Speech-to-Speech Translation for a Real-world Unwritten Language

Anonymous ACL submission

Abstract

We study speech-to-speech translation (S2ST) that translates speech from one language into another language and focuses on building systems to support languages without standard text writing systems. We use English-Taiwanese Hokkien as a case study, and present an end-to-end solution from training data collection, modeling choices to benchmark dataset release. First, we present efforts on creating human annotated data, automatically mining data from large unlabeled speech datasets, and adopting pseudo-labeling to produce weakly supervised data. On the modeling, we take advantage of recent advances in applying self-supervised discrete representations as target for prediction in S2ST and show the effectiveness of leveraging additional text supervision from Mandarin, a language similar to Hokkien, in model training. Finally, we release an S2ST benchmark set to facilitate future research in this field.

1 Introduction

Speech-to-speech translation (S2ST) aims at translating speech from one language into speech in another language. S2ST technology can not only enable communication between people speaking different languages but also help knowledge sharing across the world. Conventionally, S2ST can be achieved via the concatenation of three systems: automatic speech recognition (ASR), machine translation (MT) and text-to-speech synthesis (TTS) (Lavie et al., 1997; Nakamura et al., 2006). In recent years, the advancement from end-to-end speech-to-text translation (S2T) (Bérard et al., 2016) or text-to-speech translation (T2ST) (Zhang et al., 2020; Lee et al., 2022a) have simplified the S2ST pipeline into two stages, which reduces error propagation issues and improves efficiency (Lee et al., 2022a). Most recently, researchers have built one-stage S2ST systems (Jia et al., 2019) that jointly optimize intermediate text generation and target speech generation steps (Kano et al., 2021; Jia et al., 2022b; Anonymized) or further remove the dependency on text completely (Tjandra et al., 2019; Lee et al., 2022a,b). Directly conditioning on the source speech during the generation process allows the systems to transfer non-linguistic information, such as speaker voice, directly (Jia et al., 2022b). Not relying on text generation as an intermediate step allows the systems to support translation into languages that do not have standard or widely used text writing systems (Tjandra et al., 2019; Zhang et al., 2020; Lee et al., 2022b).

While more than 40% of the languages in the world do not have text written forms[1], S2ST for unwritten languages still remains a research area with little exploration mainly due to the lack of training data. The majority of the previous work on this topic conducts experiments on datasets built from applying TTS on S2T corpora to generate synthetic target speech for model training (Tjandra et al., 2019; Zhang et al., 2020). Lee et al. (2022b) presents the first textless S2ST system trained on real S2ST data, while it only investigates translation between high-resource and similar language pairs (English↔Spanish, English↔French).

In this work, we take Taiwanese Hokkien as an example of an unwritten language and study S2ST between English (En) and Taiwanese Hokkien. Taiwanese Hokkien (hereafter Hokkien) is one of the official languages in Taiwan spoken by over 70% of the population (approximately 15.8 million people). Hokkien lacks a unitary writing system that is widely adopted by its native speakers, though a few possible writing systems exist, e.g. Chinese characters (Hanji), or romanization systems such as Pe̍h-ōe-jī (POJ) and Tâi-lô, etc. In addition, Hokkien is a tonal language that has complex tone sandhi rules (Cheng, 1968). Wang et al. (2004) investigates Mandarin-Taiwanese Hokkien S2ST with a cascaded template matching approach. In our work, we focus on En↔Hokkien, a distant language pair,

[1] https://www.ethnologue.com

and build one-stage S2ST systems.

We take advantage of the discrete unit-based S2ST approach (Lee et al., 2022a), which applies a self-supervised speech encoder to convert the target speech into a sequence of integers and translates source speech into target discrete units, to build the En↔Hokkien systems. First, to support En→Hokkien translation, we extend HuBERT-based discrete unit extraction (Hsu et al., 2021) and examine the feasibility of unit-to-waveform generation (Polyak et al., 2021) for tonal languages. Second, we leverage the unit-based speech normalization technique proposed in Lee et al. (2022b) to remove the non-linguistic variations in speech from multiple speakers. The original study takes advantage of synthetic speech generated from TTS as the reference target for normalization, while we build the normalizer with real Hokkien speech data. Last but not least, we study two S2ST model training strategies, speech-to-unit translation (S2UT) with a single decoder (Lee et al., 2022a) or a two-pass decoding process (Anonymized) that leverages Mandarin (Zh) as a written language similar to Hokkien to provide extra text supervision.

As no En↔Hokkien S2ST dataset is available, we also leverage Mandarin to assist the S2ST data creation process and create a 60-hr human annotated training set and an open benchmark set. Nevertheless, this is still a low-resource problem. To tackle the data scarcity issue, we further apply En↔Zh MT to create weakly supervised data (Popuri et al., 2022; Dong et al., 2022) and learn a joint embedding space for English and Hokkien through Mandarin to support data mining from unlabeled English and Hokkien data (Duquenne et al., 2021).

