VibeSpeaker

Vișan Ionuț

Poață Cătălin-Andrei

Vulpe Ștefan

Artificial Intelligence Master, UNSTP Bucharest

Abstract

This study focuses on building a robust multimodal system for speech emotion recognition (SER) by integrating audio, text, and spectrogram features. The proposed system leverages a pre-trained **WavLM** model for extracting deep audio features, **RoBERTa** for semantic text representation, and a custom **CNN** architecture for spectrogram processing. These modality-specific features are fused through a dense layer to enhance multimodal representation.

The model utilizes advanced preprocessing, including spectrogram normalization and resizing, while freezing certain layers of pre-trained models to prevent overfitting and optimize usage.

Experimental results demonstrate that this approach effectively captures complementary information across modalities, improving the accuracy/CCC and robustness of emotion classification/regression tasks. The system achieves this through iterative training, evaluation, an adapted configuration for each model, providing a strong baseline for multimodal SER.

1. Introduction

Speech Emotion Recognition (SER) has become a cornerstone for improving human-computer interactions, offering significant potential in creating more intuitive and adaptive systems. In applications ranging from **customer service** to **mental health** monitoring, understanding emotional states enhances **communication** and **personalization**, leading to better outcomes and user experiences.

Our work presents an advanced solution that integrates automatic **speech-to-text transcription** with a robust **emotion recognition** system, creating a unified pipeline for seamless inference. By combining these capabilities, the model enables **real-time** analysis of spoken content, being able to classify the emotion from speech audios. This dual understanding supports a wide range of use cases, including personalized virtual assistants, sentiment-aware customer interactions, and tools for emotional well-being.

The proposed system leverages **state-of-the-art** methodologies to handle the complexity of natural speech, addressing challenges like variability in emotional expression. The integration of transcription and emotion recognition provides a complete service, making it adaptable to diverse environments where real-time emotional intelligence is critical. This approach lays the groundwork for more human-centric AI applications that respond not just to what is said, but how it is expressed.

2. Related work

The development of our system builds upon recent advancements in Speech Emotion Recognition (SER), addressing key challenges such as the valence gap, class imbalance and multimodal approach through state-of-the-art methodologies exposed in the following papers:

2.1. Addressing the Valence Gap with Transformer Models

The paper *Dawn of the Transformer Era in Speech Emotion Recognition: Closing the Valence Gap* (Wagner et al., 2023) explores the transformative impact of pre-trained models like Wav2Vec2.0 and HuBERT in SER tasks. These transformer-based architectures demonstrate exceptional performance in capturing both paralinguistic and linguistic features, significantly improving the historically challenging prediction of valence. Their ability to encode acoustic and semantic content during fine-tuning formed the foundation for our audio processing pipeline, emphasizing the importance of leveraging pre-trained architectures for robust feature extraction.

2.2. Tackling Class Imbalance, Multimodal model in SER

The paper *1st Place Solution to Odyssey Emotion Recognition Challenge Task 1: Tackling Class Imbalance Problem* (Zhou et al., 2024) focuses on addressing the challenge of imbalanced emotional distributions in SER datasets. By introducing techniques such as class-weighted focal loss and ensemble learning, this work achieved state-of-the-art performance, demonstrating the potential of tailored loss functions and multi-model integration for enhancing classification accuracy. These insights guided our approach to try a multimodal architecture, ensuring improved recognition of minority classes.

3. Research

To achieve a complete version of the desired service, we conducted extensive research to find models suitable for the SER task, such as ResNet, Wav2Vec2, Hubert and WavLM. Initially, we intended to use a separate model for extracting sentiments from text, which would be integrated with the audio model.

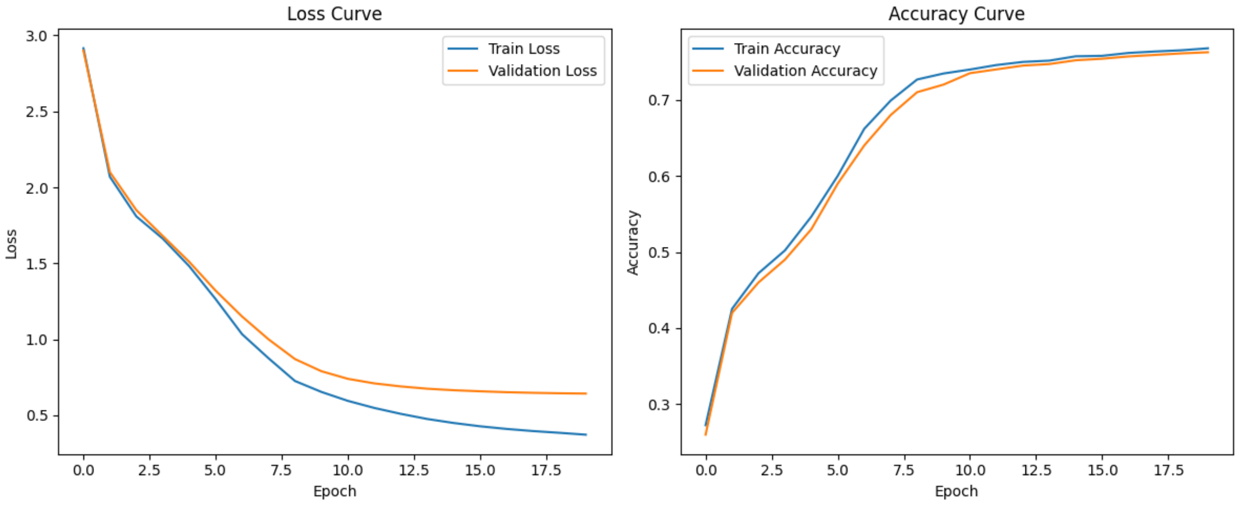
3.1. Text to Emotion

For extracting emotions from text, we implemented two separate approaches: one using the RoBERTa-base pre-trained transformer and another leveraging a BiLSTM architecture.

RoBERTa Approach

The RoBERTa-based model utilized the following configuration:

- **Tokenizer**: roberta-base
- **Text Model**: roberta-base
- **Sequence Length**: max_length=128 with truncation enabled and padding set to max_length.
- **Trainable Layers**: All layers of the RoBERTa model were fine-tuned.
- **Dropout**: A hidden layer dropout probability of 0.3 was applied to prevent overfitting.
- **Optimizer**: AdamW with a learning rate of 2e-5 and a weight decay of 0.01.
- **Scheduler**: A linear learning rate scheduler with 10% warmup steps.
- **Loss Function**: CrossEntropyLoss, suitable for multi-class emotion classification.
- **Batch Size**: 32 for both train_dataloader and test_dataloader.
- **Training**: Conducted over 20 epochs. After each epoch, the mean validation accuracy was calculated. If the validation accuracy improved, the model's state_dict, along with its processor and tokenizer, was saved as the "best model."

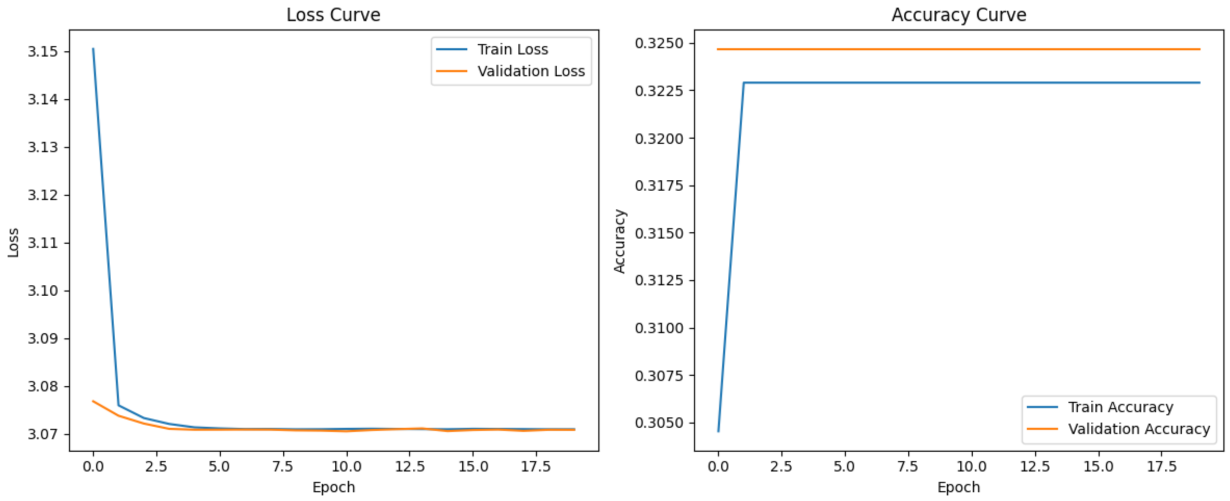


The RoBERTa approach showcased strong performance, benefiting from its pre-trained knowledge on large text corpora. However, it also demanded significant computational resources due to the training of all layers.

BiLSTM Approach

The BiLSTM-based model adopted a simpler architecture, configured as follows:

- **Tokenizer**: roberta-base tokenizer for consistent text preprocessing.
- **Text Model**: A custom BiLSTM architecture with 2 layers and a hidden dimension size of 128.
- **Sequence Length**: Same as RoBERTa (max_length=128 with truncation and padding).
- **Trainable Layers**: All parameters of the BiLSTM were trained.
- **Dropout**: A dropout rate of 0.3 was applied between BiLSTM layers to mitigate overfitting.
- **Optimizer**: AdamW with a learning rate of 2e-5 and a weight decay of 0.01.
- **Scheduler**: Linear learning rate scheduler with a warmup ratio of 10%.
- **Loss Function**: CrossEntropyLoss.
- **Batch Size**: 32 for both train_dataloader and test_dataloader.
- **Training**: Similar to the RoBERTa setup, training was conducted over 20 epochs. The mean validation accuracy was monitored, and the "best model" was saved whenever the validation accuracy improved.



While the BiLSTM approach required less computational power and was easier to train, it lacked the robust pre-trained knowledge of RoBERTa. As a result, its performance was relatively lower, especially for text inputs with subtle emotional cues.

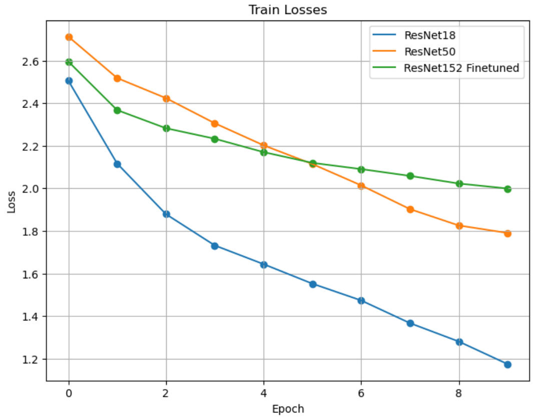
3.2. Models on full dataset

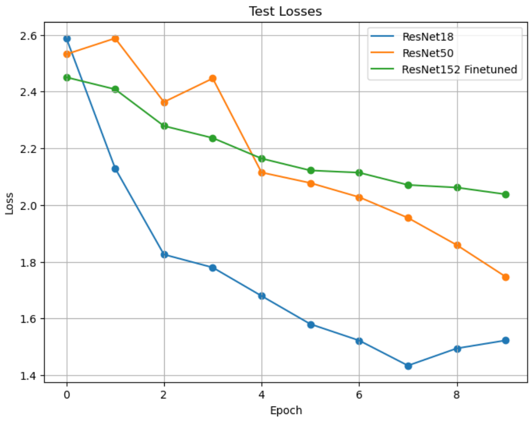
3.2.1. ResNets

We evaluated three variants of the ResNet CNN architecture on the multilingual corpus to tackle the task of emotion classification. The models tested were ResNet18, ResNet50 and ResNet152. For the ResNet18 and ResNet50 configurations, the networks were trained from scratch for 10 epochs, using the AdamW with a learning rate of *1e-4* and cross-entropy loss function. Batch size was set to 32, and a dropout of 0.3 was used in the fully connected layers at the end to add regularization. On the other hand, the ResNet152 model, was initialized as a pre-trained model on ImageNet. All convolutional layers were frozen, and only the two classification heads responsible for main emotion classes and emotion subclasses were fine-tuned using the configuration described above.

