Task 4

Compare your finetuned model from task 3 to be compared against the results from task 2a for the cv-valid-dev mp3 dataset. Describe your observations and propose a series of steps (including datasets and experiments) to improve the accuracy.

In task 2d, the generated transcribed text from the pretrained model is saved in a column of the dataset called generated_text. The finetuned model is loaded and used to generate text from the audio dataset.

The word error rates (WER) for the generated texts from the pretrained model and the finetuned model are computed and compared as shown below.

Model	Word Error Rate
Pretrained Model	0.108
Finetuned Model	0.089

It can be observed that the WER of the finetuned model is lower than the pretrained model, indicating that the finetuned model has learnt from the training process on the Common Voice dataset. However, the difference in word error rate is not very big, showing that the model could perhaps be further trained for more epochs, to allow the model to generalise more to the training data.

Some random elements are generated for visualisation.

text	pretrained_text	finetuned_text
HE'S TRYING TO TRICK YOU AGAIN	HE'S TRYING TO TREAT YOU AGAIN	HE'S TRYING TO TREAK YOU AGAIN
THIS IS FOR YOU HE SAID HOLDING ONE OF THE PARTS OUT TO THE MONK	THIS IS FOR YOU HE SAID HOLDING ONE OF THE PARTS OUT TO THE MONK	THIS IS FOR YOU HE SAID HOLDING ONE OF THE PARTS OUT TO THE MONK
THEY WERE IN AN IMMENSE SETTING SURROUNDED BY THOUSANDS OF PEOPLE SPEAKING A STRANGE LANGUAGE	THEY WERE IN AN IMMENSE SETTING SURROUNDED BY THOUSANDS OF PEOPLE SPEAKING A STRANGE LANGUAGE	THEY WERE IN AN IMMENSE SETTING SURROUNDED BY THOUSANDS OF PEOPLE SPEAKING A STRANGE LANGUAGE
SOMETIME DURING THE SECOND YEAR YOU'LL REMEMBER ABOUT THE TREASURE	SOMETIME DURING A SECOND YEAR YOU'LL REMEMBER ABOUT THE TREASURE	SOMETIME DURING THE SECOND YEAR YOU'LL REMEMBER ABOUT THE TREASURE
LOVE IS THE FALCON'S FLIGHT OVER YOUR SANDS	LOVE IS THE FALCON'S FLIGHT OVER YOUR SANDS	LOVE IS THE FALCON'S FLIGHT OVER YOUR SANDS
HE NOTICED THE UNEQUAL COOLING OF ITS SURFACE	HE NOTICED THE UNEQUAL COOLING OF ITS SURFACE	HE NOTICED THE UNEQUAL COOLING OF ITS SURFACE
IN SPITE OF THIS I STILL BELIEVED THAT THERE WERE MEN IN MARS	IN SPITE OF THIS I STILL BELIEVE THAT THERE WERE MEN IN MARS	IN SPITE OF THIS I STILL BELIEVED THAT THERE WERE MEN IN MARS
WHY DO YOU THINK I BROUGHT YOU HERE	WHY DO YOU THINK I BROUGHT YOU HERE	WHY DO YOU THINK I BROUGHT YOU HERE
MY PASSWORD HAS BEEN CHANGED WITHOUT PERMISSION	MY PASSWORD HAS BEEN CHANGED WITHOUT PERMISSION	MY PASSWARD HAS BEEN CHANGED WITHOUT PERMISSION
THE REST OF YOU GET IN HERE AND RIOT	THE RISK IF YOU GET IN HERE AND RIDE	THE REST OF YOU GET IN HERE AND WRITE

To look into the details of how different the generated text is from the pretrained and finetuned models, the number of incorrect inferences from the 2 models is calculated and compared.

