandarin CS speech translation (CS-ST), aiming to reduce the reliance on real CS speech with translation annotations. We construct synthetic CS speech from parallel text through word alignment, CS text synthesis, and text-to-speech (TTS) synthesis, using random and part-of-speech (POS)-constrained replacement strategies. The generated synthetic speech is used to train LLaMA-Omni for CS-ST without using real CS speech for training. Experimental results show that incorporating synthetic CS speech improves the BLEU score on real CS speech from 34.59 to 40.42 with random replacement and 41.18 with POS-constrained replacement, with the two strategies achieving comparable performance. Further experiments show that increasing the amount of synthetic CS data from the same parallel corpus provides no additional gains on real CS speech, suggesting a distribution gap between synthetic and natural CS speech. These results demonstrate the effectiveness of using parallel text to construct synthetic CS speech and highlight the importance of data diversity and representativeness for low-resource CS-ST.

Index Terms: code-switching, speech translation, synthetic data, Taiwanese Minnan, Mandarin

1. Introduction

Code-switching (CS) refers to the use of two or more languages within the same conversation or utterance and is common in multilingual communities worldwide. For example, in Taiwan, speakers frequently alternate between Taiwanese Minnan and Mandarin in everyday conversations. CS can occur at different structural levels, including intersentential switching, where the language switch occurs across sentence boundaries, and intrasentential switching, where two or more languages are used within the same sentence [1]. In this study, we focus on Taiwanese Minnan-dominant intrasentential CS. Recent studies have explored various code-switching speech or text processing tasks, including code-switching speech recognition (CS-ASR) [2, 3, 4, 5, 6, 7, 8], code-switching machine translation [9, 10], and code-switching speech translation (CS-ST) [11, 12]. However, existing CS systems still face two major challenges: language confusion and the scarcity of annotated CS data.

Although previous studies have developed various approaches to address language confusion [6, 7, 8, 13, 14, 15], these approaches typically rely on existing CS speech datasets, such as SEAME [16] and the ASRU 2019 Mandarin-English Code-Switching Challenge dataset [17]. For low-resource CS languages, however, obtaining sufficient CS speech data remains a major challenge in developing effective CS-ST systems.

In the absence of large-scale code-switching data, previous studies have explored the construction of synthetic code-switching text from non-code-switching resources. Some studies have utilized parallel text corpora by first obtaining correspondences between words or phrases through word alignment and then replacing selected units with their counterparts in the other language to generate synthetic code-switching sentences [9]. Other approaches have started from monolingual text and employed bilingual lexicons for lexical substitution, or further incorporated linguistic information to perform more constrained clause-level substitution for generating code-switching text [10]. For Taiwanese Minnan–Mandarin code-switching, previous work has also constructed a code-mixed corpus using parallel corpora, a bilingual dictionary, and linguistic constraints based on part-of-speech and syntactic information [18]. These studies show that non-code-switching text resources can be used to construct synthetic code-switching text for low-resource settings.

Once code-switching text is available, previous studies have further explored different approaches to generate corresponding synthetic code-switching speech. One line of work uses text-to-speech (TTS) models to synthesize speech from code-switching text. For example, previous studies have generated code-switching text from monolingual text through word translation or insertion and then used TTS to synthesize the corresponding speech, while other studies have leveraged existing code-switching data to generate additional synthetic speech using TTS models [5, 19]. Another approach is speech splicing, which first obtains word- or phone-level speech segments from monolingual recordings through forced alignment or other speech alignment methods, and then concatenates segments from different languages according to the target code-switching text [3, 4, 19]. These studies show that synthetic code-switching speech can be generated from text or monolingual speech resources using TTS or speech splicing.

Previous research has investigated code-switching across various language pairs, with established code-switching speech resources available for language pairs such as Mandarin–English and Spanish–English [16, 17, 20]. Existing studies have explored different approaches for constructing synthetic code-switching data. However, it remains unclear whether such approaches can be effectively applied to Taiwanese Minnan–Mandarin code-switching, where research and data resources remain limited, particularly for code-switching speech.

To investigate this question, we use parallel text as the primary resource to construct Taiwanese Minnan–Mandarin synthetic code-switching text through word alignment and substitution strategies. The generated code-switching text is then

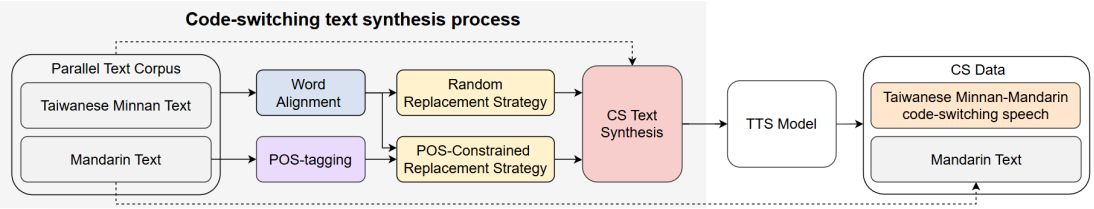


Figure 1: *The pipeline for Taiwanese Minnan–Mandarin CS speech translation data generation.*