The contributions of this work are as follows:

- We present empirical studies that consolidate various state-of-the-art techniques for S2ST that were previously studied in a controlled setup with synthetic speech and verify their effectiveness in En↔Hokkien translation, where Hokkien is a language without a widely adopted standard text writing system.
- A benchmark set on En↔Hokkien S2ST and the evaluation model for Hokkien speech will be released to encourage future research in this direction.
- To the best of our knowledge, we are the first to build one-stage S2ST systems for an unwritten language in a real-world scenario.

2 Related Work

Existing S2ST models can be categorized in several aspects. First, Jia et al. (2019, 2022a,b) directly predict spectrogram as the model output, while Lee et al. (2022a,b); Huang et al. (2022); Popuri et al. (2022); Anonymized leverage self-supervised speech model such as HuBERT (Hsu et al., 2021) to encode the target speech into a sequence of discrete units and apply knowledge from speech-to-text modeling to S2ST. Second, Jia et al. (2019, 2022b) require extra supervision from target text or phonemes during model training, while Tjandra et al. (2019); Lee et al. (2022b); Popuri et al. (2022) show the possibility of model training with speech data only. Finally, Kano et al. (2021); Anonymized concatenate multiple decoders learned with additional text targets or speech units with different granularity and perform multi-pass decoding during inference.

While the modeling choices vary, S2ST model training often faces the challenge of data scarcity. Jia et al. (2022c) applies high-quality English TTS and creates an X→En S2ST dataset with synthetic target speech for 21 languages. To create S2ST datasets with real speech, Wang et al. (2021a) aligns ASR transcripts for more than 100 language pairs, and Duquenne et al. (2022a) applies distance-based bitext mining to audio, producing a mined S2ST dataset between 17 European languages. Weakly supervised data created from TTS (Jia et al., 2022a) or a cascaded pipeline with ASR and MT models (Dong et al., 2022; Popuri et al., 2022) is often combined with the S2ST data. In addition, self-supervised pre-training with large-scale unlabeled data also effectively improves S2ST model performance (Jia et al., 2022a; Popuri et al., 2022).

3 Methods

In this section, we first present two types of backbone architectures for S2ST modeling. Then, we describe our efforts on creating parallel S2ST training data from human annotations as well as leveraging speech data mining (Duquenne et al., 2021) and creating weakly supervised data through pseudo-labeling (Popuri et al., 2022; Jia et al., 2022a).

3.1 Model architectures

As illustrated in Fig. 1, we study one model architecture that applies a single-pass decoding process and directly translates source speech to the target, and the second one relies on target text (Mandarin

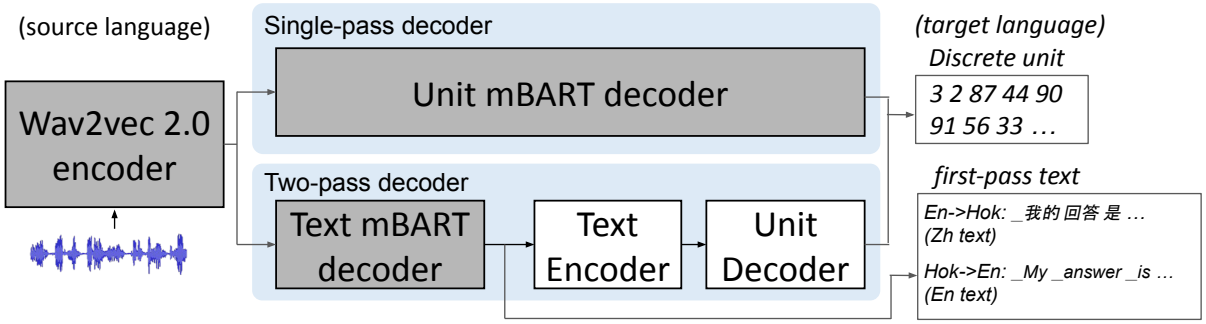

Figure 1: Model architecture of S2ST with single-pass and two-pass decoder. The blocks in shade illustrate the modules that are pre-trained. Text in italic is the training objective.

text in the case of Hokkien speech) to provide extra supervision and performs two-pass decoding. Both architectures predict discrete units as the target, and the speech encoder and text or unit decoders are pre-trained with unlabeled speech or text data.

3.1.1 Speech-to-unit translation (S2UT)

We follow the S2UT approach proposed in Lee et al. (2022a) and adopt HuBERT (Hsu et al., 2021) to convert target speech into discrete units via k-means on intermediate representation. While Hokkien→En systems can be trained on target English speech generated from single-speaker TTS to remove variations in accents from multiple speakers or noises from different recording conditions, when training En→Hokkien systems, we first apply a unit-based speech normalizer (Lee et al., 2022b) on the real Hokkien target speech. The speech normalizer is built by applying Connectionist Temporal Classification (CTC) (Graves et al., 2006) finetuning with the Hokkien HuBERT model using multi-speaker speech as input and the corresponding discrete units extracted from real Hokkien speech from a reference speaker as target.

The resulting S2ST system consists of a sequence-to-sequence S2UT model and a unit-based HiFi-GAN vocoder (Polyak et al., 2021) for unit-to-waveform conversion. For both model architectures, we pre-train the speech encoder with Conformer-based (Gulati et al., 2020) wav2vec 2.0 (Baevski et al., 2020; Popuri et al., 2022) using a large amount of unlabeled speech. To speed up model training, we replace the multi-layer convolutional feature encoder with the pre-computed 80-dimensional log-mel filterbank features. Preliminary experiments show no performance degradation with filterbank input.

