INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# HK-LegiCoST: Leveraging Non-Verbatim Transcripts for Speech Translation

*Cihan Xiao*[1], *Henry Li Xinyuan*[1], *Jinyi Yang*[1], *Dongji Gao*[1] *Matthew Wiesner*[1,2], *Kevin Duh*[1,2], *Sanjeev Khudanpur*[1,2]

[1]Center for Language and Speech Processing, [2]Human Language Technology Center of Excellence
Johns Hopkins University, Baltimore, MD, USA

{cxiao7, xli257, jyang126, dgao5, wiesner, khudanpur}@jhu.edu, kevinduh@cs.jhu.edu

## Abstract

We introduce HK-LegiCoST, a new three-way parallel corpus of Cantonese-English translations, containing 600+ hours of Cantonese audio, its standard traditional Chinese transcript, and English translation, segmented and aligned at the sentence level. We describe the notable challenges in corpus preparation: segmentation, alignment of long audio recordings, and sentence-level alignment with *non-verbatim transcripts*. Such transcripts make the corpus suitable for speech translation research when there are significant differences between the spoken and written forms of the source language. Due to its large size, we are able to demonstrate competitive speech translation baselines on HK-LegiCoST and extend them to promising cross-corpus results on the FLEURS Cantonese subset. These results deliver insights into speech recognition and translation research in languages for which non-verbatim or "noisy" transcription is common due to various factors, including vernacular and dialectal speech.

**Index Terms**: speech recognition, speech translation, corpus

## 1. Introduction

Growing demand for applications such as automatic video captioning and foreign language learning has spawned interest in improving speech-to-text translation (ST), which translates the speech of one language into the text of another language. While most work and progress has focused on high-resource languages, research on the translation of primarily spoken languages, or languages whose written forms deviate from their spoken forms, is relatively scarce. Cantonese is one such language, whose written form is often altered to appear closer to the Mandarin written form. We refer to this written form as standard Chinese and say these transcripts are ”non-verbatim” to reflect the discrepancy between what is spoken and written. This presents significant challenges for automatic speech recognition (ASR) and speech translation (ST) systems.

In this paper, we introduce HK-LegiCoST: a new corpus of Cantonese audio recordings, corresponding standard Chinese transcripts, and paired English textual translation. It contains 600+ hrs of conversational and read speech collected by the Hong Kong Legislative Council [1], mainly focusing on government policy-related inquiries and their corresponding responses, alongside discussions and debates on motions and resolutions. We first describe the challenges in converting this raw "found" resource into a large, curated and useful corpus for language technology research. We then provide automatic speech recognition (ASR), machine translation (MT), and speech translation (ST) baselines on this corpus. On the Google FLEURS [2] test set, an ASR model trained only on our training set leads to results that are comparable to Google's pre-trained and fine-tuned model results. Fine-tuning our baseline model on the FLEURS training set significantly outperforms the previously reported baseline. We also train and benchmark some standard speech translation models on our corpus. We believe this corpus will become a valuable resource for studying vernacular and dialectal speech recognition and translation due to (a) the unique linguistic features of Cantonese, (b) the non-verbatim aspect of writing Cantonese in Standard Chinese, and (c) the relatively large amount of three-way parallel data.

Section 2 provides an overview of prior research on speech translation and vernacular/dialectal speech recognition, Section 3 outlines the corpus creation pipeline, Section 4 presents our baseline experiments and their results, and Section 5 explores some of the distinctive attributes of the corpus.

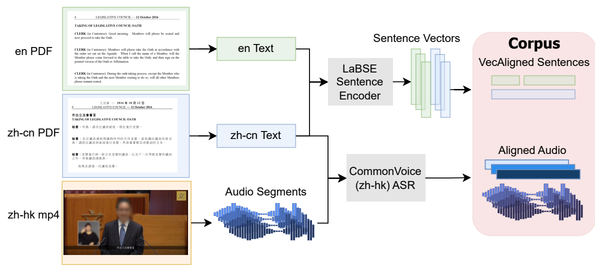


Figure 1: *The corpus creation process. The standard Chinese text (zh-cn) is a non-verbatim transcript of the Cantonese speech (zh-hk).*