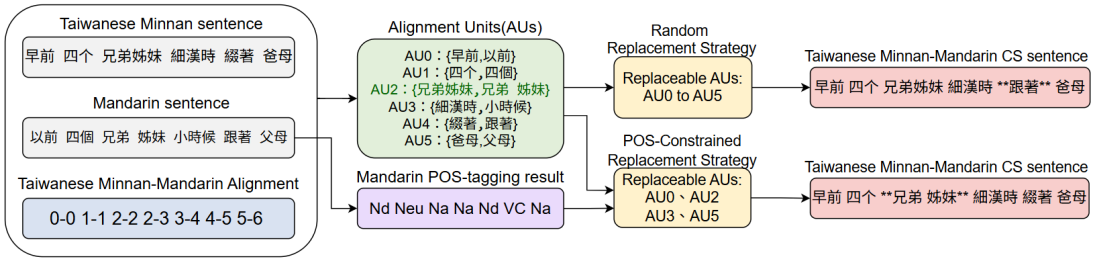


Figure 2: *Illustration of the CS text synthesis process. The example Mandarin and Taiwanese Minnan sentences mean "The four siblings used to live with their parents."*

converted into synthetic code-switching speech using a TTS model. We further investigate Taiwanese Minnan–Mandarin code-switching speech translation using a large audio-language model (LALM), specifically LLaMA-Omni [21]. The task converts Taiwanese Minnan–Mandarin code-switching speech, primarily in Taiwanese Minnan, into Chinese text. The resulting synthetic speech data are used to train the CS-ST model, which is evaluated on real Taiwanese Minnan–Mandarin code-switching speech to assess the effectiveness of synthetic CS speech. We further compare different CS text synthesis strategies and prompting designs to analyze their effects on CS-ST performance.

Experimental results show that training LLaMA-Omni with the generated synthetic CS speech training data improves ST performance on real Taiwanese Minnan–Mandarin code-switching speech, demonstrating that parallel-text-based synthetic CS speech generation can serve as a viable data construction strategy for low-resource CS-ST.

Based on these findings, the main contributions of this work are summarized as follows:

- We demonstrate the feasibility of constructing synthetic Taiwanese Minnan–Mandarin code-switching speech from parallel text and show that the resulting synthetic speech can improve CS-ST performance on real code-switching speech.
- We investigate different CS text synthesis strategies, including random and POS-constrained replacement, and compare their effects on CS-ST performance. We also investigate unified and language-specific prompting strategies for LLaMA-Omni.

2. Method

Fig. 1 shows an overview of the CS data generation pipeline. Fig. 2 illustrates the CS text synthesis process, while Fig. 3 shows the CS speech translation model.

2.1. Speech Translation Model

As shown in Fig. 3, we employ LLaMA-Omni [21] as the backbone architecture for Taiwanese Minnan–Mandarin CS-ST. Although LLaMA-Omni was originally proposed for speech-to-speech (S2S) applications, we only leverage its first-stage Speech-to-text component.

The model consists of a frozen Whisper-large-v3 [22] speech encoder, a linear speech adapter, and an LLM. During training, only the adapter and LLM parameters are updated. The acoustic features extracted by the speech encoder are projected through the adapter and fed into the LLM to generate Mandarin text.

To adapt LLaMA-Omni for Taiwanese Minnan–Mandarin CS-ST, we design two prompting strategies: **Unified** and **Language-Specific**, as shown in Table 1. The Language-Specific setting assumes prior knowledge of the input language category and selects the corresponding prompt accordingly, whereas the Unified setting applies the same prompt to all input utterances without requiring language identification. Their performance is compared in Section 5.1.

2.2. Code-Switching Data Generation Pipeline

Parallel Text Data Cleaning: We first cleaned the parallel text corpus by removing punctuation and filtering out sentences containing English characters or only a single word. The Taiwanese Minnan and Mandarin texts were then matched against the pronunciation lexicons used by the TTS models. For out-of-vocabulary (OOV) words, recursive segmentation and character variant conversion were applied to obtain lexicon-supported pronunciations. Sentence pairs containing words that remained unsupported after these procedures were removed.

Parallel Text Word Alignment: AWESOME-align [23], an unsupervised alignment model based on mBERT [24], was employed to obtain word-level alignments between Taiwanese Minnan and Mandarin. The model was further fine-tuned on the cleaned corpus to better adapt to domain-specific vocabulary. Alignment matrices that recorded the index correspondences between the two languages were subsequently extracted and used as the basis for CS text synthesis.

