

ASCEND: A Spontaneous Chinese-English Dataset for Code-switching in Multi-turn Conversation

Holy Lovenia, Samuel Cahyawijaya, Genta Indra Winata, Peng Xu, Xu Yan, Zihan Liu, Rita Frieske, Tiezheng Yu, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram E. Shi, Pascale Fung

The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

1. Introduction

- 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.

2. Corpus Collection and Annotation

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.

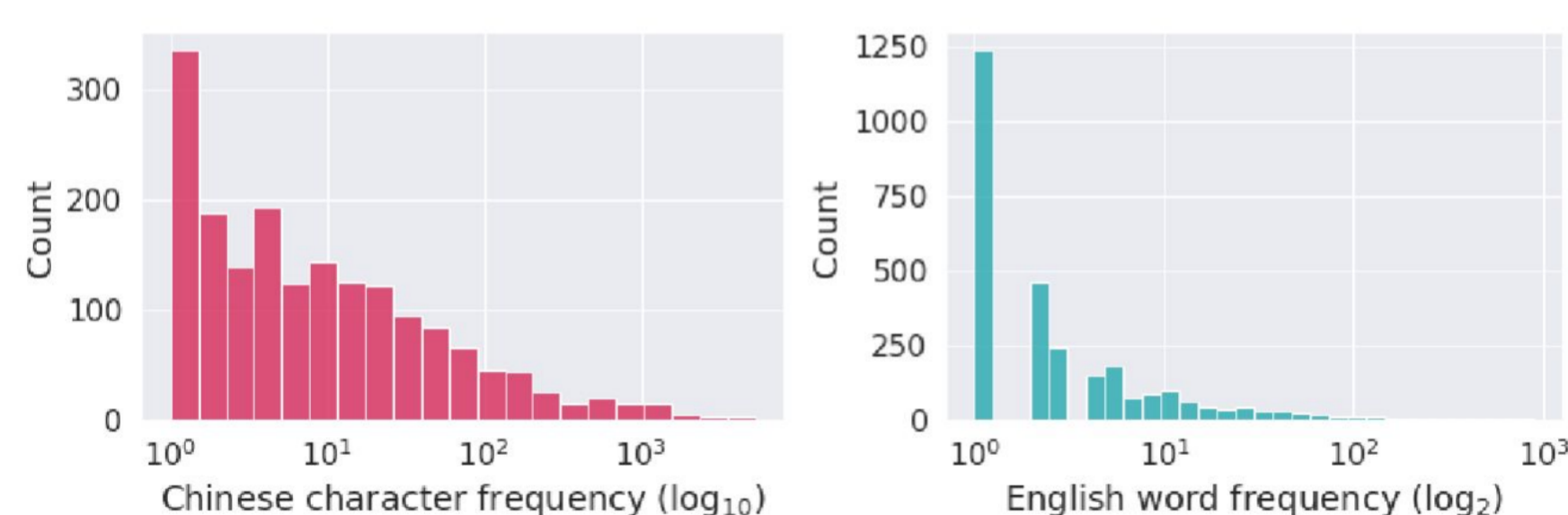
2.2. Annotation

- We split the audio into utterances based on a natural semantic boundary or a long pause.
- Manual transcription using Chinese characters and English letters.
- Annotation guidelines: see the paper.

3. ASCEND: A Spontaneous Chinese-English Dataset

3.1. Corpus profile

- 10.62 hours, 12.3K utterances, 145K tokens (unique: 1.8K Chinese + 2.8K English).
- 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).



