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

Tiezheng Yu*, Rita Frieske*, Peng Xu*, Samuel Cahyawijaya*, Tung Shadow Yiu, Holy Lovenia, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram E. Shi, Pascale Fung
Hong Kong University of Science and Technology

Contact Information:
Department of Electronic & Computer Science
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong, China

Email: tyuah@connect.ust.hk
peng.xu@connect.ust.hk
scahyawijaya@connect.ust.hk
rita.frieske@ust.hk

Introduction

Automatic speech recognition (ASR) systems take the audio as input and convert it into text. Due to the popularization of deep learning, ASR technology has grown rapidly and has led to a significant improvement in recognizing many 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. 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.

Name	Speech Type	Data source	Size [hours]	Availability
HKCAC	Spont.	Phone-in programs	8.1	Non-Public
HKCanCor	Spont.	Chat	30.0	Cvasi-Public
HKCC	Spont.	Movie	35.0	Cvasi-Public
CantoMap	Read	MapTask	12.8	Public
Common Voice zh-HK	Read	Wikipedia	96.0	Public
MDCC (Ours)	Read	Audiobook	73.6	Public

Table 1: Hong Kong Cantonese ASR corpora

Cantonese ASR Datasets

Table 1 lists the most important previous Cantonese ASR corpora with their speech type, data source, data size and availability.

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.

HKCanCor The Hong Kong Cantonese Corpus (HKCanCor) 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.

HKCC The Corpus of Mid-20th Century Hong Kong Cantonese (HKCC) 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.

CantoMap The Hong Kong Cantonese MapTask Corpus (CantoMap) aims to provide a Cantonese corpus for ASR research and also involves several controlled elicitation tasks related to the phonology and semantics of Cantonese.

Common Voice zh-HK The Common Voice zh-HK corpus 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.

Corpus Creation

Audiobook Collection

- The speech corpus of the MDCC is collected from Hong Kong Cantonese audiobook sources.
- Therefore, we apply a voice activity detection (VAD) tool to convert the original audio pieces into shorter audio utterances.
- After separation, we get 83,275 audio utterances with a total corpus size of 73.6 hours.

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.

Corpus Splitting

We randomly split the MDCC into training, validation and test sets. Table 2 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.

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

MDCC: Multi-Domain Cantonese Corpus

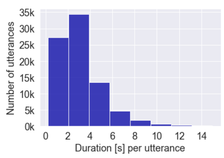

Figure 2: Distribution of the number of characters per utterance.

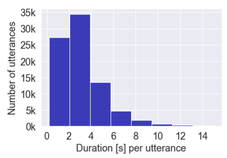

Figure 3: Distribution of the duration (in seconds) per utterance.

Figure 4: The distribution of domains of utterances.

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 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.

Experiments

The experiments compared two major Cantonese datasets, namely MDCC and Common Voice zh-HK. Furthermore, we performed experiments on the Joint dataset (MDCC+Common Voice zh-HK) to explore improvements brought by the large dataset. We have tried settings both with the SpecAugment and without SpecAugment for data augmentation.

Experimental Details

- Model: Fairseq S2T Transformer XS.
- Model details: 6 encoder layers, 3 decoder layers, multihead attention with 4 heads, loss is measured with cross-entropy with 0.1 smoothing
- We have applied SpecAugment as a state-of-the-art augmentation technique with the default Fairseq parameters.
- The results are calculated using Character Error Rate (CER) since Cantonese is a character based language.

Results

Main Results

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

- The results on both datasets were comparable: MDCC 10.15% CER and Common Voice zh-HK 8.69% CER respectively, possibly due to similar sizes of the datasets.
- Joint dataset benefitted from ordering the utterances such that MDCC was placed before Common Voice zh-HK, since its utterances were shorter and easier to learn

Conclusion and Future Work

- We review most of the previous Cantonese ASR corpora and thoughtfully analyze them.
- 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 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.
- 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.