

DEVELOPING PERFORMANT MODELS FOR TRANSLATING SPOKEN
TAIWANESE INTO SPOKEN ENGLISH USING FREE AND PUBLICLY AVAILABLE
RESOURCES

by

Eleanor M. Lin

presented to the Program of Linguistics

in partial fulfillment of the requirements

for the Major in Linguistics

Columbia College

April 2024

Acknowledgements

I am extremely grateful to Dr. William Foley, Dr. Meredith Landman, Dr. John McWhorter, and Dr. Ross Perlin for their guidance throughout the process of writing this thesis. I would also like to thank Melody Harwood for introducing me to the Coqui TTS library used in this work. Last but not least, many thanks to my fellow thesis writers in the Class of 2024 for your support.

Contents

1	Abstract	iii
2	Introduction	1
3	Literature review	4
4	Methods	11
4.1	Model architecture	11
4.2	Datasets and preprocessing	12
4.2.1	TSM speech corpora for ASR	13
4.2.2	Corpora for translating TSM to English	16
4.3	Creating the ASR module	19
4.4	Creating the MT module	21
4.5	Creating the TTS module	22
4.6	Evaluating the Translation System	24
4.7	Comparison with Chen et al. (2023)	26
5	Results	28
6	Discussion	30
6.1	ASR module error analysis	30
6.2	MT module error analysis	31
6.3	TTS module error analysis	34

7 Conclusion	37
References	39

1 Abstract

In this work, I investigate whether it is possible to build a system for translating spoken Taiwanese into spoken English, using only free and publicly available data and models. Recently, Chen et al. (2023) released the first-ever benchmark dataset and system for automatic speech-to-speech translation between Taiwanese and English. However, their dataset is not open-sourced, and their translation system shows substantial room for improvement. Given that many language communities lack the financial, data, and computational resources used by Chen et al., it is important to investigate whether a high-quality English-Taiwanese speech-to-speech translation system could be developed using only free, publicly accessible resources. Therefore, I build a system that automatically translates spoken Taiwanese to spoken English using only free, publicly available resources. While I ultimately conclude that the amount of data used in this work is insufficient to build a high-quality translation system, this finding is in itself informative for future work which seeks to build speech translation systems for Taiwanese or other low-resource languages, as it provides a baseline understanding of how much data is “enough” for building such systems. Additionally, this work produces a usable automatic speech recognition system as a sub-module of the overall translation system, demonstrating the success of a transfer learning approach to training ASR models for low-resource languages. Developing language technologies for Taiwanese advances our understanding of how to build language technologies in low-resource settings, while simultaneously providing much-needed tools to native Taiwanese speakers and Taiwanese learners.

2 Introduction

In this work, I investigate whether it is possible to build a system for translating spoken Taiwanese into spoken English, using only free and publicly available data and models. Taiwanese Southern Min (or “TSM” or “Taiwanese” for short) is a Sino-Tibetan language spoken by 13.5 million people in Taiwan (*China–Taiwan* 2023). Because TSM has not been recorded in a single, widely-used, standardized writing system for most of its history, few sources of data for developing TSM language technologies exist (Chen et al. 2023). Furthermore, due to decades of government policy that promoted Mandarin and suppressed other languages in Taiwan, TSM speakers are increasingly shifting to Mandarin, threatening the future survival of the language (Sandel, Chao & Liang 2006). While some English-speaking Taiwanese-Americans express interest in learning TSM, they face significant challenges to doing so, due to lack of language learning tools and resources (E. M. Lin 2023: 10-12). Developing language technologies for TSM could advance our understanding of how to build language technologies in low-resource settings, while simultaneously providing much-needed tools to native TSM speakers and TSM learners.

Recently, Chen et al. (2023) released the first-ever benchmark dataset and system for automatic speech-to-speech translation between TSM and English. However, much of the data used by Chen et al. is not publicly available, and their system performs relatively poorly, achieving ASR-BLEU scores (measuring how well machine translations accord with human translations) of only 7.8 (for English-to-TSM translation) and 10.0 (for TSM-to-English translation), compared to reference scores of 61.8 and 78.5, respectively (Chen et al. 2023: 4976). As highlighted by Ranathunga et al., creating more datasets and open-source models

for low-resource languages are both crucial directions that should be pursued to expand the availability of translation technologies to more diverse linguistic communities (Ranathunga et al. 2023: 229:24-229:26).

Motivated by the need for an open-source English-TSM speech-to-speech translation system that produces high-quality translations, as well as by the consideration that many language communities may not have access to the same kind of financial, data, or computational resources as Chen et al., I investigate whether it is possible to reproduce, and perhaps even improve upon, Chen et al.’s results using solely free and publicly available data and models. I build a system for automatically translating spoken Taiwanese into spoken English, using only publicly available resources. Ultimately, I find that the amount of data I use to develop my system is most likely insufficient to achieve equivalent performance to Chen et al. These findings can inform future work on building open-source systems for translating Taiwanese to English by providing evidence that such systems require, at minimum, more data than was used here to succeed. Additionally, while the translation system as a whole performs poorly, the module of the translation system that performs automatic speech recognition (transcribing Taiwanese speech) functions well enough to be used as a standalone component. This ASR module could be useful to language learners in applications such as computer-assisted pronunciation training (Golonka et al. 2014). Overall, the decent performance of the ASR module demonstrates the efficacy of a transfer learning approach to building ASR systems for low-resource settings. More generally, developing language technologies for Taiwanese advances our understanding of how to build language technologies in low-resource settings, while simultaneously providing much-needed tools to native Taiwanese speakers and Taiwanese learners.

The remainder of this thesis is organized as follows. In §3, I introduce relevant literature on automatic translation systems and sources of Taiwanese-language data. In §4, I describe my approach to collecting data and training the modules of the translation system. in §5, I report the overall performance of the system, as well as the performance of its individual modules. In §6, I discuss what types of errors the system makes, and potential reasons for these errors. I conclude in §7 with directions for future work, including proposals for improving the system's performance.

3 Literature review

Neural machine translation is the most common and successful approach for automatic translation between languages. Neural machine translation builds a translation model by teaching a *deep neural network* to translate between languages, using large datasets featuring parallel examples of the same text in multiple languages. However, such datasets are lacking for low-resource languages (Ranathunga et al. 2023: 229:5). In contrast to *high-resource languages*, *low-resource languages* are those which are less studied, potentially unwritten, and have a limited digital presence, resulting in lack of data for computational linguists and natural language processing researchers to analyze or develop language technologies with (Ranathunga et al. 2023: 229:3).

To overcome the challenge of data scarcity in building neural machine translation systems for low-resource languages, recent work has used approaches such as *transfer learning* and *data augmentation*. In *transfer learning*, a model may first be trained to translate between two high-resource languages, and then adapted for translating a low-resource language (Ranathunga et al. 2023: 229:5). Babu, Wang, Tjandra, Lakhotia, Xu, Goyal, Singh, von Platen, et al. demonstrate the effectiveness of a slightly different transfer learning approach with their XLS-R model, which first learns to create representations of speech from 128 languages, and then can subsequently be adapted for downstream tasks in low-resource languages, such as automatic speech recognition (Babu, Wang, Tjandra, Lakhotia, Xu, Goyal, Singh, von Platen, et al. 2022a). Transfer learning still requires some parallel data in the low-resource language, albeit less data than a non-transfer learning approach.

While transfer learning seeks to transfer knowledge learned from one language or task

to another, *data augmentation* involves constructing *synthetic data* for training a machine translation model (Ranathunga et al. 2023: 229:6). One method for generating synthetic data is *back-translation*, in which an existing corpus in one language is translated into either the *source language* (the language which the model is being trained to translate from) or the *target language* (the language the model is being trained to translate into). A second method for generating synthetic data is *data mining*, in which texts that are approximately semantically equivalent (rather than directly translated) from two different languages are retrieved from internet sources (Ranathunga et al. 2023: 229:7). Data augmentation relies on the existence of models that can translate between the low-resource language and another language.

While machine translation of low-resource languages using text remains a challenging task, *speech translation* for low-resource languages (i.e. generating the spoken translation of an utterance given a spoken utterance as input) is even less well-studied (Ranathunga et al. 2023: 229:4-5). But because many low-resource languages are *unwritten* (i.e. lack a widely adopted orthography), speech translation may be more useful than text translation for these languages (Lee et al. 2022, Zhang et al. 2021, Bansal et al. 2018, Scharenborg et al. 2020). However, automatically translating speech is particularly challenging because models not only must learn to align corresponding segments in a source sentence and its target translation (as in text translation), but also learn the acoustic features of the source and target languages (Lee et al. 2022: 1).

The lack of parallel data and standardized orthographies for many low-resource languages only exacerbates the speech-to-speech translation challenge, as speech-to-speech translation systems often rely on a *cascaded* approach involving an intermediate text-to-text translation

step. In the cascaded approach for speech-to-speech translation, input speech is automatically transcribed (*automatic speech recognition*) and then translated into text in the target language. Finally, speech in the target language is generated from the translated text (*text-to-speech synthesis*) (Bansal et al. 2018, Zhang et al. 2021, Lee et al. 2022, Chen et al. 2023). In addition to its reliance on learning from transcribed speech (which may not be available), cascaded speech-to-speech translation systems are computationally expensive and introduce the possibility errors being passed down from earlier stages of the translation pipeline to later stages (Lee et al. 2022: 3327, Chen et al. 2023: 4970).

Recent work on speech translation for low-resource languages has developed model architectures that condense the earlier cascaded systems into fewer steps and even bypass written representations of the speech altogether (see Table 1 for a summary). Bansal et al. (2018) train an encoder-decoder model to translate directly from Spanish speech to English text on just 160 hours of data, achieving a BLEU score of 29.4, indicating that the model achieves understandable translations (Lavie 2010: 41). Zhang et al. (2021) introduce the UWSpeech system for translating source speech into target speech in unwritten languages. Zhang et al.’s system works by converting source speech into discrete tokens representing the target speech, then synthesizes target speech from the tokens. However, Zhang et al.’s system still requires transcribed speech in written languages other than the source/target languages for training. Lee et al. (2022) introduce methods for translating from unwritten languages to unwritten languages, eliminating reliance on transcriptions. Lee et al.’s system adopts Zhang et al.’s discretization of speech into tokens, as well as utilizing transfer learning.

