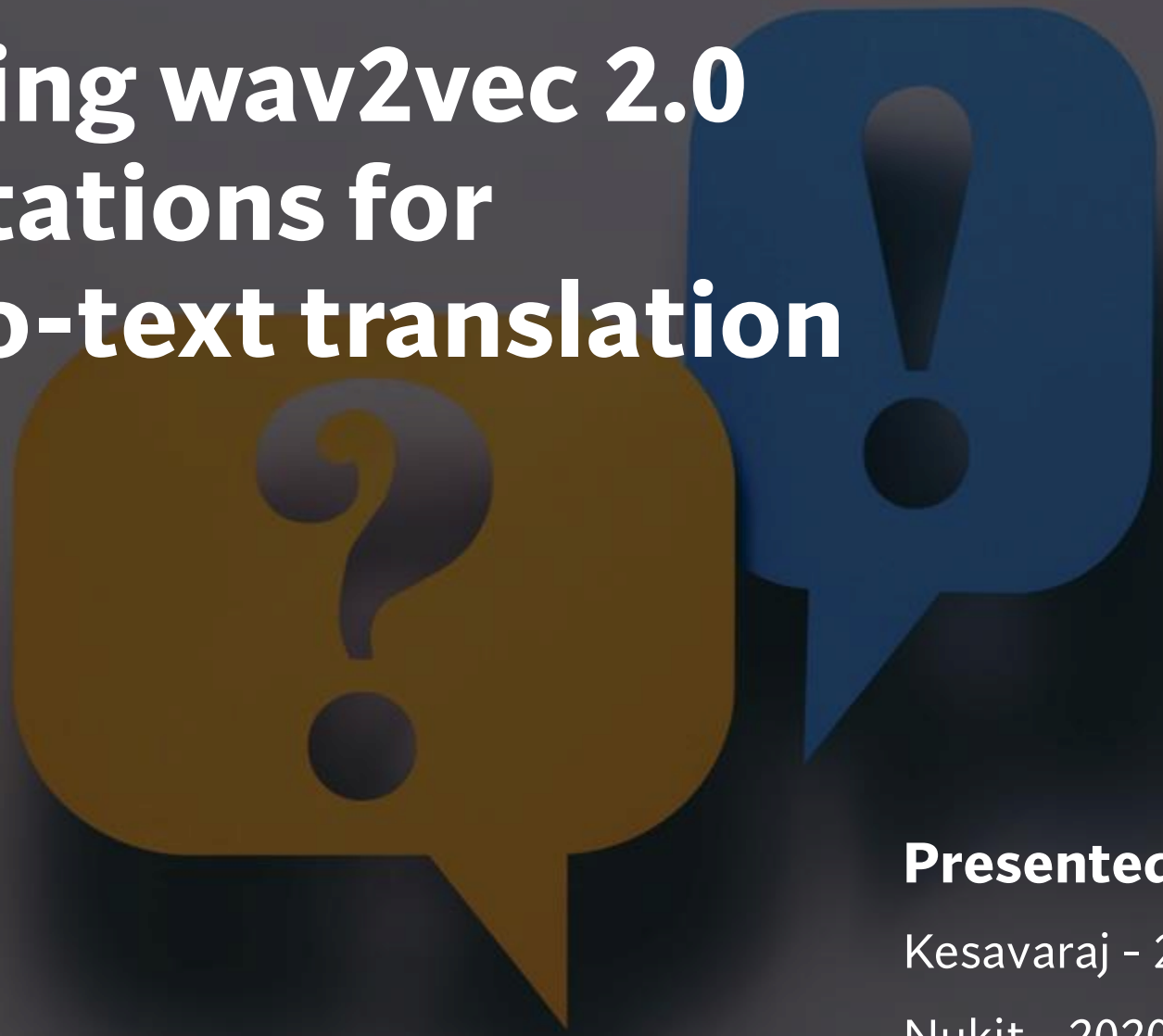
Considering wav2vec 2.0 representations for speech-to-text translation



Presented By

Kesavaraj - 2022701008

Nukit - 2020114012

Problem Definition

Description: English to Hindi speech to text translation using Wav2vec2.0

Motivation: To test the hypothesis, how wav2vec is improving the cascading speech to text translation pipeline.