Code-Switching Text Synthesis: Using the cleaned parallel corpus and alignment matrices, aligned segments were substituted to synthesize CS text, as illustrated in Fig. 2. Based on the alignment results, aligned word indices with connected alignment relationships were grouped into indivisible *Align-*

Table 1: *Prompt templates used for fine-tuning and inference with the LLaMA-Omni model in CS-ST experiments.*

Prompt Setting	**Prompt Template**
Unified	Please transcribe the following speech into Chinese text.
Language-specific	Please transcribe the following Mandarin speech into Chinese text. Please transcribe the following Taiwanese Minnan speech into Chinese text. Please transcribe the following Taiwanese Minnan–Mandarin code-switching speech into Chinese text.

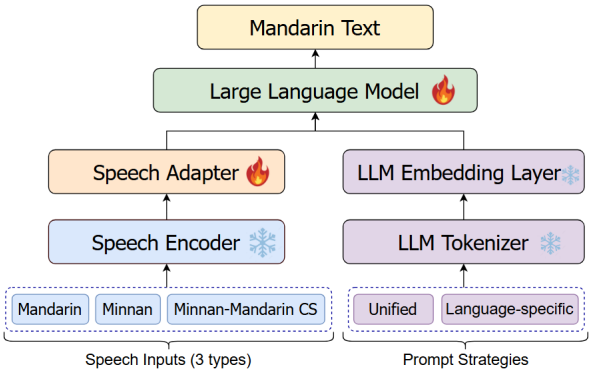


Figure 3: *Architecture of the LLaMA-Omni-based CS-ST model. The model takes Mandarin, Taiwanese Minnan, or Taiwanese Minnan–Mandarin code-switching speech as input and generates Mandarin text. We investigate two prompting strategies: unified and language-specific prompting.*

ment Units (AUs). Each AU represents a group of aligned words that are replaced as a single unit during CS text synthesis. To avoid overly short replacement segments, only AUs whose Mandarin segments contain at least two characters were retained as valid replacement candidates. Two CS text synthesis strategies were further investigated:

1. **Random Replacement Strategy**: All valid AUs were considered as replacement candidates regardless of their POS categories. For each pair of parallel sentences, a valid AU is randomly selected, and the corresponding segment in the Taiwanese Minnan sentence is replaced with the corresponding Mandarin segment.
2. **POS-Constrained Replacement Strategy**: Since nouns frequently occur at code-switching positions [25, 26], POS tagging and word segmentation are applied to Mandarin sentences, and only AUs with noun-related POS tags (Na, Nb, Nc, and Nd)[1] are retained as replaceable candidates.

To constrain the generated CS sentences to be Taiwanese Minnan-dominant, we require the number of Taiwanese Minnan characters to be greater than the number of Mandarin characters for both strategies.

Synthetic Code-Switching Speech Generation: We trained a Transformer-based speech synthesis system [27] using monolingual Taiwanese Minnan and Mandarin speech corpora. Both Taiwanese Minnan and Mandarin words are converted into IPA sequences, with their tone representations aligned across the two languages. The TTS system adopts a multi-speaker design, where speaker embeddings learned from the training speakers are used during speech synthesis. Similar multi-speaker TTS settings have been used for synthetic CS speech generation in previous studies [5]. During inference, the Taiwanese Minnan–Mandarin CS sentences generated in the CS text synthesis stage are converted into IPA sequences using the same phoneme and tone representations, and then fed into the Transformer TTS model to generate the corresponding CS speech.

3. Datasets

3.1. Word Alignment Training Data

We used iCorpus[2] as the source parallel text corpus for the proposed CS data generation pipeline. The original corpus contains 64,110 parallel sentence pairs. After data cleaning, the corpus was used to fine-tune and infer word alignments with AWESOME-align. Duplicate parallel sentence pairs were subsequently removed, resulting in 59,906 unique sentence pairs. The resulting corpus was divided into 58,000 sentence pairs for training and 1,906 sentence pairs for testing.

3.2. TTS Training Data

The TTS training data includes 156 hours of Mandarin speech from the COSPRO dataset [28], comprising 113 speakers (51 male and 62 female), and 38 hours of Taiwanese Minnan speech from the TAT dataset [29], comprising 4 speakers (2 male and 2 female).

3.3. CS-ST Training and Evaluation Data

Table 2 summarizes the real and synthetic speech datasets used for training and evaluation. The real Mandarin speech data [30] were collected from news-domain speech, while the Taiwanese Minnan speech data [31] were collected from Taiwanese local television dramas containing Taiwanese Minnan speech with corresponding Mandarin subtitles. After data cleaning, the Taiwanese Minnan data were used for Taiwanese Minnan-to-Mandarin speech translation, while the Mandarin data were used for Mandarin ASR. In addition, a real Taiwanese Minnan–Mandarin code-switching (CS) test set collected from YouTube was used to evaluate the models on naturally occurring CS speech.