Model (Citation)	Source	Target	Training data sources (hrs)	Architecture	Evaluation
Meta English-to-Taiwanese S2ST System (Chen et al. 2023)	English	TSM	Human-annotated (35), pseudo-labeled (1.5k)	Single-stage speech-to-unit system with two-pass decoding (incorporates text prediction auxiliary task)	7.8 ASR-BLEU on TAT-S2ST test set
Meta Taiwanese-to-English S2ST System (Chen et al. 2023)	TSM	English	Human-annotated (61.4), pseudo-labeled (8k), mined (8.1k)	Single-stage speech-to-unit system with two-pass decoding (incorporates text prediction auxiliary task)	10.8 ASR-BLEU on TAT-S2ST test set
Meta Spanish-to-English S2ST System (Lee et al. 2022)	Spanish	English	Human-annotated Spanish (162.5), Pseudo-labeled English (139.3), both from Fisher Spanish-English dataset (Post et al. 2013)	Single-stage speech-to-unit system with source and target text prediction auxiliary tasks	39.9 ASR-BLEU on Fisher Spanish-English test set
Meta Spanish-to-English S2ST System (Lee et al. 2022)	Spanish	English	Human Spanish speech (162.5), synthesized English speech (139.3), both from Fisher Spanish-English dataset (speech only, no transcripts used)	Single-stage speech-to-unit system with source unit prediction auxiliary task	31.8 ASR-BLEU on Fisher Spanish-English test set
Speech2S (Wei et al. 2023)	Spanish	English	Parallel human Spanish and English speech from VoxPopuli-S2S, without transcriptions (513)	Stacked speech encoder, unit encoder, and unit decoder, jointly pretrained on speech/text from 23 languages	23.3 BLEU on VoxPopuli test set, 24.4 BLEU on Europarl-ST test set

Table 1: Speech-to-speech translation models

Chen et al. (2023) incorporate the existing strategies of transfer learning and data augmentation developed for low-resource text-to-text translation, while simultaneously drawing on the recent advances in speech-to-speech translation architectures to develop the first-ever dataset and models for speech-to-speech translation between English and Taiwanese Southern Min. Chen et al. also release the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark Dataset, which I use to compare the performance of my own models against theirs.

While Chen et al. (2023) make their benchmark dataset publicly available, this dataset can only be used as a standardized measure of the performance of English-Taiwanese speech-to-speech translation systems, not as a source of training data for developing such systems. As shown in Table 2, there are few free, publicly available Taiwanese speech corpora available.

Access to the Taiwanese Across Taiwan Corpus, for example, costs thousands of dollars, placing this data out of reach for those without commensurate funding (The Association for Computational Linguistics and Chinese Language Processing n.d.).

Corpus	Transcription	Size (Hours)	Speakers	Source(s)	Intended use	Availability	Citation
Taiwanese Across Taiwan (TAT) Volume 1	Hàn-Lô, Pêh-ōe-jī, Tâi-lô	51.94	100	Read speech from native speakers	ASR	Available for (non-commercial) researchers for a fee	(Liao et al. 2022)
Taiwanese Across Taiwan (TAT) Volume 2	Hàn-Lô, Pêh-ōe-jī, Tâi-lô	52.72	100	Read speech from native speakers	ASR	Available for (non-commercial) researchers for a fee	(Liao et al. 2022)
Taiwanese Across Taiwan (TAT) MOE	Hàn-Lô, Pêh-ōe-jī, Tâi-lô	312.87	440	Read speech from native speakers	ASR	Not released	(Liao et al. 2022)
Taiwanese Across Taiwan (TAT) TTS	Hàn-Lô, Pêh-ōe-jī, Tâi-lô	40.6	4	Read speech from native speakers	TTS	Available for (non-commercial) researchers for a fee	(Liao et al. 2022)
Formosa Speech Database (ForS-Dat) Version 1.0 (Taiwanese subset)	Formosa Phonetic Alphabet	83.64	600 ¹	Read speech from native speakers	ASR	Not released	(Lyu, Liang & Chiang 2004)
Taiwanese-English Dictionary	Daī-ghî tōng-iōng, Pêh-ōe-jī, Tâi-lô	Unknown	Unknown	Maryknoll Language Service Center Taiwanese-English Dictionary, Ministry of Education Taiwanese Southern Min Dictionary	Dictionary	Free online	(<i>Taiwanese-English Dictionary</i> n.d.)
Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark	English, Hân-Lô, Pêh-ōe-jī, Tâi-lô	5.97 ²	Unknown ³	Read speech from native speakers (TAT-Voll-eval-lavalier and TAT-Voll-test-lavalier from TAT Volume 1)	Speech translation research	Free online	(Chen et al. 2023)
Taiwanese Speech Recognition Corpus	Unknown	43.78	54	Unknown	Unknown	Available for a fee	(European Language Grid 2020)
Taiwanese dramas	Chinese characters, English	23,000	Unknown	Taiwanese-language dramas	Speech-to-speech translation	Not released	(Chen et al. 2023)

Corpus	Transcription	Size (Hours)	Speakers	Source(s)	Intended use	Availability	Citation
Digital Archive Database for Written Taiwanese (2nd stage)	Hàn-Lô, Pêh-ōe-jī	Unknown	Not applicable ⁴	Poetry, prose, fiction, and play scripts from 1885 and later	Literary studies, computational linguistics research, historical preservation	Free online	(Chang & Iunn 2021, Yang et al. 2006: 308)
Shapes Game Database	Chinese characters, Pêh-ōe-jī	Unknown	6	Spontaneous task-based dialogues and read speech from native speakers	Studying spontaneous speech prosody	Not released	(Peng & Beckman 2003)
Taiwanese Southern Min Spontaneous Speech Corpus	Hàn-Lô	8	16	Monologues elicited from native speakers	Studying discourse prosody	Not released	(Wang & Fon 2013)
Taiwanese Southern Min Corpus D	English, Pêh-ōe-jī ⁹	0.63	7	Spontaneous speech elicited from native speakers during interviews	Non-commercial research	Free online	(Sun & Newman 2010)
Taiwanese Spoken Corpus	Chinese characters	Unknown ⁶	Unknown	Private conversations, public speeches, television dramas, broadcast news, folktales, call-in shows, interviews	Corpus-based register study	Not released	(Jang 1998)
SuíSiann Dataset	Chinese characters, Tâi-lô	4.73	1	Read speech	Speech synthesis research	CC BY-SA license	(<i>SuíSiann Dataset</i> 2021)
PhonBank Taiwanese Tsay Corpus	Chinese characters, Tâi-lô	330	14 children, plus experimenters and family members	Spontaneous conversation between children and caretakers	Creative Commons CC BY-NC-SA 3.0 copyright license	Free online	(Tsay 2007, 2014)
Common Voice Corpus (TSM subset: train, validation, and test splits)	Chinese characters, Tâi-lô, Pêh-ōe-jī	5.45	Unknown	Crowd-sourcing	Creative Commons CC0 license	Free online	(Ardila et al. 2020)

Table 2: Taiwanese speech corpora

As seen in Table 3, text-only Taiwanese corpora are somewhat more plentiful. In my own work, I use such corpora for developing the module of the translation system responsible for translating Taiwanese text into English text.

Corpus	Transcription	Size	Source(s)	Intended use	Availability	Citation
Taiwanese Chinese Characters	Taiwanese Chinese characters	30 songs	Song lyrics	Promote standardized Taiwanese characters	Free online	(Sih 2015a)
Taiwanese National Elementary School Textbooks	Chinese characters, Tâi-lô	349 articles	Elementary school textbooks	Unknown	Free online	(Sih 2016)
iCorpus	Chinese characters, Pêh-ōe-jī	83,544 sentences	News articles	Unknown	Free online	(Sih 2018, 2015b)
National Language Contest Southern Min Reading Competition	Chinese characters, Tâi-lô	550 articles ⁷	Southern Min Reading Competition	Language competition	Free online (partially archived by Wayback Machine)	(中華民國110年全國語文競賽 2021)
Ministry of Education Taiwanese Southern Min Dictionary	Bopomofo, Chinese characters, Tâi-lô	14,985 sentences	Example sentences from dictionary	Demonstrate usage of commonly used words	Free online	(Tang 2022, Chan 2024)
Subtitles from Public Television Service	Chinese characters	126,578 sentences	Taiwanese TV programs	Constructing a code-switched Taiwanese-Mandarin dataset	Not released	(Lu et al. 2022)
Min and Hakka Language Archives	Chinese characters, Hân-Lô, Pêh-ōe-jī	Unknown	Dramas and other literary texts, folk songs, dictionaries, medical books, news articles	Studying historical linguistics, language variation, language contact; language teaching	Free online	(Academia Sinica n.d.)
Taiwanese Written Corpus	Chinese characters	68,080 words	Editorials, popular/personal prose, academic and persuasive essays, letters, fiction	Corpus-based register study	Not released	(Jang 1998)
Bible translations	Tâi-lô	29,814 verses	National Taiwanese Bible and World English Bible	Machine learning	Free online	(Teng 2024)
Wikipedia translations	Chinese characters, Pêh-ōe-jī, Tâi-lô	536 sentences	Wikipedia article translations	Machine translation	Free online	(<i>wikipedia</i> 2023, Tiedemann 2012)

