Acoustic and Textual Data Augmentation for Code-Switching Speech Recognition in Under-Resourced Language

I-Ting Hsieh*, Chung-Hsien Wu*, † and Chun-Huang Wang†
* Graduate Program of Multimedia Systems and Intelligent Computing,
† Department of Computer Science and Information Engineering,
National Cheng Kung University, Tainan, Taiwan
E-mail: nd8081026@gs.ncku.edu.tw, {chunghsienwu, chwang512}@gmail.com,

***Abstract*— Under-resourced and code-switching speech recognition have recently received research interest, resulting in several robust acoustic and language modeling approaches. As Taiwanese and Mandarin have been popularly and widely used in Taiwan, this paper aims to address the under-resourced and code-switching issues. First, phone sharing between Taiwanese and Mandarin is employed for acoustic data augmentation to construct the acoustic models of Taiwanese speech recognizer. Regarding the lack of Taiwanese text corpus, this paper translates Mandarin corpus into Taiwanese corpus based on word-to-word translation. Moreover, additional translation rules for code-switching text are manually designed. The augmented text corpus is then used for training the code-switching language models. In the experimental results, the word error rate for code-switching speech recognition was 26.02%, which was better than that trained by the pure Taiwanese corpus.**

I. INTRODUCTION

Recently, Automatic speech recognition (ASR) related research continues to propose novel methods with the development of deep learning [1-2]. The mainstream methods achieved better performance with complex model and large corpus [3], and the languages of the corpus are almost used by most people in the world, like English or Mandarin. However, for the minority language, their corpus is not easily available. These languages are usually unofficial languages in these countries, like Taiwanese which is the Southern Min dialects spoken in Taiwan. For these languages, the under-resourced and code-switching problems are frequently encountered. Although related approaches to under-resourced and code-switching speech have been widely proposed [4-8], the characteristics between the host and guest languages should be carefully considered in constructing the code-switching recognizer. The studies about the related research between Mandarin and Taiwanese are relatively lacking [9-10]. But the characteristics of Taiwanese and Mandarin are relatively close, we use the common characteristics in pronunciation and grammar between these two languages to propose a specific method for Taiwanese speech recognition. For the under-resourced issue, besides the method of data augmentation or synthesis-based data generation for ASR [11-12], one of the most common ways is transfer learning, including weight transfer [13-14], multi-task learning [14] and domain adversarial training [15]. For the code-switching issue, there were many methods proposed, such as the hybrid CTC [5], the attention-based seq2seq model [16], the data augmentation related method [6], etc.

The main contribution of this paper is the augmentation of Taiwanese acoustic and textual data for training the Taiwanese-Mandarin code-switching speech recognition system. Taiwanese and Mandarin have similarities in pronunciation and grammar. In acoustic model training, this paper adopts phone sharing method, by combining Taiwanese and Mandarin phones for acoustic data augmentation to train the acoustic models based on the under-resourced Taiwanese corpus and resource-rich Mandarin speech corpus. Moreover, the translation from Mandarin text to Taiwanese text is based on word-to-word translation. For the augmentation of the code-switching text corpus, the manually designed translation rules considering word frequency are used to generate code-switching sentences. Finally, the generated text corpus is used for training the code-switching language model. The experiments are conducted for under-resourced and code-switching ASR. The experimental results show that the method proposed in this paper was effective.

II. DATA COLLECTION

For the preparation of the Taiwanese speech and text corpora, in this section, we will introduce the collection process of the new speech corpora and the currently available Taiwanese text corpora.