Synthetic Taiwanese Minnan–Mandarin CS speech was generated using the data generation pipeline described in Section 2.2, with two strategies: POS-constrained replacement (POS-CS) and random replacement (Random-CS). Since POS-CS imposes an additional POS constraint on replacement candidates, Random-CS yields more synthetic CS utterances. For training, four Random-CS sets and one POS-CS set were constructed, with the corresponding test sets generated using the same strategies. Detailed statistics, including the number of utterances, total duration, average utterance duration, and average

[1] https://github.com/ckiplab/ckiptagger/wiki/POS-Tags

[2] https://github.com/Taiwanese-Corpus/icorpus_ka1_han3-ji7

Table 2: *Statistics of the CS-ST Training and Evaluation Datasets*

Category	Dataset	Utterances	Hours	Average Utterance Duration (s)	Average Number of Characters	CMI (%)
Real Train	Mandarin	51,254	115.03	8.08	40.29	N/A
	Taiwanese Minnan	115,874	100.00	3.11	10.96	N/A
Synthetic Train	POS-CS	50,165	41.23	2.96	11.34	18.16
	Random-CS$_1$	53,597	43.59	2.93	11.21	18.15
	Random-CS$_2$	53,597	43.63	2.93	11.21	19.14
	Random-CS$_3$	53,597	43.63	2.93	11.21	18.22
	Random-CS$_4$	53,597	43.58	2.93	11.21	18.52
Real Test	Mandarin	1,276	2.83	7.99	41.67	N/A
	Taiwanese Minnan	2,318	2.00	3.11	10.96	N/A
	Real-CS	1,022	0.81	2.84	14.08	18.91
Synthetic Test	POS-CS	1,608	1.31	2.93	11.28	18.00
	Random-CS	1,752	1.40	2.88	11.13	18.00

number of characters, are reported in Table 2. For consistency across datasets, the average number of characters is calculated based on the Mandarin target translations rather than the source transcriptions.

To quantify the degree of language mixing in the CS datasets, we compute the Code-Mixing Index (CMI) [4, 32]. As shown in Table 2, the synthetic CS datasets have CMI values of approximately 18%–19%, comparable to the 18.91% of the real CS test set. This indicates that the synthetic and real CS data exhibit a similar degree of language mixing.

4. Experimental Setup

For the CS data generation pipeline, we used AWESOME-align[3] [23] for Taiwanese Minnan–Mandarin word alignment, with BERT-base-multilingual-cased as the backbone model. The model was fine-tuned following the unsupervised training procedure provided in the official implementation, using a fixed learning rate of 2×10^{-5} for one epoch on a single NVIDIA RTX 3090 GPU. For POS tagging, we used the CKIP toolkit[4] [33] to obtain POS tags for Mandarin text.

For speech synthesis, we trained a Transformer-based TTS model [27] using the ESPnet toolkit [34], following its official training and inference setup. The model was trained for 1,000 epochs, and the five best checkpoints according to training loss were averaged for inference. For waveform generation, we used a pretrained HiFi-GAN vocoder further fine-tuned on the COSPRO dataset [28] for 500,000 steps.

For the CS-ST experiments, we adopted LLaMA-Omni[5] [21] with Llama-3.2-1B-Instruct[6] as the backbone language model. The model was fine-tuned for 10 epochs with a batch size of 12 using the AdamW optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.95$) with a weight decay of 0.01. We used a cosine learning rate scheduler with an initial learning rate of 1×10^{-4} and a warmup ratio of 1%. Training was performed on an NVIDIA H100 GPU cluster using BF16 mixed precision.

For evaluation, character error rate (CER) was used for Mandarin ASR, while BLEU was used for Taiwanese Minnan-to-Mandarin ST and Taiwanese Minnan–Mandarin CS-ST. BLEU scores were computed using SacreBLEU with the Chinese tokenizer (`zh`) [35].

[3] https://github.com/neulab/awesome-align

[4] https://github.com/ckiplab/ckip-transformers

[5] https://github.com/BayLing-Models/BayLing-Speech

[6] meta-llama/Llama-3.2-1B-Instruct

Table 3: *Performance of unified and language-specific prompting strategies (Table 1) with Mandarin + Taiwanese Minnan + POS-CS training data. Minnan denotes Taiwanese Minnan.*

Prompt	Mandarin CER ↓	Minnan BLEU ↑	POS-CS BLEU ↑	Real-CS BLEU ↑
Unified	4.64	**68.17**	68.85	**41.18**
Language-Specific	**4.61**	67.70	**68.98**	39.62

5. RESULTS

5.1. Comparison of Prompting Strategies

Table 3 compares unified and language-specific prompting (see Table 1) under the same training configuration. The two strategies show comparable performance on Mandarin, Taiwanese Minnan, and POS-CS, with differences within 0.5 points. On Real-CS, however, unified prompting outperforms language-specific prompting by 1.56 BLEU points. One possible explanation is that language-specific prompting introduces an additional dependency on language-specific prompt patterns learned from the training data, which may generalize less effectively to the more diverse code-switching patterns in real-world speech. In contrast, unified prompting provides a consistent transcription instruction regardless of the input language composition, potentially improving robustness to real-world CS input. Given its better performance on Real-CS and the fact that it does not require language identification at inference time, we adopt unified prompting for the subsequent experiments.

