Cascade vs Pseudo-Label Speech to Text Translation from Taiwanese to Mandarin

Chenxuan Cui, Noel Chen, Kalvin Chang
Language Technologies Institute, Carnegie Mellon University
{cxcui, yunhsua3, kalvinc}@andrew.cmu.edu

Abstract

We compare two approaches for speech translation from Taiwanese to Mandarin without a direct corpus: 1) *Cascade*: use an ASR system to produce an intermediate transcription and use a Taiwanese to Mandarin machine translation (MT) system to produce the final output. 2) *End-to-end, using pseudo-data*: translate the source language transcriptions into the target language with an MT system to create an ASR dataset with pseudo-labels, then train an ST system on the pseudo dataset in an end-to-end fashion. Our results show that cascade method outperforms direct training with pseudo labels. Our code is publicly available[1][2].

1 Introduction

Taiwanese (or Taiwanese Hokkien, tâi-gí/tâi-gú, Southern Min, Min-Nan) is a language spoken by 18 million Taiwanese people to varying degrees of fluency (Directorate-General of Budget, Accounting and Statistics, Taiwan, 2020), accounting for nearly 80% of the island's population. Although the language is widely spoken, it is rarely written down by speakers of any age group. This is because most speakers are also fluent in Mandarin, the language of education and official use, so any written communication tends to happen in Standard Written Chinese (which is largely based on Mandarin). When the situation necessitates written Taiwanese, methods of writing down the language include, but are not limited to, the following:

1. **Han characters**: A set of Chinese characters specific for use in Taiwanese are recommended by Ministry of Education, Taiwan (2014). However, in more informal settings, users often substitute Taiwanese characters with phonetically similar but semantically different characters commonly used in Mandarin. The usage of these is *ad hoc* and often inconsistent among different speakers.

2. **Romanization**: Two popular romanization systems are pe̍h-oē-jī (invented by Western missionaries in the 19th century) and tâi-lô (derived from the former and promoted by the Ministry of Education in Taiwan). Although the romanization systems were designed to increase literacy in Taiwanese, few people can read them, making them more a tool for language learners than a fully fledged writing systems.

Here is a sentence written in Taiwanese Han characters and tâi-lô romanization from Liao et al. (2020). Note that Mandarin speakers would have trouble understanding this.

1. 這馬 是 暗時 八點
   tsit-má sī àm-sî peh-tiám
   'It is 8pm now'

Without a consistent writing system, it can be difficult to develop speech technologies that are grounded in written text. With a traditional automatic speech recognition system (hereinafter ASR) that outputs written Taiwanese, few people can comprehend its output transcription. It is more useful to have a speech to text translation system (hereinafter speech translation or ST) that outputs text in Mandarin, which can be read by many more users. Among the 65+ age group, 65.9% use Taiwanese as their dominant language (Directorate-General of Budget, Accounting and Statistics, Taiwan, 2020), and 10% of them are illiterate in Mandarin (Ministry of Health and Welfare, Taiwan,

[1] https://github.com/cuichenx/Taiwanese_ASR_MT

[2] https://github.com/yunhsuanchen/espnet/tree/master/egs2/nan_suisiann/asr1

2017). A Taiwanese speech to Mandarin text translation system would help elderly Taiwanese speakers interact with Mandarin interfaces such as customer support chatbots and messaging apps. The system would also help those not fluent in the language understand spoken Taiwanese, as well as aid the younger generation learn the language of their heritage.

To train an ST system end-to-end, a parallel corpus containing pairs of speech in source language and text in target language is needed. In the case of Taiwanese speech and Mandarin text, no such dataset currently exists. In this paper, we investigate two approaches for speech translation without a direct corpus: 1) *Cascade*: Use an ASR system to produce an intermediate transcription and use a Taiwanese to Mandarin machine translation (MT) system to produce the final output. 2) *End-to-end, using pseudo-data*: translate the source language transcriptions into the target language with an MT system to create an ASR dataset with pseudo-labels, then train an ST system on the pseudo dataset in an end-to-end fashion.

2 Related work

2.1 End-to-end speech translation

When a speech translation corpus exists for a source and target language pair, such as the MuST-C benchmark dataset (Di Gangi et al., 2019), training an end-to-end model is possible. Multi-task learning is often used in such end-to-end models. For example, Bahar et al. (2021) jointly optimize both the MT and ASR models, resulting in a system that outputs both transcriptions and translations and that outperforms cascade models. Indurthi et al. (2021) jointly learn the parameters for the ASR, MT, and speech translation models with a task modulation network. The network enables the learning of both task-specific features and those shared by all three tasks and eliminates the need to separately finetune each task. Le et al. (2020) propose a multilingual transformer-based joint model that uses one decoder for ASR and one for speech translation. Each decoder can attend to information from the other in what they term a dual-attention mechanism, resulting in an end-to-end model that jointly transcribes and translates.