A. Taiwanese balanced speech corpus

Since it is hard to collect a large amount of annotated Taiwanese speech data with transcription, this study tries to design the database in which the number of pronunciations is balanced for data collection. The corpus is collected based on the following rules. First, we choose Taiwanese Romanization System (TaiLo) as the transliteration system of Taiwanese. In order to consider the effect caused by the preceding and succeeding phones, the tri-phone unit is selected for designing the balanced corpus. Next, we get the text from the Taiwanese news parallel corpus [17] and the real-life corpus which has been manually translated into Taiwanese. After that, rules are designed to select balanced text by the following steps:

1. Find the phones that have the fewest occurrences in the currently selected sentences.

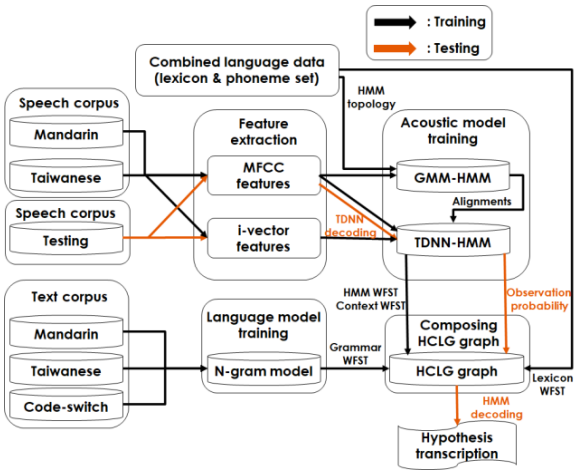

Fig. 1 The proposed system framework.

2. Find the candidate sentences which contain the phones selected in the previous step.
3. Compute the number of occurrences of each included phones in each candidate sentence. The number is then divided by the number of phones in the candidate sentence as the score.
4. Select the candidate sentence with the lowest score in the previous step.

Iterate the above steps until the number of selected sentences has been reached. In this way, 1,288 Mandarin-Taiwanese parallel sentences, in which the Taiwanese sentences are used as the transcript, and the corresponding Mandarin are used for reference. Finally, 111 speakers are invited to record the Taiwanese speech of the 1,288 Taiwanese sentences. Totally, the Taiwanese speech database contains 83,748 utterances and the duration is 70.06 hours.

B. Taiwanese-specific phone-enriched speech corpus

This corpus focuses on enriching the training sample of Taiwanese-specific phones that do not appear in Mandarin speech. The collection method is as follows:

1. List the Taiwanese-specific phones. After that, we set a threshold (i.e., 5) as the minimum number of appearances of the phones.
2. The source of speech content is also the Taiwanese news parallel corpus. Compute how many phones which have not reached the threshold in each sentence, divided by the total phone number of the sentence. The value is regarded as the score and the sentence which has the highest score is selected.

After the process, 400 sentences are selected. The total number of utterances is 9,447, and the duration is 11.31 hours.

C. Daily-life Taiwanese speech corpus

This corpus is mainly used to evaluate the accuracy of the Taiwanese speech recognizer. Since we hope that the testing utterances are close to our daily lives, the sentences in a Mandarin text corpus related to elderly care are translated into Taiwanese TaiLo sentences. After that, the sentences are recorded by 16 speakers. Finally, 509 recordings are collected, and the duration is about 0.6 hours.

D. Taiwanese text corpus

The Taiwanese text corpus, which is composed of five sub-corpora, comes from the open resources available on the websites [17-21]. This paper uses TaiLo as the transliteration system of Taiwanese, but some text corpuses are Church Romanization, so they must be converted. Thereafter, the TaiLo format of each syllable is checked. According to the Taiwanese lexicon used for experiments, word segmentation is performed on the sentences. Finally, the sentences which contain few words or high proportion of unknown words are filtered out, and the corpora from all sources are combined into a Taiwanese text corpus consisting of 649,422 sentences.

III. PROPOSED METHOD

For the implementation of code-switching ASR based on the method proposed by Mohri et al. [22], lexicon, language model, triphone data and Hidden Markov Model (HMM) in the acoustic model are saved in the form of weighted finite-state transducer (WFST). Fig. 1 shows the block diagram of the proposed system. After extracting the acoustic features of the test speech corpus, the trained Time delay neural network (TDNN) [23-24] is used to obtain the probability distribution of each HMM state, and provide the observation probability. Finally, the WFST is used to decode the path with the highest probability, which is the final recognition result.