Table 3: Taiwanese text corpora

4 Methods

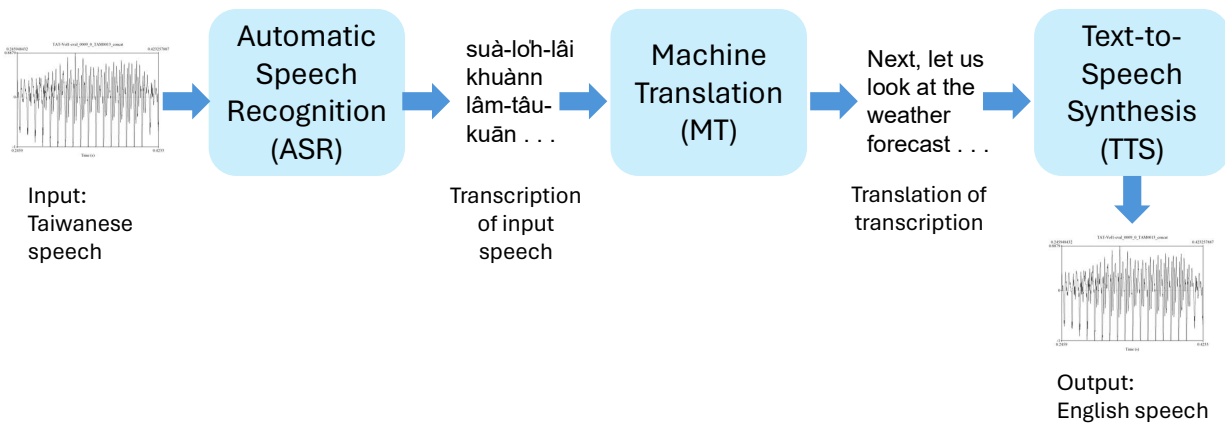
In order to develop my speech-to-speech translation system, I join together multiple modules (§4.1). I adapt existing automatic speech recognition and machine translation models (§§4.3 and 4.4) using data compiled from multiple free and publicly available sources (§4.2). I then connect the automatic speech recognition and machine translation models to an open-source text-to-speech synthesis system (§4.5). I conclude with a comparison and justification of my design choices with those of Chen et al. (2023) (§4.7).

4.1 Model architecture

The architecture of the system for translating spoken Taiwanese into spoken English is shown in Figure 1. The system takes recorded Taiwanese speech as input and outputs the corresponding spoken English translation. First, the automatic speech recognition (ASR) module of the system outputs a transcription of the Taiwanese speech. Next, the machine translation (MT) module takes the transcription generated by the ASR module as input and outputs a written English translation. Finally, the text-to-speech synthesis (TTS) module takes the written English translation as input and outputs an audio file containing the spoken English translation. I choose to structure my translation system as a “cascaded system,” breaking down the translation task into multiple simpler tasks, because I anticipate that alternative single-stage approaches to speech translation (mapping directly from input Taiwanese speech to output English speech) would require much more training data than I have access to (see Table 1 for examples of single-stage translation approaches and the amount of data used by each). In the following sections, I discuss how the ASR, MT, and

TTS modules are developed in more detail.

Figure 1: The architecture of the system for translating spoken Taiwanese into spoken English.



4.2 Datasets and preprocessing

The sources of data I use to train my translation system can be divided into transcribed recordings of spoken TSM, used to develop the ASR module for transcribing TSM (§4.2.1); and written TSM sentences paired with their written English translations, used to develop the MT module for translating TSM to English (§4.2.2).

4.2.1 TSM speech corpora for ASR

To train the module for performing automatic speech recognition, which consists of producing the correct transcription for a given segment of speech, I use several corpora listed in Table 2. All corpora used for developing the ASR model were selected based on their public availability and inclusion of romanized transcriptions of TSM speech. The Taiwanese Across Taiwan Corpus of read TSM speech from native speakers (see Table 2), though available in full only for a fee, is freely and publicly available in smaller samples that include transcriptions into Chinese characters and multiple romanization systems for the corresponding audio files (Liao 2022a,b). The SuiSiann dataset also features read TSM speech (transcribed in Chinese characters and Tâi-lô), from a single speaker. The Common Voice Corpus features read speech from many speakers, crowdsourced and validated through Mozilla’s Common Voice initiative. The Taiwanese Southern Min Corpus D includes English translations of Taiwanese speech (transcribed in Pêh-ōe-jī) from conversations with multiple native speakers. As shown in Figure 2, which illustrates the sources of the utterances included in the ASR dataset, the majority of the data for training the ASR model (69.4% of utterances) comes from the Common Voice Dataset. In total, the ASR dataset consists of 10,949 spoken utterances with a total duration of 9.57 hours.

Since TAT, SuiSiann, Common Voice, and TSM Corpus D use a variety of orthographies for transcription, preprocessing is necessary to create a unified ASR dataset, consisting of spoken TSM utterances and their corresponding transcriptions in a single orthography. I choose to convert transcriptions from TSM Corpus D to Tâi-lô. The TAT, SuiSiann, and Common Voice corpora already feature Tâi-lô transcriptions, so no conversion of orthography

Sources of utterances in ASR dataset

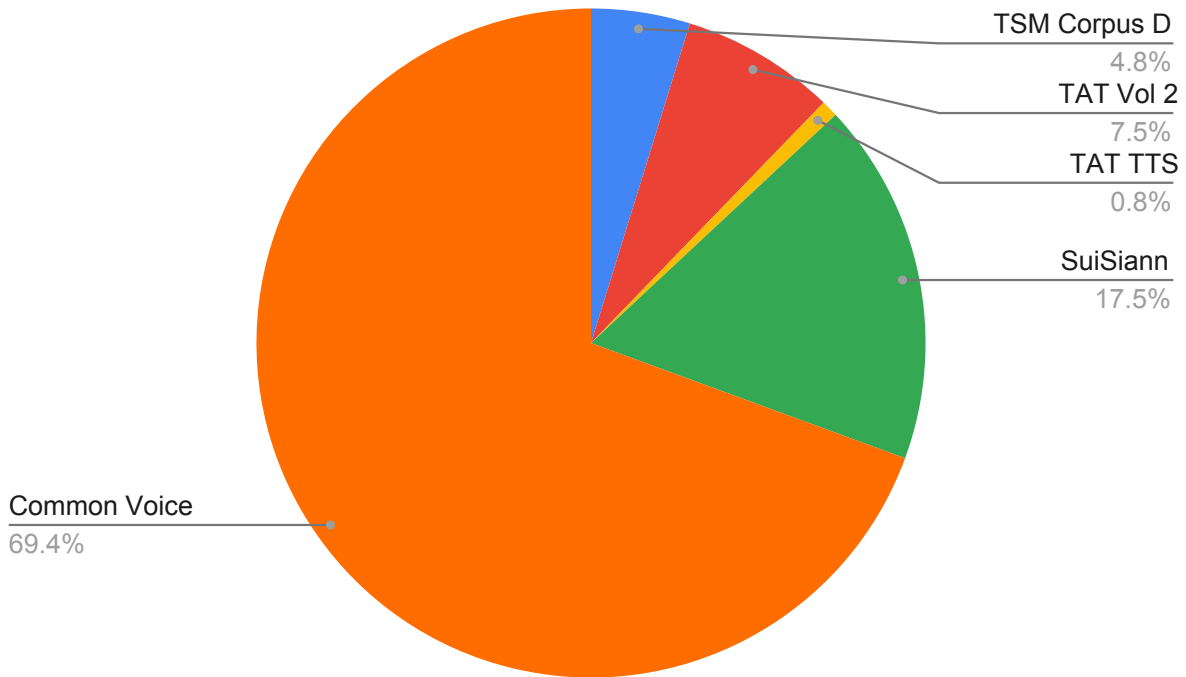


Figure 2: Sources of utterances in the dataset used to develop the automatic speech recognition module within the larger translation system.

is necessary for the transcriptions in these corpora. I choose Tâi-lô because it is the Taiwanese romanization system promoted as the standard by Taiwan’s Ministry of Education and also employed by the few TSM language-learning resources available for English speakers (e.g. the Bite-size Taiwanese podcast) (Ministry of Education 2023, Phil Lin n.d.). Additionally, Tâi-lô is a relatively shallow orthography, reflecting features such as lexical tone, voicing, and aspiration which are distinctive in TSM, so by learning to transcribe such features, the ASR module may learn to attend to these features in the input speech (Ministry of Education 2023).

After converting all transcriptions to Tâi-lô, I perform additional preprocessing to remove

special characters (e.g. punctuation) and convert all letters to lowercase. Special characters are removed because they do not correspond to pronounceable segments, while uppercase letters are converted to lowercase because capitalization does not correspond to any difference in pronunciation (von Platen 2021). I also filter out all examples that include Chinese characters, Bopomofo, or the letter “r” in their transcriptions, as these are not part of the standard Tâi-lô orthography. Filtering ensures that the ASR model will not learn to produce these non-standard characters in its transcriptions. Since the ASR module expects speech to be recorded as audio files with a sampling rate of 16000 Hz, I also resample the audio from TAT, SuiSiann, Common Voice, and TSM Corpus D to match this sampling rate (von Platen 2021).

In addition to compiling and preprocessing data to train the ASR model, it is also necessary to develop a validation set which can be used to predict how well the model will perform at transcribing TSM speech it has not seen before. The model does not learn from the examples in the validation set. Instead, the validation set is used only to test the model’s ASR capabilities throughout training, and thus monitor its progress in learning. Thus, the model’s performance on the validation set gives a decent estimate of how it will perform at transcribing other speech it has not seen before (Berrar 2019). For my validation set, I use the validation subset of Meta’s TAT Speech-to-Speech Translation Benchmark (see Table 2) (Chen et al. 2023). I preprocess the validation set speech and text in the same manner as the training set data.

4.2.2 Corpora for translating TSM to English

To develop the machine translation module (see §4.1 and §4.4) of the overall translation system, which takes in TSM text as input and outputs the corresponding written English translation, I use several corpora listed in Tables 2 and 3: Wikipedia translations, Taiwanese Across Taiwan Volume 2, Taiwanese Across Taiwan TTS, Common Voice Corpus, Taiwanese Southern Min Corpus D, SuiSiann Dataset, Ministry of Education Taiwanese Southern Min Dictionary, and Bible translations. By compiling data from these sources, I create a dataset of 57,618 translation examples for the MT module to learn from. The breakdown of the dataset by data source is illustrated in Figure 3. I included all of the datasets which I am already using to develop the ASR module of the translation system (see §2). To select additional data sources, I considered criteria including ease-of-use. For example, the Bible translation dataset and Ministry of Education dictionary dataset are both available through Hugging Face, an organization dedicated to democratizing machine learning. As a result, these two datasets are in an easy-to-work-with format downloadable directly from Hugging Face’s website (Chan 2024, Teng 2024).