Other approaches include adapters, which are lightweight task-specific trainable parameters that are tuned after freezing pretrained parameters. Le et al. (2021) first pretrain two multilingual speech translation models, one with parallel data and the other with mBART trained on non-parallel multilingual data. They then finetune adapters for specific language pairs using bilingual data to obtain bilingual speech translation systems from the multilingual model without adding an excessive number of parameters, which they argue is more efficient than full language pair-specific finetuning.

2.2 Non-end-to-end speech translation

When no direct corpus exists for some language pair, however, there are two main solutions in the literature: (Bentivogli et al., 2021):

1. *Cascade*: a two-stage pipeline consisting of an ASR model to transcribe to source language, followed by an MT model to translate to the target language. This is an intuitive and reasonably well-performing solution that has been adapted by the research community for a long time (Stentiford and Steer, 1988; Waibel et al., 1991). However, this pipeline suffers from error propagation due to the use of two separate stages, as well as the loss of speech information that could be useful in the translation. Some strategies to mitigate error propagation are outlined in section 7.

2. *Pseudo Data*: generate a pseudo corpus that contains waveform in the source language and text in the target language, and then train a direct end-to-end speech-to-text model (Bérard et al., 2016; Weiss et al., 2017).

2.3 Taiwanese to Mandarin speech translation

Taiwanese to Mandarin speech translation was one of the possible tracks (track 1) in the Formosa Speech Recognition Challenge 2020 (Liao et al., 2020), even though the competition labeled this a subtask of Taiwanese ASR. The best system for track 1 used an end-to-end model with data augmentation (Liang et al., 2021). The Challenge's official training set contained 33.3 hours of read speech from the Taiwanese Across Taiwan (TAT) corpus (Liao et al., 2020) and 73.9 hours of spontaneous speech from broadcast news, talk shows, and dramas from the proprietary Public Television Service (PTS) corpus. In addition to speech from TAT and PTS, noisier soap opera speech from Formosa TV (FTV) appears in the test set, presenting an acoustic and domain mismatch problem for teams (Liang et al., 2021). The best team remedied the issue by scraping 954.5 hours of speech from three FTV soap operas on YouTube. Because the captions are in Mandarin, a lexicon-level (word by word) translation model is applied to translate the caption to Taiwanese, thus generating pseudo-labeled data. The team did not apply a full machine translation model because they assumed the word alignment from Mandarin and Taiwanese is one-to-one and monotonic. The team also augmented the amount of data with a three-way speed perturbation and noise augmentation on the TAT and PTS data to obtain a total of 1500 hours of speech, including the soap opera data.

The acoustic model mostly matches the competition's provided baseline - a GMM and chain model (Liao et al., 2020). Their lexicon comes from the ChhoeTaigi database. As for the language model, the team first preprocessed both the Mandarin and Taiwanese Han text with a word segmentation model. For Taiwanese Han, though, they used a forward maximum word matching algorithm since there is no large-scale Taiwanese Han word segmentation model. Word-level 4-gram language models were applied on both texts. For Mandarin, a domain-general language model was trained on 2.4 billion words across multiple corpora. To alleviate the aforementioned domain mismatch, an in-domain language model was trained on pseudo-labeled text that was translated into Mandarin from Taiwanese with a statistical machine translation model (that achieved 61.4 character BLEU). The Taiwanese text was obtained from scraping closed captions of nearly all FTV dramas and a few dramas from other channels, using OCR to scrape the captions from YouTube when closed captions are not provided.

The end-to-end model achieved a 59.8% CER and 0.52, 0.43, 0.37, 0.32 BLEU for 1, 2, 3, and 4-grams, respectively. After the competition, the team also tried a cascade approach by first performing ASR with Taiwanese Han transcriptions and then translating Taiwanese Han to Mandarin. We improve upon their cascade model by:

1. applying a state-of-the-art neural end-to-end speech recognition model (Watanabe et al., 2018) instead of a Kaldi-based factored time delay neural network
2. translating from Taiwanese to Mandarin with neural machine translation (Ott et al., 2019)
3. using Taiwanese romanization instead of Taiwanese Han (the merits of which are discussed in subsection 4.1)
4. preserving the dashes in the romanization, unlike the competition's preprocessing script, which removes the dashes that connect morphemes within a word or compound and effectively de-segments the sentence into a pool of syllables
5. preserving punctuation symbols
6. working with a significantly lower amount of speech data

	# of seg.	seg. avg. length	total hours
SuiSiann	3467	4.92s	4.73h
TAT Vol2 sample	1299	7.67s	2.77h

Table 1: ASR Dataset Statistics

However, we cannot compare our results with theirs because they evaluate on an entirely different dataset. In both our experiments and theirs, the cascade model outperforms the end-to-end model.

3 Dataset

3.1 ASR Datasets

For our Cascade-ASR and End-to-end-ST model, we trained using two datasets which are publicly available: the SuíSiann Dataset (version 0.2.1) (2021) and Taiwanese Across Taiwan (TAT) Corpus-Vol2 Samples (2021). The SuiSiann dataset consists of one speaker reading lines from various situations, including news articles and conversations, in Taiwanese. They provide Taiwanese Han and tâi-lô transcription for each audio segment. The full TAT Corpus is a 300 hour reading speech dataset in Taiwanese with 200 speakers. They provide Taiwanese Han, tâi-lô, and pe̍h-oē-jī transcription for each audio segment. Unfortunately, the full TAT corpus is not publicly available, we could only access 2.7 hours of samples, which contain 4 speakers. The dataset statistics are listed in Table 1.