A. Phone set and lexicon definition

Since Taiwanese is an under-resourced language and it often occurs code-switching with Mandarin, we define a phone set for both Taiwanese and Mandarin. Among the transliteration systems for Taiwanese, TaiLo is one of the commonly used systems. In Taiwan, Zhuyin is the phonetic symbols for representation of Mandarin pronunciation. Therefore, this paper defines the corresponding phone for each symbol of the two transliteration systems. In addition, although Mandarin and Taiwanese are different languages, the International Phonetic Alphabet (IPA) is adopted for representing these symbols. As there are some phonetic symbols pronounced similarly, they can be defined as the same phone, and they are referred to as shared phones in the following, and the rest are called the Mandarin-specific and Taiwanese-specific phone, respectively. Some observations and adjustments of the phone definition are listed below:

- Some Taiwanese phones are the semivowels (such as “m” and “ng”). Since there are differences between consonants

Table 1. The defined phone set for Mandarin and Taiwanese.

Common phones							
p	ph	m	t	th	n	l	k
kh	h	ts	tsh	s	a	i	u
e	o	er	nn	ng			
Mandarin specific phones							
f	tsi	tshi	si	tsr	tsrh	sr	z
err	y						
Taiwanese specific phones							
b_T	g_T	ng_T	j_T	ms_T	ngs_T	nn_T	m_T
p_T	t_T	k_T	h_T				

and semivowels, we define the consonant "m" and the semivowel "**m**" as different phones.

- Several retroflex consonants in Mandarin, such as the phones "tsr /ㄓ/", "tsrh /ㄔ/", "sr /ㄕ/", do not exist in Taiwanese.
- There are several compound vowels exist in Zhuyin, but they are composed of shared phones. Thus, they can be split into two phones.
- Most of the vowels are shared, but some codas (also known as auslaut) only exist in TaiLo.

Based on the phone definition described above, there are totally 21 shared phones, 10 Mandarin-specific phones, and 12 Taiwanese-specific phones. Table 1 shows the defined phone set for Mandarin and Taiwanese. In order to further consider the pronunciation effect of tone, the shared and specific tones are defined according to the pitch of the tones, respectively. Table. 2 is the definition of the tones in Mandarin and Taiwanese. We consider the characteristics of tone values from the two languages to combine similar tones. Finally, after considering the tone, there are a total of 44 shared phones, 64 Mandarin-specific phones, and 77 Taiwanese-specific phones. For Taiwanese, about 40% of the phones can be shared with Mandarin.

The lexicon maps the text representation of each Taiwanese and Mandarin word to its corresponding phone sequence. On the other hand, the translation dictionary is a mapping between Mandarin and Taiwanese words. In this paper, the translation dictionary is used to generate the Taiwanese text corpus and code-switching with Mandarin via word-to-word translation. The Taiwanese dictionary of this paper is obtained from the open source provided by ChhoeTaigi [25]. Moreover, they are used in the experiment of this paper, so the TaiLo format must be further checked. After the process, 65,527 words are finally retained. After converting the TaiLo into phone sequence, and then, the Taiwanese lexicon is constructed. As for the translation dictionary, the mapping between the Mandarin words and the corresponding TaiLo words is one-to-many. To handle this situation, the word pair is set as the default translation word if the number of times in the source dictionary is the highest. Moreover, the phone sequence of the Taiwanese and Mandarin words also includes these shared phones. Finally, the Mandarin lexicon is composed of 305,938 words, and it is merged with the Taiwanese lexicon as the multilingual lexicon used in our experiments.

Table 2. The defined tone set for Mandarin and Taiwanese tones.

