Using Taigi Dramas with Mandarin Chinese Subtitles to Improve Taigi Speech Recognition

Pin-Yuan Chen, Chia-Hua Wu, Hung-Shin Lee, Shao-Kang Tsao, Ming-Tat Ko, Hsin-Min Wang
Institute of Information Science, Academia Sinica, Taiwan
{pinyuan615, hungshinlee}@gmail.com, {mtko, whm}@iis.sinica.edu.tw

***Abstract*—An obvious problem with automatic speech recognition (ASR) for Taigi is that the amount of training data is far from enough to build a practical ASR system. Collecting speech data with reliable transcripts for training the acoustic model (AM) is feasible but expensive. Moreover, text data used for language model (LM) training is extremely scarce and difficult to collect because Taigi is a spoken language, not a commonly used written language. Interestingly, the subtitles of Taigi drama in Taiwan have long been in Chinese characters for Mandarin. Since a large amount of Taigi drama episodes with Mandarin Chinese subtitles are available on YouTube, we propose a method to augment the training data for AM and LM of Taigi ASR. The idea is to use an initial Taigi ASR system to convert a Mandarin Chinese subtitle into the most likely Taigi word sequence by referring to the speech. Experimental results show that our ASR system can be remarkably improved by such training data augmentation.**

***Index Terms*—Taigi, Mandarin Chinese, speech recognition, ASR, LVCSR, data augmentation**

I. Introduction

Southern Min (Minnan or Hokkien) is a part of the Sino-Tibetan language family. Although Southern Min is often regarded as a Chinese dialect, it can also be regarded as an independent language since it is not mutually intelligible with Mandarin or other Chinese languages, such as Cantonese and Shanghainese. Southern Min is formed in the Minnan region (southern Fujian province of China) along the southeast coast of China, and is mainly distributed in mainland China (e.g., Fujian and Guangdong), Taiwan, and southeast Asian countries such as the Philippines, Singapore, Malaysia, and Indonesia. In 2019, there were approximately 48 million native speakers[1].

Taigi, also known as Taiwanese Minnan, Taiwanese Hokkien, Holo, etc., is a variant of Southern Min spoken in Taiwan. According to the 2009 Yearbook of Republic of China, about 73% of Taiwanese can speak Taigi. Since Mandarin has long been the official language of Taiwan and Taigi was suppressed to use in public for some time in the past, more and more people have given up using Taigi in communication, which has led to the loss of Taigi. According to the United Nations (Language Vitality Assessment, LVA) standard, the generational succession of Taigi is between "significant danger" of the third level and "unsafe" of the fourth level, and is closer to the third level [1]–[3]. In recent decades, although Taiwan native languages learning has been included in the compulsory courses of school education for promoting the mother tongue movement to save the loss of native languages, the overall improvement is still very limited.

In Taiwan, the development of automatic speech recognition (ASR) for Mandarin Chinese has a long history, but there is not much research on Taigi ASR. Past research can be roughly divided into two directions, one is collecting Taigi corpora [4]–[11], and the other is building Taigi ASR systems [12]–[18]. Taigi ASR research began in the late 1990s [12]–[14]. Later, there was research on Mandarin/Taigi code-switching ASR [16], [18], with a goal to deal with the situation where a small amount of Taigi is mixed in Mandarin utterances. Due to lack of Taigi training data, these Taigi or Mandarin-Taigi ASR systems are far from practical applications due to high error rates.

The big challenge in Taigi ASR is due to non-unified writing systems and lack of publicly available Taigi speech and text corpora. There are two categories of writing systems for Taigi, "Pinyin" and "Han-Luo". Pinyin was originally used to spell or transcribe Taigi speech. A representative writing system in this category is Tailo. Han-Luo is a hybrid of Han characters (Chinese characters) and Romanization. If a Taigi character cannot be represented by a suitable Chinese character, use Tailo instead. The Ministry of Education of Taiwan has defined a dictionary of common Minnan[2] to promote the Han-Luo writing system. However, most people are not familiar with any Taigi writing system because they did not formally learn to write Taigi in the school. This is the main reason why Taigi texts are extremely scarce. On the other hand, Taigi speech data is not that difficult to collect, but manual transcription is time consuming and expensive.

