SPEECH COLLAGE: CODE-SWITCHED AUDIO GENERATION BY COLLAGING MONOLINGUAL CORPORA

Amir Hussein[1*], Dorsa Zeinali[2*], Ondrej Klejch[3], Matthew Wiesner[1], Brian Yan[4], Shammur Chowdhury[5], Ahmed Ali[5], Shinji Watanabe[4], Sanjeev Khudanpur[1]

[1]Johns Hopkins University, USA, [2]Northeastern University, USA, [3] University of Edinburgh, UK, [4] Carnegie Mellon University, USA, [5] Qatar Computing Research Institute, Doha

Introduction

- Code Switching (CS) occurs when speakers use two or more languages within a sentence.
- Automatic speech recognition (ASR) struggles with recognizing CS speech due to a lack of transcribed training data, grammatical structure complexity, domain mismatch.
- Given the abundance of transcribed monolingual speech in many languages and labeled CS speech scarcity, there's a pressing need to harness monolingual resources for CS applications.
- We introduce *Speech Collage*, a data augmentation technique that constructs synthetic code-switching audio data from monolingual data.

Speech Collage

❶ For each token in a CS text (words for English and Arabic and characters for Mandarin), Speech Collage identifies a random instance of that token and combines them using overlap add

❷ Segments represent diverse speaker and audio environments

❸ **Overlap add**: Extend segments by 0.05 seconds on both sides and use overlap add with a Hamming window to mitigate discontinuity effects

❹ After splicing, utterances are then further refined with energy normalization.

❺ **Energy Normalization**

For a speech sequence X of length T, $X = \{x_t \in \mathbb{R} | t = 1, \cdots, T\}$. The average audio energy is calculated as follows:

$$e = \frac{1}{T} \sum_t x_t^2 \quad (1)$$

e is then used to normalize the utterance via

$$X' = \left\{ \frac{x_t}{\sqrt{e}} | t = 1, \cdots, T \right\} \quad (2)$$

Experimental Setup

We demonstrate efficacy of speech collage with two scenarios, both using **in-domain real CS text** and **synthetic CS text** to generate audio from monolingual data.

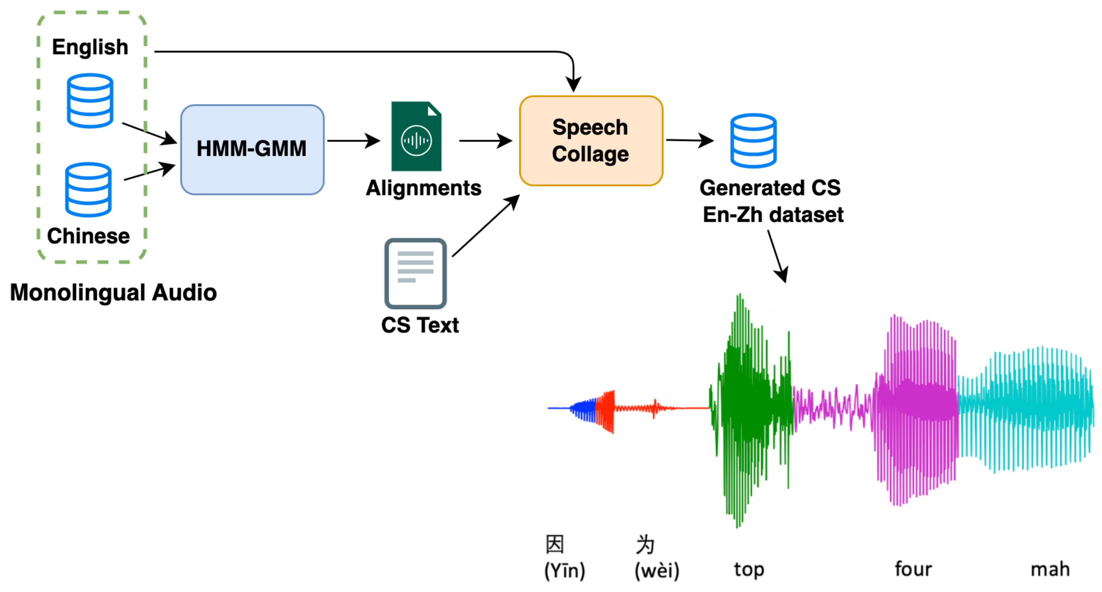

Figure 1: High level overview of the proposed Speech Collage CS generation approach.

In-domain CS Text

- The datasets used were SEAME, Tedlium3 and AISHELL-1
- Using Speech Collage, we produced 62.2 hours of code-switched Mandarin-English data.

Synthetic CS text

- Assuming no real CS text, the **Zero-shot synthetic CS text generation pipeline** is as follows: 1) Generate parallel translated text, 2) Align words, 3) Randomly swap with a 20% rate
- The datasets used were MGB-2, Tedlium3 and ESCWA
- Using speech collage, we produced 80 hours of code-switched Arabic-English data

End-to-End Speech Recognition:

- In this work, we utilized the end-to-end (E2E) ASR conformer-encoder, transformer-decoder architecture, with the ESPNET toolkit.

Results

Table 1. Comparison of the CER/WER/MER results on SEAME. **CS**: generated CS using in-domain SEAME text. **Mono**: baseline trained on monolingual data, **(Unigram, Bigram)**: generated CS using (unigram, bigram) units, **SE**: signal enhancement **SEAME-ASR**: topline model trained on SEAME.