The Wikipedia article translations, Bible translations, and TSM Corpus D translations are the only sources of data which include human-written TSM text matched with human-written English text translations. The Bible dataset consists of the same Bible verses in both TSM and English, and makes up a majority of the translation dataset (51.7%), as shown in Figure 3 (Teng 2024). TSM Corpus D includes transcriptions of TSM speech and their English free translations, both recorded by the linguists who created the corpus (Sun & Newman 2010). Wikipedia article translations are sourced from Wikipedia articles, which

are presumably also written by humans (*wikimedia* 2023, Tiedemann 2012). These human-written translation examples form the high-quality portion of the dataset for developing the MT module.

Sources of example translations

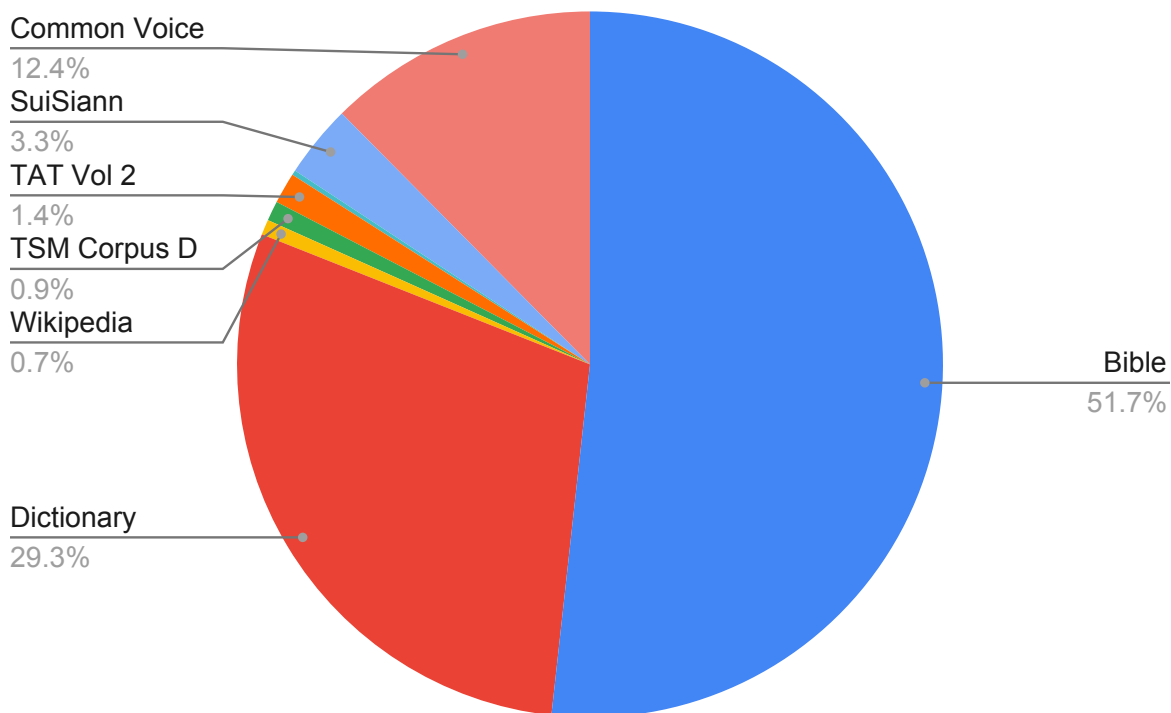


Figure 3: Sources of TSM texts and their English translations, used to develop the machine translation module of the translation system. The TAT TTS corpus is not visible on the chart but makes up 0.2% of the total training data for the MT module.

The Taiwanese Across Taiwan corpora, Common Voice Corpus, SuiSiann Dataset, and Ministry of Education dictionary dataset all lack English translations corresponding to the written examples of TSM they contain. However, since these datasets all include versions of their data written in Chinese characters, I am able to use existing models for translating Mandarin Chinese to English to create English translations for the TSM text in these

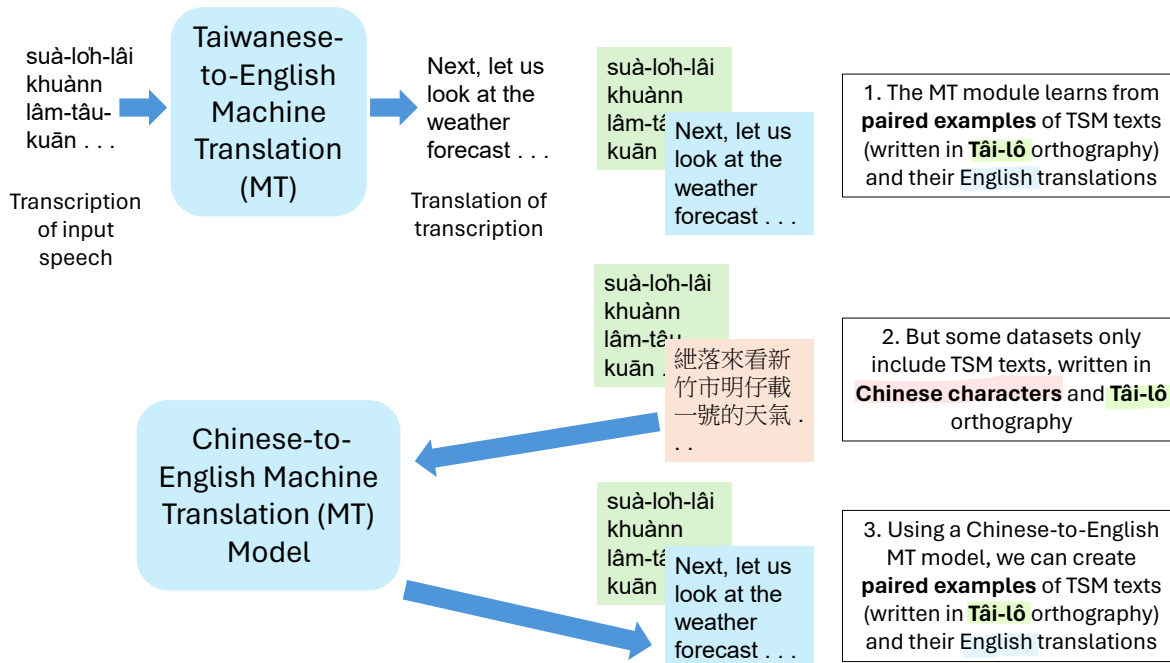


Figure 4: Process for creating synthetic data for training the MT module.

datasets. I illustrate this process of creating synthetic data in Figure 4. For Chinese-to-English translation, I use the model developed by the Language Technology Research Group at the University of Helsinki, as this model has the advantage of being freely available to use via the familiar and widely used API provided by Hugging Face machine learning organization (Tiedemann 2012, Language Technology Research Group at the University of Helsinki 2023b). Additionally, the model is relatively small, making it usable given my limited computational resources, and it achieves a BLEU score of 36.1 on Chinese-to-English translation, indicating that its translation quality ranges from “understandable to good” (Google 2024). I do not expect the English translations generated by the University of Helsinki machine translation model to be of as high quality as the human-translated examples. However, based on prior work which has found such synthetic data beneficial to developing machine

translation models, I still expect that including the synthetic data will positively benefit model training (Ranathunga et al. 2023).

In addition to the training data for teaching the MT module to perform translation, separate validation data is also required to monitor the MT module’s progress in learning. Similar to the procedure for constructing the ASR validation set (see §4.2.1), I use the 722 examples from the TAT Speech-to-Speech Translation Benchmark validation set as my own machine translation validation dataset.

4.3 Creating the ASR module

The ASR module of the translation system is responsible for taking Taiwanese speech as input, and outputting the transcription of the speech in the Tâi-lô orthography (see §4.1). To create the ASR module, I adapt Meta’s Wav2Vec2-XLS-R-300M model using a method called *fine-tuning*. In *fine-tuning*, I teach Wav2Vec2-XLS-R-300M (which has already previously been shown examples of speech from multiple languages) to transcribe TSM speech by showing it examples of TSM speech paired with their transcriptions. I choose Wav2Vec2-XLS-R-300M because its creators previously demonstrated its successful adaptation for low-resource ASR in other languages (Babu, Wang, Tjandra, Lakhota, Xu, Goyal, Singh, von Platen, et al. 2022a). I use the implementation of Wav2Vec2-XLS-R-300M and the model-training API available from Hugging Face for fine-tuning (Babu, Wang, Tjandra, Lakhota, Xu, Goyal, Singh, von Platen, et al. 2022b).

In order to set the hyperparameters which control how the model learns to perform the ASR task, I follow the guidance provided in a tutorial authored by Patrick von Platen, one of the model’s co-creators (von Platen 2021). To reduce the amount of computer memory

required for training (given my limited computational resources), I set the per-device training batch size to 8, meaning that the model sees and learns from groups of 8 examples at a time. For the same reason, I set gradient accumulation steps to 2, which allows me to compensate for the relatively small batch size. I use fp16 16-bit (mixed) precision training, which speeds up model training (Corporation 2023, von Platen 2021). To speed up training further by grouping together examples of similar length, I set the `group_by_length` parameter to “True” (von Platen 2021). Following von Platen’s example, I use a learning rate of $3e - 4$ with 500 warmup steps; this controls how quickly the model learns from the data. I fine-tune for 30 epochs, meaning that the model sees every example in the dataset 30 times. I save a checkpoint or “snapshot” of the model at the end of each epoch. (The number of epochs is again chosen according to von Platen’s example.) I select the model checkpoint with the lowest word error rate on the validation set as my final ASR model (see 4.2.1 for details on how the validation set was constructed). Training the model takes a total of 10.4 hours using a single Tesla T4 GPU to perform computations. I use the T4 GPU for training because it is available for free through the Google Colab service (*Colaboratory* n.d.).