The performance of each ResNet configuration is summarized as follows:

Model	***Accuracies***	
	Main class	***Subclass***
ResNet18	84%	70%
ResNet50	76%	58%
ResNet152	74%	51%





The results indicate that ResNet18 outperformed the other variants across both tasks, achieving the highest accuracy for both main and subclass predictions. Despite being a deeper model, ResNet152 underperformed due to its frozen convolutional layers, suggesting that fine-tuning additional layers may be beneficial. ResNet50, while achieving moderate performance, exhibited a significant drop in subclass accuracy compared to ResNet18.

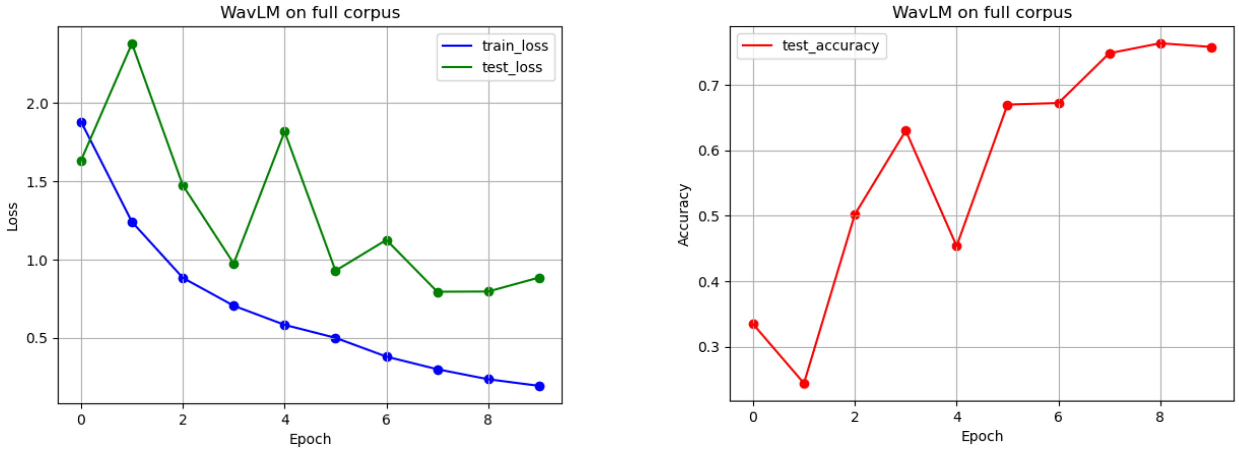
These findings highlight the trade-off between model depth and performance in our initial CNN-based experiments. Further investigation will focus on optimizing deeper models and transitioning to transformer-based architectures to improve results.

3.2.2. WavLM fine-tuning on multilingual corpus

We fine-tuned the WavLM model on the complete corpus of audio files to address the task of emotion classification. The architecture was modified by introducing a 1D adaptive average pooling layer, applied to the final hidden state of the WavLM model. Subsequently, a feed-forward network consisting of two linear layers, separated by a GELU activation function and followed by a dropout layer, was appended for emotion class prediction. During training, the feature encoder of WavLM remained frozen, ensuring that only the parameters of the added layers were updated.

The model was trained for 10 epochs using a batch size of 10 and a learning rate of *2e-4*. The AdamW optimizer was employed, incorporating a warm-up phase comprising 10% of the total training steps, followed by a linear learning rate decay schedule. To efficiently utilize computational resources, gradient accumulation was performed over five steps, allowing multiple batches to contribute to a single optimization step. These modifications and training strategies were designed to enhance the model's ability to classify emotions effectively.

The results obtained from the experiment described above can be viewed in the figure below:

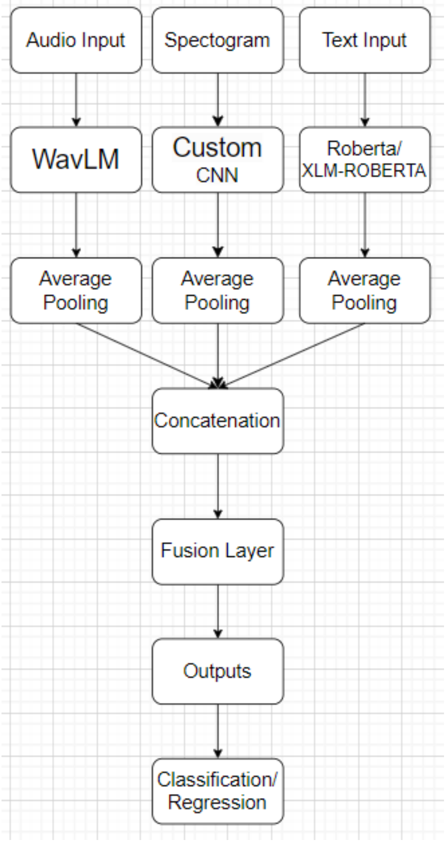


4. System Description

The proposed system integrates audio, text, and spectrogram features to achieve robust Speech Emotion Recognition (SER). For audio processing, we utilize WavLM (microsoft/wavlm-base-plus) to extract deep features from raw waveforms. Textual inputs are processed using RoBERTa (roberta-base) to capture semantic information, and spectrograms are analyzed via a CNN module with four convolutional blocks for high-level feature extraction. The fusion of these modalities is performed through a dense layer, enabling efficient emotion classification.

During training, we leverage pre-existing transcriptions provided in the dataset. However, for inference, we employ the **Whisper-large-v3-turbo** model for real-time transcription, enabling a seamless pipeline for emotion inference from audio-only inputs. This approach ensures flexibility and adaptability for real-world applications.

4.1. Model Architecture



The multimodal model used for SER is a sophisticated architecture designed for robust Speech Emotion Recognition (SER), leveraging audio, text, and spectrogram features. Below is an overview of its components:

4.1.1. Audio Processing Module

- Base Model: The model uses a pre-trained WavLM (microsoft/wavlm-base-plus) to extract deep audio features from raw waveforms.
- Pooling: An adaptive average pooling layer reduces the output to a fixed-dimensional vector, summarizing the audio features.

4.1.2. Text Processing Module

- Base Model: Pre-trained RoBERTa (roberta-base)/XLM-RoBERTa (xlm-roberta-base) is used for text feature extraction, effectively capturing the semantic meaning of transcriptions.
- Pooling: Another adaptive average pooling layer compresses the textual representations into a fixed-length vector.

4.1.3. Spectrogram Processing Module

- CNN Architecture: A convolutional neural network with four layers processes the spectrograms to extract high-level features:
    - Convolutional Layers: Each convolutional layer includes a combination of convolution, batch normalization, and ReLU activation for feature extraction and stabilization.
    - Pooling: Max-pooling is applied after each convolutional block to reduce spatial dimensions.
    - Feature Aggregation: The final adaptive average pooling layer generates a fixed-dimensional vector (256 dimensions) from the processed spectrogram.

4.1.4. Fusion Module

- Feature Fusion: The outputs from the audio, text, and spectrogram modules are concatenated into a unified feature vector.
- Fusion Layer: A dense linear layer integrates the modalities and transforms the combined features into a representation suitable for classification/regression.

4.1.5. Classification/Regression Head

- Output Layer: A fully connected layer maps the fused features to the target classes (8 emotion categories for classification, 3 dimensions for regression).

4.2. Audio Features

WavLM is a state-of-the-art pre-trained speech model built on a Transformer architecture, excelling in tasks such as speech recognition, emotion recognition, and speaker identification. Designed for robust speech representation learning, it serves as a powerful backbone for extracting high-quality audio features in our multimodal emotion recognition system.

Key Features:

- **Architecture:** WavLM employs a multi-layer Transformer encoder to capture long-term dependencies in audio data, extracting rich contextual representations directly from raw waveforms.
- **Utterance-level Training:** To enhance speaker discrimination and handle overlapping speech, WavLM incorporates utterance mixing, creating diverse and realistic training scenarios.
- **Masked Prediction:** Inspired by masked language models, WavLM masks portions of input audio, compelling the model to predict missing content and boosting feature robustness.
- **State-of-the-Art Performance:** Trained on large-scale datasets (e.g., 94k hours of speech), WavLM achieves benchmark-leading performance across numerous speech processing tasks.

WavLM-Base-Plus is a variant of WavLM with 12 Transformer layers and 94 million parameters. It balances efficiency and robust performance, leveraging masked prediction and utterance mixing to excel in speech-related tasks.

The **Wav2Vec2 Feature Extractor** complements WavLM by preprocessing raw waveforms into normalized inputs. It captures essential audio characteristics such as frequency and temporal patterns, enabling seamless integration and effective downstream task performance.

4.3. Visual Features

In our multimodal emotion recognition system, Convolutional Neural Networks (CNNs) play a pivotal role in processing spectrograms to extract high-level visual features from audio data. Spectrograms serve as a powerful representation of audio signals, capturing the frequency content over time.

Spectrogram Features:

- Mel Spectrograms: The audio signals are converted into Mel spectrograms, which represent the audio's frequency content using a scale that mimics human auditory perception.
- Normalization: Spectrograms are normalized to enhance feature extraction and reduce the impact of varying signal amplitudes.
- Resizing: The spectrograms are resized to a consistent input size suitable for CNN processing.

CNN Architecture:

- Layered Design: The CNN consists of four convolutional blocks, each comprising convolutional layers, batch normalization, ReLU activations, and pooling layers.
- Hierarchical Feature Extraction: The layered structure allows the CNN to capture local patterns (e.g., edges and textures) and high-level features (e.g., tonal and rhythmic structures).
- Dimensionality Reduction: Pooling operations progressively reduce the spatial dimensions, focusing on salient features.
- Final Aggregation: Adaptive average pooling at the final layer condenses the learned features into a compact representation for downstream tasks.

By transforming audio signals into spectrograms and utilizing CNNs for feature extraction, the system effectively captures rich and discriminative information from the audio, enhancing the emotion recognition capability.

4.4. Text Features

RoBERTa (Robustly Optimized BERT Pretraining Approach) and XLM-RoBERTa are Transformer-based models designed for advanced natural language processing (NLP) tasks. In our multimodal emotion recognition system, these models process textual inputs to extract deep contextual representations of language, complementing audio features.

Key Features:

RoBERTa:

- **Transformer Architecture**: Built with a multi-layer Transformer encoder, RoBERTa captures contextual information from text by modeling dependencies across words and phrases.
- **Pretraining**: Trained on a massive English corpus with a masked language modeling (MLM) objective, enabling nuanced understanding of language semantics.
- **Fine-Grained Tokenization**: Uses Byte-Pair Encoding (BPE) to handle rare and complex language inputs effectively.

XLM-RoBERTa:

- **Multilingual Pretraining**: Pretrained on 100+ languages, XLM-RoBERTa is highly suited for cross-lingual NLP tasks and multilingual data.
- **Masked Language Modeling**: Utilizes MLM to derive contextual representations across diverse languages.
- **Robust Tokenization**: Employs SentencePiece tokenization, making it effective for handling multilingual and low-resource text.

Text Feature Extraction:

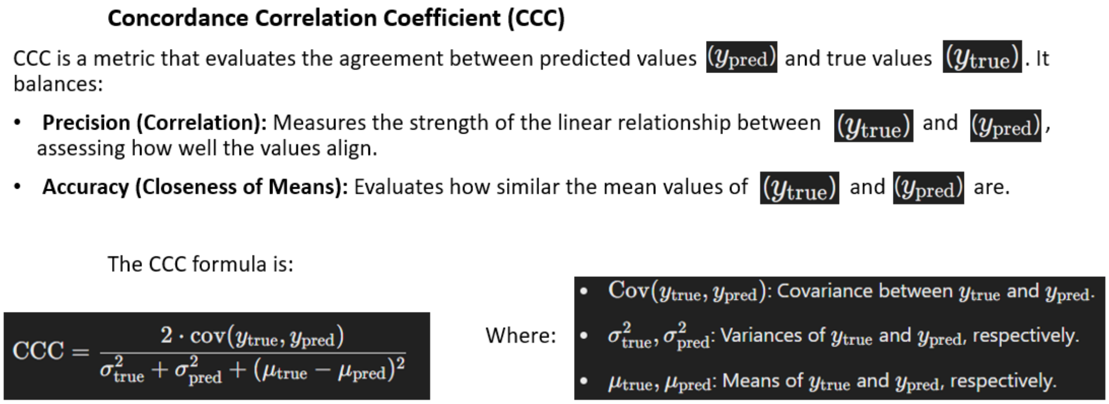
- **Input Representation**: Text is tokenized into input IDs and attention masks for proper sequence encoding.
- **Contextual Embeddings**: Both models transform tokens into rich contextual embeddings that reflect their meaning within the sentence.
- **Pooling**: An adaptive average pooling layer condenses token-level embeddings into fixed-size representations for downstream tasks.