Tone number	Mandarin tones	Taiwanese tones	
	Tone value	**Tone value**	**Shared with Mandarin tones**
0	**-**	**-**	**0**
1	**5→5**	**4→4**	**1**
2	**3→5**	**5→3**	**4**
3	**2→1→4**	**2→1**	**-**
4	**5→1**	**2**	**-**
5	**-**	**2→4**	**2**
7	**-**	**3→3**	**-**
8	**-**	**4**	**-**

B. *Acoustic model*

For acoustic model training of the TDNN, this paper adopts the approach proposed by Peddinti et al. [23]. The i-vector features which provide the offset information of different speakers on the acoustic features are appended to the Mel-frequency cepstral coefficients (MFCC) feature as the input to the TDNN.

In the training of the GMM-HMM, the preliminary alignment is obtained. For training the TDNN-HMM, the alignment information from GMM-HMM is used as the ground truth. Here, the discriminative training is adopted by maximizing the following objective function (1):

$$F_{MMI}(\lambda) = \sum_{i=1}^{I} \log(HMM(W_i)|\, O_i,\ \lambda) \quad (1)$$

$$F_{MLE}(\lambda) = \sum_{i=1}^{I} (O_i \mid HMM\ (W_i),\ \lambda) \quad (2)$$

This equation is the maximum mutual information (MMI) algorithm. W_i is the transcription of the i-th training speech data, and $HMM(W)$ denotes the training graph constructed by this transcription. Where I denotes the total number of training data, and the training graph is given based on the known transcription. The goal is to find a suitable λ, so that the sum of the log probabilities of the observed corresponding speech features is maximized.

Compared with the Maximum Likelihood Estimation (MLE) (2), MMI is under the condition of a given feature vector sequence O, and we should estimate the probability of all output state sequences which conform to the training graph. Given the model parameter λ, assuming that it is a constant in the equation, we apply the Bayes' theorem in (3):

$$F_{MMI} = \sum_{u=1}^{U} \log \frac{p(O_u\,|HMM\ (W_u),\ \lambda)}{\sum_{W \in Trainset} p\ (O_u\,|HMM\ (W),\ \lambda\ )} \quad (3)$$

Comparing with the MLE, the advantage of MMI divided by this denominator is that it not only maximizes the probability of the speech observed from referenced transcription, but

minimizes the observation probability from other wrong transcription of the training corpus.

C. Language model

The language model adopted in this paper is a typical n-gram model [26]. Taiwanese text and the real Mandarin-Taiwanese code-switching text corpus are both lacking, while the Mandarin text corpus is rich. Recently, the unsupervised machine translation approach proposed by Lample et al. [27] achieved great results in their experiments, but the two applied languages should have sufficient text corpus, respectively. Fortunately, the grammar of Mandarin is similar to Taiwanese. Therefore, this paper adopts the word-to-word translation approach and applies additional rules to translate Mandarin text corpus to Taiwanese, thereby to enrich the amount of Taiwanese corpus.

As for the generation of code-switching text corpus, in addition to converting the Mandarin word to the Taiwanese words, other rules are designed to retain some Mandarin words as the code-switching words. According to the observation, the code-switching words from Taiwanese to Mandarin are more likely to occur in the rarely used vocabulary in daily life, while the words that may not exist in Taiwanese dictionary.

IV. Experimental results

A. Experiment setup

In this paper, the speech corpora were applied to the experiment. The details of the speech corpora were shown in Table 3.

Table 3.: *Taiwanese speech dataset (hours).*

Speech corpus	Training data	Testing data
Taiwanese Balanced	70.06	
Taiwanese-Specific Phone-Enriched	11.31	
Daily-life Taiwanese		0.6
Total	81.36	0.6

In this experiment, the Kaldi speech recognition tool [28] was used as the framework. The raw speech signals were sampled at 16 kHz, and the length of each frame was 25 ms and 10 ms overlapped for feature extraction. The extracted features were 40-D MFCC and 100-D i-vector. In the training of GMM-HMM, there were 1000 mixed Gaussian components, and they were sequentially trained using Linear Discriminant Analysis Maximum Likelihood Linear Transformation (LDA_MLLT). After obtaining the alignment information through GMM-HMM, we could get more accurate alignment performance through the training of TDNN acoustic model. The number of TDNN layers was 8 and the dimension of each layer was 512. The number of epochs was 4. On the other hand, the tri-gram was used as the language model.

