Finetuning wav2vec2 on Singapore's National Corpus

By KLASSmates:
Beatrice
Johnny
Kulo
Si Hui

Table of Contents.

01 Problem Statement

02 Applications

03 What is Wav2vec2

04 Evaluation Methodology

05 Zero Shot Transcription

06 Fine Tune with National Corpus

07 Project Demo

08 Possible Future Enhancements

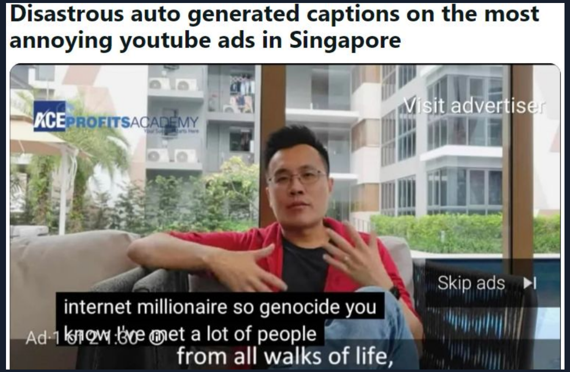
1. Problem Statement

Have you ever...

Adopted a pseudo-foreign accent, or tried to enunciate more clearly in order to make sure Siri understands you?

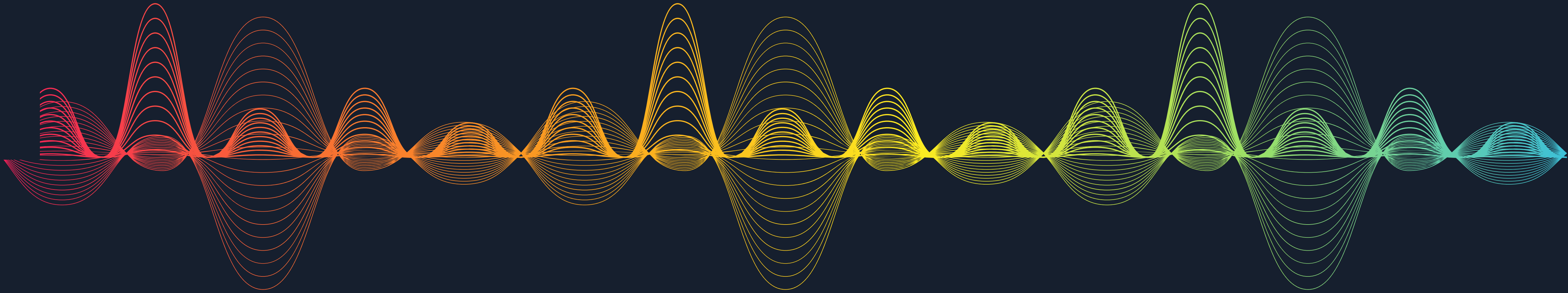
Or...

Watched a youtube video with auto-generated captions where many of the words are just wrong?

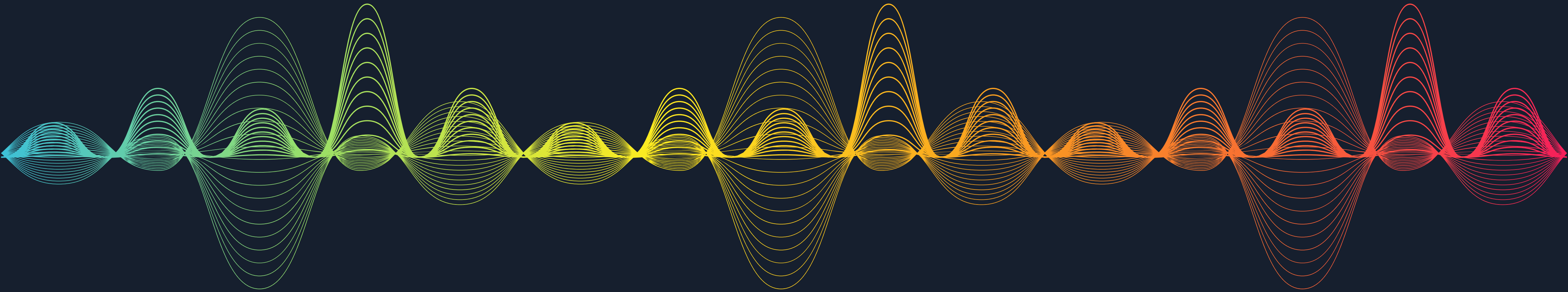


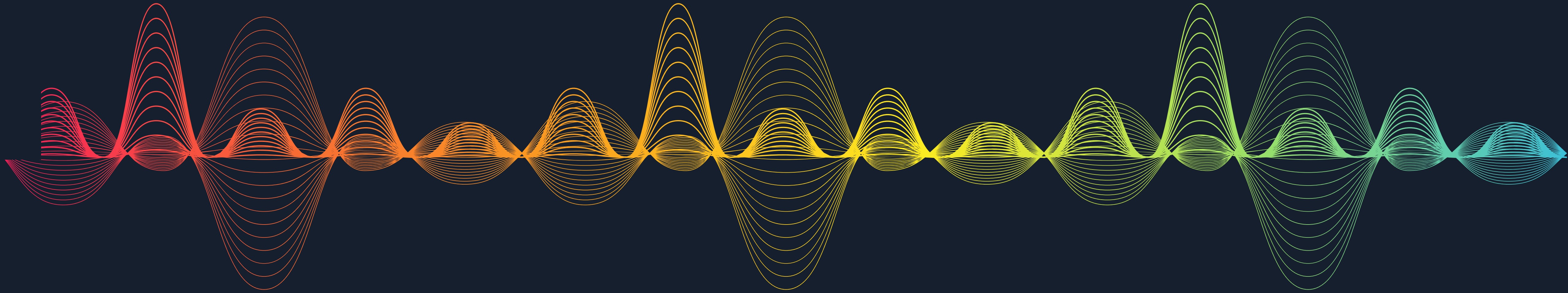
Disastrous auto generated captions on the most annoying youtube ads in Singapore

internet millionaire so genocide you
know I've met a lot of people
from all walks of life,



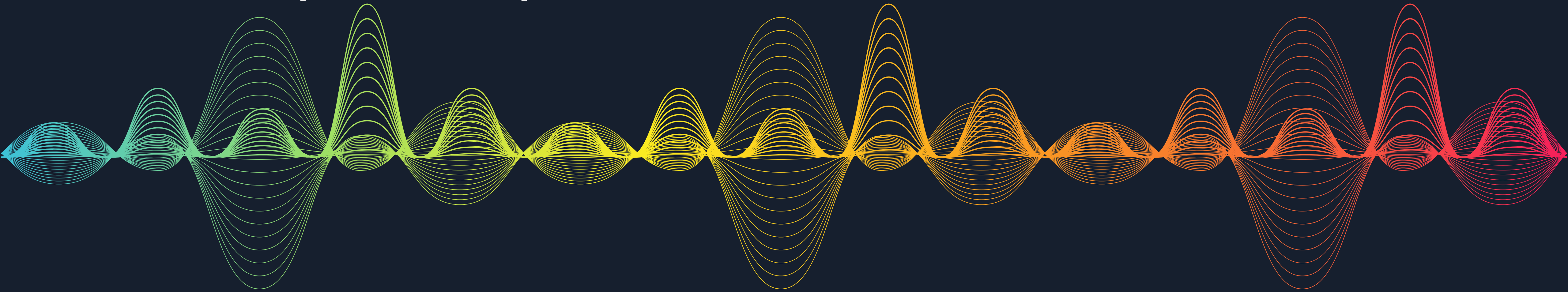
Clearly, Speech Recognition Systems are trained on foreign voices with significantly different intonation patterns from Singaporeans.





Training Audio Speech Recognition Systems with Singaporean accents:

- Improves accuracy in recognition of Singaporean intonation patterns
- Increase usability of ASR systems in Singapore
- Improves user experience



Hello! I’m...AI.PCK

I’m here to understand what talking you!

2. Applications

Enhanced Accessibility:

Assist individuals with hearing impairments or language barriers in public spaces.

Voice-activated Customer Service:

Streamline processes and provide information without human intervention.

Multilingual Communication:

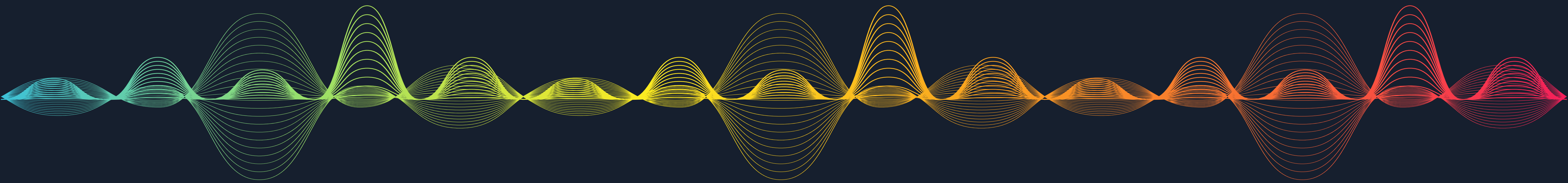
Bridge language gaps and facilitate communication in tourist destinations or international events.

Public Announcements and Alerts:

Instantly translate and convert critical information into text or voice formats.