<u>Samples where finetuned model is wrong but pretrained model is correct</u>

There are 294 samples where the finetuned model is wrong but the pretrained model is correct. 10 such samples are shown in the table below.

finetuned_text	text
THEY WERE NOON AS SEARS AND THEY WERE HELD IN FEAR BY WOMEN AND THE ALDERLY	THEY WERE KNOWN AS SEERS AND THEY WERE HELD IN FEAR BY WOMEN AND THE ELDERLY
PUT DOWN THAT SHAAR	PUT DOWN THAT CHAIR
NOW I WANT YOU TO SAME THIS AGREEMENT	NOW I WANT YOU TO SIGN THIS AGREEMENT
AND THE GIRL POINTED TO THE SOUTH INDICATING THAT IT WAS THERE THE STRANGER MAN LIVED	AND THE GIRL POINTED TO THE SOUTH INDICATING THAT IT WAS THERE THE STRANGE MAN LIVED
IN ORDER TO FIND THE TREASURE YOU WILL HAVE TO FOLLOW WITH THE OMENS	IN ORDER TO FIND THE TREASURE YOU WILL HAVE TO FOLLOW THE OMENS
YOU CAN'T BELIEVE IT'S LOT BUTTER	YOU CAN'T BELIEVE IT'S NOT BUTTER
HIRES TWO PERSONAL TRAINERS AND MAKE ONE OF THEM TREN THE OTHER ONE	HIRES TWO PERSONAL TRAINERS AND MAKE ONE OF THEM TRAIN THE OTHER ONE
HE CAME TO UNDERSTAND ITS DUCHESS AND TRICKS AND TO ACCEPT IT AS IT WAS	HE CAME TO UNDERSTAND ITS DODGES AND TRICKS AND TO ACCEPT IT AS IT WAS
JUST HANDOWING THEM MADE HIM FEEL BETTER	JUST HANDLING THEM MADE HIM FEEL BETTER
YOU SHOULD HAVE BE HERE THE ALCHEMIST ANSWERED	YOU SHOULDN'T BE HERE THE ALCHEMIST ANSWERED

It can be observed that some errors are spelling issues such as ‘seer’ being spelt as ‘sear’ or ‘elderly’ being spelt as ‘alderly’. Most of the words wrongly identified by the finetuned model are similar in sounds to the original text. There is no drastic error in the finetuned text as compared to the groundtruth text.

Samples where pretrained model is wrong but finetuned model is correct

There are 514 samples where the pretrained model is wrong but the finetuned model is correct. 10 such samples are shown in the table below.

pretrained_text	text
SHE TOLD ME TO WRITE A PLAY FOR TO NIGHT	SHE TOLD ME TO WRITE A PLAY FOR TONIGHT
IT'S CALLED THE PRINCIPLE OF FAVOURABILITY BEGINNER'S LUCK	IT'S CALLED THE PRINCIPLE OF FAVORABILITY BEGINNER'S LUCK
THE ENGLISHMAN VANISHED TOO GONE TO FAIND THE ALCHEMIST	THE ENGLISHMAN VANISHED TOO GONE TO FIND THE ALCHEMIST
HE DIDN'T KNOW THE MAN YET BUT HIS PRACTISED EYE WOULD RECOGNIZE HIM WHEN HE APPEARED	HE DIDN'T KNOW THE MAN YET BUT HIS PRACTICED EYE WOULD RECOGNIZE HIM WHEN HE APPEARED
MAYBE TO MORROW SAID THE BOY MOVING AWAY	MAYBE TOMORROW SAID THE BOY MOVING AWAY
THE BOY'S NAME WAS FANTIABBLE	THE BOY'S NAME WAS SANTIAGO
MOST METEO RITES ARE MORE OR LESS ROUNDED	MOST METEORITES ARE MORE OR LESS ROUNDED
SOMETIMES SAILS CHARLES NORVAY TO HOLD BACG SERIMO	SOMETIMES THERE'S JUST NO WAY TO HOLD BACK THE RIVER
THERE THE ALCHEMISTS SEPARATED THE DISK INTO FOUR PARTS	THERE THE ALCHEMIST SEPARATED THE DISK INTO FOUR PARTS
HE MOVED ABOUT INVISIBLE BUT EVERY ONE COULD HEAR HIM	HE MOVED ABOUT INVISIBLE BUT EVERYONE COULD HEAR HIM

Similar to the previous table, most errors are similar sounding to the original texts, such as ‘practiced’ being spelt as ‘practised’. However, there are also more gibberish texts being generated such as ‘hold bacg serimo’ or ‘fantiabble’, which indicates the pretrained model’s incapability to capture some sounds in the data. This shows an improvement in the finetuned model, since the finetuned model was able to recognise the audio data for these samples correctly.