5.2. Overall Performance and Comparison of Synthesis Strategies

Table 4 compares the performance of different training configurations and CS synthesis strategies. The configurations without synthetic CS speech (T1–T3) show limited performance on the Real-CS test set. Training on Mandarin (T1) or Taiwanese Minnan (T2) alone yields BLEU scores of 15.53 and 21.49, respectively. Combining the two monolingual training sets (T3)

Table 4: *Performance comparison of different CS synthesis strategies. M and TM denote Mandarin and Taiwanese Minnan, respectively. Random-CS and POS-CS denote synthetic CS data generated using random and POS-constrained replacement, respectively. Random-CS_1–Random-CS_4 are independently generated using the same random replacement strategy, and the average performance over T4–T7 is also reported.*

ID	Train Set	Mandarin (CER ↓)	Taiwanese Minnan (BLEU ↑)	POS-CS (BLEU ↑)	Random-CS (BLEU ↑)	Real-CS (BLEU ↑)
T1	Mandarin (M)	4.88	6.30	12.19	11.89	15.53
T2	Taiwanese Minnan (TM)	93.93	67.86	17.12	16.56	21.49
T3	M + TM	4.99	67.01	27.74	28.76	34.59
T4	M + TM + Random-CS_1	4.76	67.85	67.66	68.04	40.45
T5	M + TM + Random-CS_2	4.68	67.18	66.92	67.58	40.23
T6	M + TM + Random-CS_3	4.72	68.04	67.53	**68.61**	**41.61**
T7	M + TM + Random-CS_4	4.70	68.13	67.57	68.26	39.38
	Avg. T4–T7	4.72	67.80	67.42	68.12	40.42
T8	M + TM + POS-CS	**4.64**	**68.17**	**68.85**	64.42	41.18

Table 5: *Effects of progressively adding Random-CS data to the best-performing configuration (T8: M + TM + POS-CS) from Table 4. BLEU (↑) is reported on the POS-CS, Random-CS, and Real-CS test sets.*

Train Set		Test Set (BLEU ↑)		
Base Train Set	**Additional Random-CS Data**	**POS-CS**	**Random-CS**	**Real-CS**
M + TM + POS-CS	–	68.85	64.42	41.18
	Random-CS_1	71.27	70.63	41.21
	Random-$CS_{1,2}$	70.51	70.21	40.50
	Random-$CS_{1,2,3}$	68.84	69.33	40.19
	Random-$CS_{1,2,3,4}$	69.42	69.02	40.06

substantially improves the Real-CS BLEU score to 34.59, indicating that bilingual training improves generalization to CS speech even without explicit CS training data.

Adding synthetic CS speech further improves performance on Real-CS. Compared with the bilingual baseline T3, the four Random-CS configurations (T4–T7) achieve an average BLEU score of 40.42, while the POS-CS configuration (T8) achieves 41.18, corresponding to absolute improvements of 5.83 and 6.59 BLEU, respectively. T8 also achieves the lowest Mandarin CER (4.64%) and the highest Taiwanese Minnan BLEU score (68.17) among all configurations, indicating that the improvement on CS speech does not degrade performance on the monolingual test sets.

The difference between the two synthesis strategies on Real-CS is relatively small. T8 achieves 41.18 BLEU, compared with an average of 40.42 BLEU for T4–T7. The individual Random-CS configurations yield Real-CS scores ranging from 39.38 to 41.61 BLEU, indicating some sensitivity to the random replacement patterns used for data generation. A clearer difference is observed on the synthetic CS test sets. T4–T7 achieve an average of 68.12 BLEU on the Random-CS test set, compared with 64.42 for T8. Conversely, T8 achieves 68.85 BLEU on the POS-CS test set, compared with an average of 67.42 for T4–T7. These results suggest that performance on synthetic CS test sets is partly influenced by the correspondence between the switching patterns used for training and evaluation.

Overall, both Random-CS (T4–T7) and POS-CS (T8) substantially outperform the bilingual baseline (T3) on Real-CS, demonstrating the effectiveness of synthetic CS speech for improving translation performance on real CS speech. Although T8 achieves slightly higher Real-CS performance than the average of T4–T7, the difference is small relative to the variation among the individual Random-CS configurations.

5.3. Analysis of Synthetic Data Scaling

Based on the results in Table 4, we select T8 (M + TM + POS-CS) as the base training configuration and further investigate whether increasing the amount of synthetic CS speech provides additional gains. As shown in Table 5, adding one Random-CS dataset improves BLEU from 68.85 to 71.27 on POS-CS and from 64.42 to 70.63 on Random-CS, while the improvement on Real-CS is marginal (41.18 to 41.21).