We believe that Taigi drama can be a good resource for Taigi ASR training data. Interestingly, the subtitles of Taigi dramas in Taiwan have long been in Chinese characters for Mandarin. Such subtitles allow people who understand Mandarin but do not understand Taigi to watch Taigi dramas. Since a large amount of Taigi dramas with Mandarin Chinese subtitles are available on YouTube, we propose a method to augment the training

[1] https://www.ethnologue.com/statistics/summary-language-size-19

[2] https://twblg.dict.edu.tw/holodict_new/

data for the acoustic model (AM) and language model (LM) of Taigi ASR. The idea is to use an initial Taigi ASR system to automatically convert the Mandarin Chinese subtitles of Taigi dramas into Taigi Chinese transcripts by referring to the speech. Our experimental results show that by using the best portion of automatically generated utterance-transcript pairs to enhance AM and LM, the ASR system can be remarkably improved.

This paper is organized as follows. Section II presents our baseline Taigi ASR System. Section III explains the proposed training data augmentation method. The experimental results are discussed in Section IV. Section V gives the conclusions and future work.

II. Our Baseline Taigi ASR System

Our baseline Taigi ASR system was developed by using the Kaldi toolkit. In Kaldi, a standard recipe for GMM-DNN-HMM-based ASR generally consists of three steps. In the first step of preparation, users must provide a pronunciation lexicon and LM corresponding to the target language or output text, and specify the speech utterances and their corresponding transcripts for training and evaluation. In the second step, based on the defined speech units, the GMM-HMM system is initialized by monophone model training and extended to a context-dependent triphone system. In the last step, a more accurate DNN-based AM is developed based on the frame-level triphone alignment provided by the GMM-HMM system. The DNN-based AM is suitable for integrating higher-resolution acoustic features and i-vectors containing speaker-related information [19].

A. Pronunciation lexicon

For future convenience of developing a code-switching or multi-lingual ASR system, International Phonetic Alphabet (IPA) is used for defining speech units in our pronunciation lexicon. The word entries were collected from different sources, such as Taiwanese Dictionary of Common Minnan compiled by the Ministry of Education of Taiwan (MOE)[2], iTaigi[3], and the manual transcriptions of the speech data we collected. The number of unique words is 39,863, while the number of word entries is 146,608, since a word might have pronunciation variations, tone variations, and even completely different pronunciations. Taigi is a tonal language with extensive tone sandhi rules. Taigi syllables consist maximally of an initial consonant, a vowel, a final consonant, and a tone. Strictly speaking, there are only five tonal contours. But as in other Chinese languages, the two kinds of stopped syllables are considered two kinds of tones. In the pronunciation lexicon, the pronunciation of each word entry is annotated with syllables in canonical tones and also in colloquial tones based on the tone sandhi rules. The former is consistent with the romanized Tailo writing system, while the latter is used to model the spoken form. The number of unique syllables with tones (canonical) and without tones are 4,404 and 993, respectively. On average, each word has 3.6 ways of pronunciation in our pronunciation lexicon.

B. Acoustic model

The CGUSP [4] and TWISAS [17] databases were used to train the AM. CGUSP contains about 32 hours of Taigi read speech contributed by 600 speakers. The utterances are short phrases or isolated words. TWISAS consists of about 132 hours of speech collected from TV talk shows, news, and dramas in Taigi. About 30 hours have word-level transcription in Taigi Chinese characters and Tailo, and the remaining 102 hours only have phonetic transcription in Tailo. The specifications of front-end acoustic features, i-vectors, model structures, and training parameters follow the Kaldi recipe of `multi_cn`. The DNN-based AM was based on the CNN-TDNNF architecture. SpecAugment was used to augment the acoustic training data [20]. The lattice-free maximum mutual information (LF-MMI) criterion was applied in model training [21].

C. Language model

The trigram LM was trained by the SRILM toolkit[4] using Kneser-Ney smoothing on the text of manual transcription in Chinese characters of 30 hours of speech in the TWISAS database. There are only 417,032 words (656,261 characters). The vocabulary is limited to the 39,864 words in the pronunciation lexicon introduced in Section II-A. There is no out-of-vocabulary (OOV) issue because we have added all words in the lexicon in advance.