I evaluate my model using the word error rate because this is the standard evaluation metric for ASR systems. The word error rate is calculated as (Jurafsky & Martin 2024: 352-354):

$$\text{Word Error Rate} = \frac{\text{Insertions} + \text{Substitutions} + \text{Deletions}}{\text{Total Words in Correct Transcription}}$$

In other words, the word error rate reflects the number of words which need to be inserted, substituted for another word, or deleted from the transcription generated by the ASR model to produce the correct transcription of some speech. An ASR model which outputs per-

fect transcriptions would have a WER of 0, while an ASR model which outputs errorfull transcriptions could have a WER over 1.

4.4 Creating the MT module

The MT module of the translation system is responsible for taking the transcriptions of TSM speech generated by the ASR module as input and outputting the corresponding English translations (see §4.1). To create the MT module of my translation system, I adapt the opus-mt-mul-en model developed by the Language Technology Research Group at University of Helsinki (Tiedemann 2020, Language Technology Research Group at the University of Helsinki 2023a). The opus-mt-mul-en model has been previously trained to translate text from any one of 310 languages into English. Similar to the rationale for developing the ASR module using a multilingual model, I choose to fine-tune the multilingual opus-mt-mul-en model to translate from TSM to English under the assumption that its previous training to translate from other languages into English will aid it in learning the new task (Ranathunga et al. 2023). In addition to the benefit of being a multilingual model, opus-mt-mul-en is also freely available to use via the Hugging Face platform for machine learning, and is a relatively small model, making it an attractive choice given my limited financial and computational resources, which cannot handle training a very large model. Similar to the ASR module, the hardware used to perform the computations for training the MT module is a single Tesla T4 GPU available for free via Google Colab.

To fine-tune opus-mt-mul-en to translate Taiwanese text into English text, I train it for 3 epochs (which takes 46 minutes) on the translation data described in §4.2.2. That is, the model is shown each of the 57,618 examples of Taiwanese texts paired with their English

translations a total of 3 times. The model sees batches of 8 examples at a time and updates its internal state accordingly based on its mistakes throughout training.⁸ At the end of each *epoch* or iteration through the training data, I check the model’s progress in learning the translation task by computing its BLEU score on the validation set (see §4.2.2 for details on the validation set) and save the version of the model at that point in time. At the end of the 3 epochs, I choose the model version with the highest BLEU score on the validation set as the final model to use in the MT module of the overall translation system. I select the best model using the BLEU score because it is the long-standing standard metric used in the MT research community to evaluate translation quality, and its validity for this purpose has been verified (Reiter 2018). BLEU works by comparing a machine-generated translation with a “correct” reference translation written by a human translator. A translation which is more similar to the reference translation will receive a higher score (Papineni et al. 2002).

4.5 Creating the TTS module

The TTS module is responsible for taking the English text produced by the MT module as input, and outputting the spoken form of that text (in the form of an audio file) as output. As shown in Figure 5, the input text is first converted to phonemes, which are then fed as input to a model called Tacotron2. Using the input sequence of phonemes, Tacotron2 predicts the spectrogram corresponding to the spoken form of that input phoneme sequence (*Tacotron 1 and 2* 2021). The spectrogram generated by Tacotron2 is then used as input to a vocoder called UnivNet, which outputs an audio file containing the waveform

⁸To those in need of additional details about parameter settings to reproduce these results, it should be noted as well that I use the AdamW optimizer with a learning rate of $2e - 5$ and a linear learning rate scheduler with 0 warmup steps.

of the speech synthesized from the spectrogram as the final output. For the TTS module, I use the Tacotron2 implementation (with Double Decoder Consistency, with phonemes) available via the Coqui TTS library (Eren & The Coqui TTS Team 2021). The Coqui TTS API automatically handles grapheme-to-phoneme conversion for this model as well. As my vocoder, I use the UnivNet implementation available from Coqui TTS, which has been adapted by the Coqui TTS team to perform well at synthesizing speech from spectrograms output by the Tacotron implementation I am using (Eren & The Coqui TTS Team 2021).

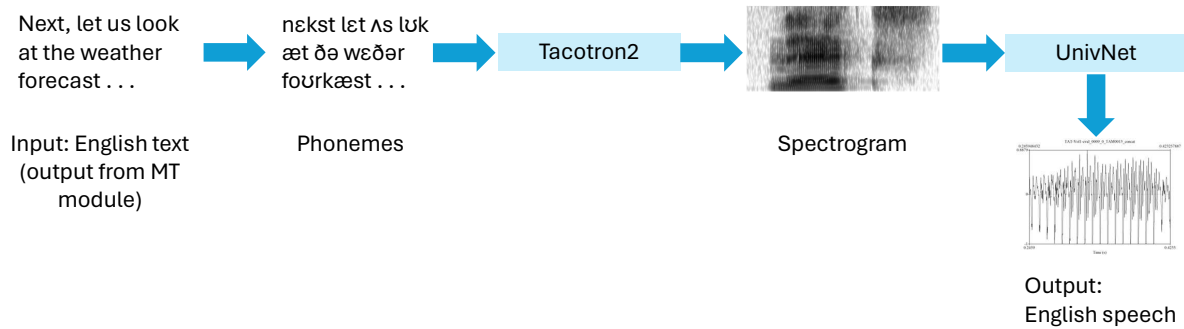


Figure 5: The TTS module of the translation system takes an English translation generated by the MT module as input and outputs the spoken form of the translation as an audio file.

Unlike the ASR and MT modules, it is difficult to evaluate the TTS module using a single

automated metric, as there is no single correct way to speak a given English sentence (Wagner et al. 2019). However, the quality of the TTS module certainly affects the overall evaluation of the translation system, which I discuss further in §4.6. To verify that the TTS module is producing clear, intelligible speech, I listen to a small random sample of speech produced by the TTS module (50 examples from the test set; see §4.6 for more details on the test set) and draw on my judgments as a native English speaker to determine if the synthesized speech clearly matches the text translations input to the TTS module. In particular, for each of the 50 TTS examples, I answer the following question with “yes” or “no”: “Does the synthesized speech contain all words and *only* those words which were included in the input text to the TTS module, in the same order as they were written in the input text?”

4.6 Evaluating the Translation System

To evaluate how well the entire translation system (ASR module, MT module, and TTS module) performs at the task of translating spoken Taiwanese into spoken English, I use the test set subset of the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark. The Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark test set consists of 686 examples of spoken Taiwanese paired with their spoken English translations, created by human translators (Chen et al. 2023). I use the 686 examples of spoken Taiwanese as input to the translation system (see §4.1), which outputs audio files containing spoken English translations of the input speech. To compare the quality of the translations generated by the translation system with the correct, reference translations provided in the benchmark test set, I follow the procedure described by Chen et al. Chen et al. provide an ASR model which can be used to transcribe the speech translations generated by the translation system.

Then, these transcriptions of the speech translation system’s speech can be compared to the human transcriptions of the correct speech translations provided in the test set, using the BLEU metric (see §4.4 for an explanation of BLEU). In order to allow comparison between my own model and that of Chen et al., it is important to use the same ASR model as they do for the evaluation. Therefore, I do so using the evaluation script available from Chen et al. at https://github.com/facebookresearch/fairseq/tree/ust/examples/speech_to_speech/asr_bleu. The evaluation process is summarized in Figure 6.

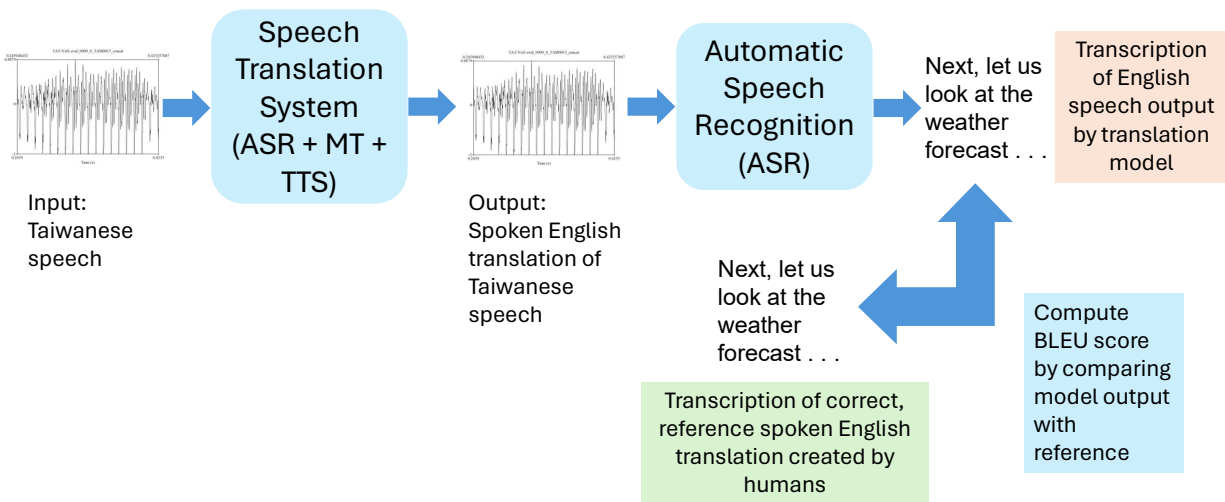


Figure 6: Process for evaluating the entire speech translation system end-to-end (ASR, MT, and TTS modules).

4.7 Comparison with Chen et al. (2023)

Chen et al. (2023) three-stage cascaded system	This work
ASR module transcribes TSM speech in Chinese characters	ASR module transcribes TSM speech in Tâi-lô
ASR module consists of Wav2vec 2.0 encoder trained on 30,000 hours of TSM speech data (not publicly accessible), concatenated with a text mBART decoder	ASR module consists of Wav2Vec2-XLS-R-300M model which has prior knowledge of multiple languages and is fine-tuned for TSM ASR on 9.57 hours of free, publicly accessible data
MT module translates TSM text written in Chinese characters into English text	MT module translates TSM text written in Tâi-lô characters into English text
MT module is a 12-layer Transformer module trained from scratch on paired English and Chinese translation examples the CCMatrix dataset	MT module is a version of the opus-mt-mul-en model fine-tuned on paired English and TSM translation examples from multiple public datasets (see §4.2.2), taking advantage of the opus-mt-mul-en model’s prior knowledge of translation
TTS module works by mapping input text to output speech “units”	TTS module works by mapping input text to output spectrograms

Table 4: Comparison of the translation system presented in this work with the three-stage translation system introduced in Chen et al. (2023).