We split the SuiSiann dataset randomly with 90-5-5 ratio for train-dev-test split and append TAT-Vol2 samples to only the training data. For our dev and test sets, we have 172 samples in each. Since SuiSiann is a ASR dataset that lacks Mandarin translation, we recruited native Taiwanese speakers to manually translate the SuiSiann test and dev sets to Mandarin, creating a few hundred audio-transcription-translation triplets for the test and dev sets.

3.2 MT Dataset

For our Cascade-MT model, we used the iCorpus dataset by Chen (2016), which is a Mandarin-Taiwanese parallel corpus. This dataset is collected from 3266 news articles and contains 83,544 sentences. They provide both tâi-lô and Taiwanese Han transcription for Taiwanese.

4 Methods

4.1 Pipelines

We compare the performance of two popular choices of speech translation: *cascade* and *pseudo label*. Schematics of the two methods from speech in source language to text in target language are shown in Figure 1.

For the *cascade* approach, we choose the tâi-lô transcription as the intermediate Taiwanese representation, because it more closely matches a fully phonemic transcription than Han characters. Doing so offloads the duty of transforming speech units to semantic units (as in the Mandarin Chinese output) to the MT stage as opposed to the ASR stage. This would likely produce better results because MT from Taiwanese to Mandarin is a fairly easy task considering the high similarity between the two languages. In addition, Taiwanese romanization is inherently morphologically segmented as morphemes in the same word are connected by a dash and words are separated by spaces (as shown in 1), whereas sentences written in Han characters do not have spaces. The word segmentation aids the ASR in detecting word boundaries. We train an MT model with iCorpus using Fairseq (Ott et al., 2019), and we train an ASR model with the training split of SuiSiann (later supplemented with the TAT sample) using ESPNet (Watanabe et al., 2018). We then run the SuiSiann test set on the two models concatenated together to generate Mandarin translations from Taiwanese speech transcriptions.

For direct end-to-end training with pseudo data, previous work (Bérard et al., 2016) have generated pseudo training data using a source language TTS model. While this is also applicable in our setting, it would introduce a new model component that makes comparison with the *cascade* method more difficult. Instead, we generate pseudo labels using the same trained Taiwanese to Mandarin MT model as above, and directly train a speech translation system on the pseudo-Mandarin labels. For two arbitrary languages, this would require a specific speech translation architecture that makes no assumption between the two languages. Using an ASR module here would be problematic because the CTC loss assumes a linear correspondence between the speech frame and output tokens. However, because Taiwanese and Mandarin have almost identical syntactic structures (due to both being Sinitic languages descended from Old Chinese), it is possible to simplify the architecture and use an ASR module here, similar to the *cascade* approach. We again use ESPNet (Watanabe et al., 2018) to train the system end-to-end. In general, for two languages with divergent syntactic properties, an ST module, such as ESPNet-ST (Inaguma et al., 2020), would be more suitable for the *pseudo label* approach.

4.2 Speech Data Augmentation

A main problem with the SuiSiann dataset was that it only has one speaker. Inspired by Schlüter and Grill (2015) and Esa et al. (2020), we apply pitch shifting to the original audio to synthesize other speakers, which also increases our training quantity at the same time. Since the speaker in SuiSiann is a woman, we synthesized three more speakers with -1, -2, -3 half steps (lower pitches) to the original pitch, expanding our dataset to 4x the original size.

4.3 Self-Supervised Learning Representations

Self-Supervised Learning Representations (SSLR) is a suite of recently proposed methods to extract speech features from the raw waveform with neural networks. They have been shown to outperform traditional spectral features consistently. In this work, we investigate two SSLR methods which trade off between number of parameters (and thus training time) and performance: TERA (Liu et al., 2021) and XLSR (Conneau et al., 2021). The latter has more than 10 times as many parameters, but remains the state of the art in many non-English ASR benchmarks.

In addition to using only SSLR features, we also explore feature fusion (Anonymous, 2021) which combines the spectral features with SSLR features with a linear projection layer, which has been shown to improve performance.