Impact on Emotion Recognition:

The advanced language modeling capabilities of RoBERTa and XLM-RoBERTa enhance the system's performance by providing detailed, context-aware text embeddings. These embeddings effectively fuse with audio features, improving multimodal emotion recognition accuracy in both monolingual and multilingual contexts.

4.5. Loss functions

In our regression system, Concordance Correlation Coefficient (CCC) Loss is used as the loss function. It is well-suited for tasks like emotion recognition, where the objective is to predict continuous values. The CCC loss measures the agreement between predicted and true continuous labels, considering both their correlation and similarity in scale. Key aspects of CCC loss include:



Concordance Correlation Coefficient (CCC)

CCC is a metric that evaluates the agreement between predicted values (y_{pred}) and true values (y_{true}). It balances:

- **Precision (Correlation):** Measures the strength of the linear relationship between (y_{true}) and (y_{pred}), assessing how well the values align.
- **Accuracy (Closeness of Means):** Evaluates how similar the mean values of (y_{true}) and (y_{pred}) are.

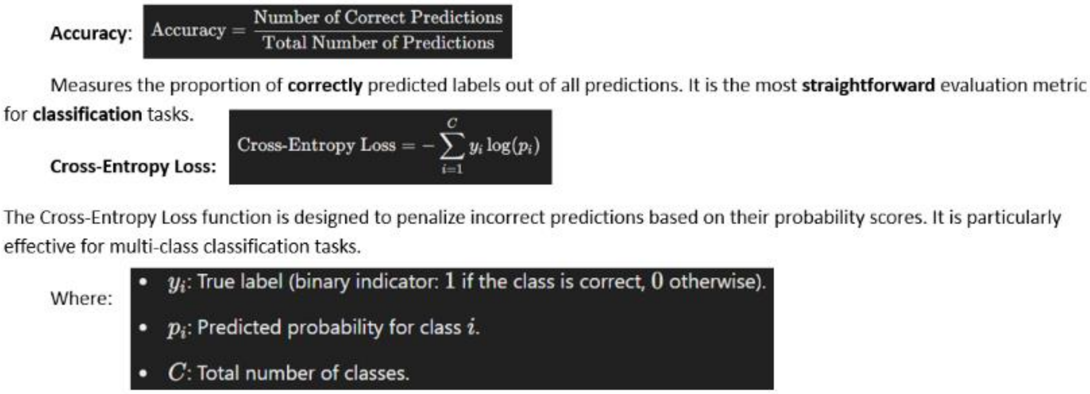
The CCC formula is:

$$\text{CCC} = \frac{2 \cdot \text{cov}(y_{\text{true}}, y_{\text{pred}})}{\sigma^2_{\text{true}} + \sigma^2_{\text{pred}} + (\mu_{\text{true}} - \mu_{\text{pred}})^2}$$

Where:

- $\text{Cov}(y_{\text{true}}, y_{\text{pred}})$: Covariance between y_{true} and y_{pred}.
- $\sigma^2_{\text{true}}, \sigma^2_{\text{pred}}$: Variances of y_{true} and y_{pred}, respectively.
- $\mu_{\text{true}}, \mu_{\text{pred}}$: Means of y_{true} and y_{pred}, respectively.

In our classification system, **Cross-Entropy Loss** is used as the loss function. It is well-suited for multi-class classification problems, such as emotion recognition. The loss function calculates the difference between the predicted probability distribution of classes and the true labels. The key aspects of this loss function include:



Accuracy: $\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}}$

Measures the proportion of **correctly** predicted labels out of all predictions. It is the most **straightforward** evaluation metric for **classification** tasks.

Cross-Entropy Loss: $\text{Cross-Entropy Loss} = -\sum_{i=1}^{C} y_i \log(p_i)$

The Cross-Entropy Loss function is designed to penalize incorrect predictions based on their probability scores. It is particularly effective for multi-class classification tasks.

Where:

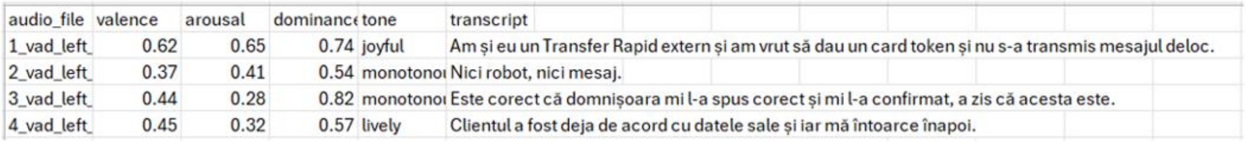
- y_i: True label (binary indicator: 1 if the class is correct, 0 otherwise).
- p_i: Predicted probability for class i.
- C: Total number of classes.

5. Experimental Setup

5.1. Regression Dataset

The dataset consists of 500 entries, each representing an audio file annotated with both acoustic and textual information. The key features include:

1. Audio File: The name of the corresponding audio file.
2. Valence, Arousal, Dominance: Continuous values representing emotional dimensions.
3. Transcript: The text transcription of the spoken content.
4. Tone: A categorical label describing the tone of the speech.



audio_file	valence	arousal	dominance	tone	transcript
1_vad_left_	0.62	0.65	0.74	joyful	Am și eu un Transfer Rapid extern și am vrut să dau un card token și nu s-a transmis mesajul deloc.
2_vad_left_	0.37	0.41	0.54	monotono	Nici robot, nici mesaj.
3_vad_left_	0.44	0.28	0.82	monotono	Este corect că domnișoara mi l-a spus corect și mi l-a confirmat, a zis că acesta este.
4_vad_left_	0.45	0.32	0.57	lively	Clientul a fost deja de acord cu datele sale și iar mă întoarce înapoi.

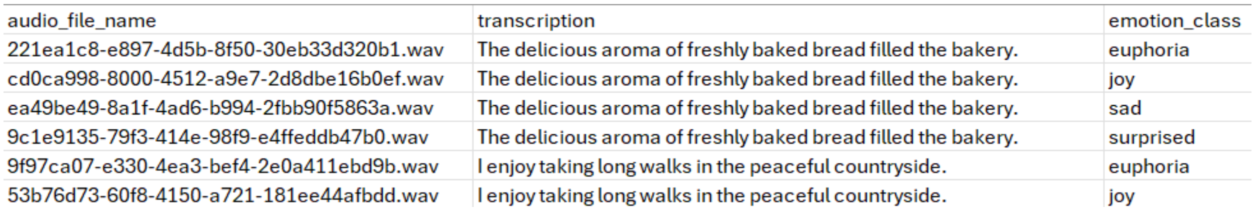
5.2. Classification Dataset

The dataset is a combination of Toronto Emotional Speech Set (TESS) and a Kaggle Speech Emotion Recognition Voice Dataset, consisting of 2468 entries, each representing an audio file annotated with emotions class and transcription. The key features include:

1. Audio File: The name of the corresponding audio file.

2. Transcription: The text transcription of the spoken content.

3. Emotion class: The corresponding emotion class of the audio file.



audio_file_name	transcription	emotion_class
221ea1c8-e897-4d5b-8f50-30eb33d320b1.wav	The delicious aroma of freshly baked bread filled the bakery.	euphoria
cd0ca998-8000-4512-a9e7-2d8dbe16b0ef.wav	The delicious aroma of freshly baked bread filled the bakery.	joy
ea49be49-8a1f-4ad6-b994-2fbb90f5863a.wav	The delicious aroma of freshly baked bread filled the bakery.	sad
9c1e9135-79f3-414e-98f9-e4ffeddb47b0.wav	The delicious aroma of freshly baked bread filled the bakery.	surprised
9f97ca07-e330-4ea3-bef4-2e0a411ebd9b.wav	I enjoy taking long walks in the peaceful countryside.	euphoria
53b76d73-60f8-4150-a721-181ee44afbdd.wav	I enjoy taking long walks in the peaceful countryside.	joy

5.3. Regression Configuration

***Audio model* =** “microsoft/wavlm-base-plus”

***Hidden_dim_audio* =** 768

***Text model* =** “xlm-roberta-base”

***Hidden_dim_text* =** 768

Hidden_dim_cnn = 256

***max_length* for tokens =** 512

***Trainable layers*:**

- audio_pooling

- text_pooling

- fusion_layers

- fc (final output layer)

***Frozen layers*:**

feature_extractor + encoder (for both Audio and Text models)

***Batch_size* =** 4 for train_dataloader/test_dataloder

***Learning_rate* =** 2.5e-4

***Criterion* =** CCCLoss

***Optimizer* =** AdamW

***Scheduler* =** linear with 0.1 warmup

***Accumulation_step* =** 1

Num_epochs = 20

After each epoch, compute the mean CCC for Valence, Arousal, and Dominance, and save the the "best model" if the mean CCC improves.

5.4. Classification Configuration

***Audio model* =** “microsoft/wavlm-base-plus”

***Hidden_dim_audio* =** 768

***Text model* =** “roberta-base”

***Hidden_dim_text* =** 768

***Hidden_dim_cnn* =** 256

***Max_length* for tokens =** 512

Trainable layers:

- audio_pooling

- text_pooling

- fusion_layers

- fc (final output layer)

Frozen layers:

feature_extractor + encoder (for both Audio and Text models)

***Batch_size* =** 8 for train_dataloader/test_dataloder

***Criterion* =** CrossEntropyLoss

***Weight_decay* =** 0.01

***Optimizer* =** AdamW

***Scheduler* =** linear with 0.1 warmup

***Accumulation_step* =** 1

Learning_rate = 1e-4

The best model is updated if validation loss decreases and validation accuracy improves compared to the previous best.

6. Speech to Text

6.1. Whisper

For training, we use transcriptions derived from the dataset, but to enable complete inference, we need a model capable of transcribing text directly from audio. For this purpose, we will use the Whisper-large-v3-turbo model.

Whisper is a state-of-the-art speech recognition model by OpenAI. Key details:

- Transcribes and translates audio into text across multiple languages.

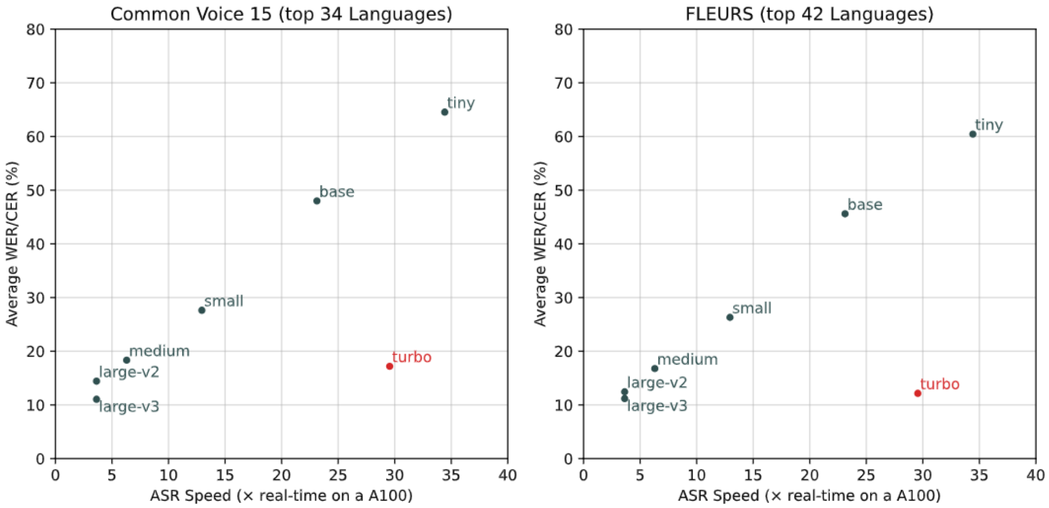
- Transformer-based with an encoder-decoder design.

