Sayint’s Submission for MUCS 2021 Code-switching task

Anand M, Brahmendra K, Sreedhar P, Sagar R

Sayint.ai, Zen3 Infosolutions (A Tech Mahindra Company)

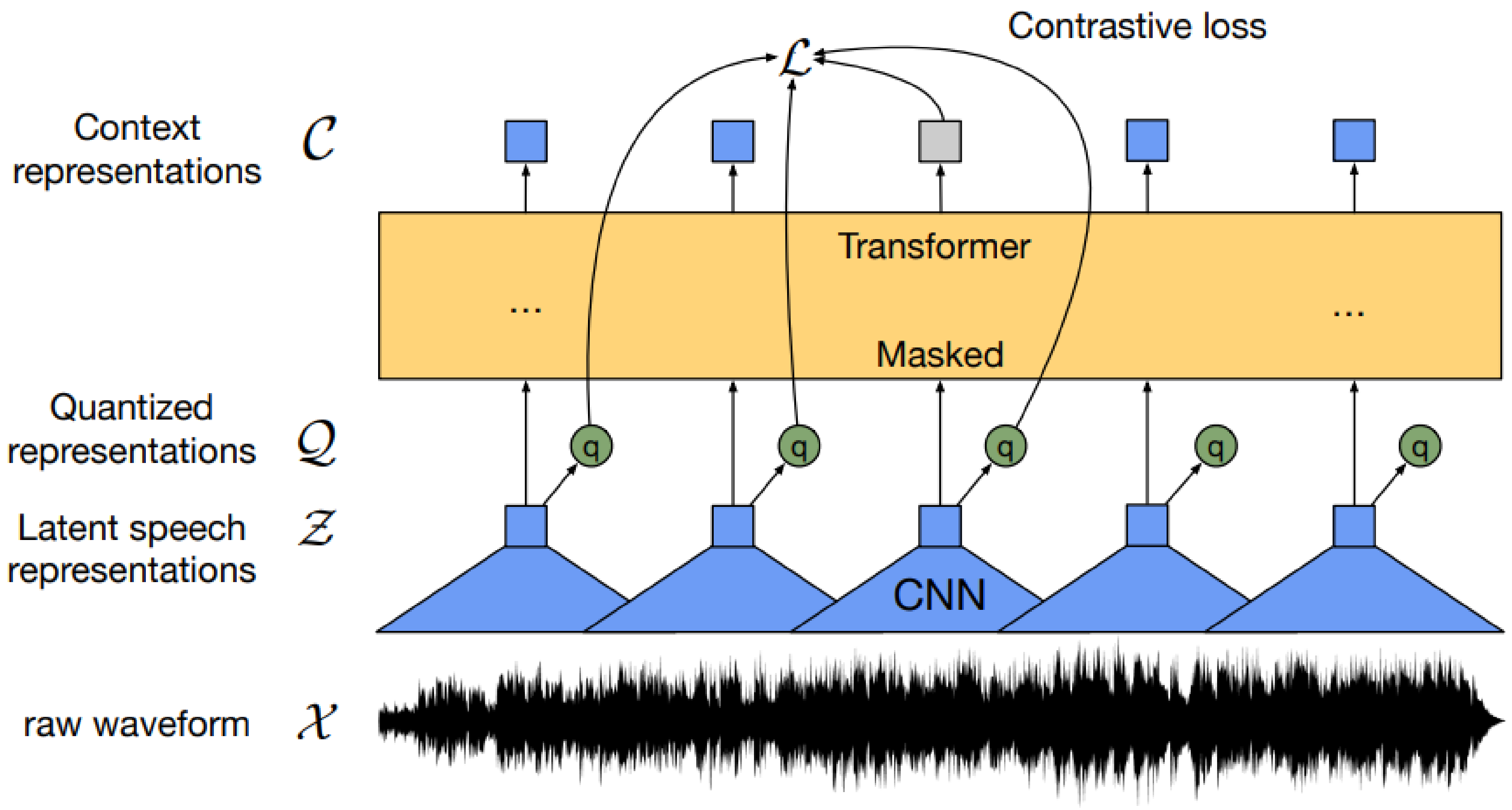
Code-Switching Task Description (Subtask 2)

- ASR for two code-switched pairs
  - Hindi-English
  - Bengali-English
- Derived from spoken tutorials
- Technical Topics
  - Programming
  - Word processing
  - GNU/Linux OS

Data Artifacts

- Symbols, punctuation and notations used/spoken inconsistently.
- English words present in both Latin as well as in the native script.
- Missing spaces between two or more words.
- Alignment issues between audio and text
- Incorrect start and end of the audio segment, resulting in words being cut off or transcription missing.
- The scripts were meant for video narrators, hence showed discrepancy between audio and text.