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

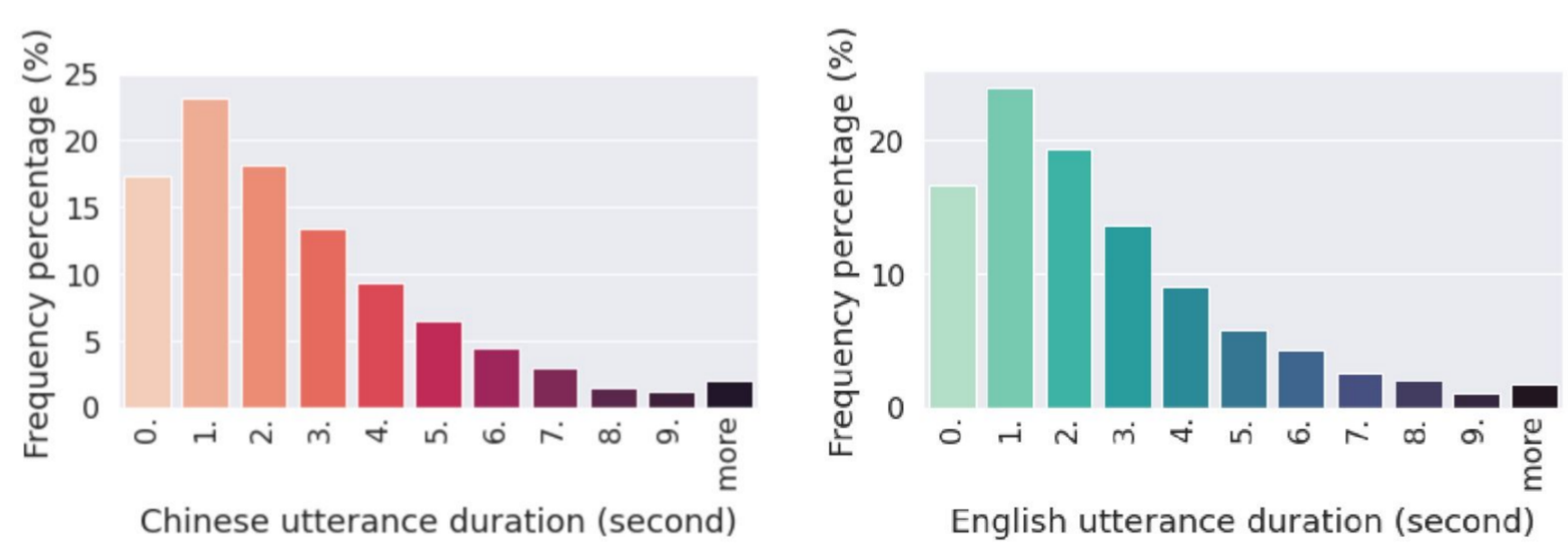
A few types of phrases come up more often than others.

Top	English phrases 1-gram	2-gram	3-gram
1	the	do you	do you think
2	you	in the	what do you
3	to	you can	how to say
4	like	kind of	in hong kong
5	and	smart phone	this kind of

Table 8: Top English 1-gram, 2-gram, and 3-gram phrases.

3.4. Inter-sentential code-switching

The language switch occurs between full utterances, hence all the utterances are monolingual.



3.5. Intra-sentential code-switching

人们like一般观众or public对这些活动的interest不是那么高

- An utterance is considered to have intra-sentential code-switching when a switch from one language to another happens within the utterance at least once.
- Language turns per utterance: mean 2.18 times, max 14 times.

Top	Language turn zh → en	en → zh
1	个 project	school 的
2	读 phd	phd 的
3	个 topic	ok 的
4	做 research	smartphone 的
5	的 major	phone 的

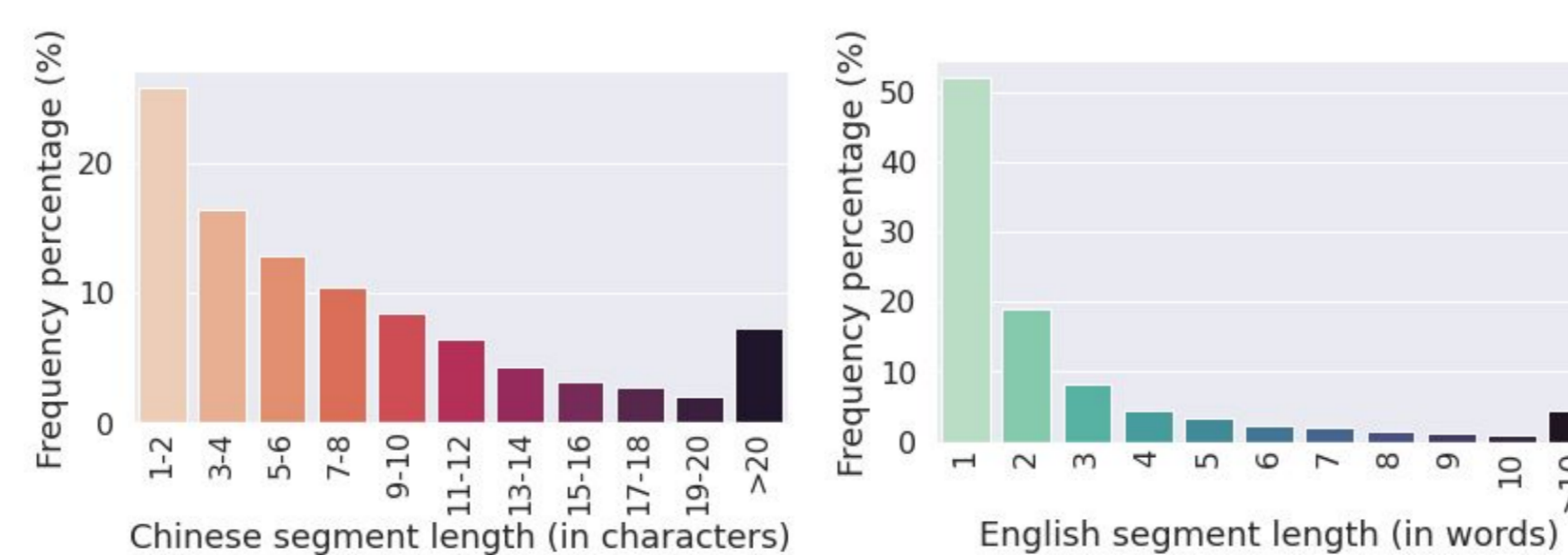
Table 9: Top 5 code-switches in language turns between Chinese and English.