## 2. Related work

### 2.1. Prior work on Speech Translation

While historically the most competitive ST systems were cascade systems constructed by attaching an ASR system to an MT system, over the past couple of years greater focus has been placed on the growing potential of end-to-end ST systems. With that comes the need for ST corpora composed of audio recordings annotated with their transcripts translated into the target language(s). Work in this area kicked off with high resource language pairs such as Spanish-English [3] or English-French [4], propelled by venues such as The International Conference on Spoken Language Translation (IWSLT). Multilinguality as well as coverage of lower resource languages soon followed: MuST-C [5] consisted of English speech aligned to transcripts in eight languages; more recent works such as CVSS [6] and FLEURS [2] saw an explosion both in the coverage of languages and in the quantity of data as measured in total speech length. FLEURS represents an important first step in that it was one of the earliest notable corpora for Cantonese speech translation; the size of its Cantonese portion is not sufficient for training a competitive speech translation system, a gap which our corpus aims to fill.

Collecting, transcribing, and translating speech is a costly and difficult process, which has led works like Europarl-ST [7] and Multilingual TedX [8] to leverage public-available, multilingual-captioned audio(visual) data, an approach that we emulate in our work.



10.21437/Interspeech.2023-2351

### 2.2. ASR, MT and ST corpora for Cantonese, other vernaculars and/or unwritten languages

While corpora on many spoken forms of standard Chinese are few and far between, a plethora of corpora for spoken Cantonese have been developed. Yu et al. [9] produced an excellent survey of existing Cantonese speech corpora, as well as a Cantonese ASR corpus comparable in size to the Cantonese section of the Common Voice corpus [10] or the Babel corpus [11].

Several other studies have drawn from the same data source – the Legislative Council (LegCo) meeting records – for translation-related studies. One of the most notable corpora for Cantonese MT is the Hong Kong Hansards Parallel Text [12], which draws from LegCo's records between 1995 and 2000. Kwong's 2021 study [13] extracted a small number of sentence pairs from LegCo meeting records to conduct a focused analysis of the linguistic characteristics of translation and interpretation of Cantonese. By contrast, we aim to support cutting-edge research in ASR and ST using a large amount of resources.

Vernacular speech corpora have grown in number in the recent past. The MGB-2 Arabic dataset [14] has served as the backbone for the eponymous MGB-2 challenge, and contained some amount of dialectal Arabic with non-verbatim transcripts. Subsequent challenges, such as the IWSLT Tunisian Arabic speech translation challenge [15], focused specifically on dialectal translation, but the transcripts were verbatim. SDS-200 [16] is a recent work that tackles the speech translation between Swiss German and Standard German.

One important characteristic of Cantonese is that it is a partially unwritten language: while it has its own writing system that captures the phonological and phonetic characteristics of spoken Cantonese, there are many situations where it is instead transcribed in standard written Chinese which does not have this property[17]. The source material of our dataset falls into the latter category. A recent study by Chen et al. [18] addressed a task in a similar setting (Hokkien ST), although both works and data in this area are still relatively scarce.

## 3. Corpus Creation

### 3.1. Data Collection and Preprocessing

The raw data of the corpus is a collection of video recordings of the Hong Kong Legislative Council's regular meetings, their corresponding transcripts, and English translations in PDF format. The Hong Kong Legislative Council, also known as HK-LegCo, is the unicameral legislature of the Hong Kong Special Administrative Region of China. The corpus was mainly derived from the recordings of the council's regular meetings from 2016 to 2021, covering a wide range of topics such as political reform, education policy, housing issues, transportation infrastructure, healthcare reform, and economic development.

To process this raw data into a usable corpus, we developed an integrated pipeline shown in Figure 1.[1] It accepts the original video recordings of each meeting, and the corresponding transcript and English translation in PDF format, and produces triplets of segmented sentence-sized audio clips, together with their transcript and translation. Each module in the integrated pipeline entails one or more technological tasks. In the following sections, we provide a comprehensive description of the technological steps involved in each of the tasks.