Model	DevMan			DevSge		
	CER-MAN	WER-EN	MER	CER-MAN	WER-EN	MER
Mono	37.2	67.4	32.9	56.7	47.5	38.4
+ SEAME-LM	36.4	65.9	32.2	55.2	46.5	37.6
+ CS-Unigram	31.5	55.3	28.4	47.5	42.2	34.4
+ CS-Unigram-SE	29.7	53.7	27.2	44.0	40.9	33.0
+ CS-Bigram-SE	**27.2**	**47.9**	**25.4**	**39.7**	**38.1**	**31.4**
SEAME-ASR (topline)	15.1	28.8	16.5	21.7	28.7	23.5

- Integrating code-switched data augmentation improves WER, surpassing monolingual training or combining with a code-switched language model.

Table 2. Comparison of the CER/WER results on ESCWA. **CS**: data generated using synthetic CS text. **Mono**: baseline trained on monolingual data, **(Unigram, Bigram)**: generated CS using (unigram, bigram) units, **SE**: signal enhancement

Model	MGB-2		TED3		ESCWA	
	CER	WER	CER	WER	CER	WER
Mono	**6.1**	**12.9**	**4.4**	**8.5**	31.1	48.7
+ CS-LM	6.3	12.5	4.6	8.7	38.0	57.0
+ CS-Unigram	6.9	14.6	5.2	10.1	24.0	42.7
+ CS-Unigram-SE	7.0	14.7	5.4	10.4	23.1	42.0
+ CS-Bigram-SE	7.0	14.7	5.2	10.2	**22.5**	**40.8**

Code-Mixing Index

To quantify amount of code-switching in an utterance, we use the code-mixing index

$$CMI = \frac{\frac{1}{2} * (N - \max_i) + \frac{1}{2} P(x)}{N} \quad (3)$$

Table 3. Comparison of the average CMI. **Mono**: baseline trained on monolingual data, **SE**: Signal enhancement **Ref**: reference, **(Uni, Bi)**: generated CS using (unigram, bigram) units.

Dataset	Ref	Mono	CS-Uni	CS-Uni-SE	Bi-SE
ESCWA	15.6	8.7	10.6	11.6	10.5
SEAME	10.4	3.3	5.4	6.2	7.3

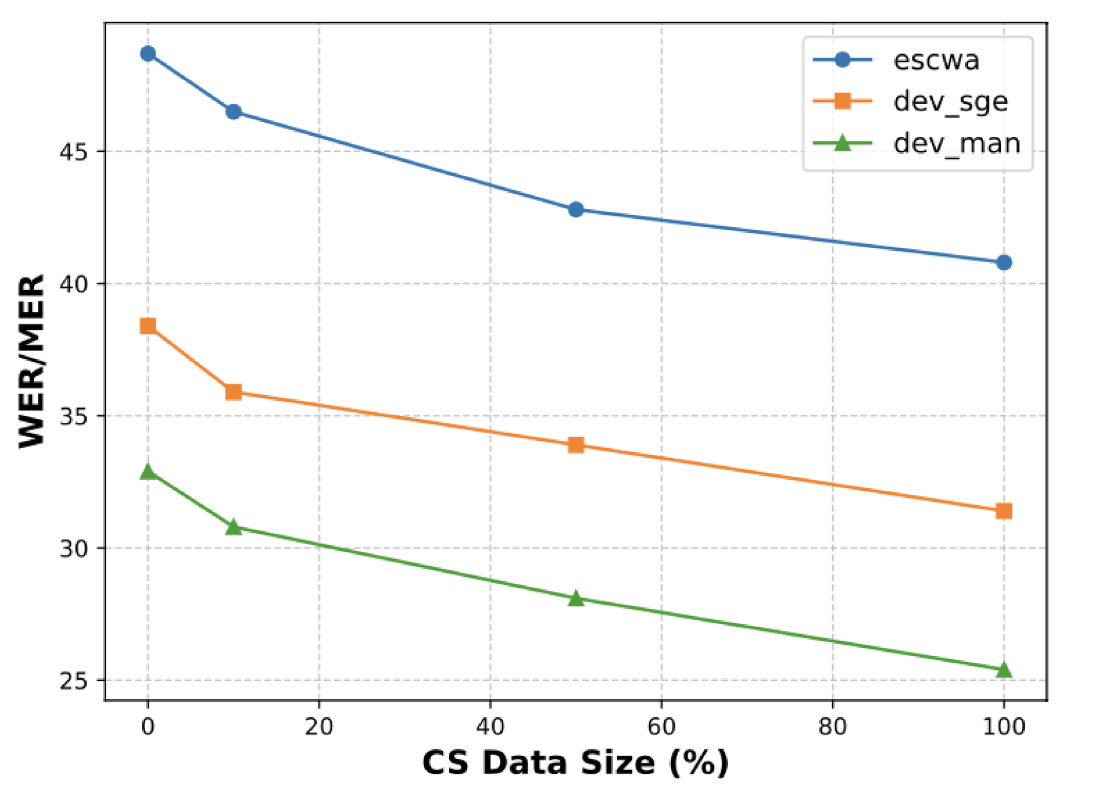

Figure 2: WER/MER at different percentages of generated CS data where 0%: represents Monolingual, 100%: represents Monolingual with all generated CS.

Discussion

- Employing CS data augmentation consistently elevates the CMI. This affirms our assumption that CS augmentation enhances the model's aptitude for code-switching
- As shown in Figure 2, as the percentage of generated CS data increases, the rate of improvement in WER/MER decreases. This suggests that with more data, further gains can be expected, albeit at a diminishing rate.
- We anticipate that further enhancements in audio quality will further bridge performance gaps.