Wav2Vec 2.0



Jointly learning contextualized speech representations and inventory of discretized speech units

Source: Alexei Baevski, Henry Zhou, et.al., "*wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*", https://arxiv.org/abs/2006.11477

Unsupervised Cross-Lingual Representation Learning for Speech Recognition

When training on L languages, multilingual batches are formed by sample speech samples from a multinomial distribution, $(p_l)_{l=1,\ldots,L}$

where $p_l \sim \left(\frac{n_l}{N}\right)^{\alpha}$,

n_l being the number of pretraining hours of language l

N being the total number of hours, and

A being the upsampling factor.

Fine-tuning XLSR-53

- XLSR-53 is a large model trained (pretrained) on 53 languages and 56000hrs of Multilingual Librispeech, Common Voice, and BABEL
- We finetune XLSR-53 on the two code-switched pairs and train with a CTC loss.
- The training data was augmented (9 times) using WavAugment with time dropout(max 100ms), pitch and speed perturbation(+/- 10%).
- Post the initial 10000/4000 (BE/HE) updates, every 10 epochs , the augmentation process was rerun and the previous data overwritten.

Language Modelling

- N-gram LM using KenLM
- 5-gram LM
  - Base LM: Built from training set transcripts alone
  - Augmented LM: Built from training and test set transcripts. English words present in one language (Hindi/Bengali) but not in the other (Bengali/Hindi) was added to the the corresponding transcripts. Some text from the English subtitles from spoken tutorials were also added.

Test Set Results

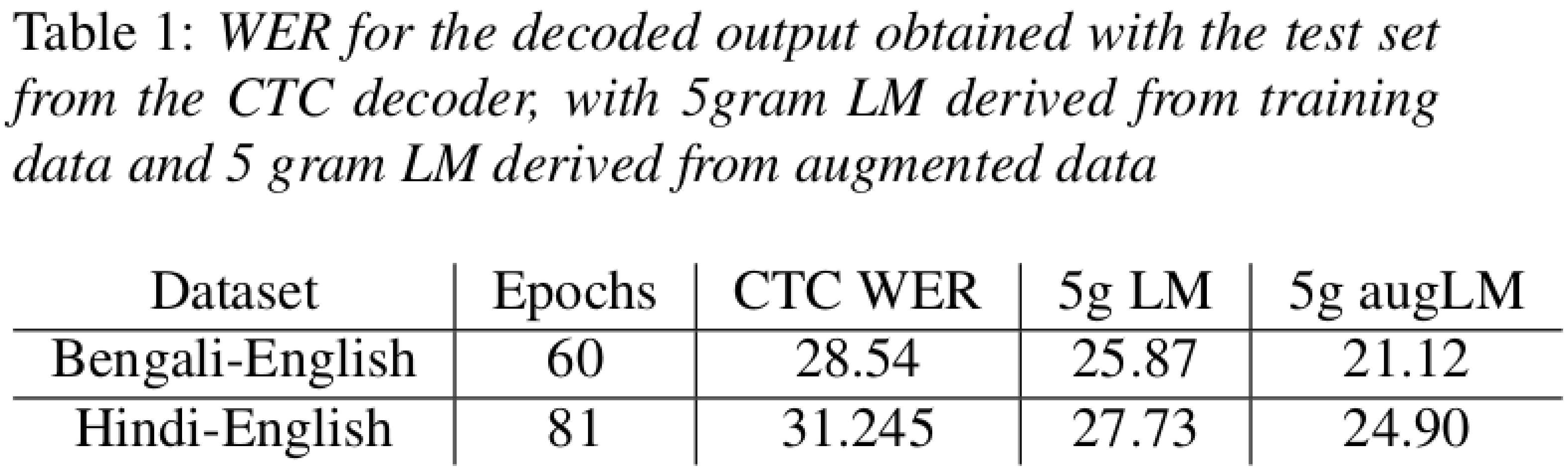


Table 1: *WER for the decoded output obtained with the test set from the CTC decoder, with 5gram LM derived from training data and 5 gram LM derived from augmented data*

