Automatic Speech Recognition Datasets in Cantonese:
A Survey and New Dataset

Tiezheng Yu[⋆], Rita Frieske[⋆], Peng Xu[⋆] [†], Samuel Cahyawijaya[⋆],
Cheuk Tung Shadow Yiu, Holy Lovenia, Wenliang Dai, Elham J. Barezi
Qifeng Chen, Xiaojuan Ma, Bertram E. Shi, Pascale Fung

The Hong Kong University of Science and Technology
{tyuah, peng.xu, scahyawijaya}@connect.ust.hk
rita.frieske@ust.hk

Abstract

Automatic speech recognition (ASR) on low resource languages improves the access of linguistic minorities to technological advantages provided by artificial intelligence (AI). In this paper, we address the problem of data scarcity for the Hong Kong Cantonese language by creating a new Cantonese dataset. Our dataset, **M**ulti-**D**omain **C**antonese **C**orpus (MDCC), consists of 73.6 hours of clean read speech paired with transcripts, collected from Cantonese audiobooks from Hong Kong. It comprises philosophy, politics, education, culture, lifestyle and family domains, covering a wide range of topics. We also review all existing Cantonese datasets and analyze them according to their speech type, data source, total size and availability. We further conduct experiments with Fairseq S2T Transformer, a state-of-the-art ASR model, on the biggest existing dataset, Common Voice zh-HK, and our proposed MDCC, and the results show the effectiveness of our dataset. In addition, we create a powerful and robust Cantonese ASR model by applying multi-dataset learning on MDCC and Common Voice zh-HK.

Keywords: Speech Corpus, Hong Kong Cantonese, Automatic Speech Recognition System

1. Introduction

Automatic speech recognition (ASR) systems take the audio as input and convert it into text (Malik et al., 2021). Due to the popularization of deep learning, ASR technology has grown rapidly and has led to a significant improvement in recognizing many languages. For instance, ASR systems in English (Zhang et al., 2020; Xu et al., 2021; Baevski et al., 2020) have been able to achieve a below 2% word error rate (WER) on the LibriSpeech (Panayotov et al., 2015) corpus. A similar trend is also observed in research on Chinese ASR (Li et al., 2019; Winata et al., 2020a; Zhang et al., 2020), exemplified by the improvement in the ASR model performance on the Aishell-1 (Bu et al., 2017) corpus from 18.7% character error rate (CER) down to 6.84% within just two years. However, many languages (e.g., Gujarati, Hindi, Bengali, Amharic and Cantonese), including those that feature code-switching, are still lacking resources, and the performance of ASR systems in these languages is unsatisfactory (Winata et al., 2021; Khare et al., 2021; Lovenia et al., 2021). Therefore, many methods in ASR have also been introduced and have shown promising results (Wang et al., 2021; Lin and Chen, 2020; Winata et al., 2020b; Winata et al., 2020c). Many of these achievements are due to the utilization of the most recent deep neural network architectures and high-performance parallel computing graphics cards. However, since deep learning techniques require large amounts of training data, the creation of ASR datasets is essential for model performance. Moreover, creating a speech recognition

corpus will accelerate the development of ASR systems in corresponding languages.

Although around 88.9% of Hong Kong's population are native Cantonese speakers, the Cantonese language is still struggling with a shortage of resources for building ASR systems. We present the most important speech resources in Cantonese in Table 1, showing their speech type used for building the dataset, data source, total size and availability of the dataset. From the table, we can see that not all of them are suitable to build robust ASR systems.

To fill the research gap, we introduce a multi-domain ASR read corpus called **M**ulti-**D**omain **C**antonese **C**orpus (MDCC) for ASR research in Cantonese. Our corpus consists of 73.6 hours of clean read speech collected from various Hong Kong Cantonese audiobook sources. It contains philosophy, politics, education, culture, lifestyle, and family domains, and covers a wider range of topics than most of the other corpora. In addition, we perform experiments using a state-of-the-art ASR framework, Fairseq S2T Transformer (Wang et al., 2020), on the two of the largest available datasets (MDCC and Common Voice zh-HK). The model achieves 10.15% CER on the test set of our corpus, which indicates the effectiveness of our dataset. We also use joint training to create a more powerful and robust model for Cantonese ASR [1].