Objectives:

- Building ASR using GMM-HMM
- Extracting learned speech representations using wav2vec
- Downstreaming to ASR task
- Trying different data for MT (domain and non-domain)

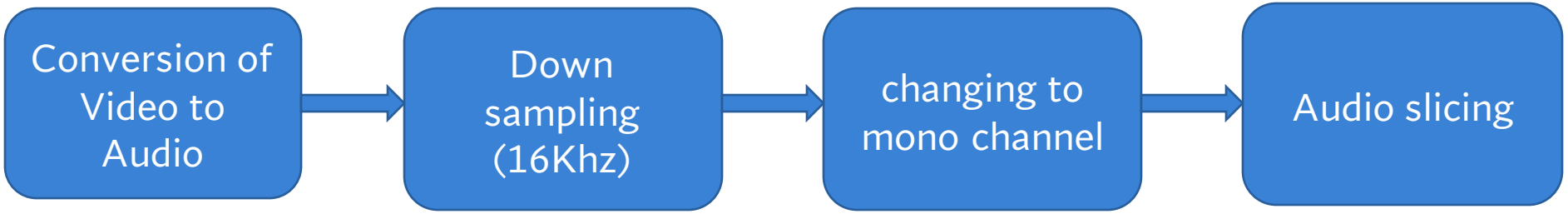
Work Done

Data preprocessing:

Data: Swayam Video lectures

- Total Duration:12.36 Hours
- Train: 9.88 Hours
- Test: 2.48 Hours

Transcript: We have seen how to implement data structures such as, stacks, queues and heaps using the



Kaldi Training

Data Preparation: wav.scp, spk2utt, text, lexicon, non_silence_phones

Features: MFCC --> CMVN (to reduce the effect of speaker variability)

Acoustic Model : GMM-HMM

- **Monophone training**: No contextual information
- **Triphone Training** : phoneme variant in the context of two other (left and right) phonemes

Wav2Vec2.0

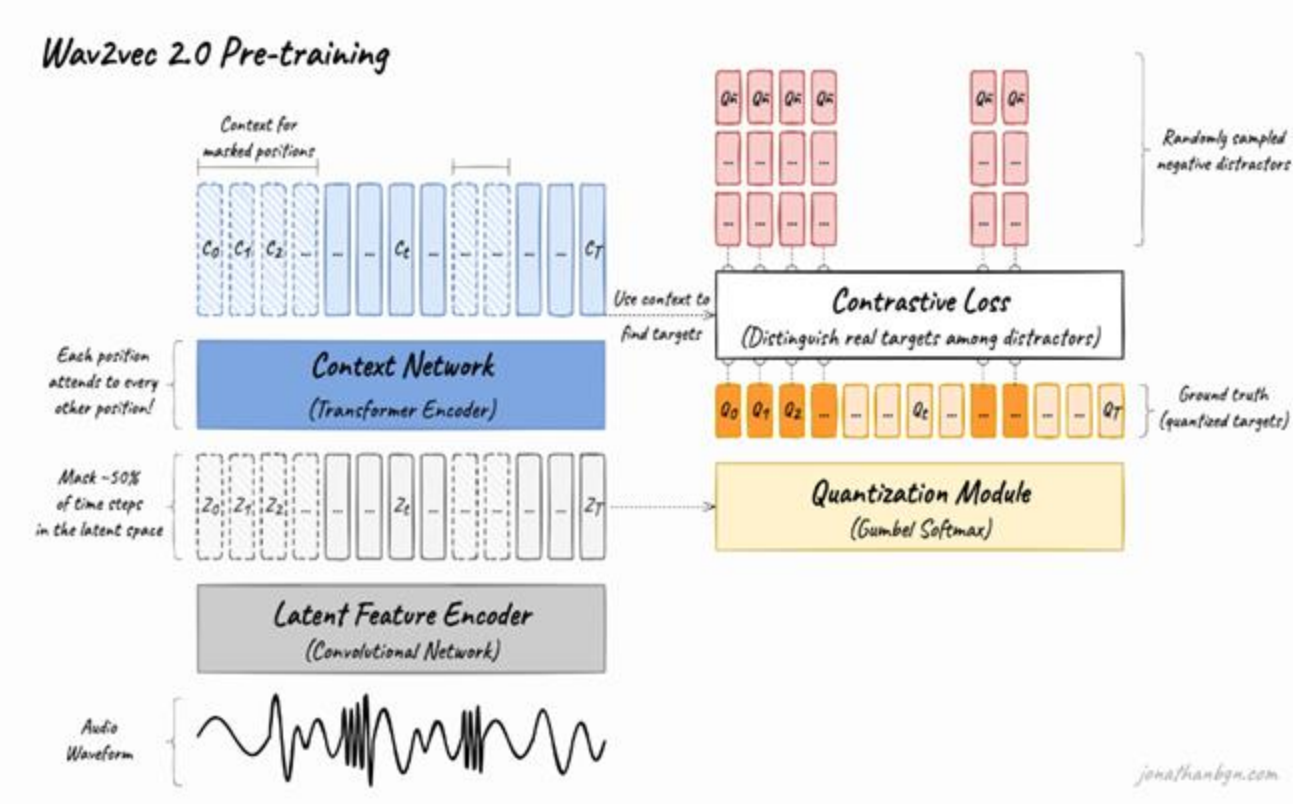
(Transformer encoder for speech)

Feature encoder: To reduce the dimensionality of the audio data, converting the raw waveform into a sequence of feature vectors.