III. Augmenting Training Data with Taigi Dramas

Our idea is to use an initial Taigi ASR system to convert a Mandarin Chinese subtitle of Taigi drama into a Taigi Chinese transcript by referring to the speech. To do so, we need a Taigi-Mandarin bilingual dictionary. By referring to the dictionary, a Mandarin Chinese subtitle can be converted into a Taigi word graph by replacing each Mandarin Chinese word with the possible translated Taigi Chinese words. Then, the Taigi word graph is further converted into a Taigi word/phone graph by associating each word with possible pronunciations in IPA phone sequences. Finally, the AM of our baseline Taigi ASR system is used to find the path in the Taigi word/phone graph that aligns the speech segment best. In this way, the Mandarin Chinese subtitle can be converted into the most likely Taigi Chinese transcript and IPA phone sequence. The process will be described in detail later in this section.

A. Preparation beforehand

1) Speech in Taigi drama: Thousands of Taigi drama episodes with manually added Mandarin Chinese subtitles can be found on YouTube. We collected 9 TV series,

[3] https://itaigi.tw

[4] http://www.speech.sri.com/projects/srilm

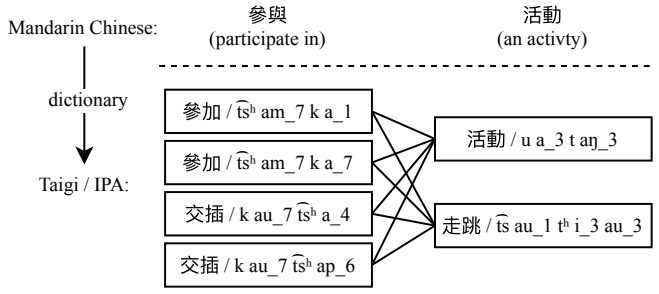


Fig. 1. An example Taigi word/phone graph generated from a Mandarin Chinese word sequence. In the IPA sequence, each vowel contains an IPA symbol followed by a tone code.

totaling about 900 hours. The audio track of each video clip was divided into shorter utterances based on the time code of the close caption. Some time codes do not align text and speech perfectly.

2) Taigi-Mandarin bilingual dictionary: The dictionary used in this work is not only a pronunciation lexicon introduced in Section II-A. Each word entry in the dictionary includes a Taigi word in Chinese characters, its pronunciation in IPA, and its synonyms in Mandarin Chinese. Therefore, each Mandarin Chinese word in a subtitle can be translated into multiple possible Taigi Chinese words, thereby converting the subtitle into a Taigi word graph.

3) Initial acoustic model: The AM of our baseline Taigi ASR system can be used to find the path in the word/phone graph that aligns the speech segment best. We can compute the average of the frame log-likelihood scores as a confidence score for data selection. That is, the model can measure the degree of coincidence between the speech segment and the most likely phone sequence (and word sequence) to some extent.

B. Word/Phone graph construction

The missing piece to augment training data with Taigi dramas is accurate Taigi transcriptions. One intuitive method is to train a machine translation system to translate a subtitle in Mandarin Chinese into possible Taigi Chinese transcripts. However, the machine translation system is not available due to lack of Taigi-Mandarin parallel text corpus. What we do in this work is to simply translate each Mandarin Chinese word in the subtitle into possible Taigi Chinese words. In this way, we can convert the Mandarin Chinese subtitle into a Taigi Chinese word graph. To do so, we only need a Taigi-Mandarin bilingual dictionary, which has been constructed by adding Mandarin Chinese synonym words to each Taigi word in our lexicon. We used Jieba [5] with our lexicon to tokenize Mandarin Chinese text into words. As mentioned in Section II-A, each Taigi word has 3.6 types of pronunciation on average in our lexicon. Therefore, the Taigi word graph can be further extended to a Tagi word/phone graph by associating each word with possible pronunciations in IPA phone sequences. A simple example Taigi word/phone graph generated in this way is illustrated in Fig. 1.

[5] https://github.com/fxsjy/jieba

Algorithm 1: Generation of the Taigi Chinese word sequence conditioned on the speech and Mandarin Chinese word sequence.

Input: G: Taigi word/phone graph; l: sequence length (number of words); p_0: initial word sequence; S: speech
Output: optimal word sequence p^*; optimal confidence score s^*

```
for position t=1 to l do
    for each candidate c ∈ G_t do
        p = p_0[0 : t − 1] + c + p_0[t + 1 : l + 1]
        s = ForcedAlignment(S, p)
        if s > s* then
            s* = s
            c* = c
        end if
    end for
    p* += candidate c*
end for
s* = ForcedAlignment(S, p*)
```

C. Forced alignment

We used a pre-trained AM to align a speech utterance with a given phone sequence and obtain a confidence score. The main idea is to find out the best phone sequence, which has the highest score among all possible phone sequences in the word/phone graph.