Typically, a regular meeting of the Council lasts for about half a day, with a video released for each meeting session. We start off by converting each video recording to WAV format, resampling it to 16KHz, and segmenting it based on metadata, i.e. the topic-level timestamps associated with each video indicating the start time of a certain topic (or sometimes speaker), yielding audio clips whose length generally spans from 5 to 30 minutes. As the raw recordings contain visual information such as the speaker's lip movements and sign language that corresponds to the speech, we plan to release a version of the corpus with visual information in a subsequent iteration of this work.

The goal of text preprocessing is to filter out irrelevant information in the PDF transcripts and segment the full document into shorter sections. The raw text was first extracted from meeting transcripts and translation files. Paragraphs with a Chinese speech marker (name plus colon in bolded) are extracted from the raw text. This process automatically divides the full document into speaker ID labeled segments, allowing for more efficient bitext and audio-text alignment.

### 3.2. Bitext Sentence Alignment

Sentence-level bitext alignment is crucial for ST corpus creation. To this end, we split each speaker-ID-labeled text segment into sentences based on punctuations, and generate a contextualized sentence embedding for each sentence using LaBSE [19], a multilingual sentence BERT model. We then used VecAlign [20] to perform alignment on the embeddings. The VecAlign algorithm works by using pre-trained sentence embeddings to compute a similarity score between sentences in the source and target languages, followed by applying a dynamic programming approach based on the Fast Dynamic Time Warping algorithm to approximately find the optimal alignment between the sentences in linear time. We utilized the alignment score generated by VecAlign to identify and eliminate inaccurate alignment results (score $> 0.627$). The threshold was obtained via a grid search with manual inspection of the filtered instances.

### 3.3. Audio-Text Sentence Alignment

#### *3.3.1. Alignment Model Training*

We performed audio-text alignment using an ASR model trained on the Cantonese language subset of the CommonVoice corpus [21] with the ESPNet toolkit [22]. Specifically, the model used an encoder-decoder architecture with a Conformer [23] encoder, which was trained under the CTC/attention hybrid multitask learning [24] framework without precomputed feature. One of the challenges of alignment model training is the discrepancy between what is written and what is spoken, i.e. non-verbatim transcripts. Therefore, we chose the CommonVoice corpus as our training data for the alignment ASR model, since its transcripts are similar to standard written Chinese. To reduce the disparity in vocabulary, we adopted a character-based tokenization approach for Chinese characters, while code-mixed English words and numbers were treated as single characters during the tokenization process. We converted the characters into jyutping, a romanization system for Cantonese, with the python-pinyin-jyutping-sentence tool,[2] for addressing the vocabulary differences between the training and target corpora. We also used k2,[3] a toolkit for integrating Finite State Automaton (FSA) and Finite State Transducer (FST) algorithms with autograd-based machine learning models, to construct WFST-based lattice and perform decoding.

#### *3.3.2. First-pass (Paragraph) Alignment*

We first conducted a first-pass alignment process that approximately aligns audio segments with corresponding sections in the transcript using an anchor-based method. To achieve this, we utilized Silero-VAD [25], a voice activity detection tool, to extract audio segments of suitable duration for decoding. Next, we created a 3-gram language model that was biased towards the target document. Using this biased language model as G graph, we applied HLG decoding to the VAD-segmented au-

[1]The code of the pipeline is available at: `https://github.com/BorrisonXiao/espnet-align/tree/master/egs2/commonvoice_align/asr1/align.sh`

[2]`https://github.com/Language-Tools/pinyin-jyutping`

[3]`https://github.com/k2-fsa/k2`

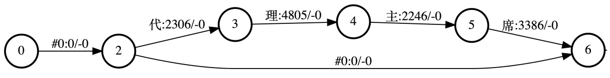


Figure 2: *A sample G graph for flexible alignment.*

dio clips. After decoding, we concatenated the decoding results for audio clips from the same meeting and aligned them with the full transcript text using KALDI's align-text tool [26]. We identified anchors, which represent reliably aligned regions of the original audio segment, based on the alignment produced by SailAlign [27], using a set of criteria including CER, number of consecutive errors, and absolute errors. We expanded the boundaries of these anchors to account for potential decoding errors and mapped the resulting reference text regions to the audio clip, generating the first-pass alignment output.