The contributions of our study to the field are threefold:

- We review existing Cantonese ASR datasets and thoroughly analyze them from various perspectives (speech type, data source, total size and availability).

[⋆] These authors contributed equally.

[†] The work was done when the author was studying at The Hong Kong University of Science and Technology.

---

[1]https://github.com/HLTCHKUST/cantonese-asr

Name	Speech Type	Data source	Size [hours]	Availability
HKCAC (Leung and Law, 2001)	Spont.	Phone-in programs	8.1	Non-Public
HKCanCor (Luke and Wong, 2015)	Spont.	Chat	30.0	Cvasi-Public
HKCC (Chin, 2015)	Spont.	Movie	35.0	Cvasi-Public
CantoMap (Winterstein et al., 2020)	Read	MapTask	12.8	Public
Common Voice zh-HK (Ardila et al., 2019)	Read	Wikipedia	96.0	Public
MDCC (Ours)	Read	Audiobook	73.6	Public

Table 1: Hong Kong Cantonese ASR corpora

- We propose a new dataset named MDCC for ASR research in Cantonese, which consists of 73.6 hours of clean read speech and covers a wide range of topics.

- We evaluate our dataset and another available Cantonese ASR dataset (Common Voice zh-HK) by using a state-of-the-art ASR model (Fairseq S2T Transformer). Furthermore, we apply multi-dataset learning approaches on the two datasets to create a powerful and robust model for the Cantonese ASR. Multi-dataset learning boosts the model's performance on both datasets. The results support the effectiveness of our dataset.

2. Cantonese ASR Datasets

Table 1 lists the most important previous Cantonese ASR corpora with their speech type, data source, data size and availability. Some other works are also related to Cantonese ASR but focus on different aspects. For instances, Cantonese multimodal (audio-visual) speech dataset for in-car command recognition (Dai et al., 2022).

HKCAC The Hong Kong Cantonese Adult Language Corpus (HKCAC) is created from spontaneous speech records from the radio phone-in programs and forums in Hong Kong. It has 8.1 hours of recordings and transcripts that contain approximately 170,000 characters. The dataset is inaccessible as no website, link, or other information is provided for retrieving the dataset.

HKCanCor The Hong Kong Cantonese Corpus (HK-CanCor[2]) is built based on spontaneous chat records. Participants were recruited for arranged recording sessions for two- or three-party chats. Later, an additional set of recordings was obtained from radio chat shows. The corpus consists of 30.0 hours of recordings, with each sample 10 minutes long. After transcription, the corpus contains around 180,000 word tokens. The transcripts of HKCanCor can be accessed from the official website, but no audio data are provided.

HKCC The Corpus of Mid-20th Century Hong Kong Cantonese (HKCC[3]) is constructed based on Cantonese films from Hong Kong in the 1950s and 1960s. HKCC has two phases, and we only introduce the first-phase corpus since the second phase's report has not been released. There are 21 movies in the first-phase corpus, and each movie is about 100 minutes long. The corpus has, in total, about 200,000 character tokens. The details of the HKCC dataset can be found on the official website, but cannot be retrieved due to limited access control.

CantoMap The Hong Kong Cantonese MapTask Corpus (CantoMap[4]) aims to provide a Cantonese corpus for ASR research and also involves several controlled elicitation tasks related to the phonology and semantics of Cantonese. The design of the corpus follows the general setup used for the HCRC MapTask corpus (Anderson et al., 1991). The corpus includes a total of 12.8 hours of recordings and transcripts of forty speakers. The CantoMap dataset is publicly available in the GitHub repository.

Common Voice zh-HK The Common Voice zh-HK corpus[5] is a massive-multilingual collection of transcribed speech collected and validated via Mozilla's Common Voice initiative. The speakers are required to read sentences from Wikipedia and the annotators verify each sentence. We use 96.0 hour split of verified Cantonese utterances in our experiments. The detailed data statistics are shown in Section 5. The dataset is available on the Common Voice website.

Although each of the existing corpora has advantages, not all of them are suitable for developing Cantonese ASR systems. None of the corpora except Common Voice zh-HK are large enough for data-intensive ASR model fine-tuning. Furthermore, even for Common Voice zh-HK, empirical experiments based on recent deep learning models are limited. To fill this research gap, we propose MDCC to enrich ASR data resources in Cantonese. Furthermore, we implement a state-of-the-art ASR model and report its performance on the Common Voice zh-HK dataset and MDCC.