A greedy algorithm was applied to find the best phone sequence. For computational efficiency, we neither perform forced alignment on all possible phone sequences in the word/phone graph nor perform speech recognition under the constraint of the word/phone graph. Instead, the best phone sequence in each word position is independently found out in order, given that the phone sequences in other word positions are arbitrarily selected and fixed. As a result, each word position of a sentence is sequentially filled with a Taigi word with a relatively high confidence score. The detailed procedure is shown in Algorithm 1. The initial word sequence p_0 is the concatenation of the phone sequence of the first word candidate at each word position in the word/phone graph. Its length is $l+2$ for a sentence with l words. The start and end of the sentence are marked with symbols $< s >$ and $< /s >$, respectively. Therefore, $p_0[0:0] =< s >$ while $p_0[l+1:l+1] =< /s >$. Finally, the Taigi word sequence (with the corresponding phone sequence) of each utterance derived in this way is used to perform forced alignment to produce a more accurate log-likelihood score for the utterance-transcript pair.

It is worth noting that different kinds of tonal variations often occur in spoken Taigi. For simplicity, only one tonal variation of a Taigi word is randomly selected as the repre-

sentative. We believe that the phonetic difference is more important than the tonal difference to distinguish different Taigi words. In fact, for a phone, the acoustic scores between its tonal variants are only slightly different, since the tonal information is incorporated into the phonetic decision tree. For instance, the Mandarin Chinese word "參與" corresponds to 2 Taigi words "參加" (/ t͡sʰ am_7 k a_1/ and /t͡sʰ am_7 k a_7/) and "交插" (/k au_7 t͡sʰ a_4/, /k au_7 t͡sʰ ap_6/ and /k au_7 t͡sʰ ap_8/). The 2 pronunciation variations of "參加" only differ in tones. We will randomly select one for evaluation. For "交插", we will also randomly select one of the two pronunciations /k au_7 t͡sʰ ap_6/ and /k au_7 t͡sʰ ap_8/.

D. Interface for Taigi transcription correction

The score associated with an utterance-transcript pair can help distinguish not only the recording quality of the speech utterance but also the correctness of the automatically transcribed Taigi word sequence with the speech content. The utterance-transcript pairs with higher confidence scores can be used as the AM training data, while the text can be used to train the LM. In addition, the pairs with higher scores are easier to correct by human transcribers. Therefore, we designed a web-based interface, so that human transcribers can conveniently and effectively correct the automatically generated Taigi word transcription. As shown in Fig. 2, the interface displays the Chinese word sequences in both Mandarin and Taigi, and the confidence score. The audio player can play the previous, current, and next speech utterances. The transcribers can obtain contextual information by playing the previous and next utterances, which help them confirm the content of the current utterance, and is also useful for judging the correctness of segmentation (which is based on the time code of the subtitle). The transcribers can edit the Mandarin Chinese subtitle and the automatically generated Taigi Chinese transcript, and give some special tags to each utterance, such as "trimmed position", "discarded", and "to be checked". As mentioned above, it is more efficient to correct transcripts with fewer errors. Therefore, the interface can provide transcribers with utterance-transcript pairs in descending order of confidence score. With this function, the transcribers can check and correct the transcripts of utterance-transcript pairs with high scores first, stop at any utterance-transcript pair with a lower score, and switch to another drama episode. We found that this can indeed improve the efficiency of corpus annotation, thereby getting more annotated utterances in the same period. The confirmed and corrected results will be stored in the database.

IV. Experiments

Several ASR experiments were performed to evaluate different versions of training data in terms of the character error rate (CER) with the fixed network structure



Fig. 2. The interface for correcting the automatically generated Taigi transcription.

introduced in Section II. In this section, we first describe the databases, and then present the experimental results.