In Table 4, I highlight important differences between the multi-stage translation system introduced in this work and that of Chen et al. (2023). To the best of my knowledge, this study and that of Chen et al. are the only two extant studies on automatically translating spoken Taiwanese into spoken English. While both studies introduce translation systems that follow the same three-stage architecture (ASR module, MT module, and TTS module), they differ in the component models and data sources used to build each module. Importantly, whereas Chen et al. rely on much data that is not freely or publicly accessible, the translation system introduced in this work is developed solely using freely and publicly accessible data sources, as the goal of this study is to investigate whether such a system can succeed.

Furthermore, my work makes greater use of transfer learning to transfer the knowledge already contained in multilingual models to the tasks of ASR and MT for TSM.

Another important difference between this study and Chen et al. (2023) is the use of the Tâi-lô orthography, rather than Chinese characters, to represent TSM speech in the ASR and MT modules. I choose Tâi-lô instead of Chinese characters because of the myriad problems with writing TSM in Chinese characters. Many TSM morphemes do not have an agreed-upon Chinese character by which they can be represented (A. Lin 1999: 7–9). Furthermore, as discussed previously in §4.2.1, the Tâi-lô romanization system is the standard promoted by the Taiwanese Ministry of Education and adopted by American TSM language learners, and Tâi-lô reflects distinctive features such as lexical tone, voicing, and aspiration. Therefore, building a system that works with Tâi-lô (rather than Chinese character) transcriptions makes sense in terms of working with the most standardized orthography possible, using the orthography employed by language educators/learners themselves, and using an orthography that conveys useful information about the sounds of the language.

5 Results

Using the evaluation procedure described in §4.6, the translation system as a whole achieves a BLEU score of 0.24 on the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark test set. For comparison, Chen et al. achieve a BLEU score of 10.0 using their own three-stage cascaded translation system, which consists of an ASR module, MT module, and speech synthesis module, similar to the system presented in this work. The low BLEU score of my translation system indicates that it does not generate good translations. In §6, I discuss how problems within the individual modules contribute to the poor performance of the overall system.

The ASR module selected for use in the translation system achieves a word error rate (WER) of 0.666 on the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark validation set. While the WER appears to be high, in §4.3 I argue (using qualitative analysis of transcriptions generated by the ASR module) that the ASR module actually performs much better than the high WER would initially suggest. The ASR module performs well enough that I release it to the public at <https://huggingface.co/emlinking/wav2vec2-large-xls-r-300m-tsm-asr-v6>. Anyone can access the model through an easy-to-use web interface, as shown in Figure 7. By clicking on the “Browse for file” or “Record from browser” buttons, users can easily input audio to the model, then press “Compute” to get the transcription of their audio.

The MT module selected for use in the translation system achieves a BLEU score of 2.55 on the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark validation set. The low BLEU score indicates that the MT module generates poor translations. I examine

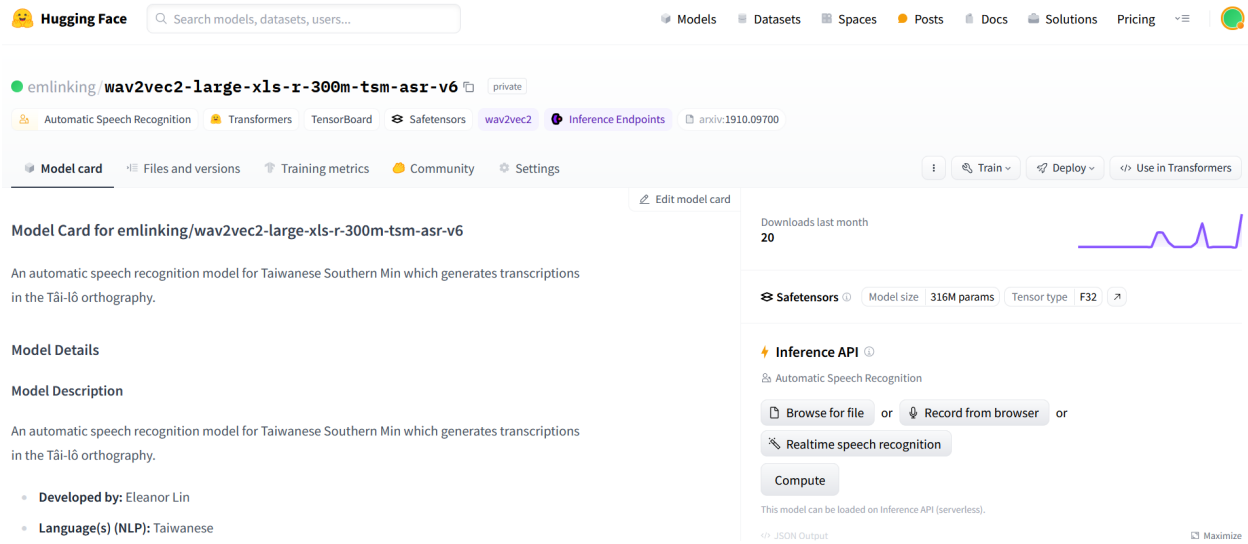


Figure 7: The webpage where members of the general public can easily access the ASR module.

examples of these translations in more detail in §6.2, as well as suggest reasons for the poor performance of the MT module. The MT module seems to be the greatest contributing factor to the overall poor performance of the speech translation system.

Of the 50 examples of speech synthesized by the TTS module selected for manual quality checking as described in §4.5, the majority (31/50 examples, or 62%) were synthesized correctly.⁹ As I argue in §6.3, cases where the TTS module did not synthesize speech correctly seem to be triggered by problems with the input text, which ultimately stem from the poor performance of the MT module that generates the text translations which are input to the TTS module.

⁹“Correctly” means that the answer to the question “Does the synthesized speech contain all words and only those words which were included in the input text to the TTS module, in the same order as they were written in the input text?” was “Yes” (see §4.5).

6 Discussion

While the ASR module does make some errors in transcribing TSM speech, its transcription quality is actually much better than initially suggested by its high WER of 0.666 (§6.1). The MT module, on the other hand, generates bad translations, which significantly harm the overall performance of the speech translation system (§6.2), including by triggering downstream errors in the TTS module (§6.3).

6.1 ASR module error analysis

By examining the transcriptions generated by the ASR model for several speech examples from the validation set, I find that several common types of errors are responsible for the high WER (see Figure 8). The single largest contributing factor to the high WER is probably the misplacement of hyphens (“-”), which mark syllable boundaries within words. In Figure 8, we see that sometimes the ASR model fails to transcribe a hyphen when there should be one, whereas in other cases, the ASR model transcribes a hyphen where there should not be one. In cases where the model adds an erroneous hyphen to connect two words which are correctly transcribed in all other respects, the two words will be counted as a single, incorrect word during computation of the WER. In cases where the model omits a hyphen needed to join two syllables of a single word, the two syllables will be counted as two incorrect words during computation of the WER. Hence hyphenation errors in the transcriptions generated by the ASR model probably contribute much to the high WER.

In addition to misplaced hyphens, as seen in Figure 8, the ASR model also makes occasional errors in transcribing nasals and nasalization (indicated by the letters *m*, *n*, *ng*, *nn*)

and tones on vowels (indicated by diacritics) (Ministry of Education 2023). In particular, the model sometimes transcribes a nasal with the wrong place of articulation (e.g. *ng* instead of *m* in the first example in Figure 8) or transcribes a nasalized vowel instead of a vowel followed by a nasal (e.g. *-ann* instead of *-an* in the first example in Figure 8). The former error is probably due to the fact that nasals share their manner of articulation, resulting in acoustic similarities which confuse the ASR model; similarly, the latter error is probably due to the acoustic similarity between a nasalized vowel and a sequence of vowel plus nasal stop. Errors in transcription of tone can be explained by the fact that TSM features extensive tone sandhi, by which the underlying tones on syllables are transformed into new surface forms (Cheng 1968). Since the underlying (and not the realized) tones are reflected in the Tâi-lô orthography, the ASR model is faced with the substantial challenge of learning which underlying tones to transcribe, given only their surface realizations. Considering the limited data available to train the model and the challenges of transcribing a tonal language, the ASR model does substantially better (as illustrated by the qualitative comparison in Figure 8) than its high WER initially suggests.

6.2 MT module error analysis

The MT module achieves a BLEU score of 2.55 on the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark validation set, indicating poor performance at the translation task. Figure 9 shows some examples of the kinds of translations the MT module produces, sourced from the validation and test subsets of the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark. While the model occasionally produces decent translations,

Figure 8: Examples of predicted and true transcriptions for several speech samples from the TSM ASR validation set. Errors are highlighted in red.

True: iōng kiân-ê ñg-tiám-guā-tsing kú ñ-tsai ē kàu-bē
Predicted: iōng-kiânⁿê ñ^gtiám^guah^htsing-kú ñ-tsai [█]kau-bē

True: hōo-tsuí líng-kí-kí ak tī a-khîm kah gín-á ê thâu-bin tsham sin-khu
Predicted: hōo-tsuí líng[█]kí-kí ak-tī a-khîm kai-gín-á ê thâu-bin tsham sing-khu

True: sián-mih teh bô tshîng-san siû-tsuí lóng mã tiòh tshîng iû-íng-san iû-íng-khò khì tī iû-íng-tī siû-ah
Predicted: siánⁿ-mih te[█] bô tshîng-san siu-tsuí lóng mã tiò[█] tshîng iû-íng[█]sa[█] iû^u-íng[█]khò[█] khì tī iû-íng-tī siû-ah

most of the time translations fail to convey key information, or even convey a meaning not intended in the source language (Taiwanese). For example, an utterance which should translate to “Next, let us look at the weather forecast in Sin-tik City for tomorrow on the first” is translated instead as the overly vague “Let’s see what’s going on in the morning,” omitting key information such as the date. In another example in Figure 9, “you are the master of tea tasting” is instead translated as “the tea’s changed your mind,” conveying a different meaning than the source text intended. Overall, we see that the MT module’s translations are usually not good enough to be useful.