3. Corpus Creation

This section describes the creation of our MDCC. We first introduce our approach to collect and pre-process Cantonese audiobooks and then present the methods used to annotate transcripts.

---

[2] http://compling.hss.ntu.edu.sg/hkcancor/
[3] http://202.45.36.235/hkcc/

[4] https://github.com/gwinterstein/CantoMap
[5] https://commonvoice.mozilla.org/zh-HK/datasets

Type	Words
Unified Writing	呢, 喇=> 呢 - (Question particle) 呢邊, 呢便=> 呢邊 - Over here / there; 裏面, 裡面=> 裏面 - Inside;
Important Words	啦, 喇, 喎 - (Question particle) 係 - Yes 後 - Later / After / Afterward 幾 - Some / A few / Several 並 - And 併 - Combine 徵 - Recruit / Ask for 關於 - About 過程 - Process 盡快 - As soon as possible 部份 - Partially 咁樣 - So / Such / Like this 嗰個 - That one 其實 - Actually 讀書 - Study

Table 2: The unified writing and important words for the attention of annotators.

3.1. Audiobook Collection

The speech corpus of the MDCC is collected from Hong Kong Cantonese audiobook sources. The corpus contains various audiobooks covering different topics (e.g., philosophy, politics, education, culture, lifestyle and family). Most of the books are in the literary form of Cantonese. However, some are written in a formal written form that is never used as spoken language, and therefore is not applicable to ASR systems. To remove these books from the dataset, we hire native Cantonese speakers to check all the audiobooks and filter them out manually.

Every chunk of the downloaded audiobook varies from 40 minutes to 2 hours, which does not fit the optimal size for ASR systems. Therefore, we apply a voice activity detection (VAD) tool to convert the original audio pieces into shorter audio utterances. The VAD tool can classify a chunk of audio data as being voiced or unvoiced, and we split the original audio samples at the unvoiced parts. After separation, we get 83,275 audio utterances with a total corpus size of 73.6 hours.

3.2. Annotation

To ensure cost-efficiency with optimal quality, we annotate all the utterances in two phases. We first conduct an automatic annotation with the Google Cloud Speech-to-Text API and then improve the quality of the automatic transcripts by hiring native Cantonese speakers to correct them manually.

Google Cloud Speech-to-Text. Google Cloud Speech-to-Text is an API that converts speech into text and is powered by Google's AI technologies. [6] We apply this API to produce the initial transcripts of the utter-

---

[6]https://cloud.google.com/speech-to-text/

去其他地方啦 - Go to other places	
我有記憶第一個冬天就係咁樣過去喇 - The first winter that I can remember was gone	
好多依然都係旗袍做校服 - Many of them still wear cheongsams as school uniforms	
由上海坐船到天津高級艙裏面亦都好多老鼠 - There are also many mice in premium cabins of a ship going from Shanghai to Tianjin	
我自己覺得好似跑得風一樣咁快 - I feel like running as fast as the wind	

Table 3: Several representative sentences from the MDCC.

ances. More specifically, we use the default model with the language set to Hong Kong Cantonese (yue-Hant-HK). The API returns a transcript with a confidence score for each utterance. These automatically generated transcripts accelerate the hand-correction process significantly.

Proofreading of the Transcription. Since Google Cloud Speech-to-Text is not entirely accurate, we hire native Cantonese speakers to hand-correct the errors in the transcripts it generates. During proofreading, the annotators are required to adjust the transcripts and take notes for each utterance according to our guidelines. Table 2 gives the list of words that the annotators need to pay attention to. The words listed as "unified writing" mean they share the same semantics, and we replace them with a unified word, while those listed as "important words" need the special focus to annotate them accurately. One type of important words is question particles indicating interrogative sentences in Cantonese. Table 2 shows samples of question particles the annotators need to focus on.

When taking notes, the annotators adhere to the following guidelines: 1) If the audio contains pure music, the annotators mark the label "(music)" in the file name of its transcript. 2) If the utterance contains one or several sentences with background music or noise, the annotators mark the label "(music)" before each sentence in the transcript. 3) The annotators use {} symbols to enclose words they are uncertain about, for example, {梁佳佳}，我是{}人.