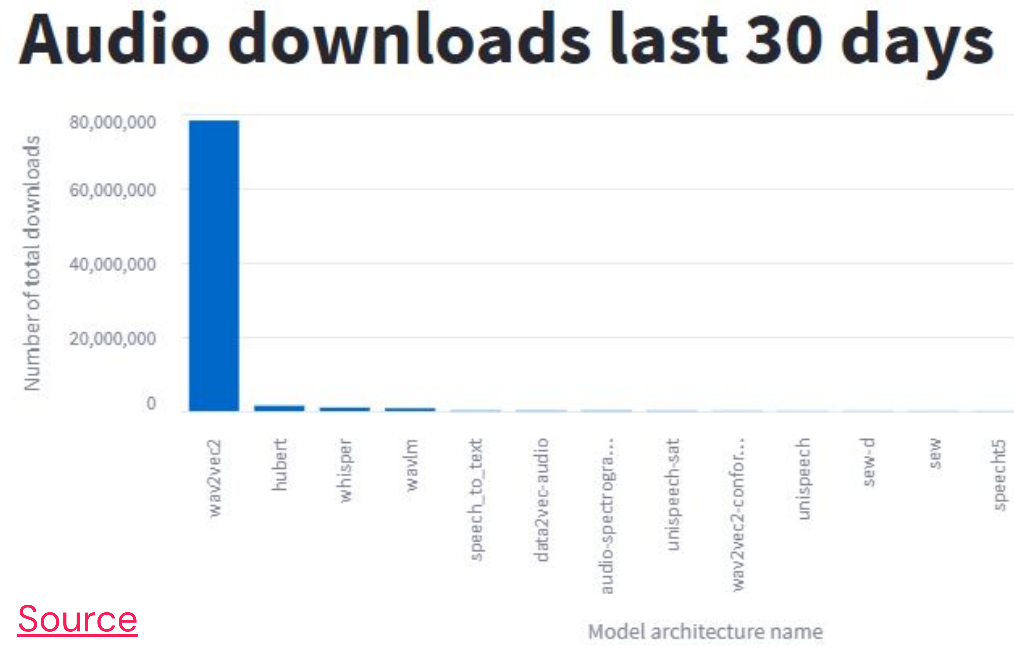
Hands-free Interaction:

Enable convenient and safe interaction in vehicles, public transport systems, and smart home appliances.

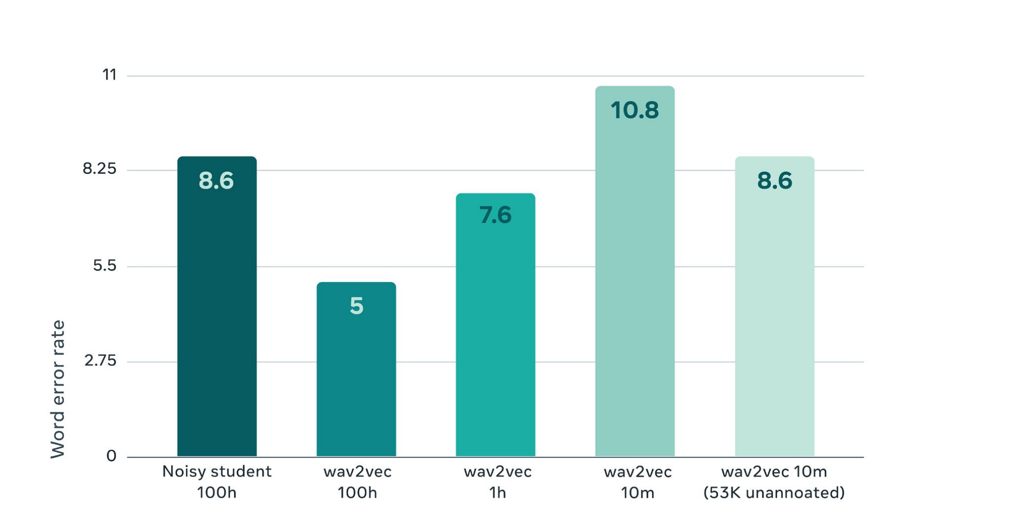


3. What is Wav2vec2?

- Pretrained model for Automatic Speech Recognition (ASR) by Meta in 2020
- Self-supervised learning with 960 hours of unannotated speech data from the LibriSpeech benchmark
- Mask feature vectors before passing them to a transformer network
- Using as little as 10 minutes of labeled data, Wav2Vec2 yields a word error rate (WER) of ~10%



Source



Source

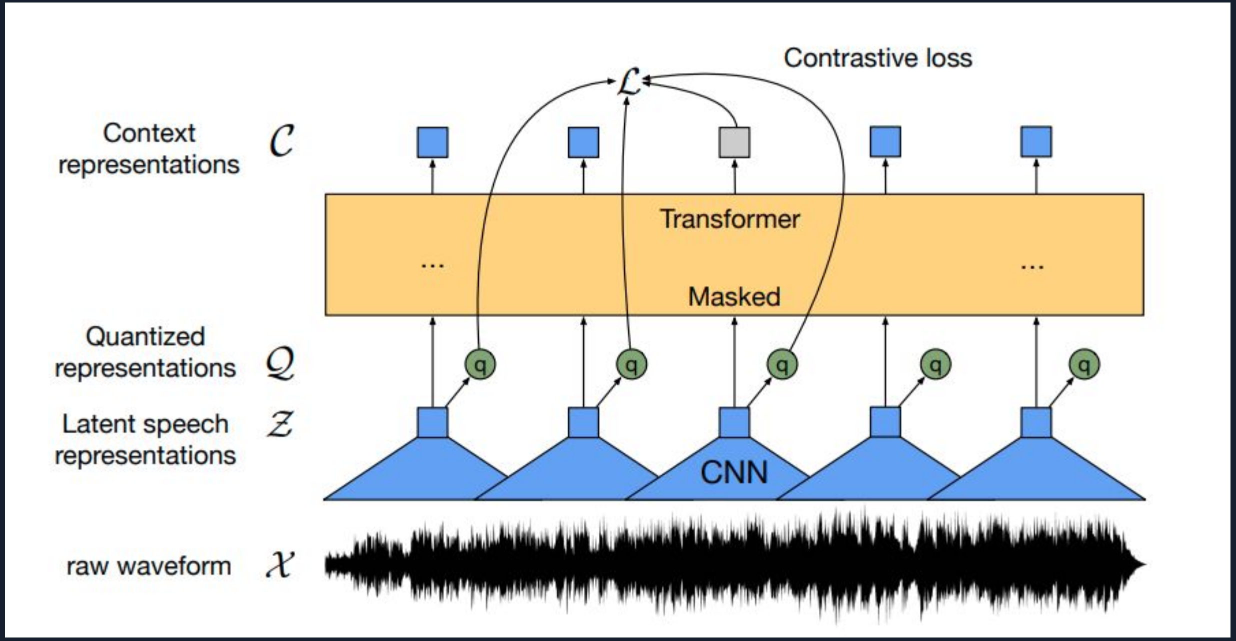
3. What is Wav2vec2? - Model Architecture

Audio data passed to a multi-layer 1-d Convolutional neural network to generate audio representations of 25ms each

Latent representation is fed into the transformer with half the representations masked

Objective: to predict these masked vectors from the output of the transformer

Quantized representation is for efficient contrastive loss calculation



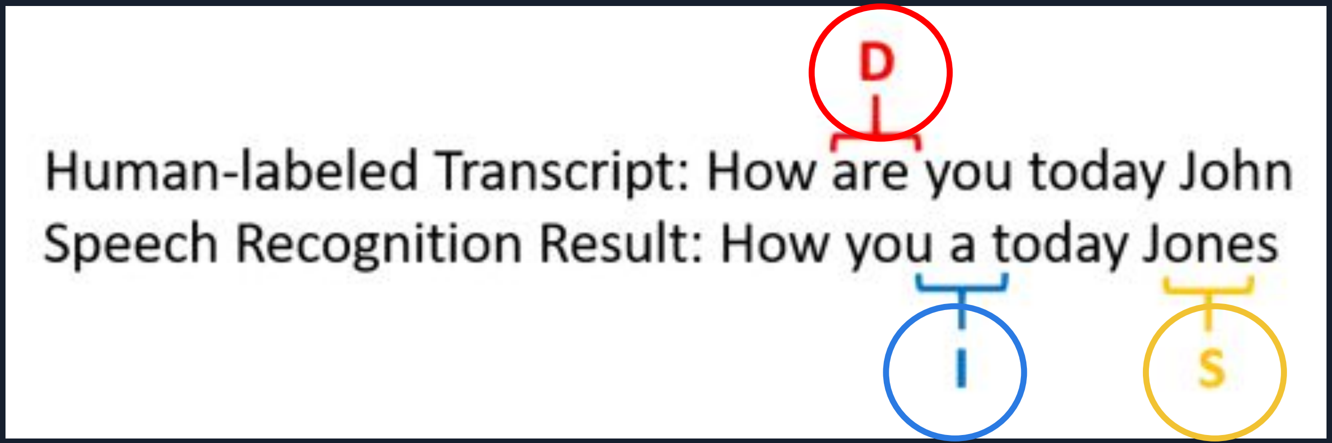
Source

4. Evaluation Methodology

Industry Standard: Word Error Rate (WER)

$$WER = \frac{I + D + S}{N} \times 100$$

I: number of words inserted by ASR engine
D: number of words deleted
S: number of words substituted
N: Total number of words in human-labeled transcript