A. Data

As described in Section II, the AM of our baseline Taigi ASR system was trained with the CGUSP and TWISAS databases. The LM was trained with the Taigi transcriptions in Chinese characters of 30 hours of speech in the TWISAS database. In this work, we propose using Taigi dramas with Mandarin Chinese subtitles to augment the training data of AM and LM. A total of 900 hours of Taigi drama episodes from 9 TV series were collected. The audio track of each drama episode was extracted, and each speech segment was aligned with the most likely Taigi word and phone sequences converted from the Mandarin Chinese subtitle based on the process introduced in Section III. We selected 2 TV series to form the test sets. The automatically generated transcriptions were manually corrected via the interface shown in Fig. 2. Test set A consists of 0.85 hours of speech (1,819 utterances / 8,591 words / 13,365 characters) from one TV series. Test set B consists of 1.74 hours of speech (2,870 utterances / 15,298 words / 25,905 characters) from another TV series. These two TV series were chosen as the test data because they have less available data. The data from the remaining 7 TV series were used as the training data. There are about 870 hours of speech in total. The automatically generated transcriptions of 19.59 hours of speech were manually corrected. This data set is denoted as Manual-tr. The counterpart data set without manual correction is denoted as Auto-tr. In addition, we sorted the utterance-transcript pairs from the drama episodes not involved in the Manual-tr set in descending order according to their confidence scores, and selected the top 20 and 40 hours of speech to form two training sets: Auto-20h and Auto-40h. This method of data selection is for fair comparison of performance with the Manual-tr training set. The Manual-tr and Auto-20h sets are roughly the same size. Furthermore, we also selected the automatically

TABLE I
STATISTICS OF TRAINING SPEECH FOR ACOUSTIC MODELING.

Data	#Utterances	#Hours
CGUSP	116,461	32.44
TWISAS	199,591	132.23
Manual-tr	34,903	19.59
Auto-tr	34,903	19.59
Auto-20h	37,910	20.00
Auto-40h	71,114	40.00

TABLE II
STATISTICS OF TRAINING TEXTS FOR LANGUAGE MODELING.

Data	#Words	#Characters
TWISAS$_t$	417,032	656,261
Manual-tr$_t$	178,763	293,086
Auto-tr$_t$	152,595	237,474
Auto-20h$_t$	164,632	267,732
Auto-20%$_t$	1,283,724	2,039,424
Auto-40%$_t$	2,607,646	4,096,830
Auto-60%$_t$	3,906,011	6,087,558
Auto-80%$_t$	5,140,971	7,955,597
Auto-100%$_t$	6,203,256	9,533,479

generated transcriptions of the top 20% to 100% pairs to form the LM training sets, namely Auto-20%$_t$, Auto-40%$_t$, ..., and Auto-100%$_t$. Note that for each training set, the speech data with transcriptions were used to train the AM, while the transcriptions in Chinese characters were used to train the LM. But Auto-x%$_t$ was only used to train the LM. The summary of the training data is presented in Tables I and II.

B. Effect of adding manually transcribed training data

In this experiment, we compared three systems, namely the baseline system, the system trained with more LM training data (+Manual-tr$_t$), and the system trained with more training data for AM (+Manual-tr) and LM (+Manual-tr$_t$). As shown in Table III, it is clear that the CER can be reduced by adding more LM training data (cf. columns 3 and 2, A0L1 vs. Baseline), and can be further reduced by adding more AM training data (cf. columns 4 and 3, A1L1 vs. A0L1). The result confirms the effectiveness of adding training data manually transcribed from dramas downloaded from YouTube. The CERs of Test A are always higher than those of Test B because the former contains more emotional speech and has a higher speaking rate.