In addition, for the English transcriptions or Arabic numerals, the annotators needed to do the following: 1) capitalize the first word of each complete English sentence; 2) capitalize proper nouns (e.g., names of people, countries and regions); 3) keep all other common English words lowercase except as set out in 1) and 2); 4) do not add a space between letters of common acronyms such as CCTV, VIP, etc, but do so between letters of unusual acronyms, and capitalize them, for example, 手機型號X L 係幾多。; 5) add a space between English and Cantonese words; 6) convert all Arabic numbers to Cantonese based on pronunciation, for example, the
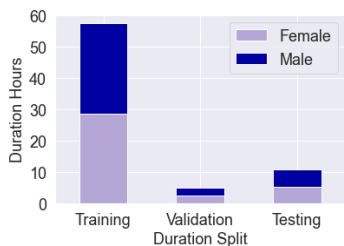
Figure 1: Gender split of the training, validation and test sets per hour of recorded audio.

Gender	# Sample				Duration (hr)			
	Train	Valid	Test	Total	Train	Valid.	Test	Total
Female	29,224	2,541	5,606	37,371	28.67	2.52	5.39	36.58
Male	35,896	3,122	6,886	45,904	28.86	2.54	5.61	37.01
Total	65,120	5,663	12,492	83,275	57.53	5.05	11.01	73.59

Table 4: Breakdown of the training, validation, and test splits in the MDCC by number of samples, gender, and the duration of the utterances.

Arabic number 60 can be converted into 六零 or 六十 based on different pronunciations.

With the proofreading, the CER of the Google Cloud Speech-to-Text result reaches 25%, proving that the annotators corrected numerous errors. Meanwhile, it is worthwhile to do annotation in two phases since most of the automatically generated transcripts from Google Cloud Speech-to-Text are in fact correct. In Table 3, we showcase several representative utterances. As we can see, most of the utterances contain a complete sentence, and the length of the utterances vary.

3.3. Corpus Splitting

We randomly split the MDCC into training, validation and test sets. Table 4 shows the detailed corpus splits which covers 65,120 utterances for training (57.53 hours), 5,663 for validation (5.05 hours), and 12,492 for testing (11.1 hours) respectively. As is shown in Figure 1, we balance the total duration of each gender's audio data within each split.

4. MDCC: Multi-Domain Cantonese Corpus

In this section, we analyze our Multi-Domain Cantonese Corpus (MDCC) from the perspectives of data statistics, domain and text.

4.1. Data Statistics

The MDCC consists of 73.6 hours of Cantonese scripted speech from Cantonese native speakers, with a balanced gender ratio of 50.29% male and 49.71% female voice talents. The corpus is divided into 83,275 audio files, each containing one utterance. The MDCC includes a total of 998,366 Cantonese characters, with each utterance being approximately 11.99 characters long. As shown in Figure 2, the length of each utterance varies from a single character to as many as 80 characters. Of these utterances, 89.85% are less than 23 characters, and the number of utterances decreases rapidly as the length of the utterance increases. Few utterances reach a length of more than 50 characters.

The duration of each utterance is between 0.22 to 15.0 seconds. Moreover, the average duration of an utterance is 3.18 seconds. As we can see in Figure 3, the duration distribution is balanced, and most of the utterances are
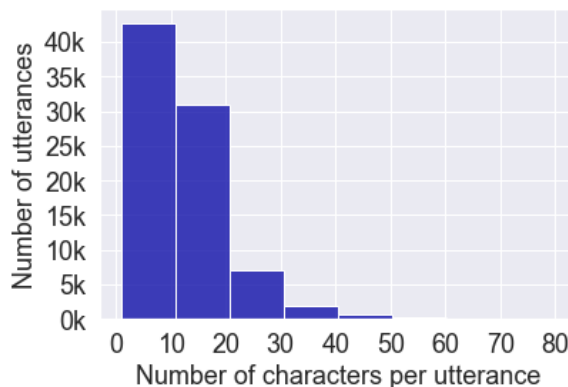