Human-labeled Transcript: How are you today John
Speech Recognition Result: How you a today Jones

$$WER = \frac{1 + 1 + 1}{5} \times 100 = 60\%$$

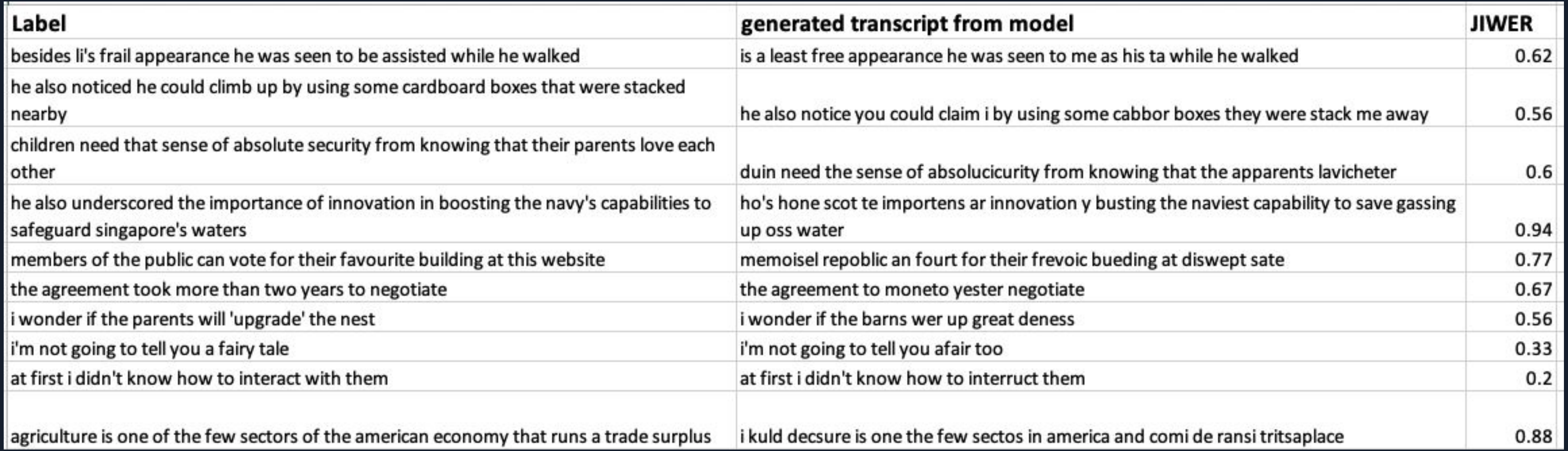
5. Source of Data

- National Speech Corpus by IMDA
- Total Size: 1.2TB
- > 2000 hours of recorded speech on Singaporean English
- Data used in project:
    - 1300 samples (each audio clip ~3s - 5s): 1-3 hrs of labelled data
    - 10 speakers
    - 70:30:30 Train, Test, Validation split

5. Performance of Zero Shot Transcription

Test samples used: 300

Selected Results:



Label	generated transcript from model	JIWER
besides li's frail appearance he was seen to be assisted while he walked	is a least free appearance he was seen to me as his ta while he walked	0.62
he also noticed he could climb up by using some cardboard boxes that were stacked nearby	he also notice you could claim i by using some cabbor boxes they were stack me away	0.56
children need that sense of absolute security from knowing that their parents love each other	duin need the sense of absolucicurity from knowing that the apparents lavicheter	0.6
he also underscored the importance of innovation in boosting the navy's capabilities to safeguard singapore's waters	ho's hone scot te importens ar innovation y busting the naviest capability to save gassing up oss water	0.94
members of the public can vote for their favourite building at this website	memoisel repoblic an fourt for their frevoic bueding at diswept sate	0.77
the agreement took more than two years to negotiate	the agreement to moneto yester negotiate	0.67
i wonder if the parents will 'upgrade' the nest	i wonder if the barns wer up great deness	0.56
i'm not going to tell you a fairy tale	i'm not going to tell you afair too	0.33
at first i didn't know how to interact with them	at first i didn't know how to interruct them	0.2
agriculture is one of the few sectors of the american economy that runs a trade surplus	i kuld decsure is one the few sectos in america and comi de ransi tritsaplace	0.88

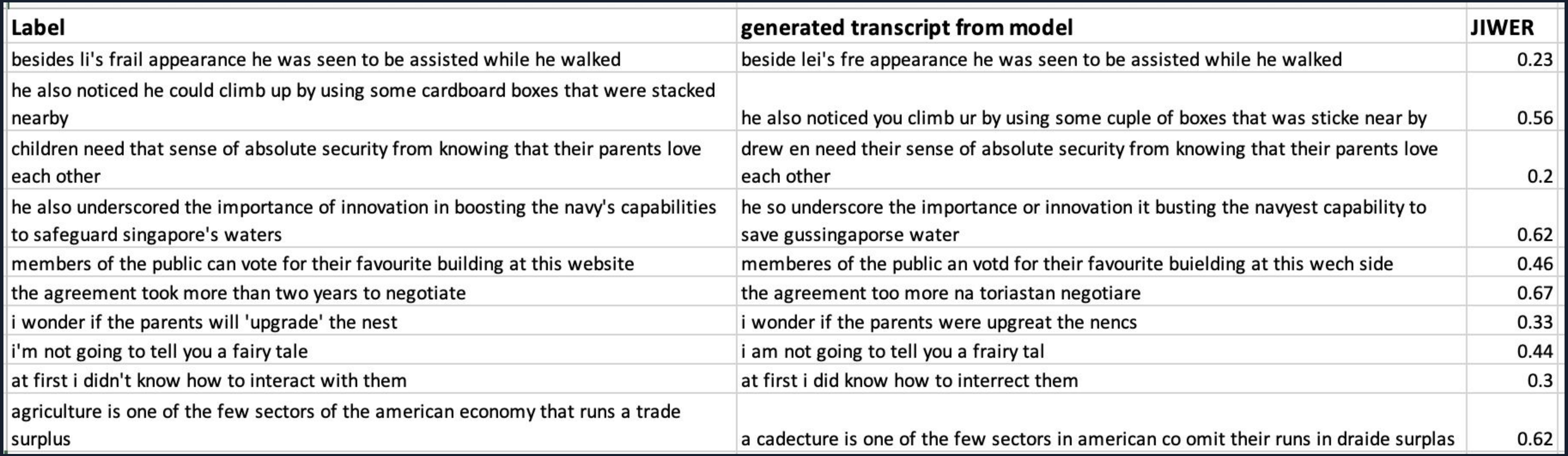
Average WER: 27%

6. Performance of Model with Fine Tuning

Train Samples used: 700, Validation Samples used: 300

Test Samples used: 300

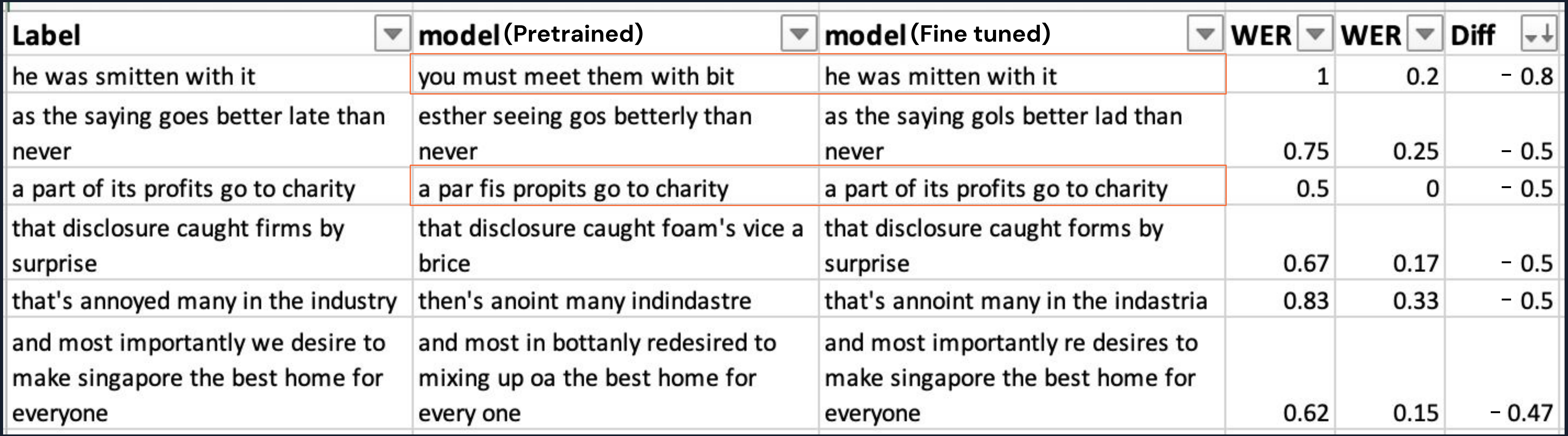
Selected Test Results:



Label	generated transcript from model	JIWER
besides li's frail appearance he was seen to be assisted while he walked	beside lei's fre appearance he was seen to be assisted while he walked	0.23
he also noticed he could climb up by using some cardboard boxes that were stacked nearby	he also noticed you climb ur by using some cuple of boxes that was sticke near by	0.56
children need that sense of absolute security from knowing that their parents love each other	drew en need their sense of absolute security from knowing that their parents love each other	0.2
he also underscored the importance of innovation in boosting the navy's capabilities to safeguard singapore's waters	he so underscore the importance or innovation it busting the navyest capability to save gussingaporse water	0.62
members of the public can vote for their favourite building at this website	memberes of the public an votd for their favourite buielding at this wech side	0.46
the agreement took more than two years to negotiate	the agreement too more na toriastan negotiare	0.67
i wonder if the parents will 'upgrade' the nest	i wonder if the parents were upgreat the nencs	0.33
i'm not going to tell you a fairy tale	i am not going to tell you a frairy tal	0.44
at first i didn't know how to interact with them	at first i did know how to interrect them	0.3
agriculture is one of the few sectors of the american economy that runs a trade surplus	a cadecture is one of the few sectors in american co omit their runs in draide surplas	0.62