Further increasing the amount of Random-CS data does not yield additional improvements. As the number of Random-CS datasets increases from one to four, Real-CS performance decreases from 41.21 to 40.50, 40.19, and 40.06 BLEU with two, three, and four datasets, respectively, despite consistently high performance on the synthetic POS-CS and Random-CS test sets. These results suggest that the benefit of synthetic CS data may saturate as more synthetic data are added. Moreover, the different trends observed between the synthetic and real CS test sets indicate a potential distribution gap between synthesized and naturally occurring CS speech.

A possible explanation is that Random-CS_1–Random-CS_4 are generated from the same parallel sentence pairs and differ primarily in their randomly selected replacement positions. Consequently, increasing the amount of synthetic data introduces additional switching patterns but limited new linguistic content. Overall, these results suggest that simply scaling synthetic CS speech derived from the same source may not continuously improve performance on real CS speech, and that fur-

Table 6: *Translation examples from the POS-CS, Random-CS, and Real-CS test sets. Baseline is T3 in Table IV (M+TM), whereas Proposed is the best-performing configuration in Table V, trained with M+TM, POS-CS, and Random-CS$_1$.*

Test Set	System	Mandarin Output and Literal English Gloss
POS-CS	Reference	民眾前往日本[CS]旅遊的人數攀升 *The number of people going to Japan[CS] for travel is increasing.*
	Baseline	民眾將只有日本女友的人手攀升 *The manpower of people who will only have Japanese girlfriends is increasing.*
	Proposed	民眾前往前往日本旅遊的人數攀升 *The number of people going to going to Japan for travel is increasing.*
Random-CS	Reference	用來培訓航艦搭載[CS]戰鬥機的飛行員 *Used to train carrier-based[CS] fighter pilots.*
	Baseline	今來陪伴行人幸好正道基的非行員 *Now come to accompany pedestrians; fortunately, the non-pilots of the proper-way base.*
	Proposed	用來培訓航艦搭載戰鬥機的飛行員 *Used to train carrier-based fighter pilots.*
Real-CS	Reference	中國的新疆伊犁[CS]地區最近因為雨下不停的關係 *Due to the recent continuous rainfall in the Ili region of Xinjiang[CS], China.*
	Baseline	心腦醫治邱最近因為雨下不利的關係 *Due to the recent unfavorable rainfall in "Xinnao Yizhiqiu," a nonsensical location phrase.*
	Proposed	中國的新疆一二地區最近因為雨下不利的關係 *Due to the recent unfavorable rainfall in the "One-Two" region of Xinjiang, China.*

Note: Reference denotes the ground-truth Mandarin translation of the input speech. Boxed spans marked with [CS] correspond to code-switched portions of the input speech. Red underlined text indicates errors in model outputs relative to the reference. Italicized English text provides literal glosses of the Mandarin reference and model outputs; repetitions and semantic anomalies are intentionally preserved.

Table 7: *Comparison of ST performance and estimated SNR between original and TTS-synthesized speech on a paired Real-CS subset. The synthesized speech uses the same transcripts and reference audio from the corresponding unseen speakers.*

Speech	BLEU ↑	SNR (dB)
Original Real-CS	**44.82**	16.59
TTS Real-CS	44.02	**19.01**
Δ (TTS − Original)	−0.80	+2.42

ther improvements may require greater data diversity or better matching between synthetic and real CS speech rather than simply increasing the quantity of synthetic data.

5.4. Qualitative Analysis of Translation Results

To further examine the effect of synthetic code-switching (CS) speech on translation quality, Table 6 presents translation examples from the POS-CS, Random-CS, and Real-CS test sets. In this comparison, the baseline model is trained on monolingual Mandarin and Taiwanese Minnan data, while the proposed model additionally incorporates synthetic CS data (POS-CS and Random-CS$_1$).

The baseline model exhibits lexical and semantic errors, whereas the proposed model generally produces outputs that more closely match the Mandarin references. Specifically, in the POS-CS example, the proposed model successfully recovers the core meaning but retains a repetition error. In the Random-CS example, it corrects the errors produced by the baseline and generates the exact reference translation. In the Real-CS example, the proposed model substantially improves the overall translation but still produces an error in the place name "Yili".

These examples show that training with synthetic CS speech can improve the completeness and accuracy of the generated translations. However, the remaining error in the Real-CS example suggests that the model may still have difficulty handling specific lexical items in naturally occurring CS speech. This observation is consistent with the potential distribution gap between synthetic and naturally occurring CS speech, suggesting that increasing the diversity and representativeness of synthetic training data may further improve model robustness.