3.1.2 Single-pass decoding S2UT

Lee et al. (2022a) proposes to use a single unit decoder, which can be trained with standard cross-entropy loss. Following Popuri et al. (2022), we apply mBART training (Liu et al., 2020), a denoising autoencoder trained with monolingual text in multiple langauges, using discrete units extracted from unlabeled speech with consecutive duplicate units removed, and use the pre-trained decoder to initialize the unit decoder. During decoding, we perform beam search with the unit decoder.

3.1.3 Two-pass decoding S2UT: UnitY

UnitY model (Anonymized) also performs speech-to-unit translation, while it includes a target text decoder and a target text to target unit encoder-decoder and incorporates an auxiliary target text prediction task during training. All the modules are trained jointly. In En→Hokkien direction, we use Mandarin as the target text due to its proximity to Hokkien and abundance in text data. We follow Anonymized to apply R-Drop (Wu et al., 2021) regularization during training as well as initializing the target text decoder with a text mBART model (Liu et al., 2020) pre-trained on the combination of En and Zh monolingual text data.

3.2 Training data

In the following sections, we describe three different efforts on creating parallel En↔Hokkien data for model training.

3.2.1 Supervised human annotated data

Since En↔Hokkien bilingual speakers are scarce, we use Mandarin as a pivot language during the data creation process whenever possible. We sample from the following data sources and adopt different strategies to create human annotated parallel data: (1) Hokkien dramas, which include Hokkien speech and aligned Mandarin subtitles (2) Taiwanese Across Taiwan (TAT) (Liao et al., 2020), a Hokkien read speech dataset containing

transcripts in Tâi-lô and Hanji, and (3) MuST-C v1.2 En-Zh S2T data (Cattoni et al., 2021).

We ask Zh-En bilinguals to translate the subtitles of the Hokkien dramas into English to create Hokkien→En S2T data. For the TAT dataset, we leverage a small group of En↔Hokkien bilinguals to translate the Hokkien speech and transcripts directly into English text. For MuST-C, we ask Zh-Hokkien bilinguals to translate the Mandarin text into a mix of Tâi-lô and Hanji script and then record the Hokkien speech[2]. The non-standardized script helps to improve the fluency and accuracy of the recorded Hokkien speech, while no Hokkien transcripts are used during S2ST training.

In the end, we build S2ST training sets, where the En→Hokkien set is from MuST-C. For Hokkien→En training, we apply an English text-to-unit (T2U) model (Lee et al., 2022b), which is a sequence-to-sequence Transformer model trained on English characters as input and units extracted from the corresponding speech as target, on the English text collected for Hokkien dramas and TAT, as well as the English transcriptions provided in MuST-C, to convert the text into units.

3.2.2 Mined data

To build a shared embedding space for Hokkien and English speech and text data for performing speech-to-text or speech-to-speech mining at scale, we again take advantage of Mandarin text as the bridge between the two languages. First, to encode En and Zh text in the same embedding space, we apply the method proposed in Duquenne et al. (2022b) to finetune `XLM-R LARGE` (Conneau and Lample, 2019) to fit LASER (Artetxe and Schwenk, 2019) English text space using Zh-En parallel MT data. Then, we minimize the mean squared error (MSE) loss between the max-pooled output of the learned text encoder and that of a speech encoder using aligned Hokkien speech and Mandarin or English text[3]. The text encoder is fixed during speech encoder training, where the latter is initialized with Conformer-based wav2vec 2.0 pre-trained with Hokkien speech, and this process further encodes the Hokkien speech, Mandarin and English text in the same embedding space. Similarly, we also leverage the fixed text encoder to train an En speech encoder using speech and text pairs from En ASR data. In the end, we create a shared embedding space for En speech and text, Mandarin text, and Hokkien speech, which supports En text and Hokkien speech or En speech and Hokkien speech mining based on cosine similarity.

[2]The annotators pointed out that it is easier to leverage both systems, which is another evidence of Hokkien lacking a commonly adopted text writing system.

[3]A subset of the Hokkien dramas data has English subtitles.

	Data source	Source speech (hrs)	Target speech (hrs)
Hokkien→En	Hokkien dramas	5.8*	synthetic
	TAT	4.6 (74M, 86F)	synthetic
	MuST-C	51 (8M, 14F)	synthetic
En→Hokkien		35*	51 (8M, 14F)

Table 1: Statistics of the human annotated training sets. (M: male, F: female, *: no gender information available)

		# samples	Duration (hrs)	# speakers
Dev	En	722	1.62	10 (5 M, 5 F)
	Hokkien		1.46	10 (8 M, 2 F)
Test	En	686	1.47	10 (5 M, 5 F)
	Hokkien		1.42	10 (3 M, 7 F)

Table 2: Statistics of the TAT-S2ST benchmark set. (M: male, F: female)

3.2.3 Weakly supervised data

We take advantage of cascaded systems to create weakly supervised data from ASR and S2T data (Popuri et al., 2022; Dong et al., 2022). For En→Hokkien, we apply En→Zh MT on the En ASR transcriptions, followed by a Zh→Hokkien text-to-unit-translation (T2UT) model, which is a Transformer-based sequence-to-sequence model trained with Mandarin characters as input and the corresponding Hokkien normalized units as targets. For Hokkien→En, we apply the Zh→En MT model on the Hokkien drama Mandarin subtitle, followed by En T2U to create pseudo-labeled data.