Context network: The Transformer encoder

Quantization module: Automatically learn discrete speech units

Contrastive loss: Pre-training objective



Source

Mid Results

Audio: 

Actual transcript:

and remove the element at the **head** of the queue using the function remove q. And for

Kaldi output: (WER : 29.79 %)

and remove the element at the **end** of the queue using the function remove queue and for

Wav2Vec2.0 (pre trained) output: (WER : 87.66%)

AND REMADE THE ELEMENT OF THE HAPER TETOU ESIN EFAUCTION REMOVED TO I'M F

ASR comparison

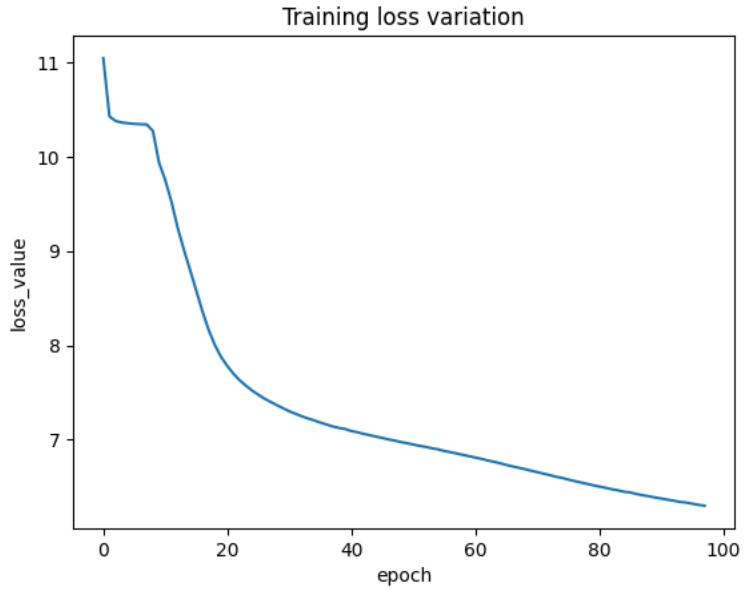
Audio: 

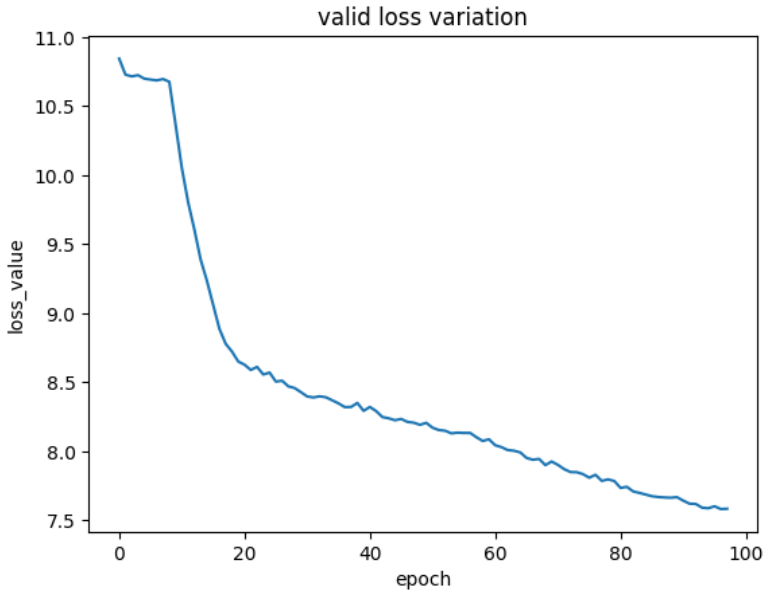
Transcript: you would have for instance a function like say push define and it will have two parameters

ASR	Training Data Duration		WER(%)	Examples
Kaldi	10 hours		29.79	you would have for instance a function like say push defined and it will have to parameters
Wav2Vec 2.0	Pretrained		87.66	youwould have for instance o fashin let se push be fine and we have pua paron ebos
	Finetuned	15 mins	62.70 %	yo wold have for instance a function like say push define and it wil haveto parameters
		1 hour	50.13 %	you would have for instance a function in length say push difan and it will have popater micus
		2 hours	39.8 %	you would have for instance a function like say push define and it will have w parameters

Training of MT model

- Sub-word tokenization to handle out of vocabulary
- Seq2seq with attention based





Post processing analysis

ASR	Kaldi (10 hours training)				Wav2Vec2 (2 hours training)			
ASR output	Actual		Post-processed		Actual		Post-processed	
WER %	29.79		29		39.8		32.4	
MT	Domain	Non-Domain	Domain	Non-Domain	Domain	Non-Domain	Domain	Non-Domain
BLEU	0.72	0.39	0.74	0.4	0.65	0.30	0.69	0.33

Post processing features:

1. Removing of repetitions.
2. Removing filled pauses (detecting uh,um although not very efficient since text)
3. Since wav2vec has no LM, using edit distance to edit minor changes when the transcription was almost correct.