- Takes mel spectrograms as input for audio processing.

- Known for robustness in noisy environments and diverse accents.

- Available in multiple sizes (e.g., small, medium, large) for different performance and resource needs.

- Used for transcription, translation, and real-time ASR tasks.



Whisper-large-v3-turbo maintains the encoder-decoder architecture of the original Whisper model but introduces specific optimizations:

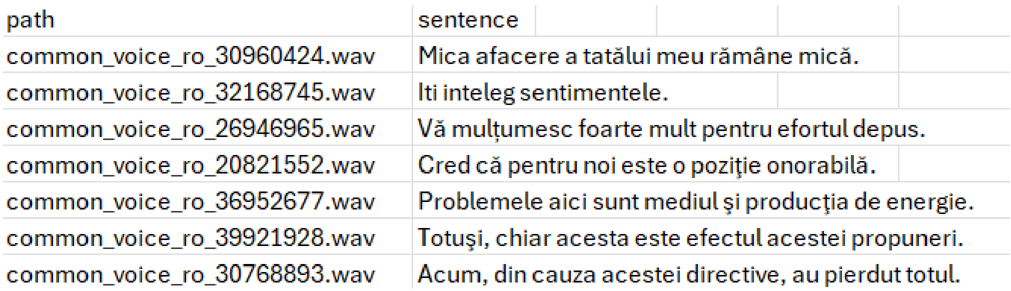
Encoders: The model uses a single encoder stack consisting of 32 transformer layers, responsible for processing the audio input (Mel spectrogram).

Decoders: Unlike the original Whisper-large-v3 with 32 transformer layers in the decoder, the turbo version reduces the decoder layers to just 4. This significantly improves transcription speed with minimal impact on accuracy.

6.2. Whisper-large-v3-turbo_ro

To improve the performance of our transcription model, we will fine-tune Whisper-large-v3-turbo using the validated segments of the Common Voice 19 dataset.

It consists of 17,965 audio files with transcriptions in Romanian.



path	sentence
common_voice_ro_30960424.wav	Mica afacere a tatălui meu rămâne mică.
common_voice_ro_32168745.wav	Iti inteleg sentimentele.
common_voice_ro_26946965.wav	Vă mulțumesc foarte mult pentru efortul depus.
common_voice_ro_20821552.wav	Cred că pentru noi este o poziţie onorabilă.
common_voice_ro_36952677.wav	Problemele aici sunt mediul şi producţia de energie.
common_voice_ro_39921928.wav	Totuşi, chiar acesta este efectul acestei propuneri.
common_voice_ro_30768893.wav	Acum, din cauza acestei directive, au pierdut totul.

WER/CER on test set before fine-tuning:

WER - **17,12**%

CER – **4,88**%

6.2.1. Fine-tuning Configuration

Trainable layers = All

Batch_size = 2 for both dataloaders

Optimizer = AdamW

Scheduler = linear with 0.1 warmup

Weight_decay = 0.3

Gradient_accumulation_steps = 8

Epochs = 10

Learning_rate = 1e-6

Dropout:

Encoder – [5, 20, 21] = 0.2, [6, 7, 8, 21, 22, 28, 29, 30] = 0.1

Decoder – [1, 2, 3] = 0.2

The model is saved if it achieves improvements in **all three metrics simultaneously**:

Test Loss: Current test loss is lower than the best recorded loss.

WER: Current Word Error Rate is lower than the best recorded WER.

CER: Current Character Error Rate is lower than the best recorded CER.

6.3. Whisper-large-v3-turbo_eng

To improve the performance of our transcription model, we will fine-tune Whisper-large-v3-turbo using the classification dataset presented at 4.2. .

WER/CER on test set before fine-tuning:

WER - **10,32**%

CER – **3,34**%

6.3.1. Fine-tuning Configuration

Trainable layers = All

Batch_size = 2 for both dataloaders

Optimizer = AdamW

Scheduler = linear with 0.1 warmup

Weight_decay = 0.2

Gradient_accumulation_steps = 8

Epochs = 10

Learning_rate = 4e-6

Dropout:

Encoder – [20] = 0.2, [21, 22, 29, 30] = 0.1

Decoder – [1] = 0.2

The model is saved if it achieves improvements in all three metrics simultaneously:

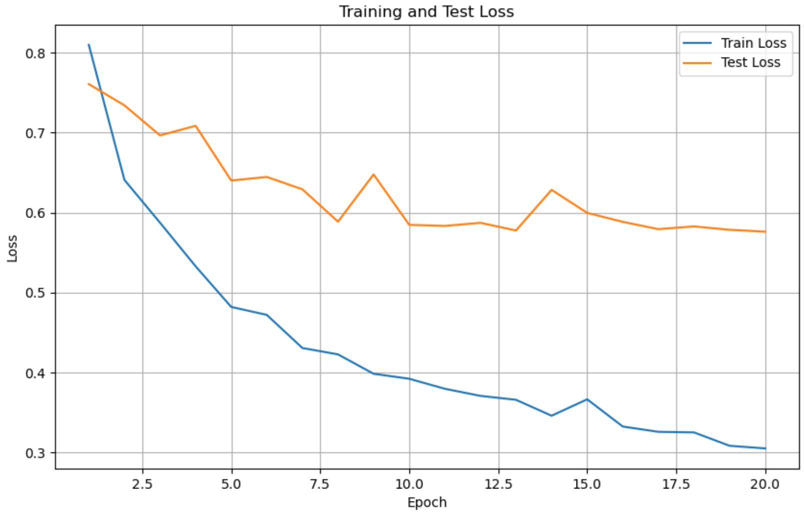
Test Loss: Current test loss is lower than the best recorded loss.

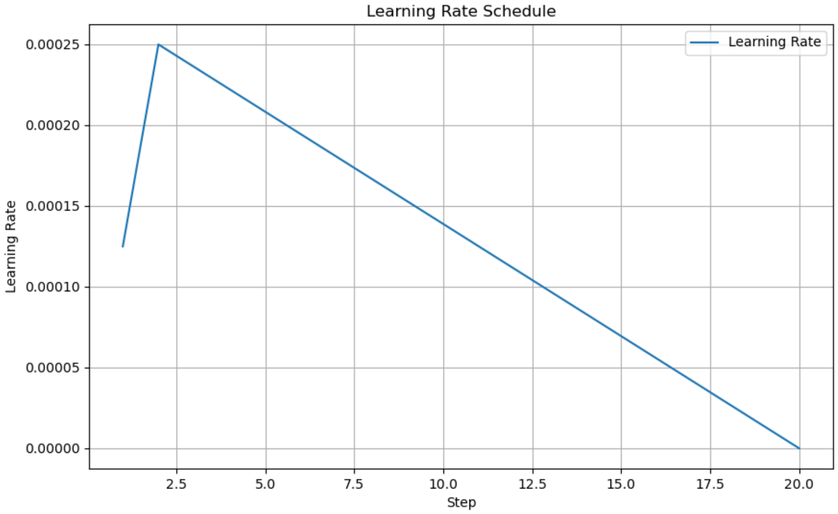
WER: Current Word Error Rate is lower than the best recorded WER.

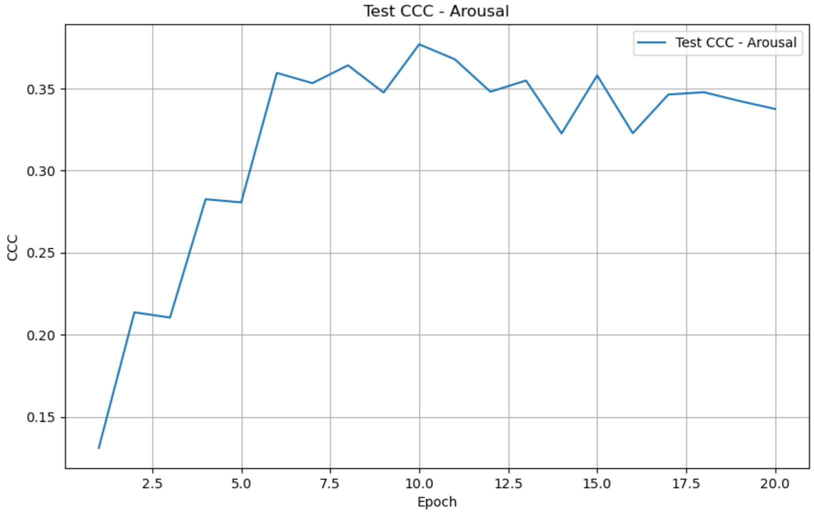
CER: Current Character Error Rate is lower than the best recorded CER.

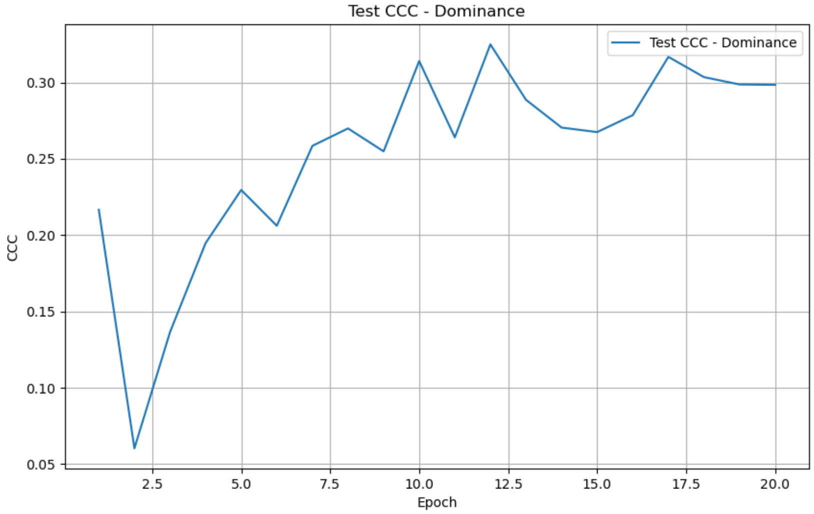
7. Results

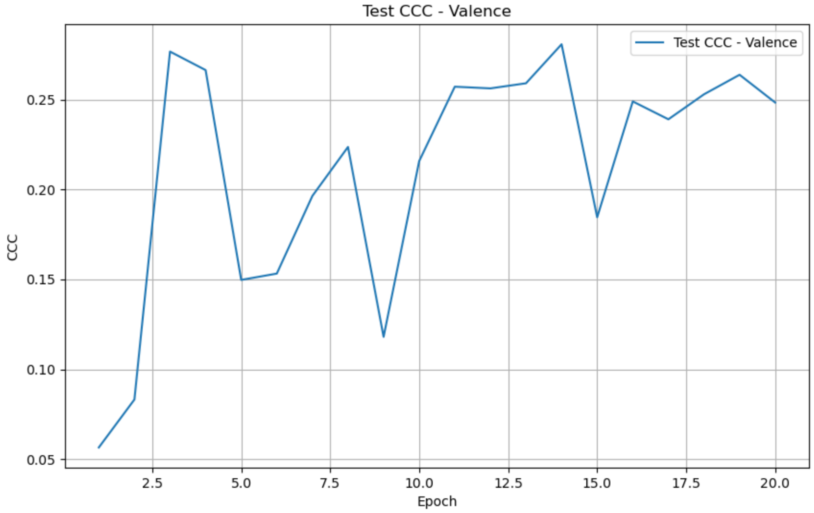
7.1. Regression model






Test CCC - Arousal
Test CCC - Arousal
CCC
0.15
0.20
0.25
0.30
0.35
Epoch
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Test CCC - Dominance
Test CCC - Dominance
CCC
0.05
0.10
0.15
0.20
0.25
0.30
Epoch
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