4 Experimental Setup

In this section, we describe the data, model training details, as well as baseline systems and the evaluation protocol. All experiments are conducted using `fairseq` (Ott et al., 2019).

4.1 Data

4.1.1 Supervised human annotated data

We carry out the annotation process in Sec. 3.2.1, and Table 1 summarizes the statistics of the training data. In the end, we create a 61.4-hr human annotated training set for Hokkien→En, and 35-hr for En→Hokkien. We do not combine the synthetic English speech created for Hokkien→En with the real En→Hokkien S2ST dataset during training.

4.1.2 TAT-S2ST: En↔Hokkien S2ST evaluation dataset

As a part of the effort on creating human annotated data, we also create an En↔Hokkien S2ST benchmark set to facilitate future research in the field. The English text translation we collect for the TAT dev and test sets are proofread first, and we recruit native speakers to record the English text translations, producing En↔Hokkien parallel speech data. Table 2 shows the statistics of this benchmark set. While Hokkien does not have a standardized and widely adopted writing system, TAT provides Tâi-lô transcripts, which is a standardized romanization system for Hokkien, which can be leveraged as reference text in evaluation (Sec. 4.4).

4.1.3 Mined data

We train the En and Zh joint text encoder on CCMatrix (Schwenk et al., 2019), the Hokkien speech encoder on Hokkien dramas, and the English speech encoder on English ASR data from Common-Voice (Ardila et al., 2020), CoVoST-2 (Wang et al., 2021b), Europarl-ST (Iranzo-Sánchez et al., 2020), MuST-C (Di Gangi et al., 2019), Voxpopuli (Wang et al., 2021a) and Librispeech (Panayotov et al., 2015). The learning rate is set to 10^{-4}, with an inverse square root schedule. The maximum number of tokens is set to 640k (equivalent to 40 seconds with 16kHz sampling rate), with a maximum number of sentences set to 32. We train the models with 48 GPUs for 60k steps.

With the trained text and speech encoders, we perform data mining between Hokkien speech from Hokkien dramas and English Common Crawl text, and between the former and Librivox English audio[4]. We post-process the mined data in order to have a maximum of 20% overlap between any two audio segments. In the end, we obtain 8.1k-hr Hokkien→En S2T mined data and 197-hr En↔Hokkien S2ST mined data. The difference in the volume is mainly due to the domain mismatch in audiobooks from Librivox and Hokkien dramas.

4.1.4 Weakly supervised data

For En→Hokkien, we apply En→Zh MT on the combination of the English transcripts from Librispeech (Panayotov et al., 2015) and TED-LIUM3 (Hernandez et al., 2018), totaling 1.5k-hr of English speech. The En→Zh MT model is a 12-layer Transformer model trained on CCMatrix (Schwenk et al., 2019) using disjoint BPEs for En and Zh encoded by the `sentencepiece` toolkit (Kudo and Richardson, 2018), each of size 32768. We use 16 GPUs, a batch size of 14,336 tokens and a learning rate of 10^{-3} during training.

The Zh→Hokkien T2UT model following the En→Zh translation step is trained on Hokkien dramas and the aligned Mandarin subtitles. We filter out speech containing Mandarin code-switching by applying Mandarin ASR and computing the Levenshtein distance between the ASR output and the subtitles, as well as short sentences with less than three characters, resulting in 1k-hr Hokkien speech for training.

For Hokkien→En, we apply Zh→En MT on the Mandarin subtitles from 8k-hr Hokkien drama data, followed by an En T2U trained on LJSpeech (Ito and Johnson, 2017). The Zh→En MT is trained with the same setup as En→Zh MT.

4.2 Model training

4.2.1 Hokkien HuBERT units

To encode En target speech, we use the multilingual HuBERT model, the k-means quantizer and the unit vocoder released from Lee et al. (2022b). Below we focus on how we build Hokkien units and the corresponding unit-based speech normalizer and unit vocoder.

We train a Hokkien HuBERT model using the combination of 10k-hr Mandarin speech from WenetSpeech (Zhang et al., 2022) and 2k-hr Hokkien speech from the combination of Hokkien dramas, TAT and 600-hr of Hokkien speech with various accents in addition to Taiwanese Hokkien, licensed from SpeechOcean[5]. When modeling Hokkien speech as discrete units, we empirically find that combining Mandarin with Hokkien speech during HuBERT training allows the units to better capture the tones and produce higher-quality speech output in the unit-to-waveform conversion stage.