The poor performance of the MT module is most likely due to several factors, including *domain mismatch*. *Domain mismatch* occurs when the data used to train a model come from


Better translations	Correct/Reference translation	Model-generated translation (BLEU score)
	Teacher Lâu, how many teachers are there in our school?	Teacher, please, how many teachers do we have at our school? (BLEU score: 26.58)
	The most important thing about swimming is to overcome the fear of water, and learning to hold your breath under water is the first step.	The main thing about drinking water is not to be afraid of water, but to drink it first. (BLEU score: 8.69)
	Oh! Lâu Hiân, you are the master of tea tasting.	Oh, God, the tea's changed your mind! (BLEU score: 4.40)
	Next, let us look at the weather forecast in Sin-tik City for tomorrow on the first.	Let's see what's going on in the morning. (BLEU score: 2.48)
	It was not until the Taiwanese language movement in the 1980s, which fought for the right to speak Taiwanese and awakened a mother tongue awareness, that the unresolved issue of the 'Taiwanese Language Debate' from the 1930s could be continued.	Until the end of the year, the Chinese language has been used to describe the language of Taiwan, and it has been adopted since the end of the language, and it has not been able to do so since the year "Taiwan Protests." (BLEU score: 2.42)
Worse translations		

Figure 9: Examples of translations of varying qualities generated by the MT module.

a different domain with different properties than the data the model is applied to (Shen et al. 2021). In my case, a large part of the training data for the MT module comes from Bible translations (51.7%, see §4.2.2), but the Taiwanese Across Taiwan Speech-to-Speech Translation Benchmark that the model is applied to and tested on consists of language spoken about topics from day-to-day life. As a result, the style and topics of much of the training data differ significantly from the style and topics of the testing data. Shen et al. have shown that this kind of mismatch can be particularly problematic for developing MT

models for low-resource languages, which is exactly the situation for my MT module.

Another reason for the MT module’s poor performance is insufficient quantity and quality of training data. The MT module was trained on just 57,618 examples, compared to the 38,000,000 examples used by Chen et al. to develop the MT module for their three-stage speech translation system. Additionally, of the 57,618 examples in my MT training data, 46.6% are synthetic data created by applying a Chinese-to-English translation model to translate Chinese character transcriptions of the Taiwanese texts into English (see §4.2.2 for more details). The English translations generated in this matter are lower quality than human translations, as seen in Figure 10. For example, some translations (such as the example in Figure 10 referencing “Egypt Protests 2011” and the example “I don’t know what I’m talking about”) are completely off-topic and do not accurately convey the original content that was written in TSM. Others are overly repetitive (e.g. the example repeating “every class” in Figure 10). The MT module may learn incorrect information from the errors in the synthetic data. However, while access to higher quality data during training would improve the MT module, including the synthetic data during training likely still improves the MT module’s performance compared to not including this data, as some synthetic examples are correct and contain useful knowledge that the the model can learn from (e.g. the **green** examples in Figure 10).

6.3 TTS module error analysis

Based on a random sample of 50 examples from the test set, the TTS module synthesizes speech correctly the majority of the time (62%). The cases where the TTS module does not synthesize speech correctly are almost entirely triggered by poor quality inputs from the MT

don't know what I'm talking about, but I don't know what I'm talking about, but I
don't know what I'm talking about, but I don't know what I'm talking about, but I
know what I'm talking about, what I'm talking about, what I'm talking about, what
I'm talking about, what I'm talking about, what I'm talking about, what I'm talking
about, what I'm talking about, what I'm talking about, what I'm talking about, what
I'm talking about, what I'm talking about, what I'm talking about, what I'm talking
about, what I'm talking about, what I'm talking about, what I'm talking about,
what I'm talking about, what I'm talking about, what I'm talking about, what I'

The TTS module has trouble synthesizing speech, given the highly repetitious text in (1) as input. In particular, the speech synthesized by the TTS module includes fewer repetitions of the phrase “what I'm talking about” than are included in the input text in (1). The solution here would be to improve the performance of the MT module, such that the TTS module would not have to deal with such highly repetitious and incorrect input translations in the first place.

7 Conclusion

In this work, I investigated whether it is possible to build a system for translating spoken Taiwanese into spoken English using only free and publicly available resources. Drawing on the Taiwanese Across Taiwan Corpus, SuiSiann dataset, Common Voice Corpus, Taiwanese Southern Min Corpus D, Wikipedia translations, dictionary translations, and Bible translations, I built a three-stage translation system consisting of ASR, MT, and TTS modules. I evaluated the system as a whole, as well as its individual modules, using standard metrics (word error rate for ASR and BLEU scores for MT) as well as manual analysis of translation results. I also analyzed the types of errors made by the ASR, MT, and TTS modules, and reasons for these errors.

Ultimately, I found that less than 10 hours of transcribed Taiwanese speech was sufficient to build a decent ASR module, for standalone use or use in a larger translation system. This demonstrates the efficacy of the transfer learning approach in which an existing multilingual model is adapted for ASR in a low-resource language. The ASR module is released for public use through a user-friendly webpage.¹⁰ The ASR module still has room for improvement, however, particularly with respect to transcribing hyphens in the correct positions. A potential solution to reduce the ASR module’s WER would be to have the ASR module only transcribe consonants, vowels (with tone-marking diacritics), and whitespaces. A separate module could then add hyphenation to the transcription produced by the ASR module. Sadeghian, Schaffer & Zahorian (2017) successfully used this approach of coupling an ASR model with a punctuation model in their own work in a low-resource setting.

¹⁰<https://huggingface.co/emlinking/wav2vec2-large-xls-r-300m-tsm-asr-v6>

While the development of the ASR module was relatively successful, on the other hand, I found that the 57,618 training examples used in this work were insufficient in quantity and quality to build a good MT module for the translation system. Future work should incorporate more of the free and publicly available data sources listed in Table 3 for training to develop a better MT module. To improve the quality of synthetic data used for training, future work may also explore other open-source models as alternatives to the particular Chinese-to-English translation model used in this work for creating synthetic data. Additionally, the data augmentation method of *self-training* may aid the development of a good MT module: (1) the MT model can be trained from available paired TSM and English translation data, then (2) applied to extant TSM data that lacks English translations, and (3) trained again on both the original data sources and the synthetic data it has itself generated (Shen et al. 2021: 1324). By leveraging more and higher quality synthetic data and the self-training method, it may still be possible in the future to develop a decent MT module for translating TSM to English using only free and publicly available data.

This work showcases both the challenges and promises of developing language technologies for low-resource languages. Employing transfer learning with multilingual models can mitigate limitations on data availability. Existing open-source models can be adapted for new languages. Such considerations are crucial for language communities who seek to shape the development of technologies for their own languages in low-resource scenarios.

References

- Academia Sinica. N.d. *Min and hakka language archives*. <https://minhakka.ling.sinica.edu.tw/bkg/>.
- Ardila, Rosana et al. 2020. Common voice: a massively-multilingual speech corpus. English. In Nicoletta Calzolari et al. (eds.), *Proceedings of the twelfth language resources and evaluation conference*, 4218–4222. Marseille, France: European Language Resources Association. <https://aclanthology.org/2020.lrec-1.520>.
- Babu, Arun, Changan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, et al. 2022a. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. interspeech 2022*, 2278–2282. DOI: [10.21437/Interspeech.2022-143](https://doi.org/10.21437/Interspeech.2022-143).
- Babu, Arun, Changan Wang, Andros Tjandra, Kushal Lakhota, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, et al. 2022b. *Wav2vec2-xls-r-300m*. Available at <https://huggingface.co/facebook/wav2vec2-xls-r-300m> (2024/03/26).
- Bansal, Sameer et al. 2018. Low-resource speech-to-text translation. English. In *Proceedings of interspeech 2018* (Proc. Interspeech 2018), 1298–1302. Interspeech 2018 ; Conference date: 02-09-2018 Through 06-09-2018. ISCA. DOI: [10.21437/Interspeech.2018-1326](https://doi.org/10.21437/Interspeech.2018-1326). <http://interspeech2018.org/>.
- Berrar, Daniel. 2019. Cross-validation. In Shoba Ranganathan et al. (eds.), *Encyclopedia of bioinformatics and computational biology*, 542–545. Oxford: Academic Press. DOI: <https://doi.org/10.1016/B978-0-12-809633-8.20349-X>. <https://www.sciencedirect.com/science/article/pii/B978012809633820349X>.

- Chan, Ching Yi. 2024. *Qrtt1/dictionary_of_frequently_used_taiwan_minnan*. https://huggingface.co/datasets/qrtt1/dictionary_of_frequently_used_taiwan_minnan.
- Chang, Miao-Hsia & Ún-gián Iúnn. 2021. A corpus-based study of directives in taiwanese southern min. *Concentric* 47(2). 300–336.
- Chen, Peng-Jen et al. 2023. Speech-to-speech translation for a real-world unwritten language. In Anna Rogers, Jordan Boyd-Graber & Naoaki Okazaki (eds.), *Findings of the association for computational linguistics: acl 2023*, 4969–4983. Toronto, Canada: Association for Computational Linguistics. DOI: 10.18653/v1/2023.findings-acl.307. <https://aclanthology.org/2023.findings-acl.307>.
- Cheng, Robert L. 1968. Tone sandhi in taiwanese. *Linguistics* 6(41). 19–42.
- China–Taiwan*. 2023. <https://www.ethnologue.com/country/TW/>. Accessed 2023-10-13.
- 中華民國110年全國語文競賽. 2021. <https://web.archive.org/web/20230125144320/https://language110.eduweb.tw/Module/Question/Index.php>.
- Colaboratory*. N.d. <https://research.google.com/colaboratory/faq.html>.
- Corporation, NVIDIA. 2023. *Train with mixed precision*. Available at <https://docs.nvidia.com/deeplearning/performance/mixed-precision-training/index.html> (2024/03/26).
- Eren, Gölge & The Coqui TTS Team. 2021. *Coqui TTS*. Version 1.4. DOI: 10.5281/zenodo.6334862. <https://github.com/coqui-ai/TTS>.
- European Language Grid. 2020. *Taiwanese speech recognition corpus (desktop)*. <https://live.european-language-grid.eu/catalogue/corpus/2147/download/>.