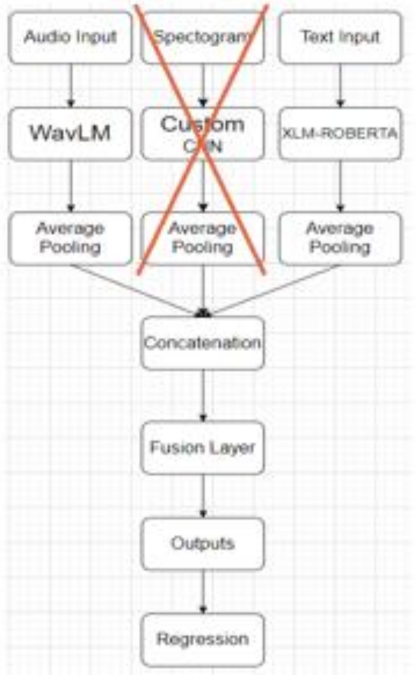


Best CCC for Valence: 0.2809 at Epoch 14

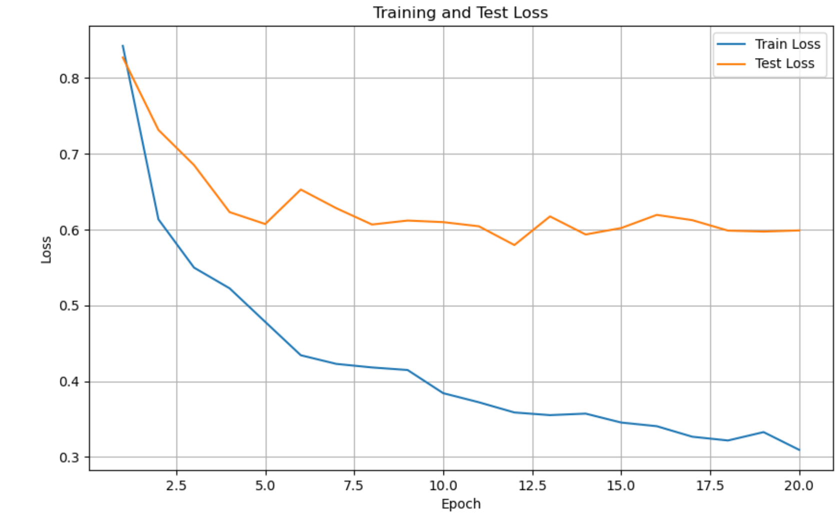
Best CCC for Arousal: 0.3769 at Epoch 10

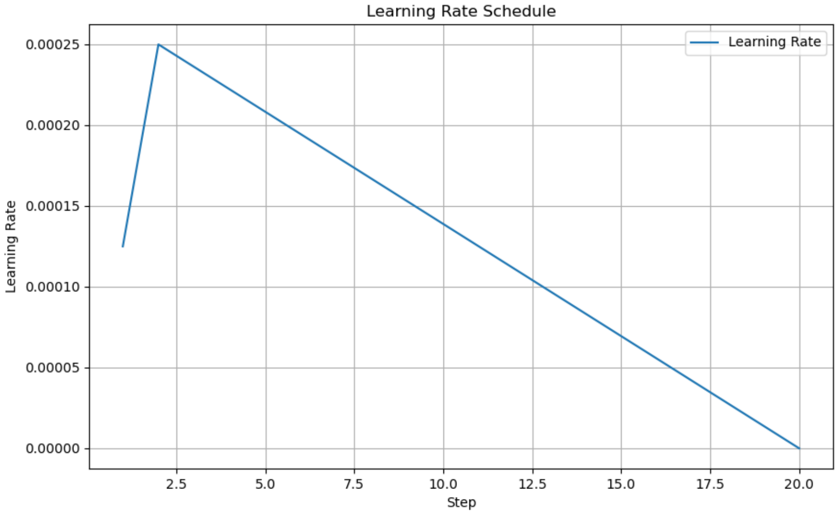
Best CCC for Dominance: 0.3249 at Epoch 12

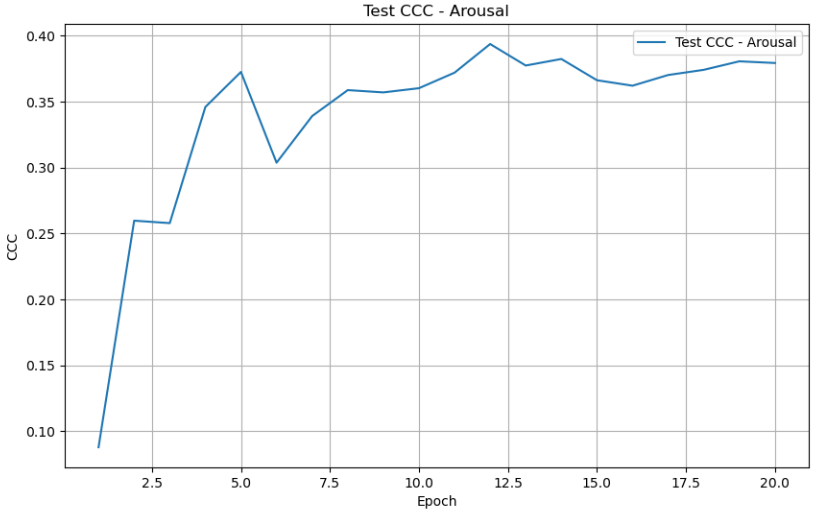
For comparison, using the same configuration but relying on audio features extracted with WavLM and text features extracted with Roberta, as illustrated in the following figure:

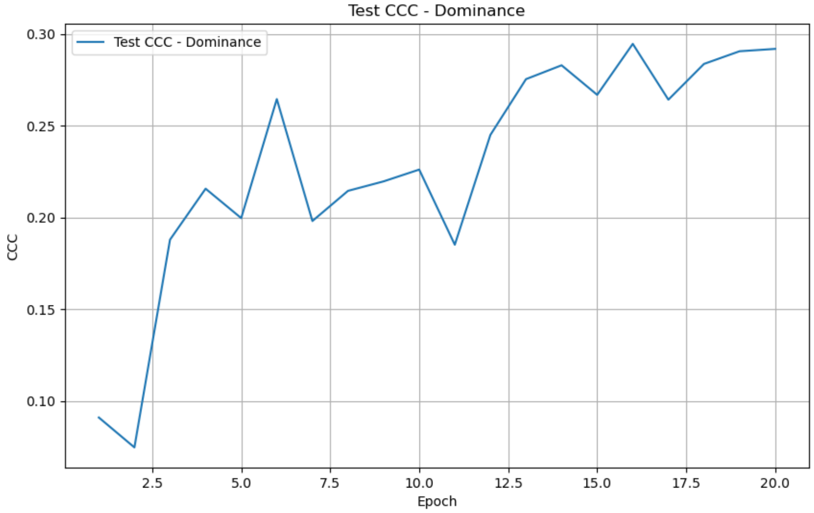


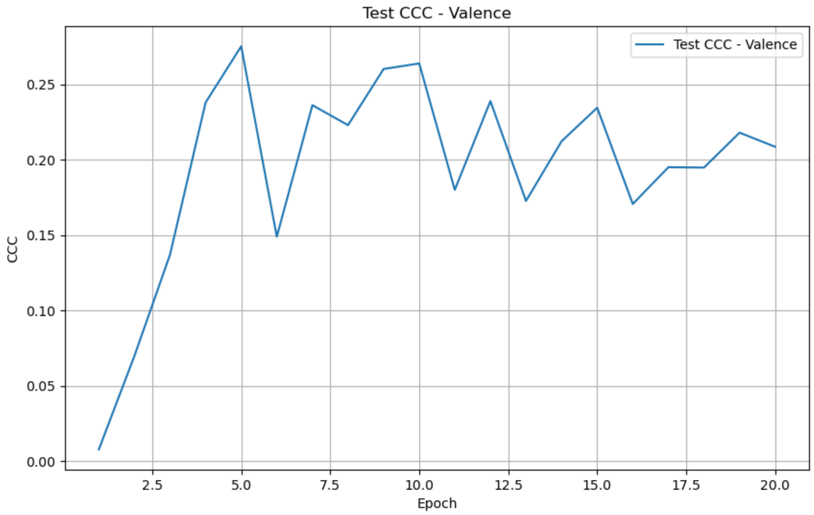
We get the following results:






Test CCC - Arousal
Test CCC - Arousal
CCC
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Epoch
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0


Test CCC - Dominance
Test CCC - Dominance
CCC
0.10
0.15
0.20
0.25
0.30
Epoch
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0

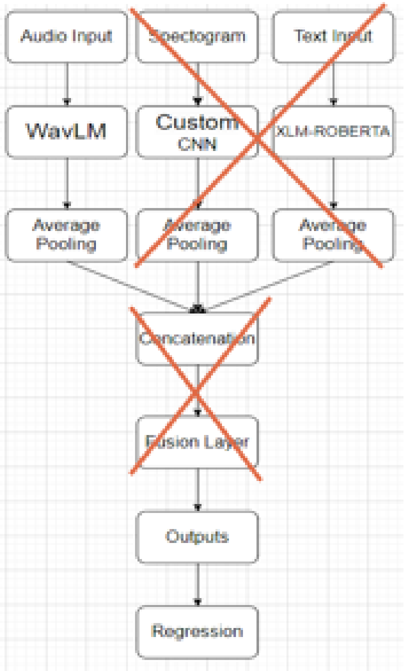


Best CCC for Valence: 0.2754 at Epoch 5

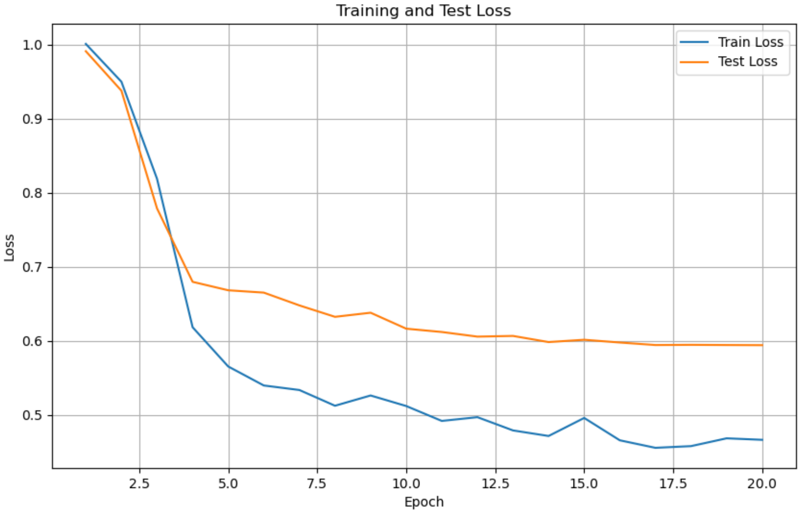
Best CCC for Arousal: 0.3937 at Epoch 12

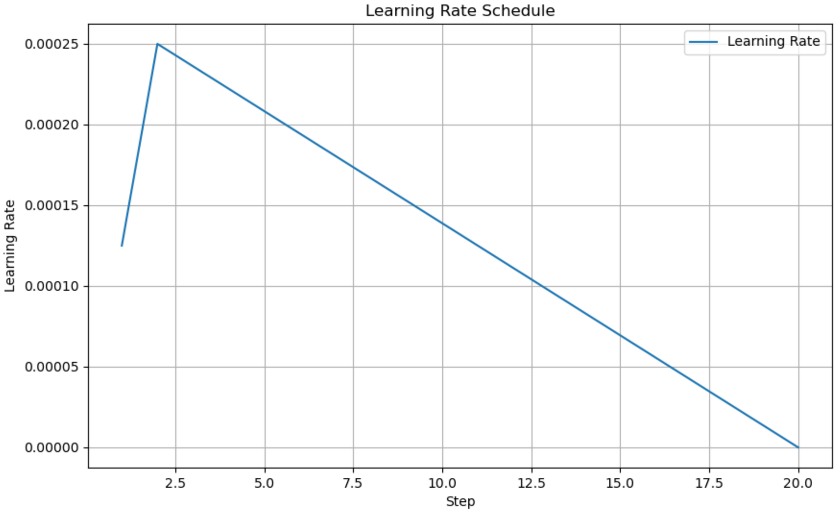
Best CCC for Dominance: 0.2947 at Epoch 16