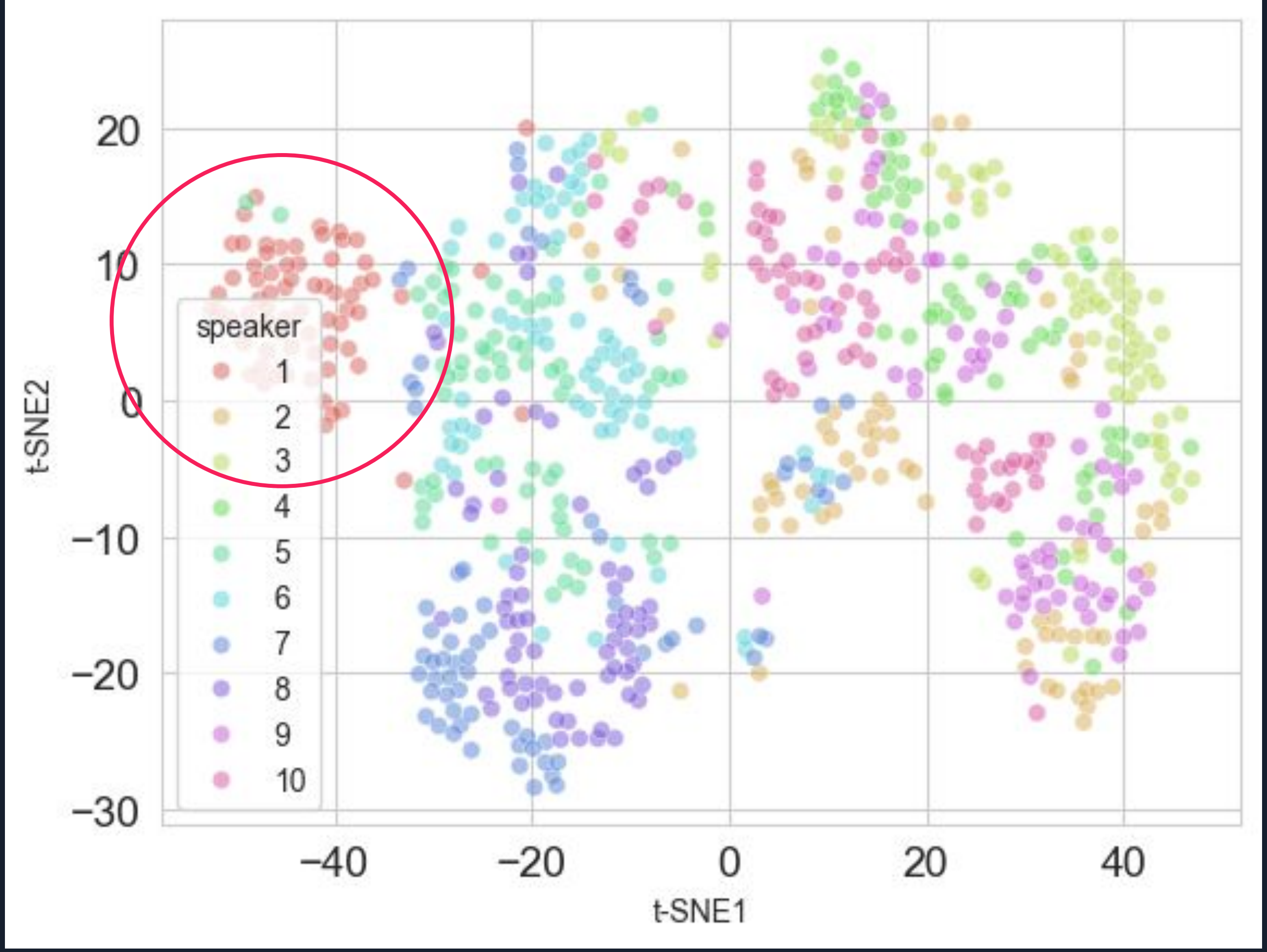
Average WER: 19%

Comparison



Label	model (Pretrained)	model (Fine tuned)	WER	WER	Diff
he was smitten with it	you must meet them with bit	he was mitten with it	1	0.2	- 0.8
as the saying goes better late than never	esther seeing gos betterly than never	as the saying gols better lad than never	0.75	0.25	- 0.5
a part of its profits go to charity	a par fis propits go to charity	a part of its profits go to charity	0.5	0	- 0.5
that disclosure caught firms by surprise	that disclosure caught foam's vice a brice	that disclosure caught forms by surprise	0.67	0.17	- 0.5
that's annoyed many in the industry	then's anoint many indindastre	that's annoint many in the indastria	0.83	0.33	- 0.5
and most importantly we desire to make singapore the best home for everyone	and most in bottanly redesired to mixing up oa the best home for every one	and most importantly re desires to make singapore the best home for everyone	0.62	0.15	- 0.47

Error Analysis

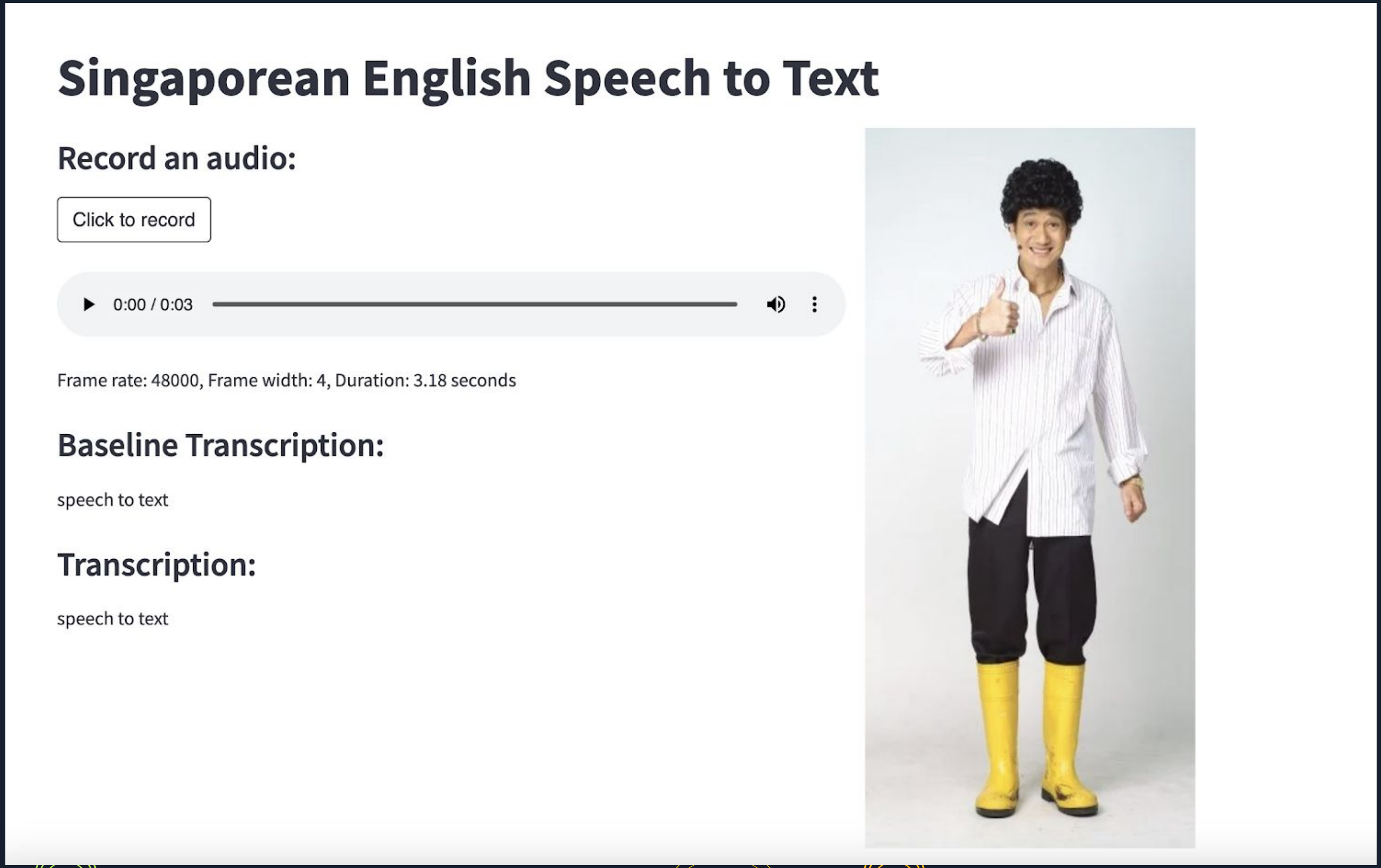


Speaker	Average WER (pretrained)	Average WER (fine tuned)
1	0.54	0.36
2	0.38	0.23
3	0.20	0.15
4	0.34	0.28
5	0.19	0.17
6	0.19	0.20
7	0.16	0.14
8	0.17	0.10
9	0.23	0.11
10	0.31	0.21

Project Demo

AI.PCK

7. Project Demo



Singaporean English Speech to Text

Record an audio:

Frame rate: 48000, Frame width: 4, Duration: 3.18 seconds

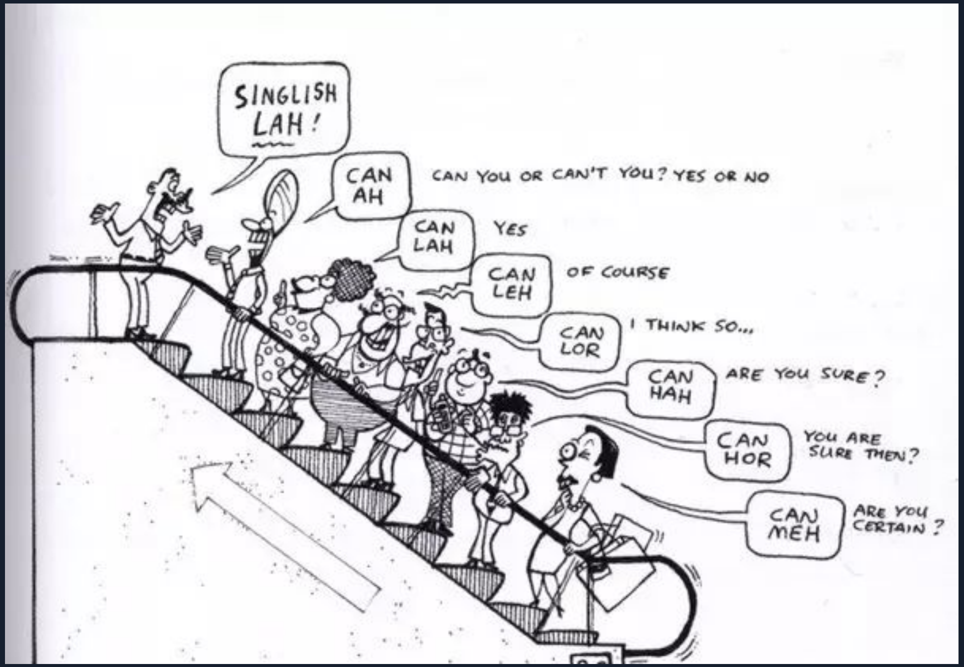
Baseline Transcription:

speech to text

Transcription:

speech to text

8. Possible Future Enhancements



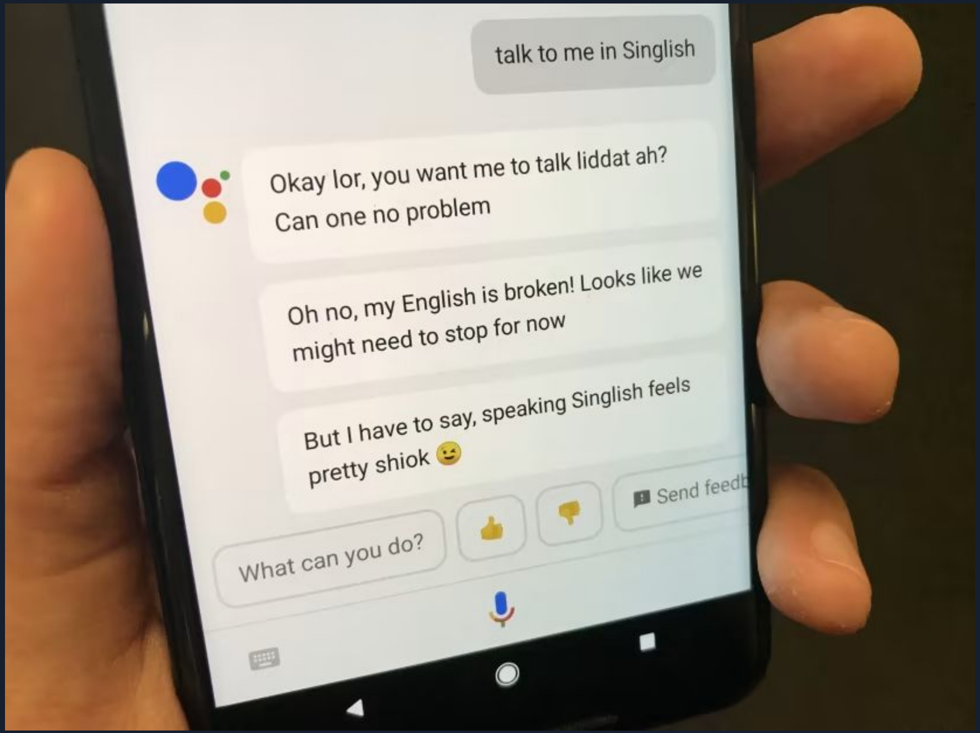
Importance of Singlish

- Singlish cuts across racial differences and thus functions as a marker of a distinct, multi-ethnic Singaporean identity
- Essential part of local culture and heritage

Problem with Singlish Transcription:

- Cannot capture the nuances in meanings

Google Assistant Example



talk to me in Singlish

Okay lor, you want me to talk liddat ah? Can one no problem

Oh no, my English is broken! Looks like we might need to stop for now

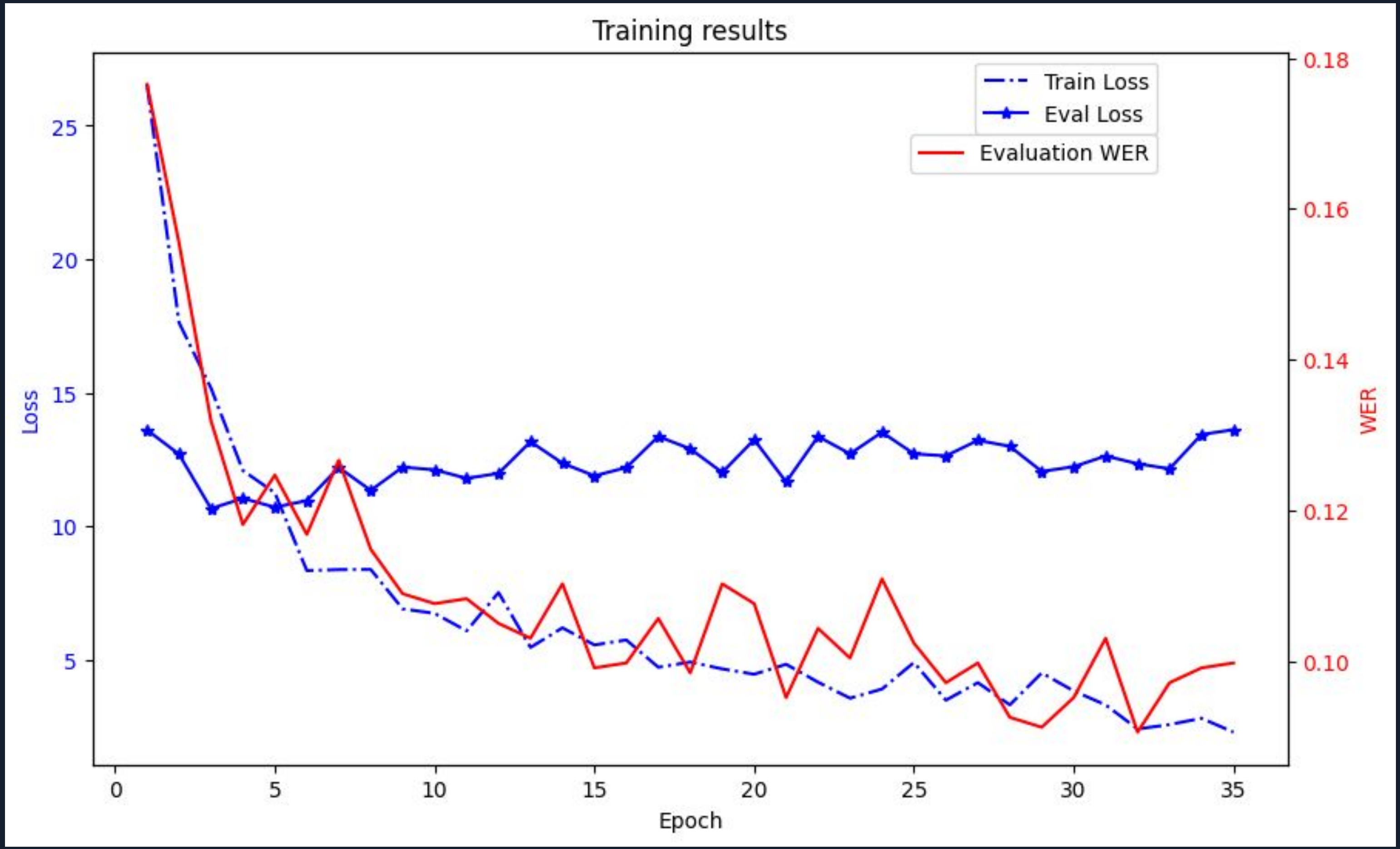
But I have to say, speaking Singlish feels pretty shiok 😉

Thank you!

Do you have any questions?