The HuBERT model is of the `BASE` architecture and pre-trained for three iterations following Hsu et al. (2021); Lakhotia et al. (2021). In the beginning of each iteration, we randomly sample 300-hr Mandarin and Hokkien speech, respectively, for k-means clustering, and apply temperature sampling to balance the amount of speech from the two languages during training. We use $T = 20$,

[4] https://librivox.org/api/

[5] https://en.speechocean.com/

and the probability of sampling from a language l is $\tilde{p}_l = \frac{p_l^{\frac{1}{T}}}{\sum_i p_i^{\frac{1}{T}}}$, where $p_i = \frac{n_i}{\sum_j n_j}$, and n_i is the number of samples from a language. No extra language information is required during pre-training. In each iteration, model weights are randomly initialized and optimized for 400k steps. We use $K = 2500$ with features from the 12-th layer of the model from the third iteration for extracting Hokkien units.

The Hokkien speech normalizer is trained on 2-hr speech from TAT. We select speaker *THF022* as the reference speaker, i.e. the normalization target, and create speech pairs by sampling from other speakers reading the same content in TAT. We use mask probability of 0.5, mask channel probability of 0.25 and learning rate of 3×10^{-5} and train for 25k updates. Finally, the Hokkien unit-based HiFi-GAN vocoder is trained on the TTS subset of the TAT dataset, which contains a total of 36 hours of clean speech from two male and two female speakers, following the training procedure in Lee et al. (2022a).

4.2.2 Wav2vec 2.0 encoder

We pre-train the Conformer En wav2vec 2.0 `LARGE` encoder (Baevski et al., 2020) with the Libri-light corpus (Kahn et al., 2020), which contains around 54k hours of read speech audio. The encoder is trained with a batch size of 2.1-hr for 1M updates, with 32k warmup steps and a peak learning rate of 5×10^{-4}. For masking, we sample a probability of 0.065 of all time-steps to be starting indices and mask the subsequent 10 time steps. For the Hokkien wav2vec 2.0 encoder, we pre-train it with 30k-hr Hokkien drama data using the same hyper-parameters as the En wav2vec 2.0 encoder.

4.2.3 Single-pass decoding S2UT

The Hokkien unit mBART is trained with 30k-hr Hokkien dramas and 10k-hr Mandarin data from WenetSpeech. The model is trained on 64 GPUs with a batch size of 3072 units, learning rate of 3×10^{-4} with Adam and 10k warmup steps. The model is trained with 500k updates with dropout 0.1. We use the En unit mBART released by Popuri et al. (2022) for training Hokkien→En models.

With the pre-trained wav2vec 2.0 encoder and the unit mBART decoder, we follow the best fine-tuning strategy in Popuri et al. (2022), where the whole encoder and the LayerNorm and both encoder and self attention in the decoder are finetuned with the parallel S2ST data. The models are trained on 32 GPUs with a batch size of 160k tokens. We used 0.1 dropout for all models and 0.2 LayerDrop (Fan et al., 2019). The models are trained using Adam optimizer with 3×10^{-4} learning rate, 10k warmup steps an 50k maximum updates.

4.2.4 Two-pass decoding S2UT: UnitY

The text mBART model is pre-trained on the combination of Mandarin and English text data from CC-100 (Conneau et al., 2019), Newscrawl (Akhbardeh et al., 2021), Leipzig Corpora (Goldhahn et al., 2012), NewsCommentary (Tiedemann, 2012). There are 2B English sentences and 230M Mandarin sentences. We learn BPE of size 65536 jointly on both languages and apply temperature sampling with $\frac{1}{T} = 0.7$ during training.

We combine the pre-trained wav2vec 2.0 encoder, the text mBART decoder, and two randomly initialized Transformer layers for the text encoder and the unit decoder, respectively, to build the UnitY model. We train our two-pass models on 16 GPUs with a batch size of 120k tokens, dropout 0.1 for all models except for the human annotated data only setup where we use dropout 0.3. We use LayerDrop (Fan et al., 2019) 0.1 and label smoothing 0.1, and train the model with a learning rate of 5×10^{-4}, 2k warmup steps, and a maximum update of 50k steps. The weight on the auxiliary loss from the text decoder is set to 8.0.

4.3 Baselines

We build two-stage and three-stage cascaded baseline systems for both En↔Hokkien directions. The two-stage cascaded system consists of a source speech (En or Hokkien) to target text (Mandarin or En) end-to-end S2T model and a target text to target speech unit T2U model (T2UT in the case of Zh→Hokkien). The three-stage cascaded system further breaks down the En→Zh S2T model into En ASR followed by En→Zh MT, and the Hokkien→En S2T model is split into a Hokkien→Zh S2T step and a Zh→En MT step.

All the speech encoders for the En ASR and S2T models are initialized with wav2vec 2.0 (Sec. 4.2.2). The text decoders of S2T models are initialized with the text mBART (Sec. 4.2.4). We use the En↔Zh MT models, the En T2U model and the Zh→Hokkien T2UT model described in Sec. 4.1.4 for building the cascaded systems.