B. Evaluation of under-resourced ASR

One of the goals of this paper was to improve the accuracy of speech recognition in under-resourced language, hence the methods such as shared phones were proposed. In this section, the experimental results were compared and analyzed. Regarding the metric of the performance, we evaluated the speech recognizers by word error rate (WER), syllable error rate (SER), and phone error rate (PER). Eq. (4) was the equation of the WER.

$$WER(\hat{W},\ W) = \frac{SID\ (\hat{W},\ W)}{C\ (W)} \tag{4}$$

W is the reference transcription, and $\hat{W}$ denotes the hypothesis transcription. *SID* denotes the total number of substitutions, insertions and deletions, and it is divided by C which was the number of words in the reference transcription W. In this experiment, the proposed method was not only compared with the baseline, but compared with the speech recognition model which used the transfer learning method. Therefore, after training the TDNN with the Mandarin speech data, the parameters of the lower layers were transferred. In a general acoustic model, the lower layers usually learn more basic acoustic features. Since the Mandarin corpus is sufficient, the basic acoustic features of the bottom layers are relatively robust. Due to the robust acoustic features from the Mandarin corpus, the Taiwanese corpus only focus on the training of Taiwanese features. After that, the Taiwanese speech training dataset described in the above section was used to fine-tune the TDNN. Regarding Mandarin corpus, this experiment adopted KING-ASR-044 and KING-ASR-360 and others for training [29-32]. The total duration was 2883 hours. The experimental results were listed in Table 4.:

Table 4.: *Error rate (%) with weight transfer.*

Mandarin corpus	Layer	WER	SER	PER
120 hours from KING-ASR-044 & KING-ASR-360	3	28.86	18.04	10.92
	4	**28.19**	17.93	10.90
	5	28.59	**17.83**	**10.66**
	6	29.72	18.60	11.20
All	4	**27.64**	**16.46**	**9.75**

Table 4. shows that the best result was obtained when transferring the parameters of the lower 4 layers. The experiment used both the Taiwanese speech and the same 120 hours Mandarin speech to train the acoustic model. In order to compare the models under the same conditions of dataset, the 28.19% WER was taken as the representative of the weight transfer methods. On the other hand, the baseline speech recognizer was trained only using the Taiwanese speech, and the used lexicon only contained Taiwanese vocabulary and phones. Besides, since the experiment in this section did not consider code-switching. Therefore, the training of the language model only used the Taiwanese text data described in the previous section, and the results of these speech recognition models are shown in Table 5.

The results in Table 5 show that comparing with the baseline speech recognizer, the proposed shared phones (SP) method could improve the recognizer by training together with the

Mandarin corpus. Moreover, the error rate could be reduced more compared to the weight transfer method.

C. Evaluation of code-switching ASR

In previous section, the experimental results showed that by phone sharing, the Taiwanese and Mandarin speech data could be trained together and the error rate of the word to phone level was reduced. Based on this multilingual acoustic model. In the experiment, the host language was Taiwanese, and the guest language was Mandarin. For the code-switching speech, since the intra-sentential and inter-sentential switching might occur, the training textual data contained pure Taiwanese, pure Mandarin, and the code-switching sentences with the ratio of 2:1:1. After the language model was trained by the above text data, the code-switching recognizer was composed of the language model, shared phones acoustic model and lexicon. In order to evaluate the performance, some code-switching speech data were recorded. There were 8 speakers and 204 utterances, totaling 0.25 hours. Each sentence contained at least 1 Mandarin vocabulary. The results of this experiment were illustrated in Table 6.