- Language turns within the utterances cause an utterance to be composed of multiple monolingual segments.

Top	Chinese segments 1-char	2-char	English segments 1-word	2-word
1	的	就是	ai	smart phone
2	啊	然后	phd	social media
3	是	所以	ok	hong kong
4	对	这个	so	i think
5	吗	那个	and	it's like

Table 10: Top 5 short monolingual segments in intra-sentential code-switching.

- An intra-sentential code-switching utterance typically comprises 1.75 Chinese segments and 1.38 English segments.
- ASCEND speakers tend to talk in longer Chinese segments (mean per segment 7.96 chars) then switch to a shorter English segment (2.96 words) in between.
 - This is expected, considering that all the speakers' first language is Chinese.



3.6. ASCEND release

ASCEND is publicly available for download at huggingface.co/datasets/CAIRE/ASCEND.

4. Baseline Experiment

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

Pre-training language	Validation		Test	
	MER (%)	CER (%)	MER (%)	CER (%)
Chinese	30.37	25.72	27.05	22.69
English	35.77	28.07	28.72	22.78
Multilingual	35.30	28.68	29.35	24.31

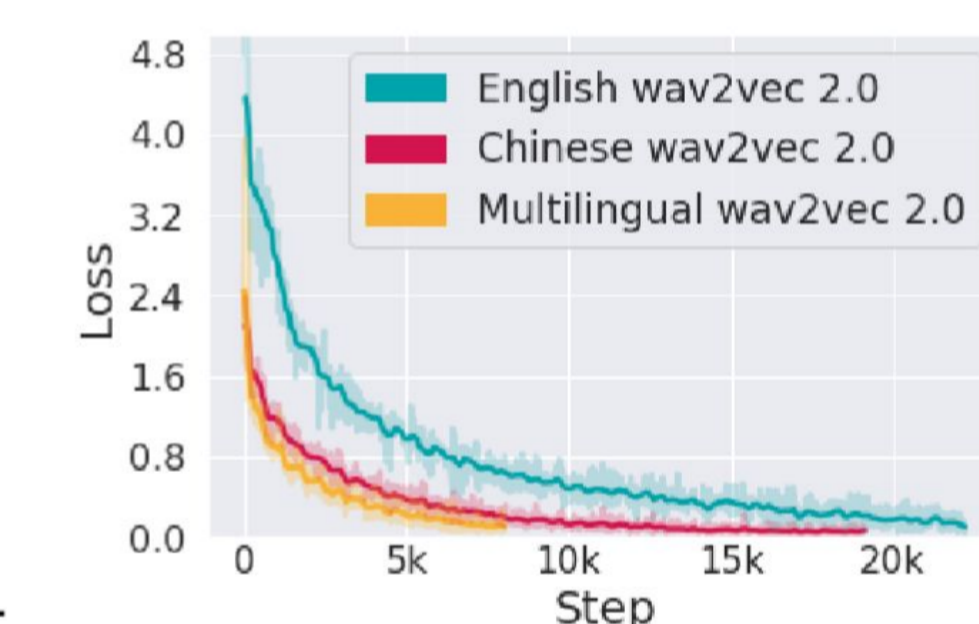


Figure 10: Loss on ASCEND train set in the baseline experiments.

Table 12: Baseline experiment results on ASCEND validation and test set. **Bold** denotes the best performance over different models.

- 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.