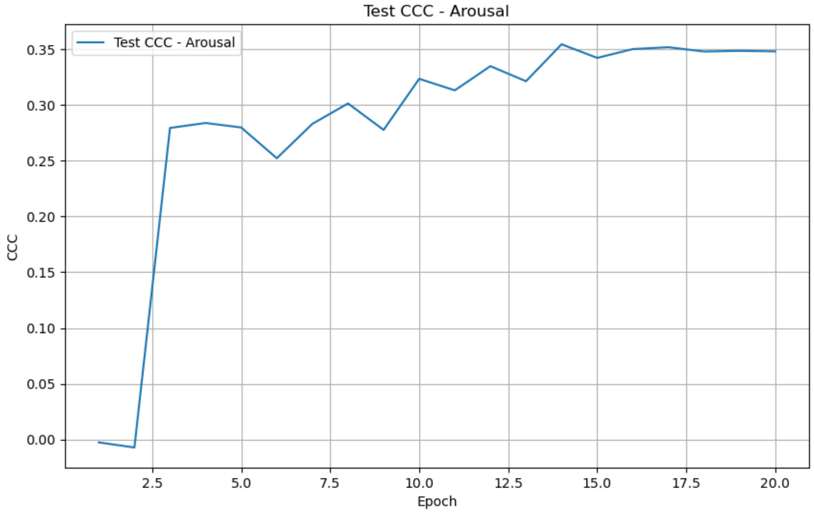
For a better comparison, using the same configuration but relying on audio features extracted with WavLM, as illustrated in the following figure:

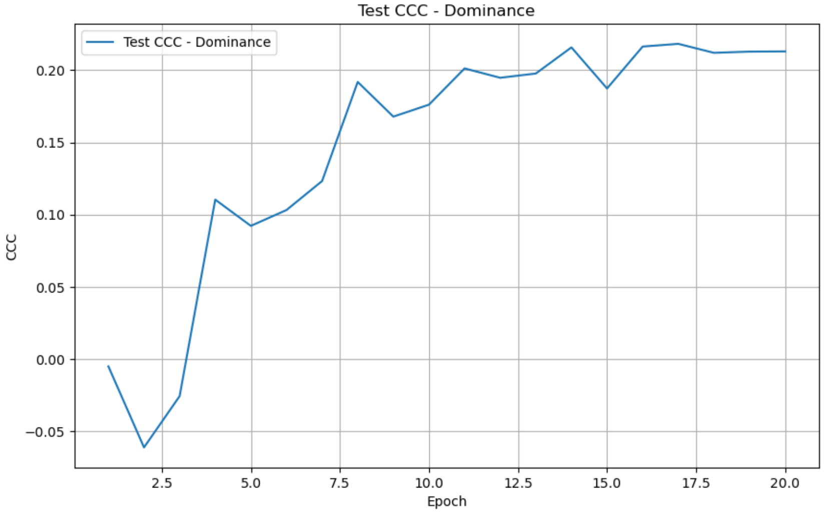


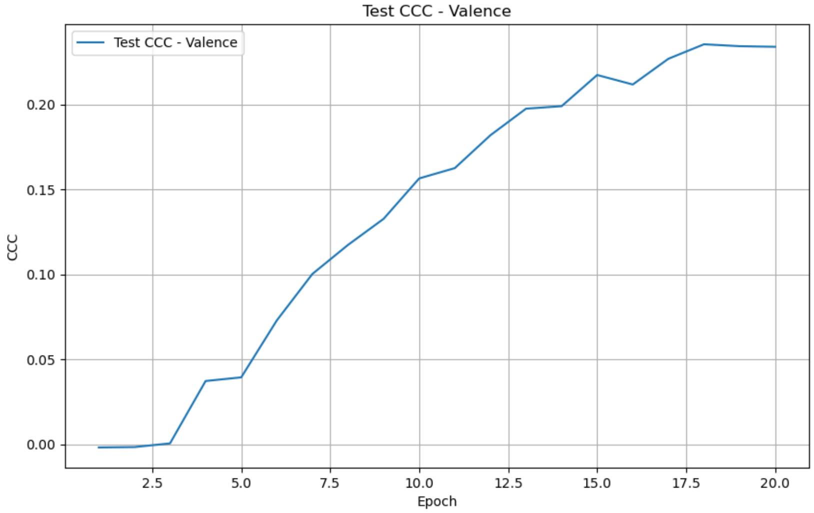
We get the following results:






Test CCC - Arousal
Test CCC - Arousal
CCC
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epoch


Test CCC - Dominance
Test CCC - Dominance
CCC
0.20
0.15
0.10
0.05
0.00
−0.05
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
Epoch

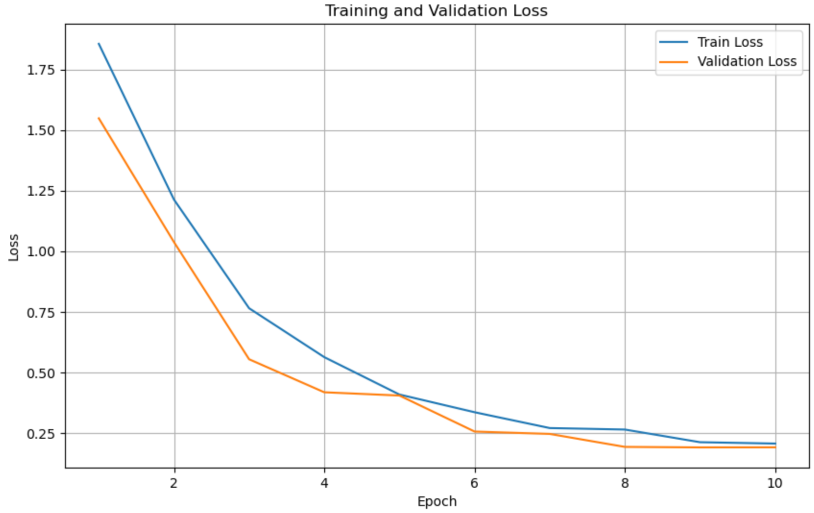


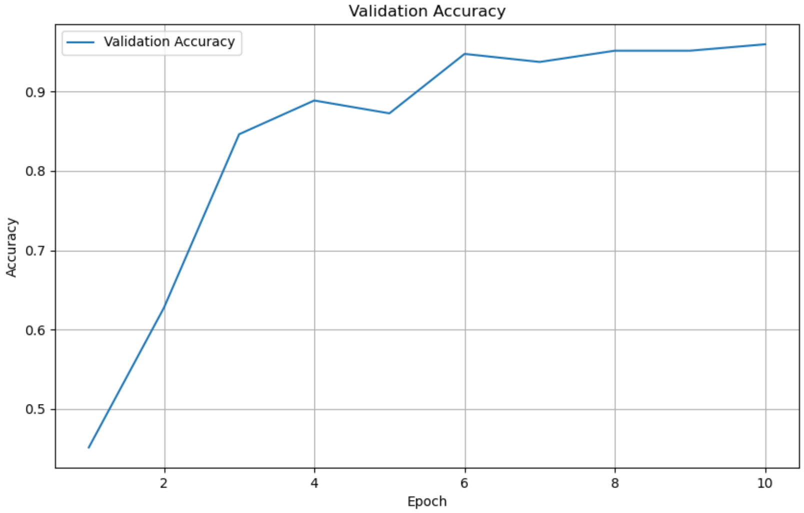
Best CCC for Valence: 0.2354 at Epoch 18

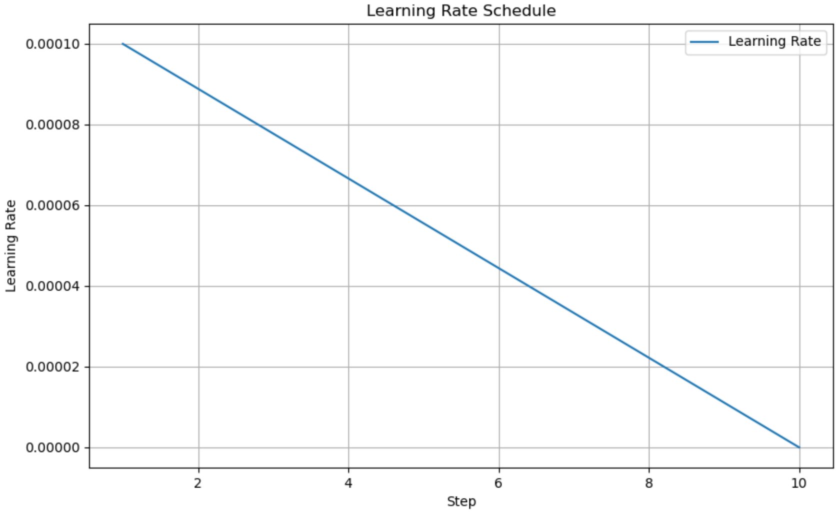
Best CCC for Arousal: 0.3544 at Epoch 14

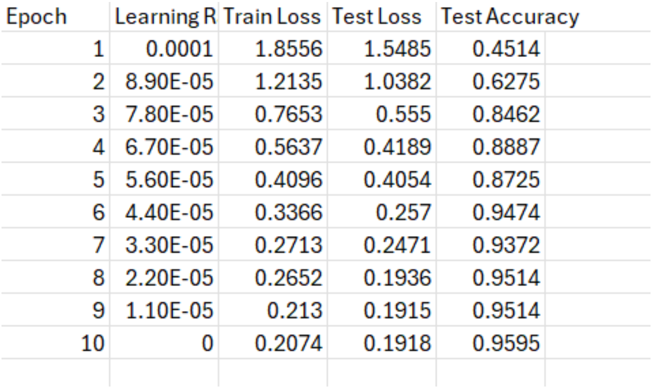
Best CCC for Dominance: 0.2183 at Epoch 17

7.2. Classification model





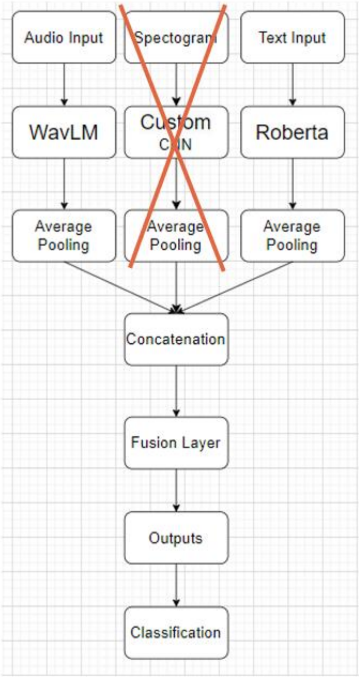




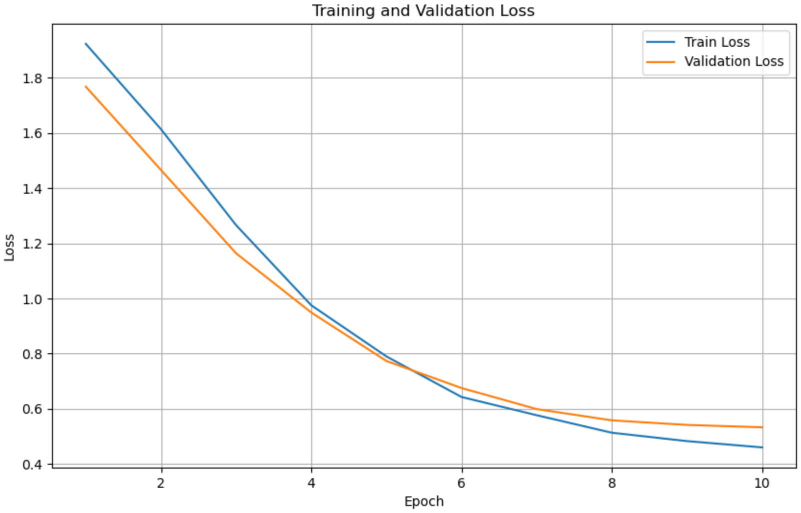
Epoch	Learning R	Train Loss	Test Loss	Test Accuracy
1	0.0001	1.8556	1.5485	0.4514
2	8.90E-05	1.2135	1.0382	0.6275
3	7.80E-05	0.7653	0.555	0.8462
4	6.70E-05	0.5637	0.4189	0.8887
5	5.60E-05	0.4096	0.4054	0.8725
6	4.40E-05	0.3366	0.257	0.9474
7	3.30E-05	0.2713	0.2471	0.9372
8	2.20E-05	0.2652	0.1936	0.9514
9	1.10E-05	0.213	0.1915	0.9514
10	0	0.2074	0.1918	0.9595

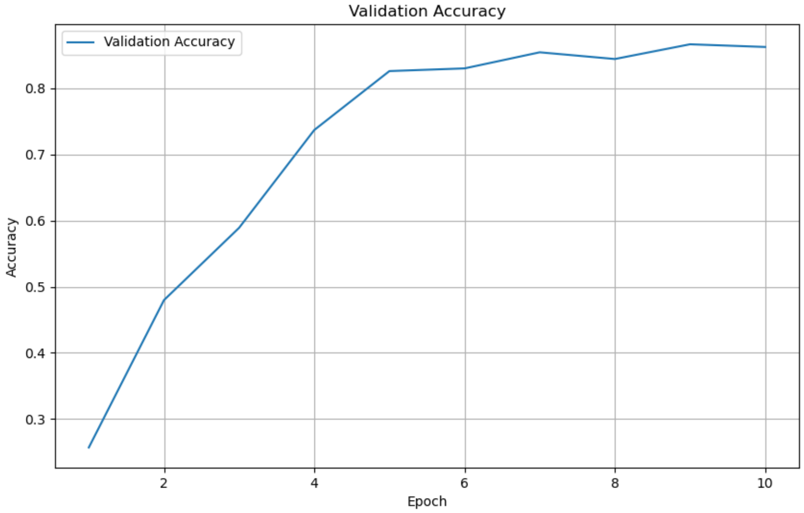
The model is saved at epoch 10 with an accuracy of 95.95%.