Samples where both pretrained and finetuned models are wrong

There are 1324 samples where both the pretrained model and the finetuned model are wrong. 10 such samples are shown in the table below.

text	finetuned_text	pretrained_text
THE OLD MAN LEAFED THROUGH THE BOOK AND FELL TO READING A PAGE HE CAME TO	THE OLD MAN LEAVED THROUGH THE BOOK AND FELL TO READING A PAGE HE CAME TO	THE OLD MAN LEAPED THROUGH THE BROOK AND FELL TO READING A PAGE HE CAME TO
WE'RE IN NO HURRY THE CHIEF ANSWERED	WE ARE IN NO HURRY THE CHIEF ANSWERED	WE ARE IN NO HURRY THE CHIEF ANSWERED
JUDGE DEBRA SENT ME	JUDGED DEBORA SENT ME	JUDGED DEBORAH SET ME
AT HIS SIDE WAS THE YOUNG ARAB THE BOY HAD SPOKEN WITH EARLIER	AT HIS SIDE WAS A YOUNG HARABS THE BOY HAD SPOKEN WITH HALIER	AT HIS SIDE WAS A YOUNG HARABSY BOY HAD SPOKEN WITH HERLIER
YOU'RE IN LUCK YOU TWO THE FAT ARAB SAID	YOU'RE IN LUCK YOU TWO THE FAD HELLAB SAID	YOU'RE UNLUCKY TOO THE FAD HAROD SAID
THINK I'LL GO HOME AND SEE WHAT THE FAMILY IS DOING	THEN YOU CI'LL GO HOME AND SEE WHAT THE FAMILY'S DOING	THENK I'LL GO HOME AND SEE WHAT THE FAMILY IS DOING
MY REVIEW OF THE SUN ONE STAR	WHYT AR YOU HEAR OF THE SUN ONE STAR	MARIVE OF THE SUN ONE STA
I DON'T WANT MRS DOUGLAS	I DON''T WANT MR TOLIGUST	I DON'T WANT MISSUS DOULAS
GET A LOAD OF THIS	GET THE LOAD OF THIS	GET THE LOAD OF THIS
THE RIVER BABBLED INANELY TO ITSELF	THE RIVER BABBLED INAMELY TO ITSELF	THE RIVER BABBLED NAMELY TO ITSELF

In certain instances, even though both finetuned text and pretrained text are wrong, the finetuned text is closer to the groundtruth text than the pretrained text. For example, the finetuned model recognised ‘inanely’ as ‘inamely’, as compared to the pretrained model which recognised it as ‘namely’. There are also some instances where both models are incorrect such as recognising the word ‘arab’ as ‘hellab’ or ‘harod’. Some of these errors seem to require more contextual knowledge to provide the correct audio transcription.

Samples where the finetuned model is wrong

There were 1618 samples where the finetuned model is wrong.