- Golonka, Ewa M. et al. 2014. Technologies for foreign language learning: a review of technology types and their effectiveness. *Computer Assisted Language Learning* 27(1). 70–105. DOI: 10.1080/09588221.2012.700315. <https://doi.org/10.1080/09588221.2012.700315>.
- Google. 2024. *Evaluating models*. <https://cloud.google.com/translate/automl/docs/evaluate#interpretation>.
- Jang, Shyue-Chian. 1998. *Dimensions of spoken and written taiwanese: a corpus-based register study*. English. Copyright - Database copyright ProQuest LLC; ProQuest does not claim copyright in the individual underlying works; Last updated - 2023-02-21. (Doctoral dissertation). 186. <http://ezproxy.cul.columbia.edu/login?url=https://www.proquest.com/dissertations-theses/dimensions-spoken-written-taiwanese-corpus-based/docview/304435800/se-2>.
- Jurafsky, Dan & James H. Martin. 2024. *Speech and language processing*. Stanford University.
- Language Technology Research Group at the University of Helsinki. 2023a. *Helsinki-nlp/opus-mt-mul-en*. <https://huggingface.co/Helsinki-NLP/opus-mt-mul-en>.
- Language Technology Research Group at the University of Helsinki. 2023b. *Helsinki-nlp/opus-mt-zh-en*. <https://huggingface.co/Helsinki-NLP/opus-mt-zh-en>.
- Lavie, Alon. 2010. Evaluating the output of machine translation systems. In *Proceedings of the 9th conference of the association for machine translation in the americas: tutorials*.
- Lee, Ann et al. 2022. Direct speech-to-speech translation with discrete units. In Smaranda Muresan, Preslav Nakov & Aline Villavicencio (eds.), *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 3327–

3339. Dublin, Ireland: Association for Computational Linguistics. DOI: 10.18653/v1/2022.acl-long.235. <https://aclanthology.org/2022.acl-long.235>.
- Liao, Yuan-Fu. 2022a. *Taiwanese across taiwan corpus*. <https://sites.google.com/speech.ntut.edu.tw/fsw/home/tat-corpus>.
- Liao, Yuan-Fu. 2022b. *Taiwanese text-to-speech corpus*. <https://sites.google.com/speech.ntut.edu.tw/fsw/home/tat-tts-corpus>.
- Liao, Yuan-Fu et al. 2022. Taiwanese across taiwan corpus and its applications. In *2022 25th conference of the oriental cocosda international committee for the co-ordination and standardisation of speech databases and assessment techniques (o-cocosda)*, 1–5. DOI: 10.1109/0-COCOSDA202257103.2022.9997977.
- Lin, Alvin. 1999. Writing taiwanese: the development of modern written taiwanese. *Sino-Platonic Papers* (89).
- Lin, Eleanor M. 2023. Toward a sociolinguistic profile of taiwanese americans. https://emlinking.github.io/files/Toward_a_Sociolinguistic_Profile_of_Taiwanese_Americans.pdf. Working paper.
- Lu, Sin-En et al. 2022. Exploring methods for building dialects-Mandarin code-mixing corpora: a case study in Taiwanese hokkien. In Yoav Goldberg, Zornitsa Kozareva & Yue Zhang (eds.), *Findings of the association for computational linguistics: emnlp 2022*, 6287–6305. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics. DOI: 10.18653/v1/2022.findings-emnlp.469. <https://aclanthology.org/2022.findings-emnlp.469>.
- Lyu, Ren-yuan, Min-siong Liang & Yuang-chin Chiang. 2004. Toward constructing a multilingual speech corpus for taiwanese (min-nan), hakka, and mandarin. In *International*

journal of computational linguistics & chinese language processing, volume 9, number 2, august 2004: special issue on new trends of speech and language processing, 1–12.

Ministry of Education, Republic of China (Taiwan). 2023. 臺灣閩南語羅馬字拼音方案. Available at https://language.moe.gov.tw/result.aspx?classify_sn=42&subclassify_sn=446&content_sn=7 (2024/03/26).

Papineni, Kishore et al. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics*, 311–318.

Peng, Shu-hui & Mary E Beckman. 2003. Annotation conventions and corpus design in the investigation of spontaneous speech prosody in taiwanese. In *Isca & iee workshop on spontaneous speech processing and recognition*.

Phil Lin, Alan Chen, Phín-tsi Kí. N.d. *Bite-size taiwanese*. <https://bitesizetaiwanese.com/>.

Post, Matt et al. 2013. Improved speech-to-text translation with the fisher and callhome spanish-english speech translation corpus. In *Proceedings of the 10th international workshop on spoken language translation: papers*.

Ranathunga, Surangika et al. 2023. Neural machine translation for low-resource languages: a survey. *ACM Computing Surveys* 55(11). 1–37.

Reiter, Ehud. 2018. A structured review of the validity of bleu. *Computational Linguistics* 44(3). 393–401.

Sadeghian, Roozbeh, J David Schaffer & Stephen A Zahorian. 2017. Speech processing approach for diagnosing dementia in an early stage. In *Interspeech 2017*, 2705–2709. DOI: 10.21437/Interspeech.2017-1712.

- Sandel, Todd L, Wen-Yu Chao & Chung-Hui Liang. 2006. Language shift and language accommodation across family generations in taiwan. *Journal of Multilingual and Multicultural Development* 27(2). 126–147.
- Scharenborg, Odette et al. 2020. Speech technology for unwritten languages. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28. 964–975.
- Shen, Jiajun et al. 2021. The source-target domain mismatch problem in machine translation. In Paola Merlo, Jorg Tiedemann & Reut Tsarfaty (eds.), *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, 1519–1533. Online: Association for Computational Linguistics. DOI: 10.18653/v1/2021.eacl-main.130. <https://aclanthology.org/2021.eacl-main.130>.
- Sih, Sîng-hông. 2015a. 咱的字你敢捌 — 台語漢字. https://github.com/Taiwanese-Corpus/Linya-Huang_2014_taiwanesecharacters.
- Sih, Sîng-hông. 2015b. *Icorpus* 臺華平行新聞語料庫漢字臺羅版. https://github.com/Taiwanese-Corpus/icorpus_ka1_han3-ji7.
- Sih, Sîng-hông. 2016. 國校仔課本 <https://github.com/Taiwanese-Corpus/kok4hau7-kho3pun2>.
- Sih, Sîng-hông. 2018. 臺華新聞語料庫. <https://github.com/sih4sing5hong5/icorpus>.
- SuiSiann Dataset*. 2021. <https://suisiann-dataset.ithuan.tw/>. Accessed 2024-01-21.
- Sun, Ching Chu & John Newman. 2010. *Taiwanese southern min corpus 1.0*. https://sites.ualberta.ca/~johnnewm/TSM/Taiwanese_Southern_Min/TSM.html.
- Tacotron 1 and 2*. 2021. <https://docs.coqui.ai/en/latest/models/tacotron1-2.html>.
- Taiwanese-English Dictionary*. N.d. <https://www.mkdict.net/>.
- Tang, Audrey. 2022. 萌典網站. <https://github.com/g0v/moedict-webkit>.

- Teng, Ashley. 2024. *Atenglens/taiwanese_english_translation*. https://huggingface.co/datasets/qrtt1/dictionary_of_frequently_used_taiwan_minnan.
- The Association for Computational Linguistics and Chinese Language Processing. N.d. *Aclclp*. https://www.aclclp.org.tw/use_mat.php.
- Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, vol. 2012, 2214–2218.
- Tiedemann, Jörg. 2020. The Tatoeba Translation Challenge – Realistic data sets for low resource and multilingual MT. In *Proceedings of the fifth conference on machine translation*, 1174–1182. Online: Association for Computational Linguistics. <https://www.aclweb.org/anthology/2020.wmt-1.139>.
- Tsay, Jane S. 2007. Construction and automatization of a minnan child speech corpus with some research findings. In *International journal of computational linguistics & chinese language processing, volume 12, number 4, december 2007: special issue on speech and language processing for taiwanese minnan, hakka, and mandarin*, 411–442.
- Tsay, Jane S. 2014. A phonological corpus of l1 acquisition of taiwan southern min. In *The oxford handbook of corpus phonology*.
- von Platen, Patrick. 2021. *Fine-tuning xls-r for multi-lingual asr with hugging face transformers*. Available at <https://huggingface.co/blog/fine-tune-xlsr-wav2vec2> (2024/03/26).
- Wagner, Petra et al. 2019. Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. In *Proceedings of the 10th speech synthesis workshop (ssw10)*.

- Wang, Sheng-Fu & Janice Fon. 2013. A taiwan southern min spontaneous speech corpus for discourse prosody. *The Proceedings of Tools and Resources for the Analysis of Speech Prosody, Aix-en-Provence, France*. 20–23.
- Wei, Kun et al. 2023. Joint pre-training with speech and bilingual text for direct speech to speech translation. In *Icassp 2023-2023 ieee international conference on acoustics, speech and signal processing (icassp)*, 1–5.
- wikimedia. 2023. <https://opus.nlpl.eu/wikimedia/corpus/version/wikimedia>.
- Yang, Yun-Yan et al. 2006. *Digital archive database for written taiwanese (2nd stage)*. <http://ip194097.ntcu.edu.tw/nmt1/dadwt/pbk.asp>.
- Zhang, Chen et al. 2021. Uwspeech: speech to speech translation for unwritten languages. *Proceedings of the AAAI Conference on Artificial Intelligence* 35(16). 14319–14327. DOI: 10.1609/aaai.v35i16.17684. <https://ojs.aaai.org/index.php/AAAI/article/view/17684>.