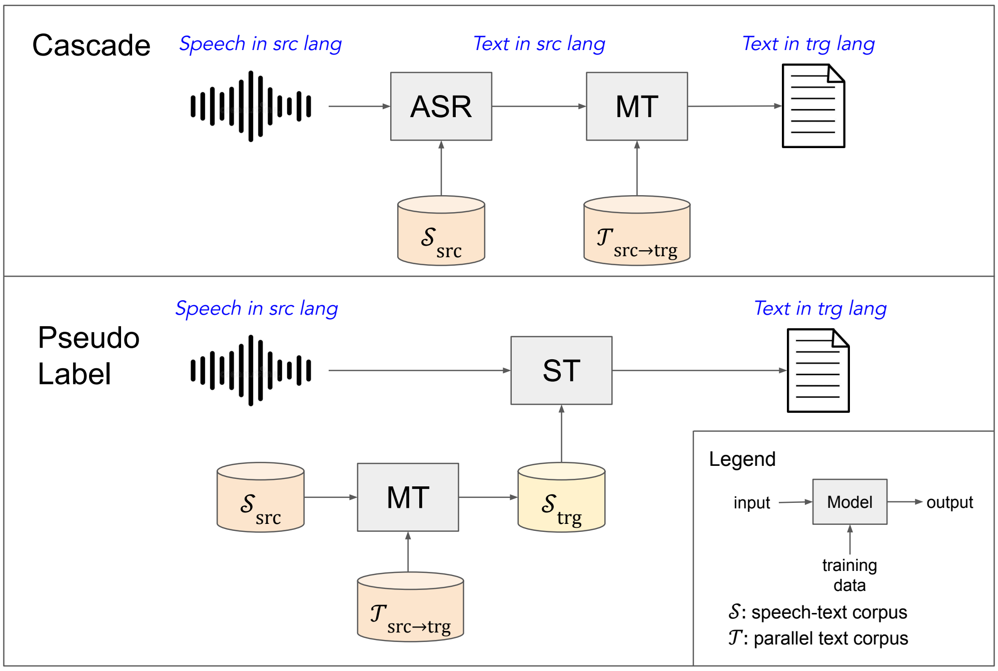
Cascade
Speech in src lang
Text in src lang
Text in trg lang
ASR
MT
$\mathcal{S}_{\text{src}}$
$\mathcal{T}_{\text{src}\rightarrow\text{trg}}$
Pseudo Label
Speech in src lang
Text in trg lang
ST
$\mathcal{S}_{\text{src}}$
MT
$\mathcal{S}_{\text{trg}}$
$\mathcal{T}_{\text{src}\rightarrow\text{trg}}$
Legend
input
Model
output
training data
$\mathcal{S}$: speech-text corpus
$\mathcal{T}$: parallel text corpus

Figure 1: Schematics of two methods of speech translation without direct data: *cascade* and *pseudo label*.

4.4 Decoding without LM

Because our dataset is small, we find that the language model trained on it can sometimes be ineffective, so decoding with the poorly trained LM can lead to worse performance. For this reason, we report ASR results for decoding with and without the trained LM in order to find the best performing configuration.

4.5 Evaluation

4.5.1 Automatic Metrics

We evaluate the final speech translation results using two popular automatic metrics: BLEU (Papineni et al., 2002) and Character Error Rate (CER, equivalent to WER for word-based languages). Neither of these are perfect metrics for speech translation in general, since BLEU focuses only on n-gram overlap without taking the semantics of the entire sentence into account, and CER assumes that there is only one possible translation for every sentence, from which it measures an error rate. For Taiwanese to Mandarin translation, since the two languages are typologically very similar, translation can be sometimes (but not always) simplified to vocabulary substitution. Therefore, BLEU and CER can give a reasonable estimate to the translation quality, unlike other language pairs where there is more freedom in translation.

For *Cascade*, we also evaluate the ASR and MT components separately. We evaluate ASR with different types of error rates on the intermediate tâi-lô romanization: Word (a multi-syllabic unit connected with dashes in tâi-lô), Character (a Latin letter, close to a phoneme segment), Syllable (equivalent to a Han character), as well as Syllable without regard of tones (atonal). We evaluate MT using BLEU and COMET (Rei et al., 2020), but it should be noted that Taiwanese is not covered by the XLM-R encoder used by COMET, so the numbers are unreliable.

4.5.2 Post-editing

Due to the nature of the datasets we use, all evaluations are done with only 1 reference translation. Therefore, to account for minor linguistic differences in the translation output, we adopt post-editing (PE), a common evaluation methodology used for translation quality assessment (Snover et al., 2006, 2009; Denkowski and Lavie, 2010; Cettolo et al., 2013; Bojar et al., 2015; Graham et al., 2016; Bentivogli et al., 2018). In PE-based evaluation, a human annotator edits the original output in places where it only differs from the reference due to linguistic variation, i.e. the output is an alternative translation that is also correct. The post-edited sentence is then evaluated with automatic metrics in the same manner as above. The benefit of PE is that it more accurately measures translation quality by penalizing only true errors made by the system while allowing for variation in translation. The disadvantage is that a human annotator needs to manually go through the system output. For this reason, we only report PE results for one *cascade* and one *pseudo-label* system. We believe that the improvement in evaluation metrics due to PE can be generalized to the rest of the results as well.

5 Experiments

In the three subsections below, we report experimental results on the ASR and MT modules separately before diving into the full speech translation pipeline.

5.1 ASR

Table 2 shows the ASR results for various configurations as described in section 4. We note several observations from these experiments.