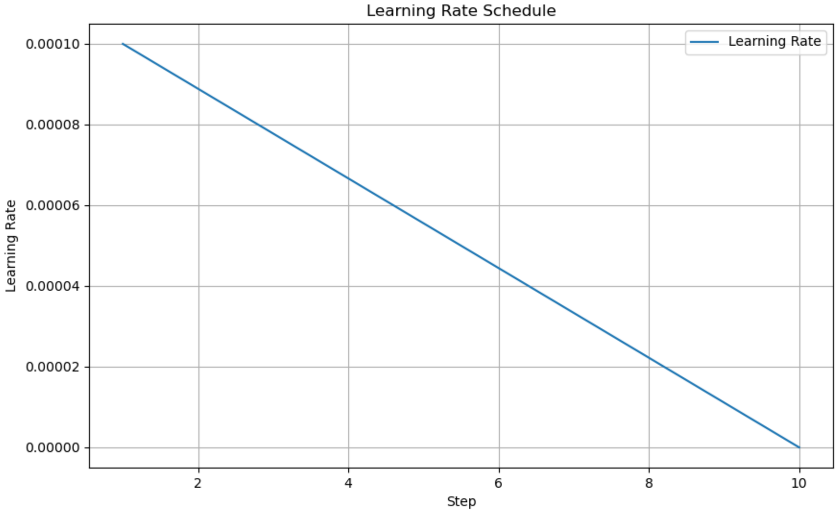
For comparison, using the same configuration but relying on audio features extracted with WavLM and text features extracted with Roberta, as illustrated in the following figure:

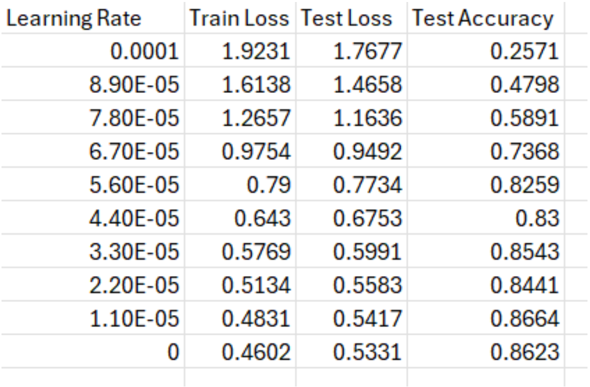


We get the following results:





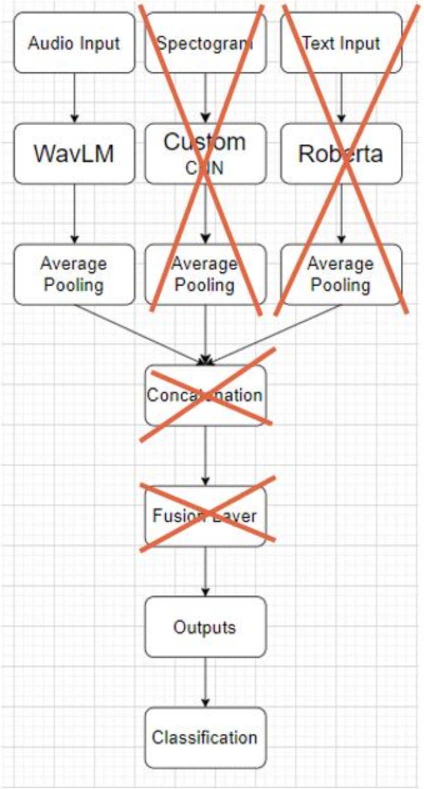




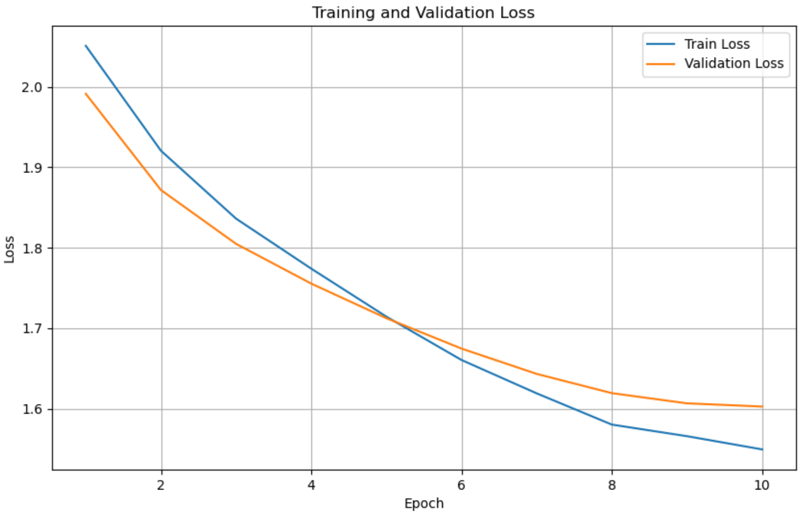
Learning Rate	Train Loss	Test Loss	Test Accuracy
0.0001	1.9231	1.7677	0.2571
8.90E-05	1.6138	1.4658	0.4798
7.80E-05	1.2657	1.1636	0.5891
6.70E-05	0.9754	0.9492	0.7368
5.60E-05	0.79	0.7734	0.8259
4.40E-05	0.643	0.6753	0.83
3.30E-05	0.5769	0.5991	0.8543
2.20E-05	0.5134	0.5583	0.8441
1.10E-05	0.4831	0.5417	0.8664
0	0.4602	0.5331	0.8623

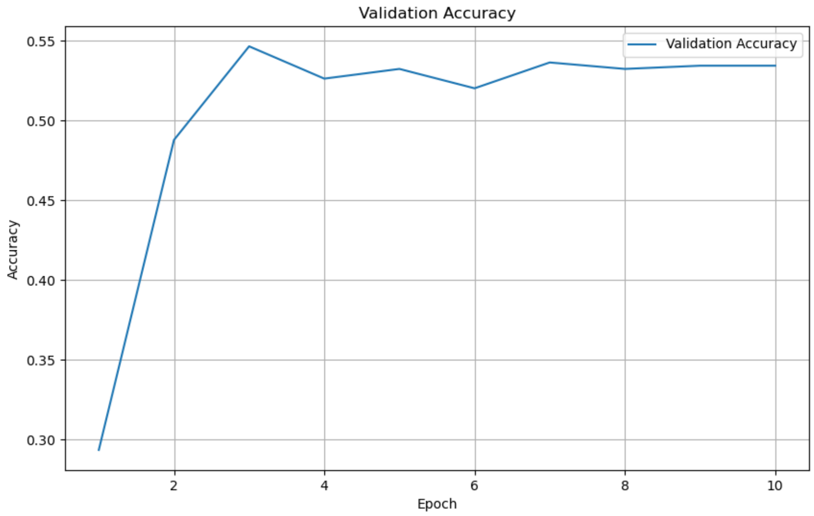
The model is saved at epoch 9 with an accuracy of 86.64%.

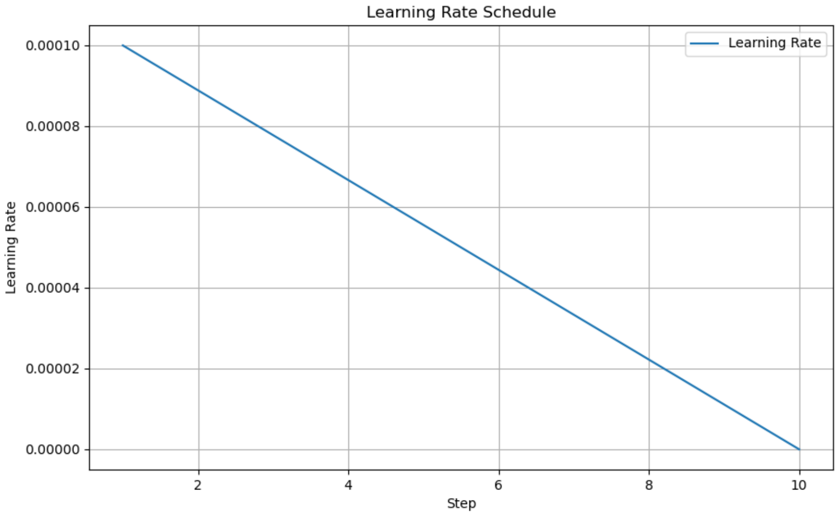
For a better comparison, using the same configuration but relying on audio features extracted with WavLM, as illustrated in the following figure:

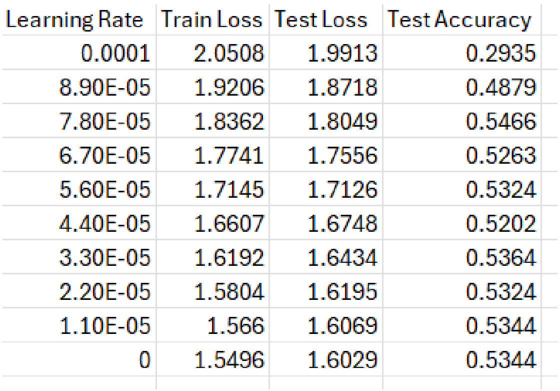


We get the following results:



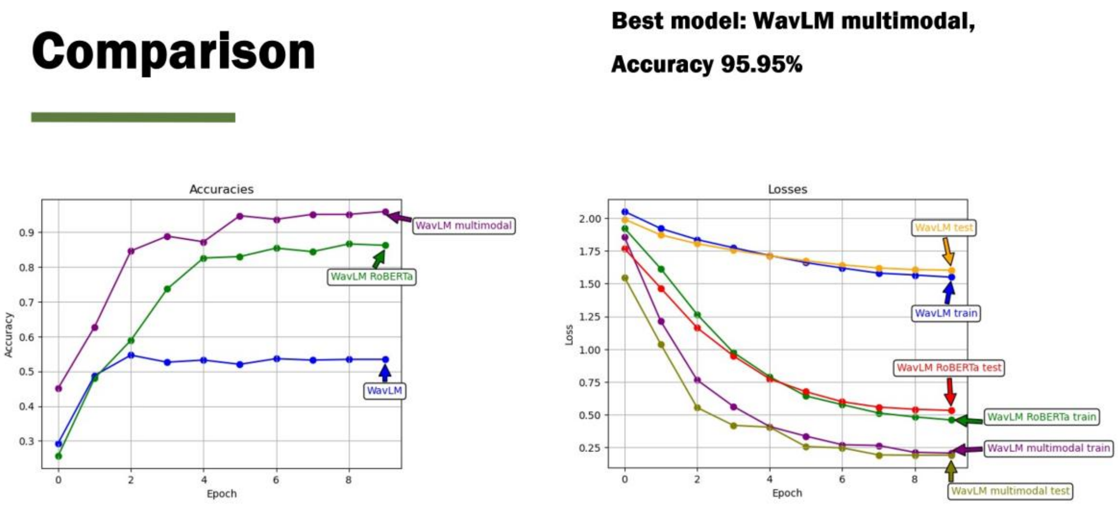






Learning Rate	Train Loss	Test Loss	Test Accuracy
0.0001	2.0508	1.9913	0.2935
8.90E-05	1.9206	1.8718	0.4879
7.80E-05	1.8362	1.8049	0.5466
6.70E-05	1.7741	1.7556	0.5263
5.60E-05	1.7145	1.7126	0.5324
4.40E-05	1.6607	1.6748	0.5202
3.30E-05	1.6192	1.6434	0.5364
2.20E-05	1.5804	1.6195	0.5324
1.10E-05	1.566	1.6069	0.5344
0	1.5496	1.6029	0.5344

The model is saved at epoch 3 with an accuracy of 54.66%.