text	finetuned_text	pretrained_text
THE AREA WAS SWIRLING IN DUST SO INTENSE THAT IT HID THE MOON FROM VIEW	MARIA WAS FILLING A DUST SO ENTENSETHAT IT HID THE MOON FROM YOU	MYRIA WAS HOLLING IN THUS SO INTLONS THAT IT HID THE MOON FROM YOU
HE PUT HIS HEADCLOTH IN PLACE AND SECURED IT WITH A RING MADE OF CAMEL SKIN	HE PUT HIS HEAT GLOWED IN BLAZE AND SECURED IT WITH A RING MADE OF CAMEL'S KIN	HE PUT HIS HAT CLOTHED IN PLACE AND SECURED IT WITH A RING MADE OF CAMEL SKIN
I WASN'T BORN YESTERDAY	I WAS IT BORN YESTERDAY	I WAS IN BORN YESTERDAY
THE BOY NOTICED THAT THE OWNER OF THE BAR STOOD NEARBY LISTENING ATTENTIVELY TO THEIR CONVERSATION	THE BOY NOTICED THA THEY ONNEAR THE BAR STOOD NEARBY LISTENING ATTENTIVELY TO THEIR CONVERSATION	THE BOY NOTICED TH THE ONER OF THE BAR STOOD NEAR BY LISTENING ATTENTIVELY TO THEIR CONVERSATION
EVIDENTLY WE'RE THE FLOOR SHOW	EVIDENTLY WHERE THE FLOOR SHOW	EVIDENTLY WHERE THE FLOOR SHOW
ARE YOU CERTAIN IT'S NOT A MONSTER	FINLY CRTIN THE CERTAN WAS	BY IS CERTAINTE SARTON ANS
THE BASKETBALL BOUNCED OFF HIS SHIELD OF TITANIUM	THE BASKETBALL BOUNCED WITH HIS SHIELD OF TITANIUM	THE BASKETBALL BOUNCED OF HIS SHIELD OF TITANIUM
COMPILING THE LINUX KERNEL CAN BE TIME CONSUMING	COMPILING THE LINNOX COLONEL COULD BE TIME CONSUMING	COMPILING THE LINNOX COLONEL CAN BE TIME CONSUMING
THIS MORNING I FOUND A CALCULATOR TAPED TO MY WII	THIS MORNING I FOUND A CALCULATOR TIPPED TO MY WII	THIS MORNING I FOUND A CALCULATOR TIPPED TO MY WEE
HE HAD COME TO THE TOWN ONLY TO FIND A WOMAN WHO COULD INTERPRET HIS DREAM	HE HAD COME TO A TOWN ONLY TO FIND A WOMAN WHO COULD INTERPRET HIS DREAM	HE HAD COME TO A TOWN ONLY TO FIND A WOMAN WHO COULD INTERPRET HIS DREAM

Looking closer into the samples where the finetuned model is wrong, some errors such as spelling ‘camel’s kin’ instead of ‘camel skin’ indicates that the model was roughly able to capture the sounds, just unable to make out the correct words. Other errors such as recognising ‘are you certain its not a monster’ as ‘finly crtin the certan was’ indicate that the model is still inadequate and requires more training to be able to be accustomed to more words.

Proposed Steps for Improvement

1. Longer Training Time:

Extend the number of training epochs to allow the model more time to learn and adapt to the nuances of the dataset.

2. Data Augmentation

Implement data augmentation techniques that specifically introduce audio variations that mimic common mishearings. This would involve creating synthetic training samples where phonetically similar words are swapped, thereby teaching the model to better discriminate between such sounds. This could potentially address the issue of the model making spelling errors in similar sounding words.

3. Contextual Training:

Incorporate language modeling to give the system better contextual understanding, which can help differentiate between similar-sounding words. Training on a large corpus of text data such as books, articles and conversations can help the model to gain better contextual knowledge for better word prediction.

4. Ensemble Methods:

Experiment with an ensemble of models or multimodal architectures to improve robustness. Different models may capture different aspects of the speech patterns and when combined, could lead to a higher overall performance.

5. Phonetic Distinction Training

Train the model to distinguish between phonemes that are commonly confused. This could involve a specialized pre-training phase that focuses solely on phonetic clarity. There are also models such as 'vitouphy/wav2vec2-xls-r-300m-timit-phoneme' on HuggingFace which uses the Wav2Vec2 model as a base for phoneme recognition. A language model trained with phonemes and ASR transcripts could learn a phonetic-aware representation that is more robust to noise and errors in transcript.