Effect of Data Size First, comparing configurations A and B, we see that error rates decrease dramatically with the addition of TAT data. This is probably because four additional speakers are included to the training set, despite providing only 2.7hr of speech. When synthetic speech augmentation is used (expanding SuiSiann by 4x), error rates can be further reduced when only spectral features are used (B and F). However, with a more powerful feature extractor such as XLSR, the effect of synthetic speech augmentation diminishes, as seen in the roughly equal performances of E and G.

SSLR Front Ends In configurations C and E, we see that XLSR (317M parameters) outperforms TERA (23M parameters) consistently across error metrics. This is reasonable since XLSR is a much larger model, but the improvement in accuracy comes at a cost of a much longer training time (roughly 3x). Comparing C and D, it appears that SSLR feature fusion does not outperform SSLR feature only, as suggested by previous work (Anonymous, 2021), at least for the TERA front end. Experiments with XLSR was not done due to time constraints.

Different Error Rates In most cases, improvement in one error rate metric corresponds to improvement in another, although exceptions do occur (e.g. E has lower atonal SER, but G has lower WER). Atonal SER is 20-30% lower than tonal SER, suggesting that roughly 20% of the syllable errors are mistakes in tone. This is an expected result due to the complex tone sandhi system in Taiwanese (see subsection 6.1).

LM during Decoding Because our dataset is small, whether or not the trained language model is used during decoding does not alter the result significantly. When only SuiSiann is used (configuration A), using LM in decoding actually degrades performance.

5.2 MT

We train the MT model with Fairseq (Ott et al., 2019), using the a BPE SentencePiece model and the `transformer_iwslt_de_en` architecture. We experimented with only one model configuration, since the output appears very reasonable to the authors, who are all proficient in Mandarin. We evaluate performance on the iCorpus test set (same domain as training data) and the SuiSiann test set, where the Mandarin transcriptions are created by Taiwanese native speakers that we recruited (see subsection 3.1. With SuiSiann, we further segmented the sentences by commas and other punctuation marks in order to improve the translation quality, since the iCorpus training set does not contain many intra-sentence punctuation marks.

Table 3 shows the results. With the same model, we observe that the performance is much higher on iCorpus because it is in-domain with the training set. By translating the transcriptions of SuiSiann, this performance is equivalent to a cascade model with a perfect ASR, so it is an upper bound of the cascade performance. The domain gap here is a problem that should be addressed to further improve the MT performance.

5.3 Full ST Pipeline

Now we experiment with the two methods of speech translation. In *cascade*, we combine each ASR configuration in Table 2 with the MT model. In *pseudo-label*, we experiment with a high performing configuration (spectral + TERA with speech aug) that is also relatively fast to train. Incorporating the feedback from the poster session, we also initialize the encoder weights with the trained ASR model to speed up training and further improve the results.

Table 4 shows the comparison of all configurations. The results show that, in their best configurations, cascade outperforms direct training with

Config	Training Data	Speech Aug	Features	LM decode	WER	CER	SER	SER (atonal)
A	SuiSiann	✗	Spectral	✓	82.9	29.5	63.5	51.9
				✗	81.2	27.5	62.9	49.7
B	SuiSiann + TAT	✗	Spectral	✓	38.3	8.8	20.1	14.2
				✗	38.1	8.7	20.1	14.0
C			Spectral + TERA	✓	34.9	7.5	17.8	12.3
				✗	35.1	7.6	17.9	12.6
D			TERA	✓	34.3	7.3	17.3	11.8
				✗	34.9	7.3	17.2	11.6
E			Spectral + XLSR	✓	33.8	**6.8**	15.5	**10.3**
				✗	33.5	**6.8**	15.4	**10.3**
F		✓	Spectral	✓	31.4	7.7	17.1	13.0
				✗	31.2	7.5	16.7	12.6
G			Spectral + XLSR	✓	**30.7**	6.9	**15.1**	11.0
				✗	31.3	**6.8**	**15.1**	10.9

Table 2: Performance of various ASR configurations on the SuiSiann test set.

test sets	BLEU	COMET
iCorpus	56.61	0.7744
SuiSiann	38.70	-0.5672

Table 3: Performance of MT model on the two test sets.

pseudo labels. Moreover, we find that the ranking of ASR performance does not match ranking of cascade performance (e.g. TERA + spectral beats spectral only in ASR but not in cascade.) The cascade methods achieve a performance that is close to the ceiling of 38.70 BLEU points, suggesting that more emphasis should be placed on improving the MT model in future work. We did not have much time to run multiple configurations of *pseudo-label*, but we suspect that the performance could be improved with more tuning. Finally, we find that post-editing does not improve the performance dramatically similar to the findings by (Bentivogli et al., 2021), so most of the errors are true model errors instead of minor linguistic differences.

6 Analysis of Model Errors

6.1 ASR

The cascade approach produces intermediate outputs that allows us to identify ASR errors. Substitution errors accounted for a majority of these. A specific breakdown of the type of the errors in configuration C on the test set is shown in Figure 2. We manually classified the errors and observed that most of the errors are, in order from most frequent to least frequent, tone errors (assigning the wrong tone to an otherwise correct syllable), punctuation errors (substituting a comma for a period or forgetting to add a comma, for instance), syllable errors (predicting the wrong graphemes in the romanization), and code-switching or loanwords.