Figure 2: Distribution of the number of characters per utterance in the MDCC.
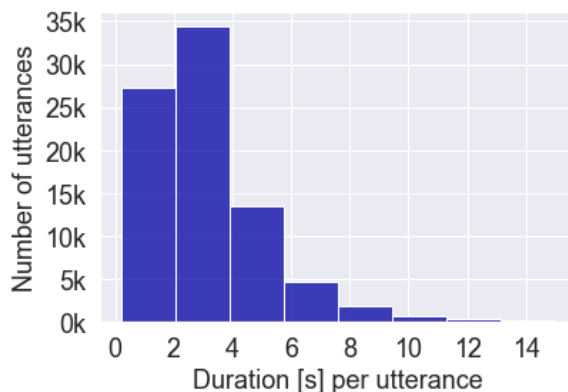

Figure 3: Distribution of the duration (in seconds) per utterance in our MDCC.

between one to nine seconds. Meanwhile, the duration distribution is generally aligned with the length distribution since longer utterances take more time for the speaker to read.

4.2. Domain Analysis

To create a general, natural and commonplace conversational ASR system, we choose a wide range of audiobook sources for the dataset. As a result, our MDCC dataset covers multiple domains, including philosophy, politics, education, culture, and lifestyle. We hire a native Cantonese speaker to read and annotate the domain for each

Top	Character/Unigram	Bigram	Trigram
1	嘅 - is / are	但係 - but	嘅時候 - the time of
2	一 - one	一個 - one	呢一個 - this one
3	係 - am / is / are	亦都 - and / also / as well	更懂得 - more clear / understand better
4	佢 - he / she / it	同埋 - with	同著作 - same book
5	我 - I / me	佢哋 - they / them	自己嘅 - myself / my own
6	有 - have / has	我哋 - we / us	嘅學生 - the student of
7	人 - person / people	咁樣 - like this	唔能夠 - cannot / unable to
8	唔 - no	就係 - that is / just like	中國人 - Chinese
9	個 - pieces	學生 - student	小王子 - little prince
10	好 - yes / good	裏面 - inside	呢一種 - this kind

Table 5: The statistics of the MDCC vocabulary: top 10 high-frequency unigrams, bigrams and trigrams.
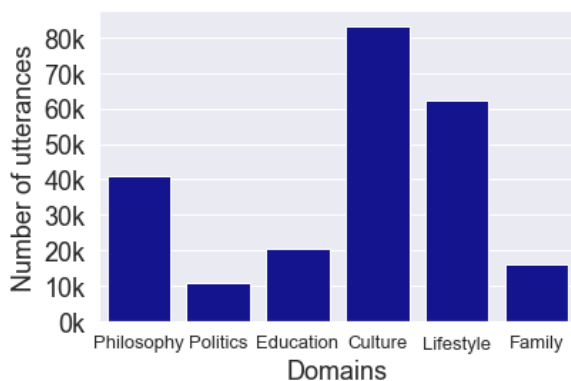

Figure 4: The distribution of domains of utterances in the MDCC. Each utterance can belong to more than one domain. Hence the total number of utterances per domain is bigger than the real total number per utterance.

Figure 5: The distribution of log character frequency in the MDCC.

audiobook. Figure 4 provides a summary of the domain distribution in the corpus. The domain of each sentence follows the domain of the audiobook that the sentence belongs to. Since an audiobook can cover multiple domains, the sum of the sentences in each domain is greater than the total number of sentences in the MDCC dataset. The culture and lifestyle domains have the most utterances in our dataset, which shows that the content of our dataset reflects people's daily lives. Besides culture and lifestyle, the philosophy domain also includes many utterances. It is worth explaining that "philosophy" here mainly refers to self-help literature. The politics, education and family domains have less data, but are still considered essential topics of the MDCC. We believe that this domain analysis can help the research community to better understand the semantic distribution of our dataset.