Dataset	Epochs	CTC WER	5g LM	5g augLM
Bengali-English	60	28.54	25.87	21.12
Hindi-English	81	31.245	27.73	24.90

Blind Test Set Results

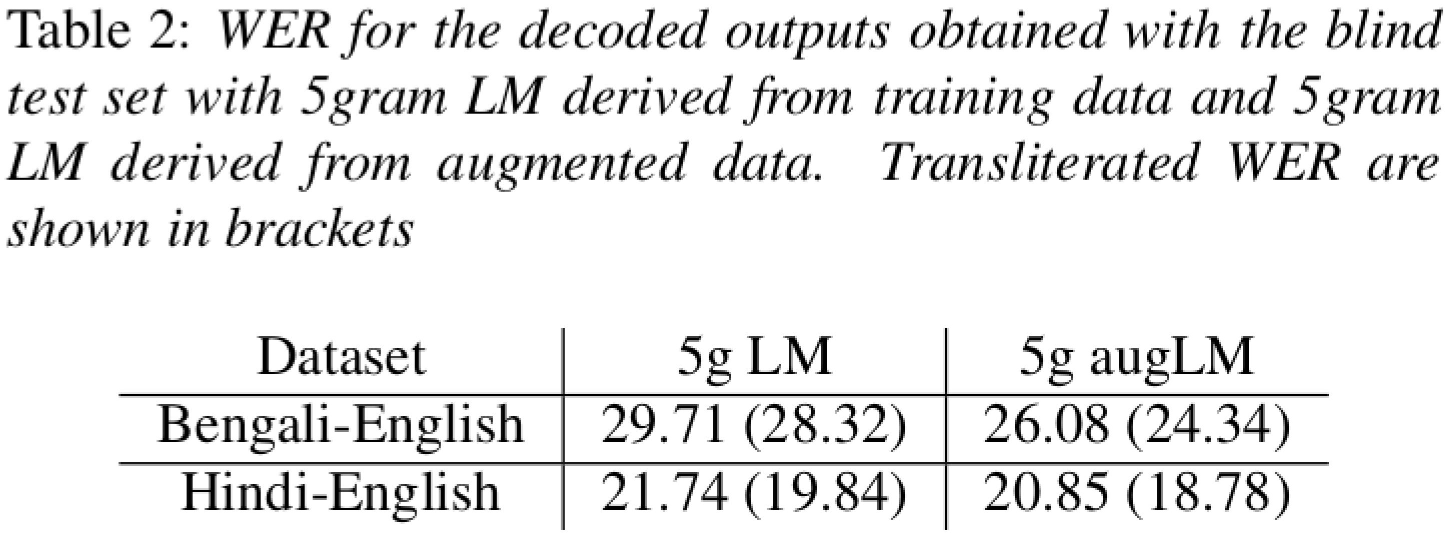


Table 2: *WER for the decoded outputs obtained with the blind test set with 5gram LM derived from training data and 5gram LM derived from augmented data. Transliterated WER are shown in brackets*

Dataset	5g LM	5g augLM
Bengali-English	29.71 (28.32)	26.08 (24.34)
Hindi-English	21.74 (19.84)	20.85 (18.78)

Discussions

- Errors primarily due to punctuations and short segments
- Missing words in ASR output at segment boundaries
  - Due to alignment issues in training data
  - Resegmentation using force aligner could have helped
- Translation of english transcripts could have improved the LM and reduced the WER further by minimizing the OOVs

About Sayint.ai

- Sayint is Tech Mahindra's conversational intelligence platform which used Speech and NLP to uncover meaninful insights from customer conversations. These insights can be used to develop, automate and/or improve key business functions/decisions and create better experiences for their customers.
- Areas we work on include
  - Speech Recognition, Speech Synthesis and Speech Enhancement
  - NLU – Intent/Entity extraction,
  - Multilingual search and information extraction
  - Text-based predictive analytics