Tone Errors Among the 128 tone errors, 55 (43%) were manually identified to be related to tone sandhi in that the ASR module is mistakenly applying a tone sandhi rule on the reference tone to arrive at the hypothesized tone or vice versa. Taiwanese tone sandhi requires speakers to change the base tone of a syllable according to rules shown in Figure 4 whenever it does not occur in the final position of a phrase (Cheng, 1968), which is frequent. [3]. For more details, refer to Appendix A.

The ASR module's tone labeling mistakes resemble those of the speakers labeled as L1 "attriters" in Chang (2012), native Taiwanese speakers who mostly use Mandarin. The author found that when a word appears in a sandhi-free context, when the base form should be used, the attriters apply the tone sandhi rule to change the word from

[3] tone sandhi chart from https://commons.wikimedia.org/wiki/File:Taiwanese_Hokkien_tones.svg

Pipeline	ASR Init	Speech Aug	Features	CER	BLEU	COMET
cascade	✗	✗	Spectral	0.5540	33.89	-0.7361
cascade	✗	✗	TERA	0.5582	33.10	-0.7606
cascade	✗	✗	Spectral + TERA	0.5821	32.47	-0.8085
cascade	✗	✗	Spectral+ XLSR	0.5490	34.89	-0.7334
cascade	✗	✓	Spectral	0.5629	33.72	-0.7826
cascade	✗	✓	Spectral + TERA	0.5418	35.11	-0.7120
+ post-edit				0.5389	36.33	-0.7044
pseudo-label	✗	✓	Spectral + TERA	0.5700	29.85	-0.8346
+ post-edit				0.5663	31.69	-0.8346
pseudo-label	✓	✓	Spectral + TERA	0.5604	31.86	-0.7857

Table 4: Performance of various full pipeline configurations on the SuiSiann test set.

base form to its tone sandhi form when it is unnecessary. Both these speakers and the ASR model did not know or at least could not correctly reproduce what the base tone was, possibly due to the tone sandhi version being more frequent. The takeaway is that tone sandhi systems similar to Taiwanese that apply to almost every character in a sentence present a challenge for ASR models.

Code-switching and Loan Words In Figure 3, we show our errors in code-switching and loan words. Out of the 172 sentences, there are a total of 31 English words and 9 Japanese loan words (e.g., mài-kù is from the Japanese word マイクロフォン *maikurofon*). Our model has 45.5 WER with Japanese loan words and 41.9 WER with English words, compared to the 34 WER on average. In the Japanese loan words, the mistakes were mainly tone errors (correct consonant and vowels but wrong accent), and transcribing the wrong consonants (p to b, m to n). The consonants that the machine confuses with each other are phonetically similar: [p] and [b] are both bilabial plosives (differing only in voicing), while [m] and [n] are both nasals. In the English words, the words our model output correct were the countries (e.g., Sudan), American president names (e.g., Obama, Clinton) and the single English characters (e.g., A, B). Single English characters are less complicated thus are easier to transcribe. American president names and country names are reasonable because our training data includes news articles and these are common words in news. Words outside of the news-domain are all transcribed incorrectly, which may be because these words were never seen by the model. Some words that can't be captured by the model include Ubuntu, Windows, H7N9, and U-Bike.

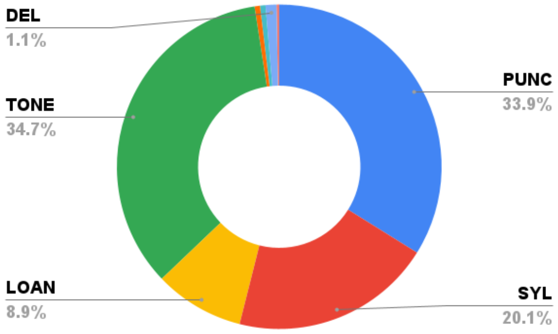

Figure 2: Types of errors in the ASR SSLR TERA fused model

6.2 MT

As discussed earlier, one problem with the cascade approach is that errors from the ASR propagate into the MT. Indeed, the ASR module's tone labeling errors discussed in the previous section propagate into the MT module, regardless of whether or not the tone error is a misapplication of tone sandhi rules or not. In tonal languages like Taiwanese, tone is contrastive, meaning that two syllables with the same phonemes but different tone will have different meanings (Chang, 2012). A manual inspection of the final outputs reveals that when ASR outputs the correct syllable but the wrong tone, the wrong Mandarin word is chosen during translation.