4.3. Common Phrases in MDCC

After a thorough analysis, we calculated a total of 998,366 Cantonese characters in the MDCC dataset. Their distribution is depicted in Figure 5. In order to have an explicit understanding of the common phrases in
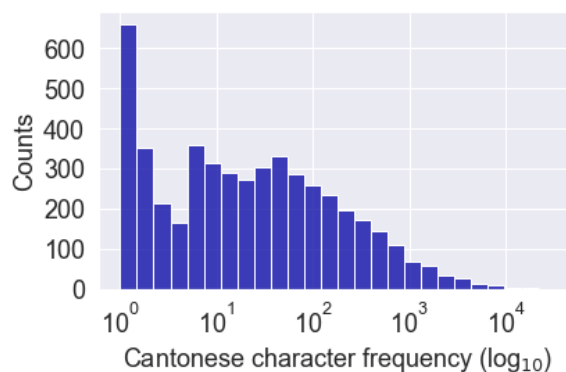
the MDCC, we report the top 10 most common n-grams in Table 5. A small proportion of the characters appear much more frequently than the others in the statistics. In detail, 14.56% of the characters in the MDCC is made up of the 10 most common characters. Meanwhile, there is also a large number of characters that appear less than 10 times in the corpus, which reflects the diversity of the text in our dataset and its compliance with Zipf's law (Yu et al., 2018).

5. Cantonese ASR using Fairseq S2T Transformer

In this section, we introduce the Fairseq S2T Transformer (Wang et al., 2020; Ott et al., 2019) model and conduct experiments on the MDCC and Common Voice zh-HK dataset. We also apply multi-dataset learning approaches to further improve the model's performance.

5.1. Fairseq S2T Transformer

The reasons for choosing Fairseq S2T Transformer are twofold: 1) The model can achieve start-of-the-art performance on LibriSpeech, the de-facto standard ASR benchmark. 2) It is friendly for training custom models for ASR, and we can easily adapt the model to Cantonese. Based on the original transformer architecture (Vaswani

Datasets	# train	# val	# test	# Total
Common Voice zh-HK	65,437	4,089	12,269	81,795
MDCC (ours)	65,120	5,663	12,492	83.275

Table 6: The statistics of the MDCC and Common Voice zh-HK, the two biggest Cantonese datasets, trained and tested with the Fairseq S2T Transformer model.

Test set/Train set	MDCC	Common Voice zh-HK	Joint
MDCC	10.15	83.42	9.38
Common Voice zh-HK	53.44	8.69	7.65
Joint	31.33	51.56	8.63

Table 7: Character error rates (%) returned by models trained on the MDCC, Common Voice zh-HK dataset and both combined datasets.

et al., 2017), Fairseq S2T Transformer proposes to add convolutional layers to the encoder (Mohamed et al., 2019), which is the optimal way of processing audio data in the form of log mel filterbanks. We use the S2T Transformer XS version for all the experiments to implement the model.

5.2. Datasets

We conduct experiments on the two largest Cantonese datasets, MDCC and Common Voice zh-HK, for a better comparison, and jointly train them to see how the performance of the dataset improves if we double its size. The split of the Common Voice zh-HK dataset follows the same ratio as our MDCC, which is 80% for training, 5% for validation and 15% for testing. The detailed information of these two datasets is shown in Table 6. In addition, the audio files are downsampled to the frequency of 16kHz, with 32-bit depth. As follow-up experiments, we join the MDCC and Common Voice zh-HK datasets into one and apply a multi-dataset training approach. We hope that this can improve the model's performance in the cross-dataset setting and increase the robustness of the model.

5.3. Implementation Details

Data pre-processing. We implement spectral augmentation (SpecAugment), a state-of-the-art audio data augmentation method, which is implemented by masking certain frequency and time values on the spectrogram (Park et al., 2019). We use SpecAugment for the Common Voice zh-HK baseline, where it shows an improvement in overall results. Furthermore, we apply cepstral mean and variance normalisation (CMVN) for all the utterances (Strand and Egeberg, 2004). In Fairseq S2T, pre-processed audio can be used directly or stored in the form of .npy files. The latter is the way in which we store features extracted from Cantonese datasets to achieve faster training. For tokenization of the transcribed data, we use the SentencePiece tokenizer (Kudo and Richardson, 2018) with unigram subword tokenization (Kudo, 2018) and an 8,000-word vocabulary. The vocabulary covers 99.95% of the characters in the MDCC (the default coverage for character-based languages).