4.4 Evaluation

To evaluate the translation quality, we compute ASR-BLEU on the TAT-S2ST evaluation set (Sec. 4.1.2) by applying ASR on the generated speech and computing 4-gram BLEU against the reference text using SACREBLEU (Post, 2018). We use an open-sourced En ASR model[6] when evaluating Hokkien→En systems. For En→Hokkien systems, we build an ASR model to transcribe Hokkien speech into Tâi-lô. The Hokkien ASR is initialized with a w2v-BERT (Chung et al., 2021) LARGE model pre-trained on 10k-hr Mandarin speech from WenetSpeech and 30k-hr Hokkien speech from Hokkien drama, followed by finetuning with CTC loss on 480-hr Hokkien speech and Tâi-lô scripts from TAT (Liao et al., 2020). Each Tâi-lô syllable is split into initial and final with tone as the target. The resulting Hokkien ASR model achieves 6.8% syllable error rate (SER) on the TAT-Vol1-test-lavalier set. To evaluate En→Hokkien translation quality, we compute syllable-level ASR-BLEU.

To evaluate the naturalness of the speech output, we collect mean opinion scores (MOS) ranges from 1 (the worst) to 5 (the best) from human listening tests. Each item is labeled by three annotators.

5 Results

5.1 Single-pass vs. two-pass decoding

We first study the model architecture choice in both En↔Hokkien directions. Table 3 summarizes the results. We include ASR-BLEU from the target reference speech as a indication of the effect from the unit vocoder and the ASR errors (row 7). We start from training on human annotated data, and it results in very low BLEU score in both directions (row 3, 5), indicating that pre-training, including wav2vec 2.0 and unit or text mBART, is not enough for building a S2ST system under low-resource for distant language pairs. With extra supervision from text, the UnitY model works slightly better than single-pass S2UT by 3.7 BLEU in Hokkien→En (row 3 vs. 5).

We then combine the human annotated data with weakly supervised data. Both systems achieve significant gain (6.2-7.5 BLEU) in both directions, indicating the effectiveness of combining self-supervised pre-training and data augmentation with weakly supervised data in low-resource S2ST for a distant language pair.

In addition, we find that UnitY outperforms single-pass S2UT in Hokkien→En direction (row 4 vs. 6) by 2.9 BLEU. However, in En→Hokkien, UnitY is merely 0.4 BLEU higher than single-pass S2UT. The larger impact from the additional text supervision in Hokkien→En may be due to the fact that the target text and speech are of the same language, or the larger amount of training data available. As the focus of this work is to present a data creation and model training strategy, we leave the investigation to future work.

For the cascaded baselines, the two-stage system is worse than the three-stage system in both En↔Hokkien directions (row 1 vs. 2). Our best one-stage system performs similarly to the best cascaded systems (row 2 vs. 6).

For MOS, the cascaded systems and single-stage S2UT systems have similar naturalness in both En→Hokkien and Hokkien→En directions.

5.2 Mined data

In this section, we study how to leverage mined Hokkien→En S2T and En↔Hokkien S2ST data.

5.2.1 Leveraging mined En↔Hokkien S2ST in En→Hokkien direction

In Table 4, we show the results of leveraging the mined En↔Hokkien S2ST data in En→Hokkien direction. In order to train the UnitY model, we apply Hokkien→Zh S2T to generate pseudo-labeled Mandarin text for the mined Hokkien speech as the auxiliary task target.

We first train both one-stage models with mined data and the human annotated data. While the single-pass decoding S2UT model still yields very low BLEU score (row 8), the UnitY model achieves 4.8 BLEU improvement with the extra 197-hr of mined S2ST data (row 5 vs. 10), showing that noisy Mandarin text generated from pseudo-labeling still provides useful signals in model training. We then further combine with weakly supervised data but do not see significant gain with the additional mined data (row 4 vs. 9, 6 vs. 11). Note that the size of mined data is only 13% of the total amount of weakly supervised data we have. As discussed in Sec. 4.1.3, the limited amount of mined data available is mainly due to the domain mismatch issue. In the future, we plan to explore mined data from more similar domains and aim to increase the amount of data for better S2ST performance.

[6] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

Table 3: Dev / test ASR-BLEU on TAT-S2ST dataset. MOS results are reported with 95% confidence interval. (*: synthetic Hokkien speech is generated by applying unit vocoder on the normalized units extracted from the ground truth Hokkien speech, while synthetic En speech is generated by applying En T2U followed by the unit vocoder on the ground truth En text. **: Human annotated TAT data (2-hr) is not included in the training data of Hokkien→Zh S2T system due to lack of Mandarin translation.)

		En→Hokkien					Hokkien→En				
		Training data		ASR-BLEU		MOS	Training data		ASR-BLEU		MOS
ID	Model	Human (35-hr)	Weakly (1.5k-hr)	Dev	Test	Test	Human (61.4-hr)	Weakly (8k-hr)	Dev	Test	Test
Cascaded systems:											
1	Three-stage	✓	✓	8.9	7.5	3.54 ± 0.05	✓**	✓	10.7	**10.0**	3.22 ± 0.06
2	Two-stage	✓	✓	8.4	6.9	3.52 ± 0.05	✓	✓	11.4	8.1	3.09 ± 0.06
Single-stage S2UT systems:											
3	Single-pass decoding	✓	✗	0.1	0.0	-	✓	✗	0.1	0.1	-
4	Single-pass decoding	✓	✓	8.6	7.4	3.58 ± 0.05	✓	✓	8.1	7.1	3.06 ± 0.06
5	Two-pass decoding (UnitY)	✓	✗	1.0	0.3	-	✓	✗	4.2	3.8	-
6	Two-pass decoding (UnitY)	✓	✓	**9.3**	**7.8**	3.69 ± 0.05	✓	✓	**11.8**	**10.0**	3.15 ± 0.06
7	Synthetic target*	✗	✗	61.9	61.8	3.85 ± 0.05	✗	✗	76.4	78.5	3.24 ± 0.05

Table 4: Results of En→Hokkien models trained with mined En↔Hokkien S2ST data. We report dev / test ASR-BLEU on TAT-S2ST dataset.