Appendix

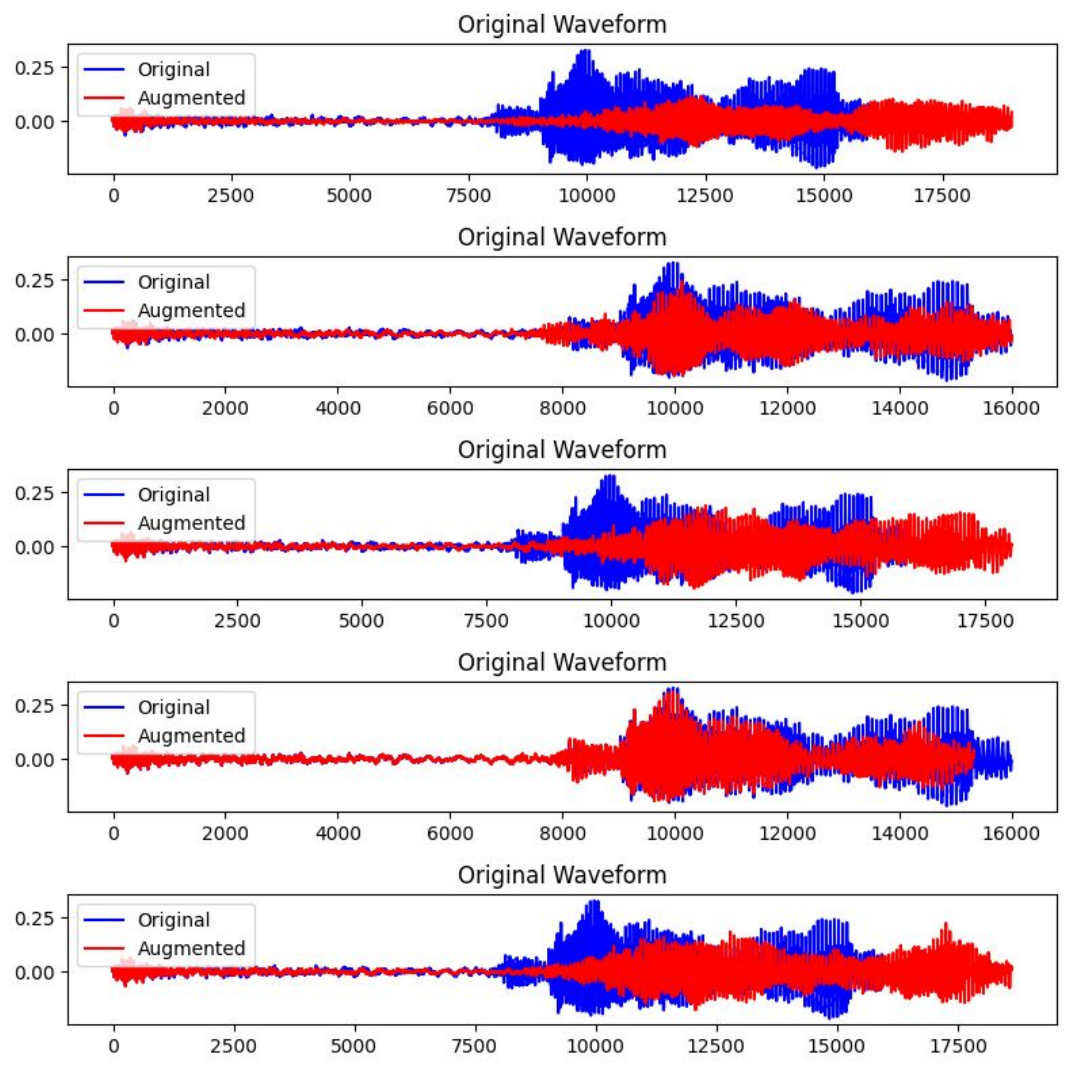
Training loss

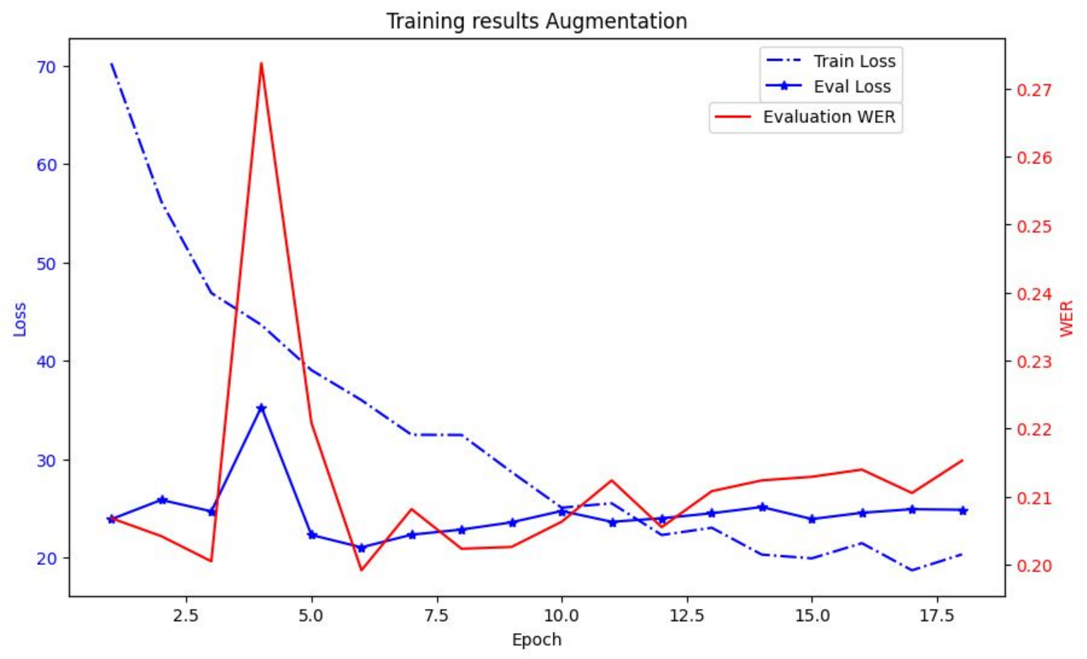


Data Set

Speaker	Gender	Pitch
1 (e.g. 000160006.WAV)	Female	Middle
2 (e.g. 000980001.WAV)	Male	Low
3 (e.g. 001760028.WAV)	Male	Low
4 (e.g. 001980017.WAV)	Male	Low
5 (e.g.002390014.WAV)	Female	Middle
6 (e.g. 002790045.WAV)	Female	High
7 (e.g. 005280036.WAV)	Female	Middle
8 (e.g. 008700063.WAV)	Female	High, nasal
9 (e.g. 010180060.WAV)	Male	Low/Middle
10 (e.g. 100010398.WAV)	Female	High

Audio augmentation





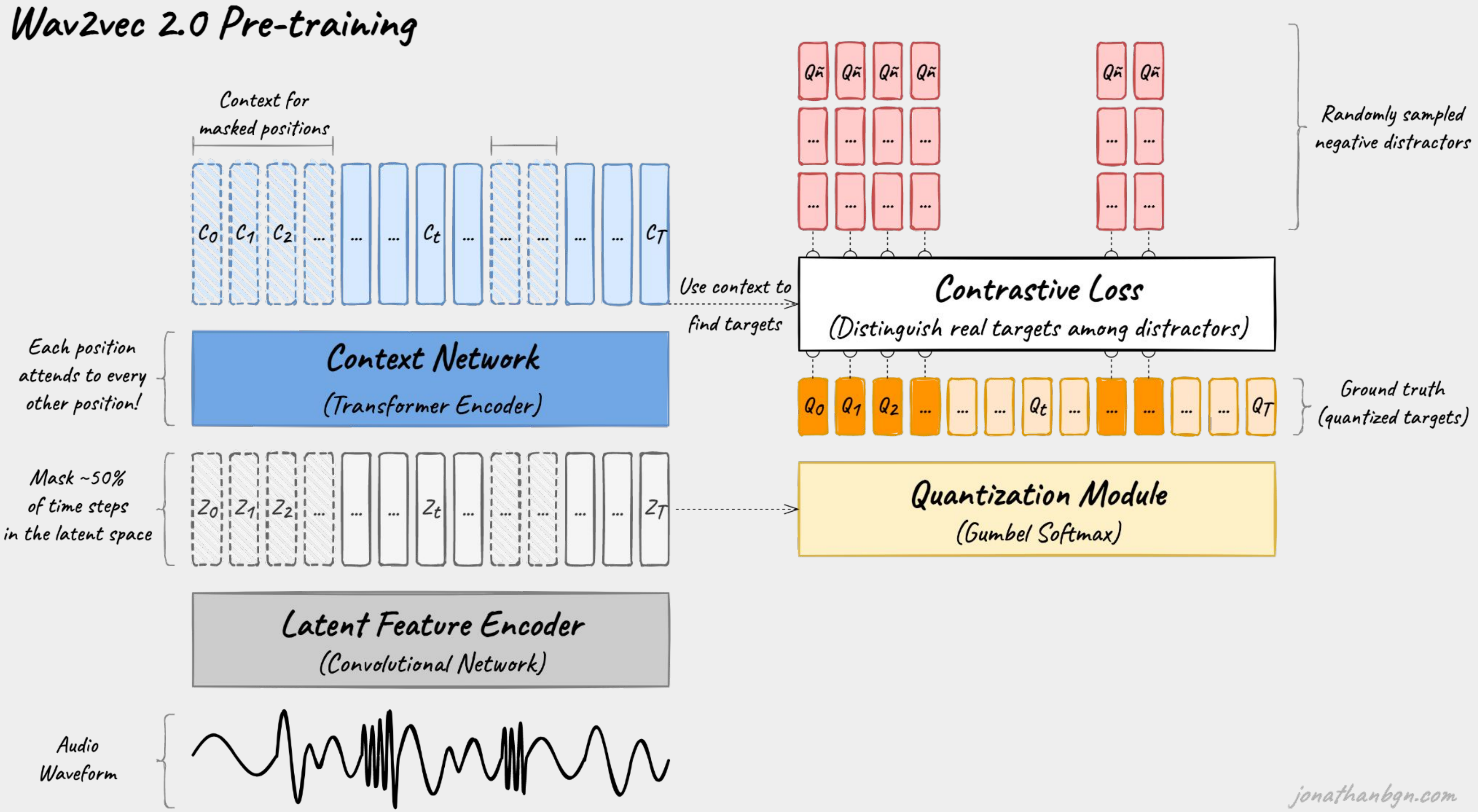
-Augmentations applied:

- time_strech
- pitch_shift
- noise

Appendix: What is Wav2vec2?: Feature Extractor

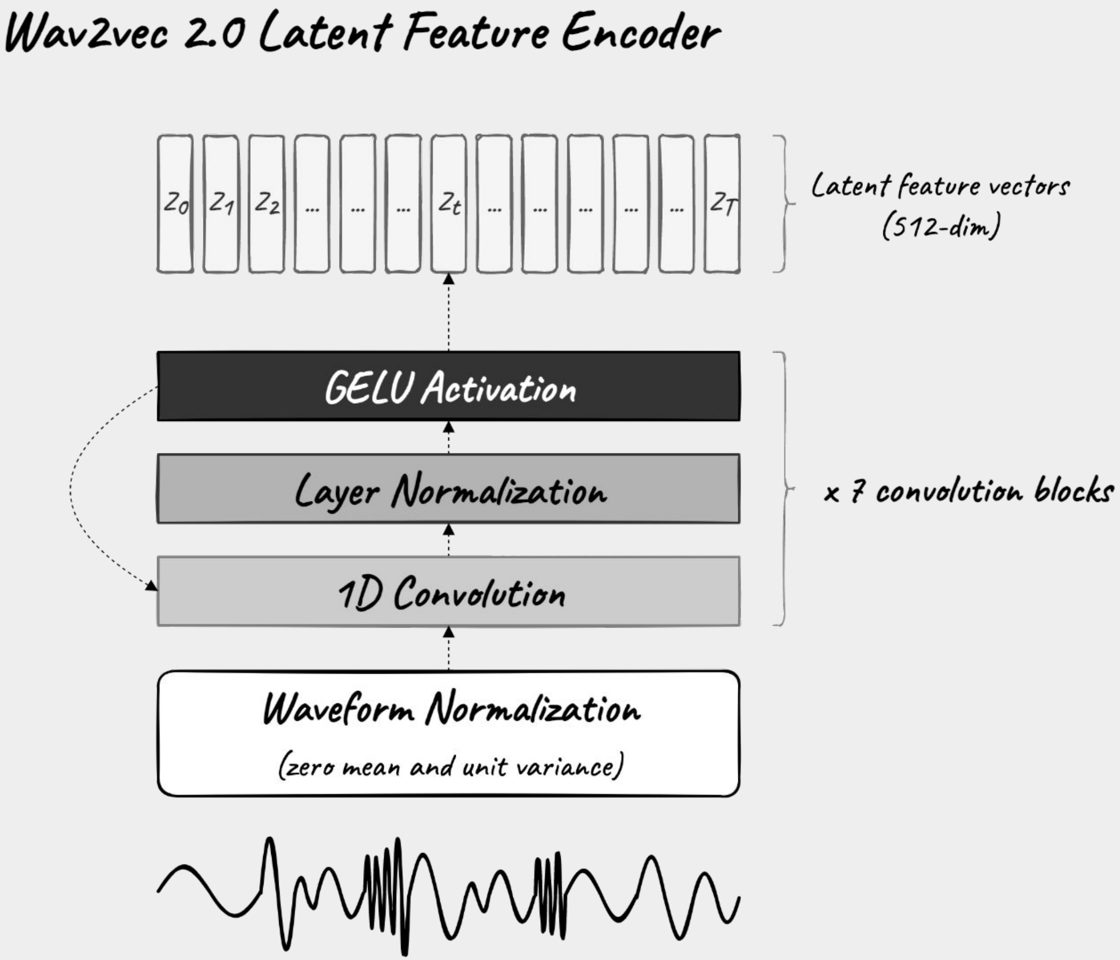
- sampling rate = how many data points of the speech signal are measured per second
- sampling rate of the data that was used to pretrain the model should match the sampling rate of the dataset used to fine-tune the model
  - up or downsample the speech signal to match the sampling rate
- Pretrained on 16kHz
- A stack of CNN layers that are used to extract acoustically meaningful - but contextually independent - features from the raw speech signal. This part of the model has already been sufficiently trained during pretraining and don’t require finetuning