5.5. Effect of TTS Synthesis on Real-CS Performance

To isolate the effect of speech synthesis, we conduct an ablation study on a paired subset of Real-CS. We keep the transcripts and speaker identities fixed, and replace the original recordings with speech synthesized using reference audio from the corresponding unseen speakers. As shown in Table 7, replacing the original speech with TTS-synthesized speech reduces BLEU only slightly, from 44.82 to 44.02, while increasing the estimated SNR from 16.59 to 19.01 dB. The small BLEU difference suggests that TTS synthesis itself has a limited effect on ST performance. Therefore, the performance gap between synthetic and real CS test sets is more likely associated with differences in linguistic content and code-switching patterns than with synthesis

quality alone.

6. Conclusions

This paper investigated the feasibility of constructing synthetic Taiwanese Minnan–Mandarin code-switching (CS) speech from parallel text for low-resource code-switching speech translation (CS-ST). The proposed data generation pipeline combines word alignment, CS text synthesis, and text-to-speech (TTS) synthesis to generate synthetic CS speech, which is then used to train LLaMA-Omni for Taiwanese Minnan–Mandarin CS-ST.

Compared with the "M + TM" baseline, the BLEU score on the Real-CS test set increases from 34.59 to 40.42 with random replacement and to 41.18 with POS-constrained replacement, corresponding to absolute improvements of 5.83 and 6.59 BLEU points, respectively. The two synthesis strategies achieve comparable performance on Real-CS, with POS-constrained replacement showing slightly better performance. Further scaling experiments show that increasing the amount of synthetic CS data derived from the same parallel corpus does not lead to additional improvements on Real-CS, despite strong performance on the synthetic CS test sets. This difference suggests a potential distribution gap between synthetic and naturally occurring CS speech. Furthermore, our TTS ablation study shows that the acoustic synthesis process itself has only a limited impact on performance, suggesting that the remaining gap may be more closely related to differences in linguistic content and code-switching patterns.

Overall, these results demonstrate that parallel text is an effective resource for constructing synthetic CS speech when real CS-ST data are limited. They also indicate that improving the diversity and representativeness of synthetic CS data may be more important than simply increasing the amount of synthetic data derived from the same source. Future work will investigate more diverse parallel text sources and more representative CS patterns to reduce the mismatch between synthetic and real CS speech.

7. Acknowledgments

This work was supported by the National Science and Technology Council of Taiwan under Grant NSTC 115-2634-F-001-006.

8. Generative AI Use Disclosure

ChatGPT was used to assist with parts of the code implementation and to edit and polish the manuscript. AI-assisted code was reviewed and tested by the authors. The research methodology, experimental evaluation, analysis, and conclusions were conducted by the authors.

9. References

[1] S. Poplack, "Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching," *Linguistics*, pp. 581–618, 1980.

[2] T. Nguyen and H.-D. Tran, "AsyncSwitch: Asynchronous text-speech adaptation for code-switched ASR," in *Proc. ASRU*, 2025, pp. 1–7.

[3] ——, "Can we train ASR systems on code-switch without real code-switch data? case study for Singapore's languages," in *Proc. Interspeech*, 2025, pp. 753–757.

[4] A. Hussein, D. Zeinali, O. Klejch, M. Wiesner, B. Yan, S. Chowdhury, A. Ali, S. Watanabe, and S. Khudanpur, "Speech collage: code-switched audio generation by collaging monolingual corpora," in *Proc. ICASSP*, 2024, pp. 12 006–12 010.

[5] Y. H. Yeo, Y. Hu, S. Gopal, Y. Peng, H. Liu, and E. S. Chng, "Improving code-switching speech recognition with TTS data augmentation," in *Proc. APSIPA ASC*, 2025, pp. 992–997.

[6] F. Zhang, W. Geng, H. Huang, Y. Shan, C. Yi, and H. Qu, "Boosting code-switching ASR with mixture of experts enhanced speech-conditioned LLM," in *Proc. ICASSP*, 2025, pp. 1–5.

[7] H. Wang, X. Wan, N. Zheng, K. Liu, H. Zhou, G. Li, and L. Xie, "Camel: Cross-attention enhanced mixture-of-experts and language bias for code-switching speech recognition," in *Proc. ICASSP*, 2025, pp. 1–5.

[8] S. Ye, S. Chen, X. Hu, and X. Xu, "SC-MoE: Switch conformer mixture of experts for unified streaming and non-streaming code-switching ASR," in *Proc. Interspeech*, 2024, pp. 3999–4003.

[9] J. Xu and F. Yvon, "Can you traducir this? machine translation for code-switched input," in *Proc. CACLS*, 2021, pp. 84–94.

[10] I. Tarunesh, S. Kumar, and P. Jyothi, "From machine translation to code-switching: Generating high-quality code-switched text," in *Proc. ACL-IJCNLP*, 2021, pp. 3154–3169.