C. Effect of adding automatically transcribed training data

In this experiment, we compared systems trained by adding automatically transcribed training data based on different data selection strategies. The results are shown in Table III. Comparing columns 5 and 4 (A2L2 vs. A1L1), we can see that the manually corrected training data is indeed more useful than the automatically transcribed training data. Note that their speech content is the same, and the only difference is that the transcriptions are automatically generated (Auto-tr$_t$) or manually corrected (Manual-tr$_t$). Comparing columns 5, 4, and 2, we can see that the system trained with additional automatically transcribed training data for AM (+Auto-tr) and LM (+Auto-tr$_t$) is obviously better than the baseline system (A2L2 vs. Baseline), although it is slightly worse than the system trained with additional manually corrected training data (A2L2 vs. A1L1). This is because that human transcribers actually chose the utterance-transcript pairs with higher scores to correct them via the interface shown in Fig. 2. If we compare columns 6 and 4, columns 6 and 5, and columns 6 and 2, the results indeed confirm the effectiveness of our proposed training data augmentation method. A3L3 is better than A2L2 because Auto-20h consists of the best 20 hours of utterance-transcript pairs while the data in Auto-tr is not always selected from the best utterance-transcript pairs. So far, we have confirmed the effectiveness of augmenting training data with automatically generated and manually corrected data. Therefore, we can combine the automatically generated and manually corrected data. In this case, the A1L1 system can be regarded as a stronger baseline. Comparing columns 7 and 4 (A4L4 vs. A1L1), we can see that using additional automatically generated training data can improve performance, especially for Test A, which is a more difficult test set. However, when we further increase the amount of automatically generated training data, the performance is not improved (A5L5 vs A4L4). The result indicates that our training data augmentation method still needs to be improved. Another possible improvement direction is to add more drama episodes, from which more reliable utterance-transcript pairs can be selected.

D. Effect of adding more automatically transcribed LM training data

As shown in Table III, the A1L1 system is better than our baseline system. Therefore, in this experiment, we evaluated the effectiveness of augmenting LM training data based on the A1L1 system. We compared systems trained by adding different amounts of text in the automatically transcribed training data. The results are shown in Table IV. For Test A, we can see that as the automatically generated text used to train the LM increases, the CER continues to decrease. However, for Test B, as the automatically generated text used to train the LM increases, the CER first decreases and then increases. Overall, the result once again confirms the usefulness of text automatically generated by our data augmentation method for LM training. It is worth mentioning that compared to LM training data, we can only select a much smaller proportion of AM training data to use. When using the best 20% data to train the AM, in our preliminary experiments, the CER may not necessarily decrease, or even increase.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we have proposed a framework to augment the training data of Taigi ASR with Taigi Dramas

TABLE III
CERs WITH RESPECT TO VARIOUS COMBINATIONS OF TRAINING DATA.

Model	Data for AM	Data for LM	Test A	Test B
Baseline			49.21	42.08
A0L1		+Manual-tr$_t$	44.76	37.81
A1L1	+Manual-tr	+Manual-tr$_t$	38.11	26.95
A2L2	+Auto-tr	+Auto-tr$_t$	44.35	32.81
A3L3	+Auto-20h	+Auto-20h$_t$	36.44	32.58
A4L4	+Manual-tr +Auto-20h	+Manual-tr$_t$ +Auto-20h$_t$	34.42	**26.62**
A5L5	+Manual-tr +Auto-40h	+Manual-tr$_t$ +Auto-40h$_t$	**33.99**	27.50

TABLE IV
CERs WITH RESPECT TO VARIOUS AUTOMATICALLY TRANSCRIBED TEXTS.

Model	Data for LM	Test A	Test B
A1L1		38.11	26.95
A1L6	+Auto-20%$_t$	34.19	**26.38**
A1L7	+Auto-40%$_t$	33.80	26.75
A1L8	+Auto-60%$_t$	33.66	26.68
A1L9	+Auto-80%$_t$	33.71	26.75
A1L10	+Auto-100%$_t$	**33.36**	26.85

with Mandarin Chinese subtitles. The reason behind the proposal is twofold. The first is that the lack of Taigi speech data with reliable transcripts has caused problems for the development of a practical Taigi ASR system. Text data for language model training is particularly scarce because Taigi is a spoken language and is not widely used in written form. The second reason is that there are a large number of Taigi dramas with Mandarin Chinese subtitles on YouTube. We believe that they can be used to augment the training data, if an initial Taigi ASR system is available. Our experimental results show that by using the best portion of automatically generated utterance-transcript pairs to augment the training data of AM and LM, the ASR system can be remarkably improved.

This is our first attempt to augment the training data in this way. Due to time constraints, we cannot process all the 3,000 hours of Taigi drama episodes that have been downloaded. The results reported in the paper are based on 900 hours of Taigi dramas. In the future, we will continue to process the remaining drama episodes to enrich the training data. We will also continue to collect more data from the Internet. In addition, we will try to improve the current forced-alignment-based method and gradually apply a better AM to process the data.

VI. Acknowledgement

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grant: MOST 109-2634-F-008-006.

References

[1] 陳淑嬌, "台灣語言活力研究," in 語言政策的多元文化思考, 鄭錦全, 何大安, 蕭素英, 江敏華, and 張永利, Eds. 中央研究院語言學研究所, 2007.