Appendix: Word2Vec2 Architecture


Wav2vec 2.0 Pre-training
Context for masked positions
c_0 c_1 c_2 ... ... c_t ... ... c_T
Use context to find targets
Contrastive Loss
(Distinguish real targets among distractors)
$Q_{\tilde{n}}$
Randomly sampled negative distractors
Q_0 Q_1 Q_2 ... ... Q_t ... ... Q_T
Ground truth (quantized targets)
Each position attends to every other position!
Context Network
(Transformer Encoder)
Mask ~50% of time steps in the latent space
z_0 z_1 z_2 ... ... z_t ... ... z_T
Quantization Module
(Gumbel Softmax)
Latent Feature Encoder
(Convolutional Network)
Audio Waveform
jonathanbgn.com

Appendix: Word2Vec2 Architecture

Feature Encoder

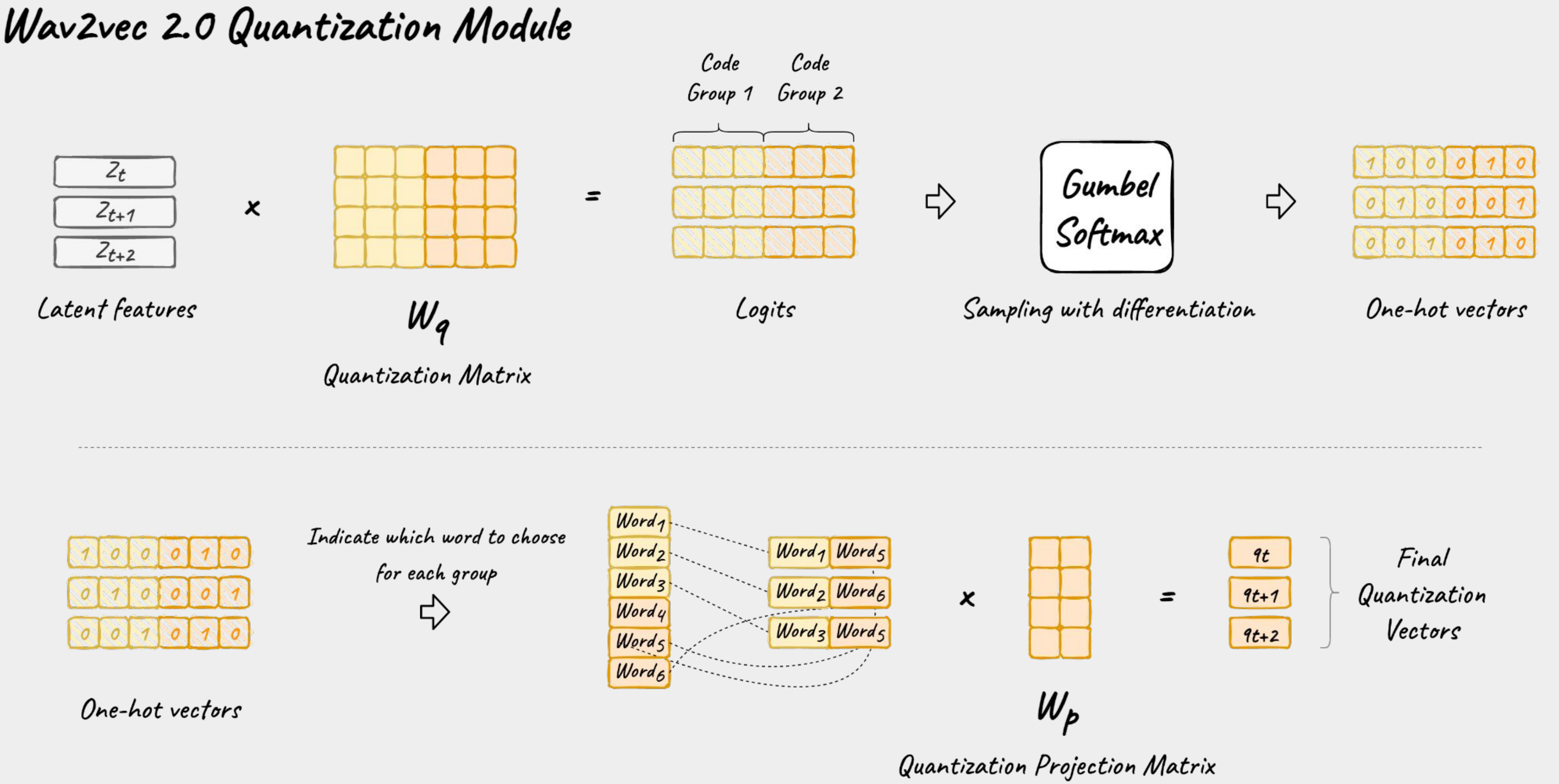

Wav2vec 2.0 Latent Feature Encoder
z_0 z_1 z_2 ... ... z_t ... ... z_T
Latent feature vectors
(512-dim)
GELU Activation
Layer Normalization
1D Convolution
x 7 convolution blocks
Waveform Normalization
(zero mean and unit variance)

