Conversational multi-turn Chinese-English code-switching corpus ASCEND

2.1. Recording procedure
- We collect 23 speakers’ speech data through casual 1-on-1 conversations.
- Each recording takes up approx. 1 hour.
- We obtain a total of 49 sessions, involving various topics to encourage code switch occurrence and vocabulary diversity.

2.2. Annotation
- We split the audio into utterances based on a natural semantic boundary or a long pause.
- Annotation guidelines: see the paper.

3. Corpus profile
- Mean per utterance: 3.1 sec, 11.78 tokens.
- Speakers: 13 females + 10 males from HK, Taiwan, and China (mean age 24, SD 2.24; mean IELTS speaking score 6.5, min 5.5).
- Speakers’ first language is Chinese.
- The language switch occurs between full utterances, hence all the utterances are monolingual.

3.2. Topic and code-switching
- Out of 49 sessions: 13 persona, 7 sports, 12 education, 4 philosophy, 13 technology.
- What topics trigger more code-switches?
  - (Intra-sentential) With popular English terms, e.g., technology and philosophy.
  - (Inter-sentential) Are familiar to the speakers, e.g., persona/self-intro.

3.3. Common English phrases in ASCEND
- A few types of phrases come up more often than others.

4. ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation
- Despite the abundance of Chinese-English corpora for code-switching, many either use read speech, which does not capture the particular actualities of spoken speech, or are no longer publicly available.
- Therefore, we introduce ASCEND, a spontaneous multi-turn conversational dialogue Mandarin Chinese-English code-switching corpus.

4.1. Code repository
The experiment code can be found at github.com/HLTCHKUST/ASCEND.

4.2. Experiment settings
- Baseline models with CTC loss. We use pre-trained wav2vec 2.0 models with 3 different initializations: 1) original/multilingual (no fine-tuning), 2) fine-tuned on English Common Voice, and 3) fine-tuned on Chinese Common Voice.
- Preprocessing. We extend the vocabulary of the pre-trained tokenizers with ASCEND-specific vocabulary from the transcription data. We normalize the audio data and apply SpecAugment to increase the models’ robustness.
- Evaluation. We generate the transcriptions with CTC decoding. As for the metrics, we use MER and CER.

4.3. Results and analysis
- The Chinese pre-trained wav2vec 2.0 model outperforms both the English and the multilingual pre-trained models.
- ASCEND’s baseline experiment yields ~28% MER and ~23% CER, which are comparable to other works on code-switching datasets. Additionally, ASCEND is also on par in terms of dataset size, tokens variety, and word distribution.
- These results indicate that ASCEND is reliable for training and evaluating Chinese-English code-switching ASR.

5. Conclusion
- We introduce ASCEND. It consists of 10.62 hours of clean spontaneous speech with a total of ~12.3K utterances. The corpus has balanced gender proportion.
- We analyze the statistics of code-switching utterances in ASCEND.
- We conduct baseline experiments with 3 variants of pre-trained wav2vec 2.0 models, achieving a best performance of 22.69% CER and 27.05% MER.