Comparing the results before and after replacing with the code-switching language model, since the Mandarin words in testing data were out-of-vocabulary (OOV), we found that the performance of ASR with Taiwanese language model was quite bad. After using the code-switching language model, about 29% WER could be achieved. The result showed that the speech recognizer had the ability to recognize the code-switching speech, since it could decode the Mandarin word.

Table 5.: *Error rate (%) of under-resourced ASR.*

ASR system	WER	SER	PER
Baseline GMM-HMM	41.23	28.80	18.27
Baseline TDNN-HMM	29.23	18.54	11.17
Transfer learning	28.19	17.93	10.90
SP-GMM-HMM	40.77	27.74	17.79
SP-TDNN-HMM	26.02	15.73	9.28

Table 6.: *Error rate (%) of code-switching ASR.*

Language model	WER	SER	PER
Pure Taiwanese	57.94	52.91	32.62
Code switching	29.05	25.53	16.01

V. Conclusion

The main goals of this paper are twofold. One is to improve the accuracy of speech recognition applied to under-resourced language, and the other is to construct the ASR to recognize the code-switching speech. In order to achieve both goals, this paper proposed shared phones and data augmentation method. Experimental results show that both goals have been achieved, but there are still some drawbacks. The main drawback is that it is bound to the languages. If we want to use the corpus of other languages, we must make the shared phones or make sure the quality of word-to-word translation dictionary for the applied languages. Therefore, in future work, it is important to find a non-manual and non-subjective method for automatically defining shared phones and translations.

VI. References

[1] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," in IEEE Signal Processing Magazine, vol. 29, no. 6, pp. 82–97, 2012.

[2] A. Graves, A. Mohamed and G. Hinton, "Speech recognition with deep recurrent neural networks," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 26-31, Vancouver, Canada, pp. 6645-6649, 2013.

[3] V. Panayotov, G. Chen, D. Povey and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 19-24, South Brisbane, Queensland, Australia, pp. 5206-5210, 2015.

[4] J. Yi, J. Tao, Z. Wen and Y. Bai, "Adversarial Multilingual Training for Low-Resource Speech Recognition, "IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), April 15-20, Calgary, Alberta, Canada, pp. 4899-4903, 2018.

[5] Z. Zeng, Y. Khassanov, V. T. Pham, H. Xu, E. S. Chng, and H. Li, "On the end-to-end solution to Mandarin-English code-switching speech recognition," arXiv:1811.00241, 2018.

[6] E. Yılmaz, H. Van den Heuvel, and D. A. Van Leeuwen, "Acoustic and textual data augmentation for improved asr of code-switching speech," in INTERSPEECH 2018 – 19th Annual Conference of the International Speech Communication Association, September 2-6, Hyderabad, India, Proceedings, 2018, pp. 1933-1937.

[7] C.-H. Wu, H.-P. Shen and C.-S. Hsu, "Code-Switching Event Detection by Using a Latent Language Space Model and the Delta-Bayesian Information Criterion," IEEE/ACM Trans. Audio, Speech, and Language Processing, VOL. 23, NO. 11, November 2015, pp. 1892-1903. (SCIE).

[8] C.-H. Wu, H.-P. Shen and Y.-T. Yang, "Chinese-English Phone Set Construction for Code-Switching ASR Using Acoustic and DNN-Extracted Articulatory Features," IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol. 22, No. 4, April 2014, pp. 858-862.

[9] H. Y. Su, "Code-switching between Mandarin and Taiwanese in three telephone conversation: The negotiation of interpersonal relationships among bilingual speakers in Taiwan," In Proc. of the Symposium about Language and Society, April, 2001.

[10] D. C. Lyu, R. Y. Lyu, Y. C. Chiang, and C. N. Hsu, "Speech recognition on code-switching among the Chinese dialects," In 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings, May, Vol. 1, pp. I-I, 2006.

[11] Y.-C. Huang, C.-H. Wu, Y.-Y. Chen, M.-G. Shie, and J.-F. Wang, "Personalized Spontaneous Speech Synthesis using a Small-Sized Unsegmented Semi-Spontaneous Speech," IEEE/ACM Trans. Audio, Speech, and Language Processing, Vol. 25, No. 5, May 2017, pp. 1048-1060.