Hyper-parameters. We use the off-the-shelf Fairseq S2T Transformer XS model, which consists of a 6-layer encoder and a 3-layer decoder with a multi-head attention mechanism with four attention heads. For the objective function, we apply a cross-entropy loss with

0.1 label smoothing. The models are trained on four GPUs with a mini-batch size of 32. We use the default settings of SpecAugment provided in the S2T bundle: frequency masking width parameter F is set to 27, the number of time and frequency masks is set to 1, with the upper bound of the time masking width 1; and the time masking width parameter T is set to 100. In our experiments, the applied SpecAugment policy does not include time warping.

Evaluation Metric. The models are evaluated on separated test sets, as is shown in Table 6. For the evaluation, we average the performance of the last 10 checkpoints of the model using a beam search with beam size 8. Since the transcribed language is character based, we use the CER rather than WER as an evaluation metric (Wang et al., 2013). The CER is calculated by adding the number of substituted, inserted and deleted characters together and dividing them by the total character count of the reference.

5.4. Results and Analysis

The CERs returned by S2T Transformer XS on the MDCC and Common Voice zh-HK datasets are comparable (10.15% and 8.69% CER respectively), possibly due to the similarities in domains and the size of the datasets themselves. We discover, however, that the datasets react differently to spectral augmentation. While SpecAugment hinder the training of the model on the MDCC, it does not influence the model trained on the Common Voice zh-HK dataset. Similarly the training on the joint dataset is hindered by adding SpecAugment. In the joint dataset we originally shuffled the training data, but the model did not converge. We conjecture that if we mix the datasets in the same batch, the model can not reach an optimal direction of gradient descent. The model benefits however from ordering the two datasets such that utterances from the MDCC are featured first and those from Common Voice zh-HK afterwards, supporting the theory that modelling long span word dependencies depends on the ordering of training data (Vazhenina and Markov, 2014). This decision is based on the fact that the MDCC data are cleaner, shorter and therefore easier to learn that those in Common Voice zh-HK.

The CER returned by the model trained on the joint dataset shows improvement in the results of both datasets and a large improvement in the out-of-domain testing scenario. Even though the datasets used by us are the largest among all Cantonese ASR datasets, they are still much smaller compared with more comprehensive

datasets such as LibriSpeech. Each of the Cantonese datasets contains fewer than 100 h of speech, while LibriSpeech alone contains 960 h of English audio (Panayotov et al., 2015). Thus, combining both datasets (i.e. MDCC and Common Voice zh-HK) is a natural step in creating strong Cantonese baselines for data-dependant deep learning models.

6. Conclusion and Future Work

In this paper, we review most of the previous Cantonese ASR corpora and thoughtfully analyze them. To address the limitations of the existing corpora, we propose a new dataset named the MDCC for the ASR research in the Cantonese language, which consists of 73.6 hours of clean read speech. We evaluate our dataset and compare it with the Common Voice zh-HK dataset using the Fairseq S2T Transformer model, and confirm that the results indicate the effectiveness of our proposed dataset. Our model trained on joint data outperforms Wav2Vec2-Large model on Cantonese dataset. [7] For future work we plan to collect data from more audiobooks to enrich our dataset. In addition, we will create new Cantonese ASR corpora from different sources such as meetings and movies. Another future work direction is performing more experiments that combine the performance of the MDCC with multilingual datasets. We believe that our dataset and analysis can pave the way for future research works on the Cantonese ASR task and ASR for other low resource languages.

7. Acknowledgements

8. Bibliographical References

Anderson, A. H., Bader, M., Bard, E. G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., et al. (1991). The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2019). Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baevski, A., Zhou, H., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Bu, H., Du, J., Na, X., Wu, B., and Zheng, H. (2017). Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pages 1–5. IEEE.

Chin, A. (2015). A linguistics corpus of mid-20th century hong kong cantonese. *Department of Linguistics and Modern Language Studies, The Hong Kong Institute of Education, Retrieved*, 23(3):2015.

Dai, W., Cahyawijaya, S., Yu, T., Barezi, E. J., Xu, P., Yiu, C. T. S., Frieske, R., Lovenia, H., Winata, G. I., Chen, Q., Ma, X., Shi, B. E., and Fung, P. (2022). Ci-avsr: A cantonese audio-visual speech dataset for in-car command recognition.