[2] 葉高華, "臺灣民衆的家庭語言選擇," 臺灣社會學刊, vol. 62, pp. 59–111, 2017.

[3] 蘇鳳蘭, "在語言保存或流失的十字路口徘徊: 台灣國中、小學生台語," 臺灣語文研究, vol. 14, no. 1, pp. 81–119, 2019.

[4] R.-y. Lyu, M.-s. Liang, and Y.-c. Chiang, "Toward constructing a multilingual speech corpus for Taiwanese (Min-nan), Hakka , and Mandarin," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 9, pp. 1–12, 2004.

[5] W.-C. Kuo, Y.-R. Wang, and S.-H. Chen, "A model-based tone labeling method for Min-nan/Taiwanese speech," in *Proc. ICASSP*, 2004.

[6] D.-c. Lyu, B.-H. Yang, R.-y. Lyu, and C.-N. Hsu, "Taiwanese TV news-to-document index system," in *Proc. IEEE CNNA*, 2005.

[7] Y.-C. Chiang, M.-S. Liang, H.-Y. Lin, and R.-y. Lyu, "The multiple pronunciations in Taiwanese and the automatic transcription a of Buddhist Sutra with augmented read speech," in *Proc. Interspeech*, 2005.

[8] M.-S. Liang, R.-y. Lyu, and Y.-C. Chiang, "Using speech recognition technique for constructing a phonetically transcribed Taiwanese (Min-nan) text corpus," in *Proc. Interspeech*, 2006.

[9] ——, "Phonetic transcription using speech recognition technique considering variations in pronunciation," in *Proc. ICASSP*, 2007.

[10] C. Jun, J. Feldmar, Y. Laprie, D. Fohr, and J. P. Haton, "Transcribing Southern Min speech corpora with a web-based language learning system," in *Proc. ICALIP*, 2008.

[11] M.-T. Ko, R.-y. Lyu, T.-y. Liao, and Y.-c. Chiang, "Development of a Taiwanese speech and text corpus," in *Proc. O-COCOSDA*, 2012.

[12] R.-y. Lyu, Y.-j. Chiang, and W.-p. Hsieh, "A large-vocabulary Taiwanese (Min-nan) multi-syllabic word recognition system based upon right-context-dependent phones with state clustering by acoustic decision tree," in *Proc. ICSLP*, 1998.

[13] R.-y. Lyu, Y.-J. Chiang, R.-J. Fan, and W.-P. Hsieh, "A large-vocabulary Taiwanese (Min-nan) speech recognition system based on inter-syllabic initial-final modeling and lexicon-tree search," in *Proc. ROCLING*, 1998.

[14] R.-y. Lyu, Y.-j. Chiang, and W.-p. Hsieh, "A new hybrid duration hidden Markov model with application to large vocabulary Taiwanese (Min-nan) word recognition," in *Proc. ISCSLP*, 1998.

[15] D.-c. Lyu, M.-S. Liang, Y.-C. Chiang, C.-N. Hsu, and R.-y. Lyu, "Large vocabulary Taiwanese (Min-nan) speech recognition using tone features and statistical pronunciation modeling," in *Proc. EUROSPEECH*, 2003.

[16] D.-c. Lyu, R.-y. Lyu, Y.-c. Chiang, and C.-n. Hsu, "Speech recognition on code-switching among the Chinese dialects," in *Proc. ICASSP*, 2006.

[17] W.-C. Kuo, C.-C. Ho, X.-R. Zhong, Z.-F. Liang, H.-M. Yu, Y.-R. Wang, and S.-H. Chen, "Some studies on Min-nan speech processing," *International Journal of Computational Linguistics & Chinese Language Processing*, vol. 12, pp. 391–410, 2007.

[18] D.-c. Lyu, R.-y. Lyu, Y.-C. Chiang, and C.-N. Hsu, "Cross-lingual audio-to-text alignment for multimedia content management," *Decision Support Systems*, vol. 45, no. 3, pp. 554–566, 2008.

[19] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Language Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[20] D. S. Park, W. Chan, Y. Zhang, C. C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "SpecAugment: A simple data augmentation method for automatic speech recognition," in *Proc. Interspeech*, 2019.

[21] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Proc. Interspeech*, 2016.