Block	Channels	Kernel width	Stride
7	512	2	2
6			
5		3	
4			
3			
2			
1		10	5

jonathanbgn.com

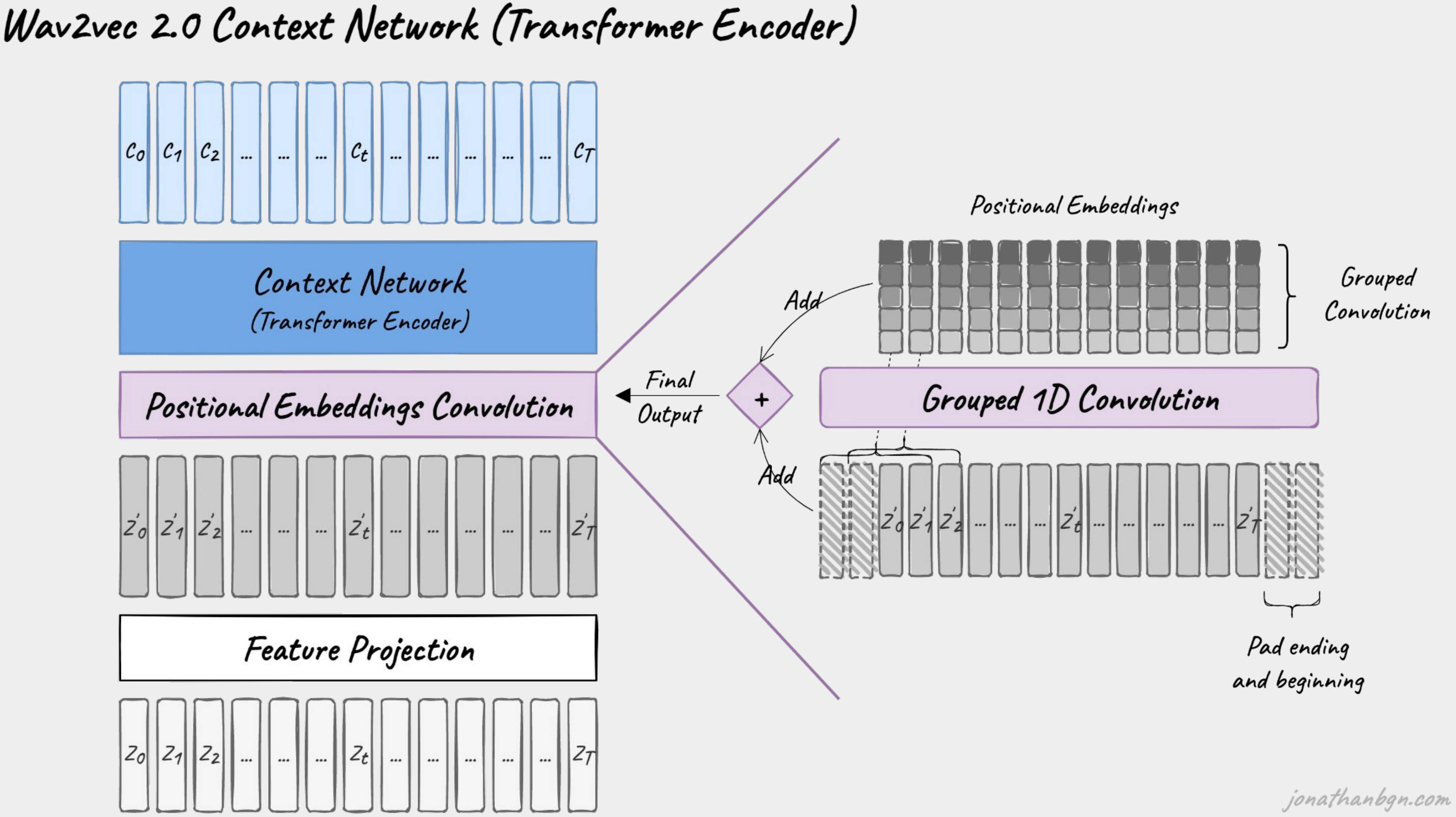
Appendix: Word2Vec2 Architecture

Quantization module



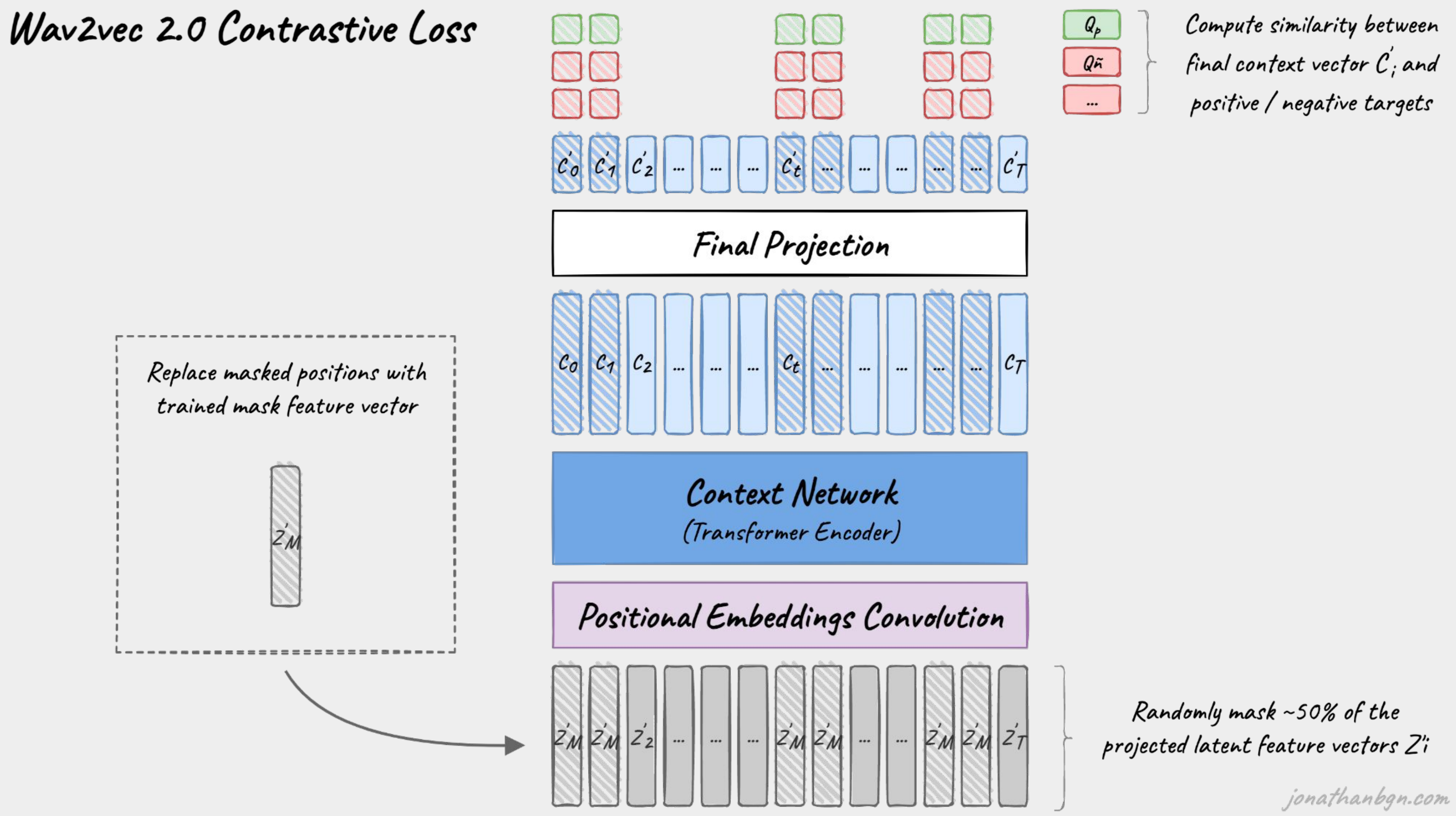
Appendix: Word2Vec2 Architecture

Context network (transformer encoder)


Wav2vec 2.0 Context Network (Transformer Encoder)
c_0 c_1 c_2 ... ... c_t ... ... c_T
Context Network
(Transformer Encoder)
Positional Embeddings Convolution
z'_0 z'_1 z'_2 ... ... z'_t ... ... z'_T
Feature Projection
z_0 z_1 z_2 ... ... z_t ... ... z_T
Final Output
+
Add
Add
Positional Embeddings
Grouped Convolution
Grouped 1D Convolution
z'_0 z'_1 z'_2 ... ... z'_t ... ... z'_T
Pad ending and beginning
jonathanbgn.com

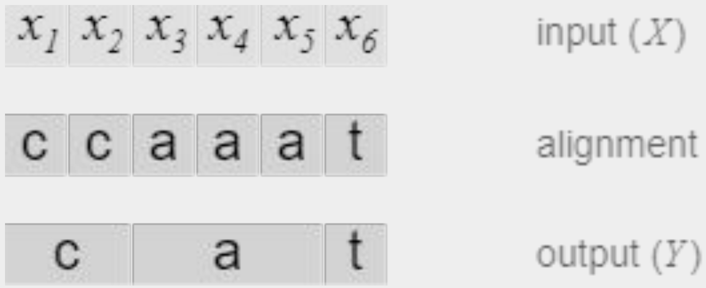
Appendix: Word2Vec2 Architecture

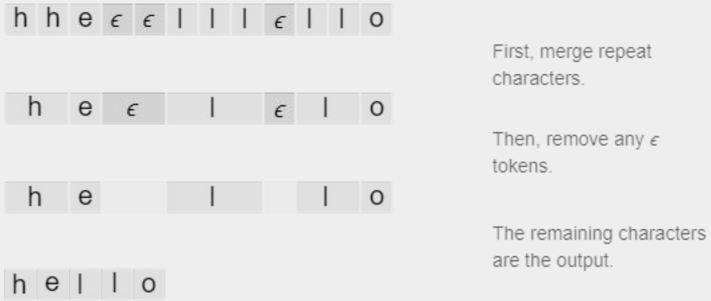
Pre-training & contrastive loss


Wav2vec 2.0 Contrastive Loss
Q_p
$Q_{\tilde{n}}$
...
Compute similarity between final context vector C'_i and positive / negative targets
C'_0 C'_1 C'_2 ... ... C'_t ... ... C'_T
Final Projection
C_0 C_1 C_2 ... ... C_t ... ... C_T
Context Network
(Transformer Encoder)
Positional Embeddings Convolution
Replace masked positions with trained mask feature vector
Z'_M
Z'_M Z'_M Z'_2 ... ... Z'_M Z'_M ... ... Z'_M Z'_M Z'_T
Randomly mask ~50% of the projected latent feature vectors Z'i
jonathanbgn.com

Finetuning via CTC

- Connectionist Temporal Classification (CTC) is an algorithm that is used to train neural networks for sequence-to-sequence problems
- Problem with finding the alignment when input sequence is much longer than output sequence
- We only feed the output matrix of the NN and the corresponding ground-truth (GT) text to the CTC loss function
- CTC uses RNN to find possible alignments of the GT text in the image and takes the sum of all scores
- Loss function: negative sum of log-probabilities

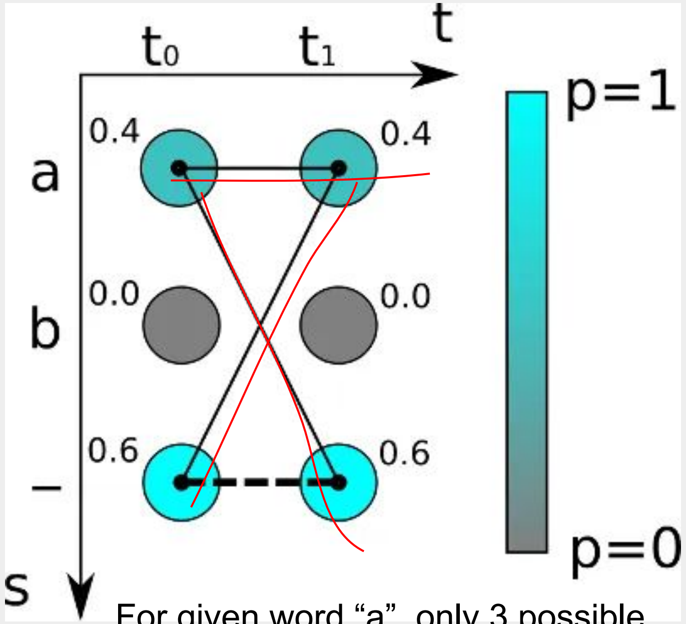




First, merge repeat characters.

Then, remove any ϵ tokens.

The remaining characters are the output.



For given word “a”, only 3 possible lines and sum of these predictions taken

Resources

Data source: National Speech Corpus, IMDA

https://huggingface.co/blog/fine-tune-wav2vec2-english

https://www.youtube.com/watch?v=t_qDAqUfqhY

https://medium.com/d-classified/speech-to-text-for-c3-ai-assistant-2d580f564a6b

https://www.researchgate.net/publication/335829066_Building_the_Singapore_English_National_Speech_Corpus