7 Conclusions and Future Work

Our results show that adding SSLR features and adding pitch shifting improves our ASR and overall ST performance. Our experiment also shows that when dealing with syntactically similar lan-

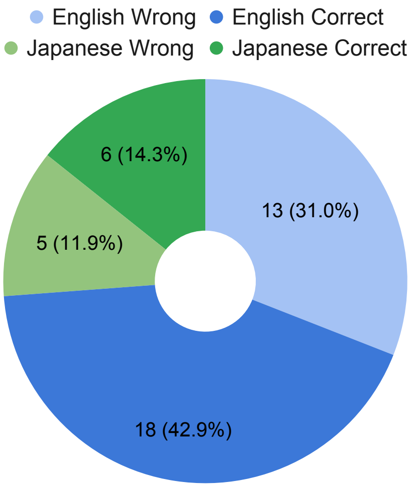

Figure 3: Correctness of English code-switching and Japanese loan words in test set.

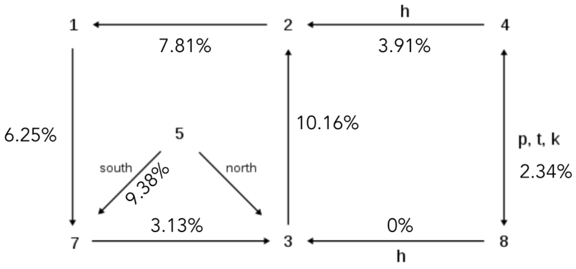

Figure 4: Tone sandhi errors.

guages in speech translation, we could treat it as an ASR task, which could also apply to other Sinitic languages.

We are aware that our results are constrained by the small size of our dataset, and that larger resources exist, e.g., the full TAT Corpus, Formosa Speech Database (ForSDat) (Lyu et al., 2004). Utilizing all the available resources could be done in the future to increase the performance as well as the robustness of our system. The improvement that pitch shifting brought shows us that other speech augmentation techniques could also be experimented with in the future.

Our cascade model suffers from error propagation due to the domain mismatch in our training data for ASR model and MT model. One way to mitigate this problem is to feed the MT model with ASR top-n or lattices, which are more informative than top-1 output (Lavie et al., 1996; Matusov et al., 2005; Rei et al., 2020). Another way is to corrupt MT training data to emulate ASR errors (Sperber et al., 2017; Gangi et al., 2019; Cheng et al., 2019).

References

Anonymous. 2021. Two front-ends, one model : Fusing heterogeneous speech features for low resource asr with multilingual pre-training.

Parnia Bahar, Tobias Bieschke, Ralf Schlüter, and Hermann Ney. 2021. Tight integrated end-to-end training for cascaded speech translation. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 950–957.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine Translation Human Evaluation: an investigation of evaluation based on Post-Editing and its relation with Direct Assessment. In *Proceedings of the International Conference on Spoken Language Translation (IWSLT)*, Bruges, Belgium.

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus direct speech translation: Do the differences still make a difference? 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 2873–2887, Online. Association for Computational Linguistics.

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*, Barcelona, Spain.

Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Barry Haddow, Matthias Huck, Chris Hokamp, Philipp Koehn, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Carolina Scarton, Lucia Specia, and Marco Turchi. 2015. Findings of the 2015 Workshop on Statistical Machine Translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, Lisbon, Portugal.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2013. Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Heidelberg, Germany.

Yufen Chang. 2012. First language attrition: An investigation of taiwanese tones and tone sandhi. doctoral dissertation.

Meng Chang Chen. 2016. icorpus dataset.

Qiao Cheng, Meiyuan Fang, Yaqian Han, Jin Huang, and Yitao Duan. 2019. Breaking the data barrier: Towards robust speech translation via adversarial stability training.

Robert L. Cheng. 1968. Tone sandhi in taiwanese. *Linguistics: An Interdisciplinary Journal of the Language Sciences*, 6(41):19–42.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised Cross-Lingual Representation Learning for Speech Recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Michael Denkowski and Alon Lavie. 2010. Choosing the right evaluation for machine translation: An examination of annotator and automatic metric performance on human judgment tasks. In *Proceedings of the Conference of the Association of Machine Translation in the Americas (AMTA)*, Denver, US-CO.

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.

Directorate-General of Budget, Accounting and Statistics, Taiwan. 2020. Taiwan 2020 population and housing census.

M. F. M. Esa, Noorfa Haszlinna Mustaffa, Hazim Omar, N. H. M Radzi, and Roselina Sallehuddin. 2020. Learning convolution neural network with shift pitching based data augmentation for vibration analysis.

Mattia Antonino Di Gangi, Robert Enyedi, Alessandra Brusadin, and Marcello Federico. 2019. Robust neural machine translation for clean and noisy speech transcripts. *CoRR*, abs/1910.10238.

Yvette Graham, Timothy Baldwin, Meghan Dowling, Maria Eskevich, Teresa Lynn, and Lamia Tounsi. 2016. Is all that Glitters in Machine Translation Quality Estimation really Gold? In *Proceedings of the International Conference on Computational Linguistics (COLING)*, pages 3124–3134, Osaka, Japan.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Yalta, Tomoki Hayashi, and Shinji Watanabe. 2020. ESPnet-ST: All-in-one speech translation toolkit. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 302–311, Online. Association for Computational Linguistics.