Post Processing Results

Transcript: fm and fn, where constructed from 1 to m and 1 to n the largest value will also be the

Kaldi

Actual : fm and **f** l will constructed from one**(1)** to m and one to n the largest value will also be the

Post Processed: fm and **fn** l will constructed from one to m and one to n the largest value will also be the

Transcript: list it does not mean that the functions that are defined for lists are actually legal for

Wav2Vec2.0

Actual : list it does not mean that the **functiones** that are defined for list are actually **lvegal** f

Post Processed: list it does not mean that the **functions** that are defined for list are actually **legal** f

Results

Transcript: you would have for instance a function like say push define and it will have two parameters

MT (Domain):

Actual : उदाहरण के लिए आपके पास एक **फ़ंक्शन** होगा जैसे पुश परिभाषित किया गया है और इसमें पैरामीटर होंगे

Prediction : आपके पास उदाहरण के लिए एक **फ़ंक्शन** होगा जैसे कि पुश परिभाषित किया गया है और यह पैरामीटर होंगे

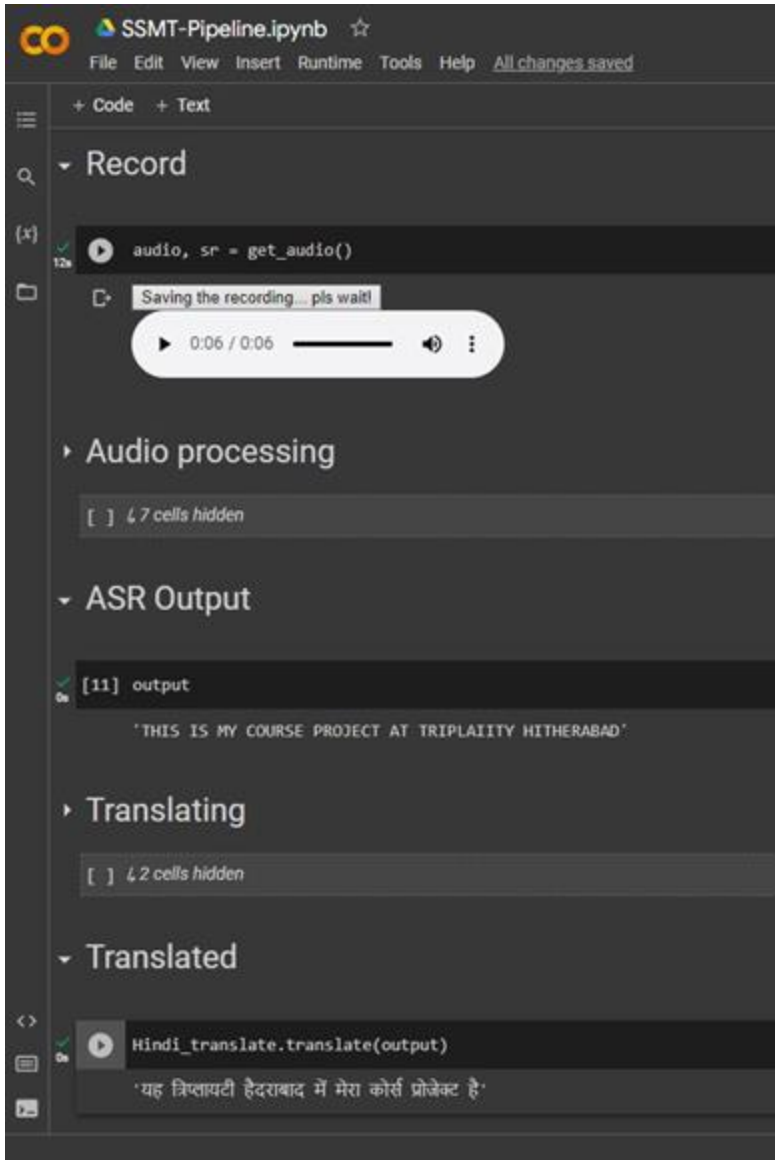
Bleu score: 0.8947368421052632

MT (Non-Domain):

Actual : उदाहरण के लिए आपके पास एक **फ़ंक्शन** होगा जैसे पुश परिभाषित किया गया है और इसमें पैरामीटर होंगे

Prediction : आपने देखा है कि इस **कार्यक्रम** के लिए **एक समारोह** का **आयोजन किया है और यह आयोजन किया ह**ै

Bleu score: 0.3157894736842105



Read mode vs Lecture mode

Thank you!