Khare, S., Mittal, A., Diwan, A., Sarawagi, S., Jyothi, P., and Bharadwaj, S. (2021). Low resource asr: The surprising effectiveness of high resource transliteration. *Proc. Interspeech 2021*, pages 1529–1533.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia, July. Association for Computational Linguistics.

Leung, M.-T. and Law, S.-P. (2001). Hkcac: the hong kong cantonese adult language corpus. *International journal of corpus linguistics*, 6(2):305–325.

Li, M., Liu, M., and Masanori, H. (2019). End-to-end speech recognition with adaptive computation steps. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6246–6250. IEEE.

Lin, W.-T. and Chen, B. (2020). Exploring disparate language model combination strategies for mandarin-english code-switching asr. In *ROCLING*.

Lovenia, H., Cahyawijaya, S., Winata, G. I., Xu, P., Yan, X., Liu, Z., Frieske, R., Yu, T., Dai, W., Barezi, E. J., and Fung, P. (2021). Ascend: A spontaneous chinese-english dataset for code-switching in multi-turn conversation.

Luke, K. K. and Wong, M. L. (2015). The hong kong cantonese corpus: design and uses. *Journal of Chinese Linguistics*, 25(2015):309–330.

Malik, M., Malik, M. K., Mehmood, K., and Makhdoom, I. (2021). Automatic speech recognition: a survey. *Multimedia Tools and Applications*, 80(6):9411–9457.

Mohamed, A., Okhonko, D., and Zettlemoyer, L. (2019). Transformers with convolutional context for ASR. *CoRR*, abs/1904.11660.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). fairseq: A fast,

---

[7]`https://huggingface.co/ctl/wav2vec2-large-xlsr-cantonese` The data was not compared in the experiment section since the huggingface uses different splits of Common Voice zh-HK corpus.

extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *Interspeech 2019*, Sep.

Strand, O. M. and Egeberg, A. (2004). Cepstral mean and variance normalization in the model domain. In *COST278 and ISCA Tutorial and Research Workshop (ITRW) on Robustness Issues in Conversational Interaction*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vazhenina, D. and Markov, K. (2014). Sequence memoizer based language model for russian speech recognition. In *SLTU*, pages 183–187.

Wang, P., Sun, R., Zhao, H., and Yu, K. (2013). A new word language model evaluation metric for character based languages. In Maosong Sun, et al., editors, *Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data*, pages 315–324, Berlin, Heidelberg. Springer Berlin Heidelberg.

Wang, C., Tang, Y., Ma, X., Wu, A., Okhonko, D., and Pino, J. (2020). Fairseq s2t: Fast speech-to-text modeling with fairseq. *arXiv preprint arXiv:2010.05171*.

Wang, D., Yu, J., Wu, X., Sun, L., Liu, X., and Meng, H. M. (2021). Improved end-to-end dysarthric speech recognition via meta-learning based model re-initialization. *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 1–5.

Winata, G. I., Cahyawijaya, S., Lin, Z., Liu, Z., and Fung, P. (2020a). Lightweight and efficient end-to-end speech recognition using low-rank transformer. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6144–6148.

Winata, G. I., Cahyawijaya, S., Lin, Z., Liu, Z., Xu, P., and Fung, P. (2020b). Meta-transfer learning for code-switched speech recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3770–3776, Online, July. Association for Computational Linguistics.

Winata, G. I., Cahyawijaya, S., Liu, Z., Lin, Z., Madotto, A., Xu, P., and Fung, P. (2020c). Learning fast

adaptation on cross-accented speech recognition. In *INTERSPEECH*.

Winata, G. I., Wang, G., Xiong, C., and Hoi, S. (2021). Adapt-and-Adjust: Overcoming the Long-Tail Problem of Multilingual Speech Recognition. In *Proc. Interspeech 2021*, pages 2451–2455.

Winterstein, G., Tang, C., and Lai, R. (2020). Cantomap: a hong kong cantonese maptask corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2906–2913.

Xu, Q., Baevski, A., Likhomanenko, T., Tomasello, P., Conneau, A., Collobert, R., Synnaeve, G., and Auli, M. (2021). Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE.

Yu, S., Xu, C., and Liu, H. (2018). Zipf's law in 50 languages: its structural pattern, linguistic interpretation, and cognitive motivation. *CoRR*, abs/1807.01855.

Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., and Wu, Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.