#### 3.3.3. *Sentence-level Alignment*

In our efforts to perform sentence-level alignments, we encountered two primary challenges. Firstly, we found that supplementary text that was not spoken during the meeting was present in the transcript, complicating the alignment process. Secondly, for longer audio segments lasting upwards of 10 minutes, decoding proved to be challenging due to memory constraints. To overcome these challenges, we implemented a sliding-window flexible alignment algorithm, which enabled us to address both issues by efficiently breaking down long segments into smaller ones of manageable lengths while simultaneously filtering out non-spoken text from the transcript.

The basic flexible alignment algorithm is a variation of the forced-alignment algorithm.[4] Specifically, the algorithm employs a linear FSA as the G graph for decoding, but includes a skip connection from the start of a sentence towards the end of the sentence with customizable weight. This approach allows the algorithm to skip a sentence that may contain irrelevant text and focus on the next sentence that is more likely to be spoken. Our approach resembles the factor transducer [28] method, which enables the decoding of sub-strings for long audio alignment by adding a skip connection from the start state to each state in the linear FSA and making each state a terminal state. However, unlike the factor transducer, our approach operates at the sentence level and permits the skipping of redundant sentences in-between.

To further improve the efficiency and performance of the algorithm, a sliding-window strategy is used to break down audio segments into shorter, overlapping sub-segments, which can be processed in parallel on a GPU. The length of the decoding graph decreased with the length of the input audio signal, leading to a reduction of the search space and therefore a significant improvement in the alignment accuracy. The algorithm has been shown to effectively filter out unspoken text and improve the accuracy of sentence-level alignments, particularly for longer audio segments.

#### 3.3.4. *Post-Filtering*

In our pipeline, we utilized ASR decoding to evaluate the quality of the alignment by measuring the CER, number of consecutive errors, and error ratio. To ensure optimal alignment results, we established a threshold by creating CER-based bins and randomly sampling 300 utterances from the corpus based on the bins. We manually labeled the samples and fine-tuned the threshold to optimize the precision of the subset and to filter out instances where the speech contained significant deviations from the transcript due to repetitions and other disfluencies.

[4]https://github.com/DongjiGao/flexible_alignment.git

Table 1: *Split Overview. The perplexity of the dev and test sets is computed based on a 3-gram language model trained on the training split. The LM used is character-level for Cantonese and word-level for English.*

| **Split** | **Hours** | **# Utts** | **# Tokens** | | **Perplexity** | |
|---|---|---|---|---|---|---|
| | | | zh-hk | en | zh-hk | en |
| train | 518 | 142K | 6.8M | 4.8M | - | - |
| dev-ASR | 41 | 11.1K | 557K | 380K | 36.7 | 394.8 |
| dev-MT | 40 | 10.7K | 519K | 371K | 39.5 | 371.5 |
| test | 10 | 2.5K | 123K | 84K | 40.3 | 471.5 |
| **Total** | 609 | 166.3K | 8M | 5.6M | - | - |

### 3.4. Corpus Splits

We partitioned the data into four major subsets: training (train), ASR development (dev-ASR), MT development (dev-MT), and test (test) sets. We ensured that the test set was disjoint in both documents and speakers with respect to the training set, while preserving the gender distribution in the training set. The splits were created using `nachos`,[5] a tool that automates the creation of held-out partitions using metadata of the units to be split. Specifically, we defined each document in the corpus as the transcript of a full meeting recording, and used the document ID, speaker ID, and speaker gender as features for `nachos`. As a result, the dev-ASR set was created with speaker disjointness from the train set, while the dev-MT set was formed with document disjointness from the train set. The test set was both speaker and document disjoint from the train set. In addition, NACHOS approximately preserved the gender and speaker distribution across all splits based on gender features. Table 1 displays the statistics of the splits. The perplexity measure suggests, as expected, that the Cantonese content in the dev-MT and test split is less similar with respect to the training set comparing to the dev-ASR split. Yet, the same pattern does not apply for the English text, presumably due to the translator switches across documents. In addition, it is worth mentioning that we partitioned the dev-ASR and dev-MT into three subsets each to reduce the training and validation cost.