[11] B. Shankar, P. Jyothi, and P. Bhattacharyya, "CoSTA: Code-switched speech translation using aligned speech-text interleaving," in *Proc. COLING*, 2025, pp. 9194–9208.

[12] B. Alastruey, M. Sperber, C. Gollan, D. Telaar, T. Ng, and A. Agarwal, "Towards real-world streaming speech translation for code-switched speech," in *Proc. CALCS*, 2023, pp. 14–22.

[13] H. Liu, H. Xu, L. P. Garcia, A. W. Khong, Y. He, and S. Khudanpur, "Reducing language confusion for code-switching speech recognition with token-level language diarization," in *Proc. ICASSP*, 2023, pp. 1–5.

[14] J. Zhao, H. Shi, C. Cui, T. Wang, H. Liu, Z. Ni, L. Ye, and L. Wang, "Adapting Whisper for code-switching through encoding refining and language-aware decoding," in *Proc. ICASSP*, 2025, pp. 1–5.

[15] Z. Fan, L. Dong, C. Shen, Z. Liang, J. Zhang, L. Lu, and Z. Ma, "Language-specific acoustic boundary learning for Mandarin-English code-switching speech recognition," in *Proc. Interspeech*, 2023, pp. 3322–3326.

[16] D.-C. Lyu, T. P. Tan, E. Chng, and H. Li, "SEAME: a Mandarin-English code-switching speech corpus in south-east Asia," in *Proc. Interspeech*, 2010, pp. 1986–1989.

[17] X. Shi, Q. Feng, and L. Xie, "The ASRU 2019 Mandarin-English code-switching speech recognition challenge: Open datasets, tracks, methods and results," in *Proc. WSTCSMC*, 2020, p. 71.

[18] S.-E. Lu, B.-H. Lu, C.-Y. Lu, and R. T.-H. Tsai, "Exploring methods for building dialects-mandarin code-mixing corpora: A case study in Taiwanese Hokkien," in *Proc. EMNLP*, 2022, pp. 6287–6305.

[19] C. Du, H. Li, Y. Lu, L. Wang, and Y. Qian, "Data augmentation for end-to-end code-switching speech recognition," in *Proc. SLT*, 2021, pp. 194–200.

[20] M. Deuchar, P. Davies, J. Herring, M. C. P. Couto, and D. Carter, "Building bilingual corpora," *Advances in the Study of Bilingualism*, pp. 93–110, 2014.

[21] Q. Fang, S. Guo, Y. Zhou, Z. Ma, S. Zhang, and Y. Feng, "LLaMA-Omni: Seamless speech interaction with large language models," in *Proc. ICLR*, 2025, pp. 57 607–57 624.

[22] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. ICML*, 2023, pp. 28 492–28 518.

[23] Z.-Y. Dou and G. Neubig, "Word alignment by fine-tuning embeddings on parallel corpora," in *Proc. EACL*, 2021, pp. 2112–2128.

[24] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[25] J. Chi and P. Bell, "Analyzing the role of part-of-speech in code-switching: A corpus-based study," in *Proc. EACL*, 2024, pp. 1712–1721.

[26] C.-H. Hsiao, "A corpus investigation of the typology of code-switching between closely related languages: Data from Mandarin–Taiwanese code-switching," *International Journal of Bilingualism*, pp. 1619–1637, 2025.

[27] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI*, 2019, pp. 6706–6713.

[28] C.-Y. Tseng, Y.-C. Cheng, and C.-H. Chang, "Sinica COSPRO and toolkit – corpora and platform of Mandarin Chinese fluent speech," in *Proc. O-COCOSDA*, 2005, pp. 23–28.

[29] Y.-F. Liao, J. S. Tsay, P. Kang, H.-L. Khoo, L.-K. Tan, L.-C. Chang, U.-G. Iunn, H.-L. Su, T.-G. Thiann, H.-K. Tiun *et al.*, "Taiwanese across Taiwan corpus and its applications," in *Proc. O-COCOSDA*, 2022, pp. 1–5.

[30] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Chinese broadcast news corpus," in *International Journal of Computational Linguistics & Chinese Language Processing, Volume 10, Number 2, June 2005: Special Issue on Annotated Speech Corpora*, 2005, pp. 219–236.

[31] P.-Y. Chen, C.-H. Wu, H.-S. Lee, S.-K. Tsao, M.-T. Ko, and H.-M. Wang, "Using taigi dramas with mandarin chinese subtitles to improve taigi speech recognition," in *Proc. O-COCOSDA*, 2020, pp. 71–76.

[32] B. Gambäck and A. Das, "Comparing the level of code-switching in corpora," in *Proc. LREC*, 2016, pp. 1850–1855.

[33] W.-Y. Ma and K.-J. Chen, "Introduction to CKIP Chinese word segmentation system for the first international Chinese word segmentation bakeoff," in *Proc. SIGHAN*, 2003, pp. 168–171.

[34] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[35] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, 2018, pp. 186–191.