		Training data			ASR-BLEU	
ID	Model	Human (35-hr)	Weakly (1.5k-hr)	Mined (197-hr)	Dev	Test
3	Single-pass decoding	✓	✗	✗	0.1	0.0
8		✓	✗	✓	0.1	0.1
4		✓	✓	✗	8.6	7.4
9		✓	✓	✓	7.2	7.3
5	Two-pass (UnitY)	✓	✗	✗	1.0	0.3
10		✓	✗	✓	5.9	5.1
6		✓	✓	✗	9.3	7.8
11		✓	✓	✓	9.0	7.7

Table 5: ASR-BLEU scores on TAT-S2ST test set from Hokkien→En UnitY models trained with mined data filtered at different thresholds (t) for the similarity score. Amount of mined data (hr) per threshold is listed.

Data combined with mined	No filter (0-hr)	t=1.08 (356-hr)	t=1.07 (2274-hr)	t=1.065 (4732-hr)	t=1.06 (8101-hr)
human (61.4-hr)	4.7	7.4	6.3	7.1	3.8
human (61.4-hr) + weakly (8k-hr)	10.0	9.9	10.7	10.5	10.8

We convert the mined Hokkien→En S2T data to S2ST data with the En T2U model and train UnitY models with the combination of human annotated data and optionally the 8k-hr weakly supervised data to examine the effect of mined data on model performance. Table 5 shows the ASR-BLEU scores on the TAT-S2ST test set with respect to different thresholds on the similarity scores of the mined pairs.

We see that adding 4.7k-hr mined S2T data ($t = 1.065$) in Hokkien→En is the most helpful and improves the model quality by 3.6 BLEU when only human annotated data is available. With 8.1k-hr mined data ($t = 1.06$), the BLEU gain drops to 0.9 BLEU. In addition, it is 5.3 BLEU lower than the UnitY model trained with human annotated data and 8k-hr of weakly supervised data (Table 3 row 6). As the Hokkien speech for both weakly supervised data and mined data come from the same Hokkien dramas dataset, the gap implies that pseudo-labeling is a generally effective data augmentation technique for low-resource scenarios, while the quality of the mined data is constrained by the content of the data available for mining. However, combining all three types of data together is still beneficial. We obtain 0.8 BLEU gain by adding 8.1k-hr mined data to the combination of human annotated and weakly supervised data.

6 Conclusions

We present the first En↔Hokkien S2ST systems, where Hokkien is an oral language that does not have standard and widely adopted text writing systems, i.e. an unwritten language. To tackle the challenges of speech translation for unwritten languages and the lack of parallel training data, we present an end-to-end study. First, we explore three options of training data creation including human annotation, weakly supervised data from pseudo-labeling and data mining. Second, we investigate two modeling choices including direct speech-to-unit translation with a single speech unit decoder and two-pass decoding that leverages extra supervision from target text. Experimental results show that leveraging a similar high-resource written language (Mandarin in the case of Hokkien) is effective in both the data creation process and model training. Finally, we release the benchmark dataset and ASR evaluation model to facilitate research in this field. In the future, we aim to expand study and establish an S2ST model building strategy that works for a diverse set of unwritten languages.

References

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Anonymized. UnitY: Two-pass direct speech-to-speech translation with discrete units (in preparation).

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *TACL*, pages 597–610.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

Alexandre Bérard, Olivier Pietquin, Laurent Besacier, and Christophe Servan. 2016. Listen and translate: A proof of concept for end-to-end speech-to-text translation. In *NIPS Workshop on end-to-end learning for speech and audio processing*.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. MuST-C: A multilingual corpus for end-to-end speech translation. *Computer Speech & Language*, 66:101155.

Robert L. Cheng. 1968. Tone sandhi in taiwanese. *Linguistics*, 41:19–42.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021. w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. *Advances in neural information processing systems*, 32.

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Qianqian Dong, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, and Yu Zhang. 2022. Leveraging pseudo-labeled data to improve direct speech-to-speech translatio. *arXiv preprint arXiv:2205.08993*.

Paul-Ambroise Duquenne, Hongyu Gong, Ning Dong, Jingfei Du, Ann Lee, Vedanuj Goswani, Changhan Wang, Juan Pino, Benoît Sagot, and Holger Schwenk. 2022a. Speechmatrix: A large-scale mined corpus of multilingual speech-to-speech translations.