## 4. Baseline Experiments and Results

### 4.1. Speech Recognition

We employed icefall to create baseline ASR models for the corpus.[6] The 100M parameters Conformer model with CTC loss was trained using the same character-level tokenization and pronunciation lexicons as detailed in Section 3.3.[7] 80-dimensional fbank features were used as input features. We trained a baseline model using only the HK-LegiCoST data and then finetuned it for 20 epochs on the FLEURS training set.

To further evaluate the robustness of the corpus, we tested the model on the yue_hant_hk subset of Google's FLEURS corpus which, similar to our dataset, contains Cantonese speech transcribed in standard simplified Chinese. We evaluated the model's zero-shot performance and performance after finetuning on the FLEURS training set. Despite being only 1/6th the size of the FLEURS baseline model, the HK-LegiCoST baseline model achieved competitive results in the zero-shot setting, while the fine-tuned model significantly outperformed the FLEURS baseline. Furthermore, the fine-tuned model suffered no performance degradation on the HK-LegiCoST test set, which indicates that the two corpora share a similar domain. These results demonstrate the robustness of our corpus from the ASR perspective.

[5]https://github.com/m-wiesner/nachos.git

[6]https://github.com/k2-fsa/icefall

[7]The code for model training and finetuning is available at https://github.com/BorrisonXiao/icefall/blob/master/egs/hklegco/ASR/train.sh

Table 2: *CER↓ of the baselines. The Conformer model is the baseline trained on the HK-LegiCoST training corpus, while the Conformer + FT model represents the finetuned version of the Conformer baseline on the FLEURS training set.*

| Model | HK-LegiCoST | FLEURS |
|---|---|---|
| FLEURS [2] | - | 37.0 |
| Conformer | 23.2 | 42.9 |
| Conformer + FT | **23.0** | **26.3** |

### 4.2. Machine Translation

Our baseline MT system was built using FAIRSEQ [29], employing a transformer-based architecture [30]. We adopted a unigram character-level tokenization method to tokenize the source (Cantonese) text and a BPE-based [31] approach to tokenize the target English text. Our baseline achieved a BLEU score of 24.9 on our test set. Table 3 compares our MT system with one of the best existing benchmarks, M2M-100 [32]. Zero-shot M2M100 performed very poorly on named entities, whereas our system performed much better in that regard; this, combined with the relatively large number of similar utterances in parliamentary proceedings, resulted in our trained-from-scratch system pulling a sizeable gap ahead of M2M100.

Table 3: *BLEU↑ of MT and Cascaded ST on HK-LegiCoST test set.*

| Input | Ours | M2M100 |
|---|---|---|
| Oracle transcript | **24.9** | 12.1 |
| ASR hypotheses | **17.3** | 7.9 |

### 4.3. Cascaded Speech Translation

We report the BLEU scores of the two cascaded systems, each of which takes the output of the ASR system as described in section 4.1 and translates with the two MT systems outlined in section 4.2. The results are shown in Table 3. The unigram-input transformer vastly outperformed the bpe-input pre-trained m2m-100 model: the latter suffered greatly from the mismatch between its input vocabulary and the output vocabulary of our ASR system.

## 5. Discussion

We observed two notable features in the proposed corpus, namely the phenomenon of word and phrase reordering introduced by non-verbatim transcript and the frequent occurrence of document-level contextual dependencies such as co-reference and named entities.

### 5.1. Text Reordering

Two major factors were identified as contributing to the phenomenon of word and phrase reordering. The difference between the grammar of Cantonese and standard written Chinese can lead to word reordering. For instance, the phrase "你走先" (you go first) in Cantonese, when transcribed to standard Chinese, is converted to "你先走" (you first go). In addition, the formalization of the speech, as illustrated in table 4, was an outcome of the transcriber's attempt to enhance the fluency of the text in standard Chinese.

### 5.2. Long Context Dependency

Due to the nature of the council meetings, it is common for the translations of the transcripts to exhibit contextual dependencies at the document level. Errors made by our baseline translation system are often attributed to instances of such contextual dependence. We speculate that a translation system that could make use of these types of contextual information would outperform our baseline system.