[12] Y.-Y. Chen, C.-H. Wu, Y.-C. Huang, S.-L. Lin, and J.-F. Wang, "Candidate Expansion and Prosody Adjustment for Natural Speech Synthesis using a Small Corpus," IEEE/ACM Trans. Audio, Speech, and Language Processing, VOL. 24, NO. 6, June 2016, pp. 1052-1065. (SCIE)

[13] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer learning for speech recognition on a budget,"

in Annual Meeting of the Association for Computational Linguistics, July 30-August 4, Vancouver, Canada, 2017.
[14] P. Ghahremani, V. Manohar, H. Hadian, D. Povey and S. Khudanpur, "Investigation of transfer learning for ASR using LF-MMI trained neural networks," IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), December 16-20, Okinawa, Japan, pp. 279-286, 2017.
[15] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," The Journal of Machine Learning Research, 17(1): 2096-2030., 2016.
[16] S. Watanabe, T. Hori, S. Kim, J. R. Hershey and T. Hayashi, "Hybrid CTC/Attention Architecture for End-to-End Speech Recognition," in IEEE Journal of Selected Topics in Signal Processing, vol. 11, no. 8, pp. 1240-1253, Dec. 2017.
[17] iCorpus 臺華平行新聞語料庫. Available: http://icorpus.iis.sinica.edu.tw 2020.03.23.
[18] 台語文語料庫蒐集及語料庫為本台語書面語音節詞頻統計. Available: http://ip194097.ntcu.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp 2020.03.23
[19] 台語文數位典藏資料庫. Available: http://ip194097.ntcu.edu.tw/nmtl/dadwt/pbk.asp 2020.03.23
[20] 新約聖經語料. Available: https://bible.fhl.net 2020.03.23
[21] 臺語國校仔課本. Available: https://github.com/Taiwanese-Corpus/kok4hau7-kho3pun2 2020.03.23
[22] M. Mohri, F. Pereira, and M. Riley, "Weighted finite-state transducers in speech recognition," in Computer, Speech and Language, vol. 16, no. 1, pp. 69–88, 2002.
[23] 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 INTERSPEECH 2016 – 17th Annual Conference of the International Speech Communication Association, September 2-6, San Francisco, USA, Proceedings, 2016, pp. 2751-2755.
[24] V. Peddinti, D. Povey, and S. Khudanpur, "A time delay neural network architecture for efficient modeling of long temporal contexts," in Sixteenth Annual Conference of the International Speech Communication Association, September 6-10, Dresden, Germany, 2015.
[25] ChhoeTaigi 找台語：台語字詞資料庫. Available: https://github.com/ChhoeTaigi/ChhoeTaigiDatabase 2020.03.23
[26] A. Stolcke, "SRILM - an extensible language modeling toolkit," in Proc. ICSLP, September 16-20, Denver, Colorado, USA, pp. 901–904, 2002
[27] G. Lample, M. Ott, A. Conneau, L. Denoyer, and M. A. Ranzato, "Phrase-based & neural unsupervised machine translation," arXiv:1804.07755, 2018.
[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi speech recognition toolkit," in ASRU, December 11-15, Waikoloa, HI, USA, 2011.
[29] King-ASR-044. Available: http://en.speechocean.com/datacenter/details/78.html 2020.03.23
[30] King-ASR-360. Available: http://en.speechocean.com/datacenter/details/349.html 2020.03.23
[31] H.-M. Wang, B. Chen, J.-W. Kuo, and S.-S. Cheng, "MATBN: A Mandarin Mandarin broadcast news corpus," International Journal of Computational Linguistics and Mandarin Language Processing, vol. 10, no. 2, pp. 219–236, 2005.
[32] TCC-300. Available: http://www.aclclp.org.tw/use_mat_c.php 2020.03.23.