Paul-Ambroise Duquenne, Hongyu Gong, Benoît Sagot, and Holger Schwenk. 2022b. T-Modules: Translation modules for zero-shot cross-modal machine translation. *arXiv preprint arXiv:2205.12216*.

Paul-Ambroise Duquenne, Hongyu Gong, and Holger Schwenk. 2021. Multimodal and multilingual embeddings for large-scale speech mining. *Advances in Neural Information Processing Systems*, 34.

Angela Fan, Edouard Grave, and Armand Joulin. 2019. Reducing transformer depth on demand with structured dropout. *arXiv preprint arXiv:1909.11556*.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on Machine learning*, pages 369–376.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang.

2020. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: twice as much data and corpus repartition for experiments on speaker adaptation. In *International conference on speech and computer*, pages 198–208. Springer.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*.

Rongjie Huang, Zhou Zhao, Jinglin Liu, Huadai Liu, Yi Ren, Lichao Zhang, and Jinzheng He. 2022. TranSpeech: Speech-to-speech translation with bilateral perturbation. *arXiv preprint arXiv:2205.12523*.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchís, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A multilingual corpus for speech translation of parliamentary debates. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 8229–8233. IEEE.

Keith Ito and Linda Johnson. 2017. The LJ speech dataset.

Ye Jia, Yifan Ding, Ankur Bapna, Colin Cherry, Yu Zhang, Alexis Conneau, and Nobuyuki Morioka. 2022a. Leveraging unsupervised and weakly-supervised data to improve direct speech-to-speech translation. *arXiv preprint arXiv:2203.13339*.

Ye Jia, Michelle Tadmor Ramanovich, Tal Remez, and Roi Pomerantz. 2022b. Translatotron 2: High-quality direct speech-to-speech translation with voice preservation. In *International Conference on Machine Learning*, pages 10120–10134. PMLR.

Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022c. CVSS corpus and massively multilingual speech-to-speech translation. *arXiv preprint arXiv:2201.03713*.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *Proc. Interspeech 2019*, pages 1123–1127.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-light: A benchmark for ASR with limited or no supervision. In *ICASSP*.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. 2021. Transformer-based direct speech-to-speech translation with transcoder. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 958–965. IEEE.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al. 2021. Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. 1997. JANUS-III: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 99–102. IEEE.

Ann Lee, Peng-Jen Chen, Changhan Wang, Jiatao Gu, Sravya Popuri, Xutai Ma, Adam Polyak, Yossi Adi, Qing He, Yun Tang, et al. 2022a. Direct speech-to-speech translation with discrete units. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3327–3339.

Ann Lee, Hongyu Gong, Paul-Ambroise Duquenne, Holger Schwenk, Peng-Jen Chen, Changhan Wang, Sravya Popuri, Juan Pino, Jiatao Gu, and Wei-Ning Hsu. 2022b. Textless speech-to-speech translation on real data.

Yuan-Fu Liao, Chia-Yu Chang, Hak-Khiam Tiun, Huang-Lan Su, Hui-Lu Khoo, Jane S Tsay, Le-Kun Tan, Peter Kang, Tsun-guan Thiann, Un-Gian Iunn, et al. 2020. Formosa speech recognition challenge 2020 and Taiwanese across Taiwan corpus. In *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 65–70. IEEE.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Satoshi Nakamura, Konstantin Markov, Hiromi Nakaiwa, Gen-ichiro Kikui, Hisashi Kawai, Takatoshi Jitsuhiro, J-S Zhang, Hirofumi Yamamoto, Eiichiro Sumita, and Seiichi Yamamoto. 2006. The ATR multilingual speech-to-speech translation system. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):365–376.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for

sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an ASR corpus based on public domain audio books. In *ICASSP*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021. Speech resynthesis from discrete disentangled self-supervised representations.

Sravya Popuri, Peng-Jen Chen, Changhan Wang, Juan Pino, Yossi Adi, Jiatao Gu, Wei-Ning Hsu, and Ann Lee. 2022. Enhanced direct speech-to-speech translation using self-supervised pre-training and data augmentation. *arXiv preprint arXiv:2204.02967*.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191.

Holger Schwenk, Guillaume Wenzek, Sergey Edunov, Edouard Grave, and Armand Joulin. 2019. CCMatrix: Mining billions of high-quality parallel sentences on the WEB. *CoRR*, abs/1911.04944.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2019. Speech-to-speech translation between untranscribed unknown languages. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 593–600. IEEE.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. CoVoST 2 and massively multilingual speech translation. In *Interspeech*, pages 2247–2251.

Jhing-Fa Wang, Shun-Chieh Lin, Hsueh-Wei Yang, and Fan-Min Li. 2004. Multiple-translation spotting for Mandarin-Taiwanese speech-to-speech translation. 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*, pages 13–28.

Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, Tie-Yan Liu, et al. 2021. R-Drop: Regularized dropout for neural networks. In *Proceedings off NeurIPS*, volume 34, pages 10890–10905.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, et al. 2022. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6182–6186. IEEE.

Chen Zhang, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, and Tie-Yan Liu. 2020. UWSpeech: Speech to speech translation for unwritten languages. *arXiv preprint arXiv:2006.07926*.