Sathish Indurthi, Mohd Abbas Zaidi, Nikhil Kumar Lakumarapu, Beomseok Lee, Hyojung Han, Seokchan Ahn, Sangha Kim, Chanwoo Kim, and Inchul Hwang. 2021. Task aware multi-task learning for speech to text tasks. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7723–7727.

Alon Lavie, Donna Gates, Marsal Gavalda, Laura Mayfield, Alex Waibel, and Lori Levin. 1996. Multilingual translation of spontaneously spoken language in a limited domain. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2020. Dual-decoder transformer for joint automatic speech recognition and multilingual speech translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3520–3533, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Hang Le, Juan Pino, Changhan Wang, Jiatao Gu, Didier Schwab, and Laurent Besacier. 2021. Lightweight adapter tuning for multilingual speech translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 817–824, Online. Association for Computational Linguistics.

Hong-Bin Liang, Chung-Yi Li, and Hung-Yi Lee. 2021. The ntu asr system for formosa speech recognition challenge 2020. In *Speech Signal Processing Workshop*.

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

Andy T. Liu, Shang-Wen Li, and Hung yi Lee. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

Ren-Yuan Lyu, Min-Siong Liang, and 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*, pages 1–12.

Evgeny Matusov, Hermann Ney, and Ralph Schluter. 2005. Phrase-based translation of speech recognizer word lattices using loglinear model combination. In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pages 110–115, San Juan, Puerto Rico.

Ministry of Education, Taiwan. 2014. A 700 taiwanese word list in recommended han characters.

Ministry of Health and Welfare, Taiwan. 2017. Taiwan 2017 report of the senior citizen condition survey.

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*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Jan Schlüter and Thomas Grill. 2015. Exploring data augmentation for improved singing voice detection with neural networks. In *ISMIR*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the Conference of the Association for Machine Translation of the Americas (AMTA)*, pages 223–231, Cambridge, US-MA.

Matthew Snover, Nitin Madnani, Bonnie J Dorr, and Richard Schwartz. 2009. Fluency, Adequacy, or HTER?: exploring different human judgments with a tunable MT metric. In *Proceedings of the Workshop on Statistical Machine Translation (WMT)*, pages 259–268, Athens, Greece.

Matthias Sperber, Jan Niehues, and Alex Waibel. 2017. Toward robust neural machine translation for noisy input sequences. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 90–96, Tokyo, Japan. International Workshop on Spoken Language Translation.

Fred W.M. Stentiford and Martin G. Steer. 1988. Machine translation of speech. *British Telecom Technology Journal*, 6(2):116–122.

SuíSiann Dataset (version 0.2.1). 2021. https://suisiann-dataset.ithuan.tw/.

Taiwanese Across Taiwan (TAT) Corpus-Vol2 Samples. 2021. https://drive.google.com/drive/folders/10mme5o3-tm$_b us8b9u - wsgwqcjj3g9x-$.

Alex Waibel, Ajay N Jain, Arthur E McNair, Hiroaki Saito, Alexander G Hauptmann, and Joe Tebelskis. 1991. Janus: a speech-to-speech translation system using connectionist and symbolic processing strategies. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 793–796, Toronto, Canada.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. ESPnet: End-to-End Speech Processing Toolkit. In *Proc. Interspeech 2018*, pages 2207–2211.

Ron J. Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Z. Chen. 2017. Sequence-to-sequence models can directly translate foreign speech. In *INTERSPEECH*.

A Tone sandhi analysis

The base tone is the one used in the romanization, so the reference uses the base tone. Each directed edge in Figure 4 represents one tone sandhi rule where the source indicates the base tone; the destination indicates the tone sandhi version. Each edge is labeled with the percentage of the tone errors related to that rule. We do not display tone alternations that are not valid tone sandhi combinations. We count both directions of the edge. For instance, if the hypothesized tone is 2 when the reference is 3, in the direction of the directed edge, the model has indeed heard tone 2 (a correct application of tone sandhi by the speaker) but has not learned that the base form, the one that should be used in the transcription, is 3. In the other direction, if the model hypothesizes tone 3 when the reference was 2, going in the opposite direction of the tone sandhi arrow, it is because the model properly learned identified that the base form is actually 2 (realized as tone 1 perhaps) but was hypercorrecting and applied tone sandhi to change it to tone 3. We count both cases (and both directions of the tone sandhi edge) because the model's mistake appears to be tied to the same tone sandhi rule.

B Contributions

Chenxuan:

- Pre-processed iCorpus Data and trained MT model with Fairseq.
- Trained pseudo-label models using ESPNet
- Wrote post-processing scripts to generate cascade results

Noel:

- Pre-processed SuiSiann dataset into KALDI format
- Organized data preparation scripts and trained ASR models in ESPNet
- Added speech augmentation
- Added SER and SER(atonal)scripts for ASR evaluation

Kalvin:

- Pre-processed the TAT-Vol2 samples into KALDI format
- Converted tones in tâi-lô to number form with a Python script
- Conducted Literature review into linguistic knowledge for Taiwanese (for tone analysis) and related works for speech translation.