Table 4: *An example of data triplet in the corpus, where the differences between Cantonese and standard written Chinese are highlighted in blue, the results of lexical replacement for linguistic fluidity enhancement are highlighted in red, and the reordered text resulting from formalization is underlined.*

| | |
|---|---|
| Cantonese Speech | 佢 係我 在 上 水 石 湖 墟 辦 事 處經常來的街坊 |
| Jyutping | keoi5 hai6 ngo5 zoi6 soeng5 seoi2 sek6 wu4 heoi1 baan6 si6 cyu3 ging1 soeng4 loi4 dik1 gaai1 fong1 |
| Reference Transcript | 她是經常到訪我在上水石湖墟的辦事處的街坊 |
| Jyutping | taa1 si6 ging1 soeng4 dou3 fong2 ngo5 zoi6 soeng5 seoi2 sek6 wu4 heoi1 dik1 baan6 si6 cyu3 dik1 gaai1 fong1 |
| English Translation | She is a local resident who has made frequent visits to my office in Shek Wu Hui in Sheung Shui |

## 6. Conclusions

We created the HK-LegiCoST corpus for the research of vernacular speech recognition and translation. The original data was collected from publicly-available meeting recordings and transcripts from the Hong Kong Legislative Council. Our corpus, with 518 hours of Cantonese speech and 142k sentence pairs, is (to our knowledge) one of the largest for both Cantonese ASR and Cantonese-English speech translation. We report some of the linguistic characteristics in our corpus, namely text reordering and long context dependency. Results from our baseline experiments show that our corpus is validated and of great value for the research of ASR and ST. In addition, our approaches successfully address some of the top challenges for creating a corpus from scratch, especially for segmenting and aligning long recordings with non-verbatim transcripts. We are currently in the process of going through the formality of releasing our corpus to the public, as we aim to ensure that it is made available in accordance with established protocols and procedures.

## 7. Acknowledgements

We are grateful to the Legislative Council of the Hong Kong Special Administrative Region for making their meeting proceedings and bilingual transcripts publicly available, which has allowed us to use them as a resource for academic research. We thank Neha Verma and Weiting Tan from Johns Hopkins University for helpful early discussions.

## 8. References

[1] [Online]. Available: https://www.legco.gov.hk/tc/index.html

[2] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: few-shot learning evaluation of universal representations of speech," in *SLT 2022, Doha, Qatar, January 9-12, 2023*. IEEE, 2022, pp. 798–805. [Online]. Available: https://doi.org/10.1109/SLT54892.2023.10023141

[3] M. Post, G. Kumar, A. Lopez, D. Karakos, C. Callison-Burch, and S. Khudanpur, "Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus," in *IWSLT 2013*, Heidelberg, Germany, Dec. 5-6 2013. [Online]. Available: https://aclanthology.org/2013.iwslt-papers.14

[4] A. C. Kocabiyikoglu, L. Besacier, and O. Kraif, "Augmenting

librispeech with French translations: A multimodal corpus for direct speech translation evaluation," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA), May 2018. [Online]. Available: https://aclanthology.org/L18-1001

[5] M. A. Di Gangi, R. Cattoni, L. Bentivogli, M. Negri, and M. Turchi, "MuST-C: a Multilingual Speech Translation Corpus," in *NAACL 2019*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 2012–2017. [Online]. Available: https://aclanthology.org/N19-1202

[6] Y. Jia, M. Tadmor Ramanovich, Q. Wang, and H. Zen, "CVSS corpus and massively multilingual speech-to-speech translation," in *LREC 2022*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6691–6703. [Online]. Available: https://aclanthology.org/2022.lrec-1.720

[7] J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan, "Europarl-st: A multilingual corpus for speech translation of parliamentary debates," in *ICASSP 2020*, 2020, pp. 8229–8233.

[8] E. Salesky, M. Wiesner, J. Bremerman, R. Cattoni, M. Negri, M. Turchi, D. Oard, and M. Post, "The multilingual tedx corpus for speech recognition and translation," 08 2021, pp. 3655–3659.

[9] T. Yu, R. Frieske, P. Xu, S. Cahyawijaya, C. T. Yiu, H. Lovenia, W. Dai, E. J. Barezi, Q. Chen, X. Ma, B. Shi, and P. Fung, "Automatic speech recognition datasets in Cantonese: A survey and new dataset," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 6487–6494. [Online]. Available: https://aclanthology.org/2022.lrec-1.696

[10] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC 2020*. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222. [Online]. Available: https://aclanthology.org/2020.lrec-1.520

[11] M. Harper, "The iarpa babel multilingual speech database," 2011.

[12] X. Ma, *Hong Kong Hansards Parallel Text LDC2000T50. Web Download*. Philadelphia: Linguistic Data Consortium.

[13] O. Y. Kwong, *Probing a Two-Way Parallel T&I Corpus for the Lexical Choices of Translators and Interpreters*. Singapore: Springer Singapore, 2021, pp. 101–132. [Online]. Available: https://doi.org/10.1007/978-981-16-4918-9_5

[14] A. Ali, P. Bell, J. Glass, Y. Messaoui, H. Mubarak, S. Renals, and Y. Zhang, "The mgb-2 challenge: Arabic multi-dialect broadcast media recognition," 12 2016, pp. 279–284.

[15] A. Antonios, B. Loc, L. Bentivogli, M. Z. Boito, B. Ondřej, R. Cattoni, C. Anna, D. Georgiana, D. Kevin, E. Maha *et al.*, "Findings of the iwslt 2022 evaluation campaign." in *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*. Association for Computational Linguistics, 2022, pp. 98–157.

[16] M. Plüss, M. Hürlimann, M. Cuny, A. Stöckli, N. Kapotis, J. Hartmann, M. A. Ulasik, C. Scheller, Y. Schraner, A. Jain, J. Deriu, M. Cieliebak, and M. Vogel, "SDS-200: A Swiss German speech to standard German text corpus," in *LREC*. Marseille, France: European Language Resources Association, Jun. 2022, pp. 3250–3256. [Online]. Available: https://aclanthology.org/2022.lrec-1.347

[17] R. Xiang, H. Tan, J. Li, M. Wan, and K.-F. Wong, "When Cantonese NLP meets pre-training: Progress and challenges," in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*. Taipei: Association for Computational Linguistics, Nov. 2022, pp. 16–21. [Online]. Available: https://aclanthology.org/2022.aacl-tutorials.3

[18] P.-J. Chen, K. Tran, Y. Yang, J. Du, J. Kao, Y.-A. Chung, P. Tomasello, P.-A. Duquenne, H. Schwenk, H. Gong, H. Inaguma, S. Popuri, C. Wang, J. Pino, W.-N. Hsu, and A. Lee, "Speech-to-speech translation for a real-world unwritten language," 2022. [Online]. Available: https://arxiv.org/abs/2211.06474

[19] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, and W. Wang, "Language-agnostic bert sentence embedding," 2020. [Online]. Available: https://arxiv.org/abs/2007.01852

[20] B. Thompson and P. Koehn, "Vecalign: Improved sentence alignment in linear time and space," in *EMNLP-IJCNLP*. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 1342–1348. [Online]. Available: https://aclanthology.org/D19-1136

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *LREC 2020*, 2020, pp. 4211–4215.

[22] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-end speech processing toolkit," in *Proceedings of Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456

[23] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," 2020. [Online]. Available: https://arxiv.org/abs/2005.08100

[24] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," 2016. [Online]. Available: https://arxiv.org/abs/1609.06773

[25] S. Team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[26] 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 *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[27] A. Katsamanis, M. Black, P. Georgiou, L. Goldstein, and S. Narayanan, "Sailalign: Robust long speech-text alignment," 01 2011.

[28] P. J. Moreno and C. Alberti, "A factor automaton approach for the forced alignment of long speech recordings," in *2009 ICASSP*, 2009, pp. 4869–4872.

[29] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 48–53. [Online]. Available: https://aclanthology.org/N19-4009

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., vol. 30. Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[31] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725. [Online]. Available: https://aclanthology.org/P16-1162

[32] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Celebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, and A. Joulin, "Beyond english-centric multilingual machine translation," *J. Mach. Learn. Res.*, vol. 22, no. 1, jan 2021.