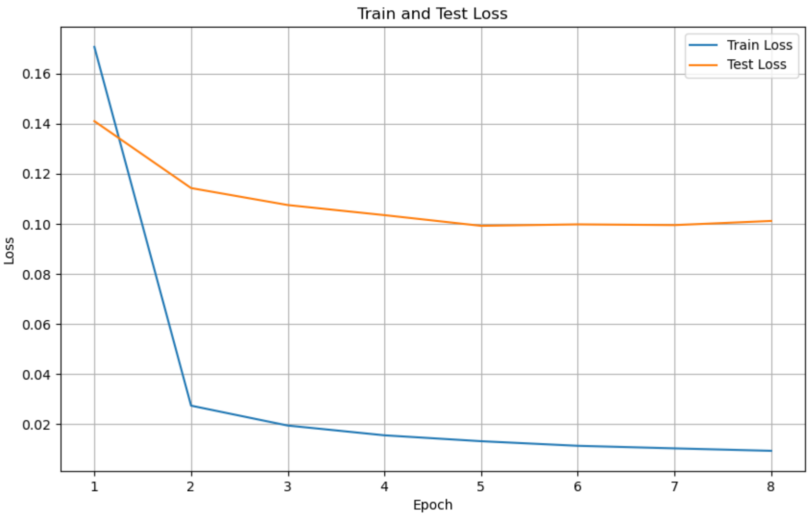


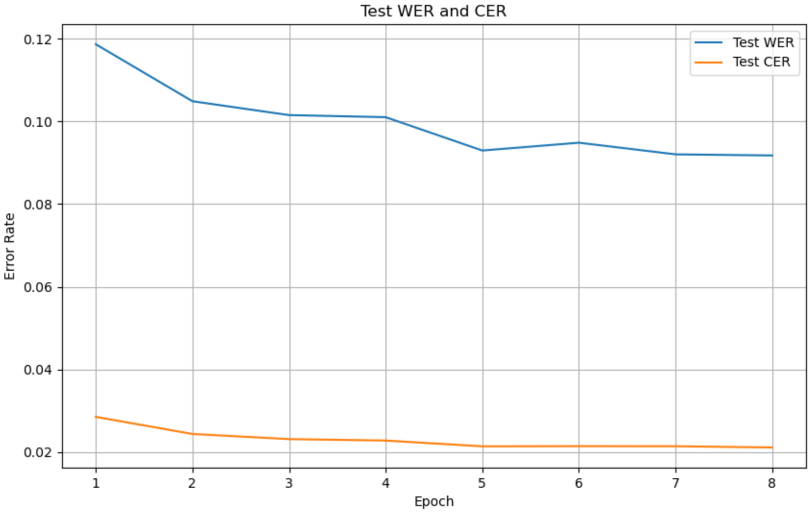
Comparison

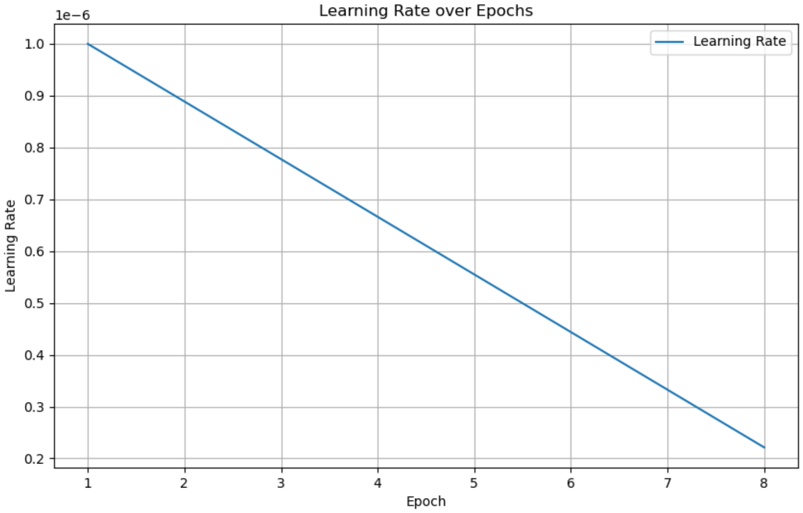
Best model: WavLM multimodal,

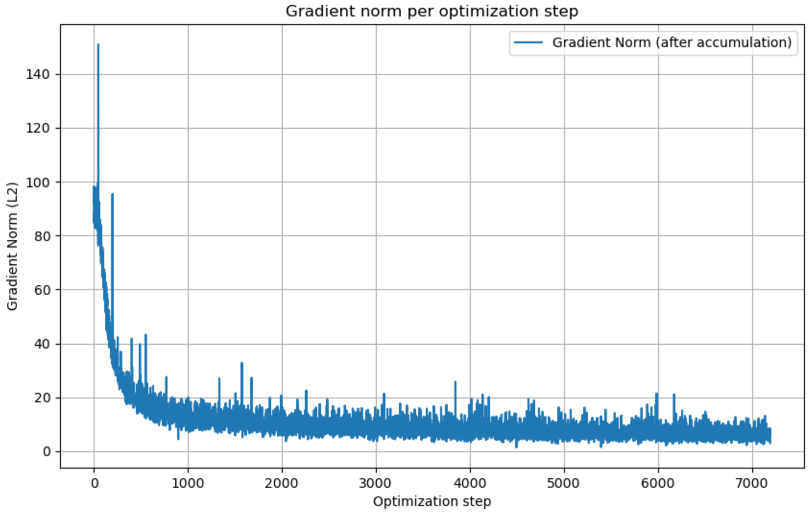
Accuracy 95.95%

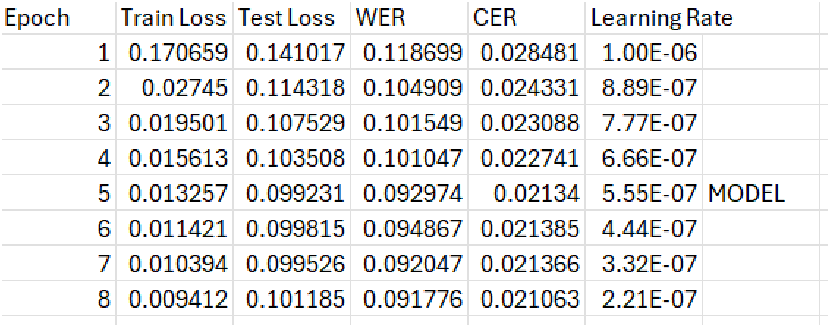
7.3. Whisper-large-v3-turbo_ro fine-tuning




Test WER and CER
Test WER
Test CER
Error Rate
0.12
0.10
0.08
0.06
0.04
0.02
Epoch
1
2
3
4
5
6
7
8


Learning Rate over Epochs
1e−6
Learning Rate
Learning Rate
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
Epoch
1
2
3
4
5
6
7
8

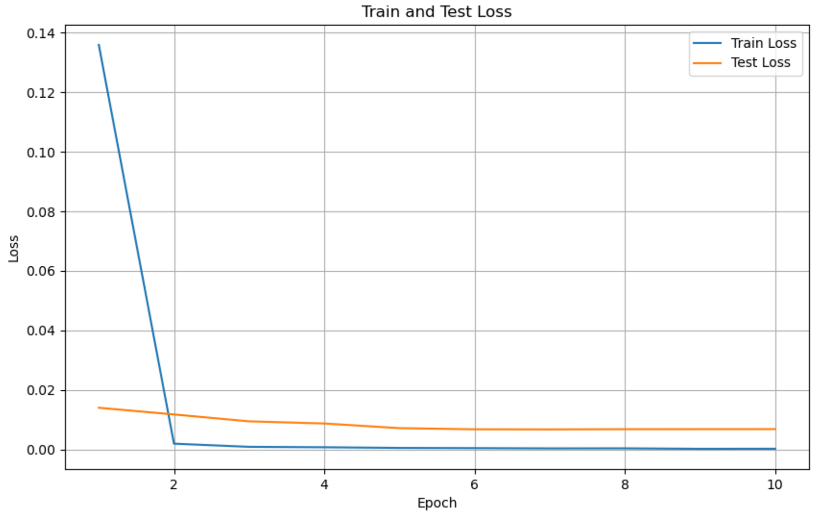


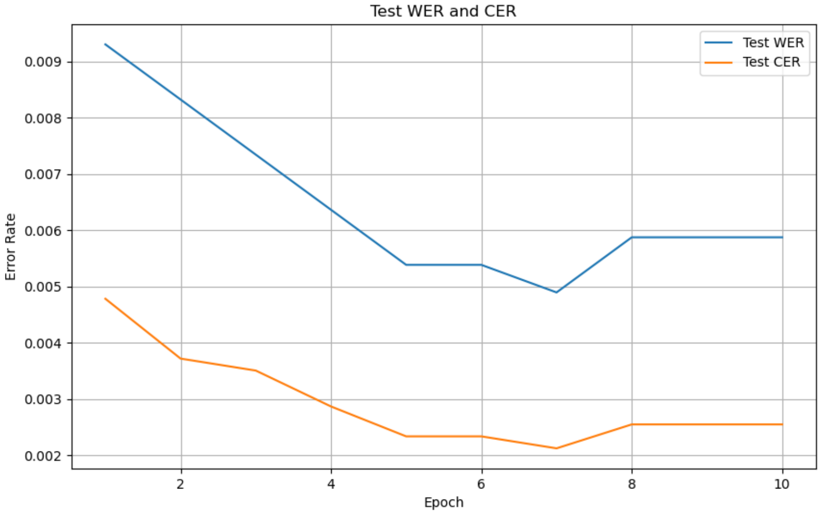


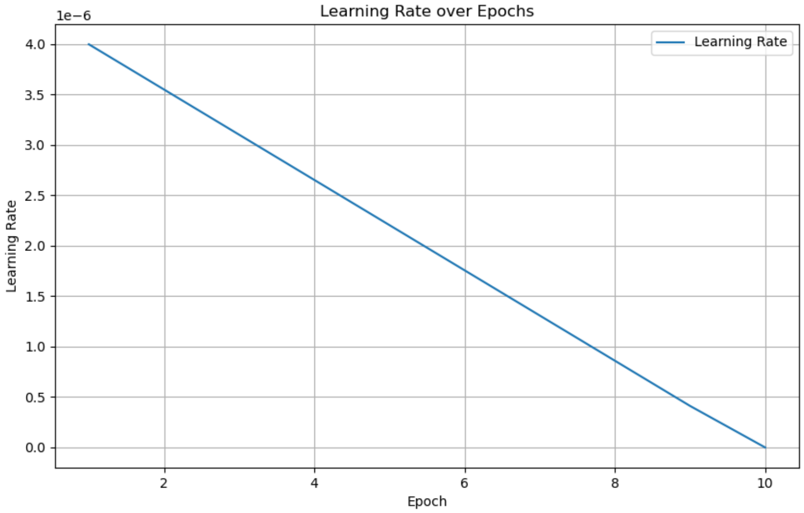
Epoch	Train Loss	Test Loss	WER	CER	Learning Rate	
1	0.170659	0.141017	0.118699	0.028481	1.00E-06	
2	0.02745	0.114318	0.104909	0.024331	8.89E-07	
3	0.019501	0.107529	0.101549	0.023088	7.77E-07	
4	0.015613	0.103508	0.101047	0.022741	6.66E-07	
5	0.013257	0.099231	0.092974	0.02134	5.55E-07	MODEL
6	0.011421	0.099815	0.094867	0.021385	4.44E-07	
7	0.010394	0.099526	0.092047	0.021366	3.32E-07	
8	0.009412	0.101185	0.091776	0.021063	2.21E-07	

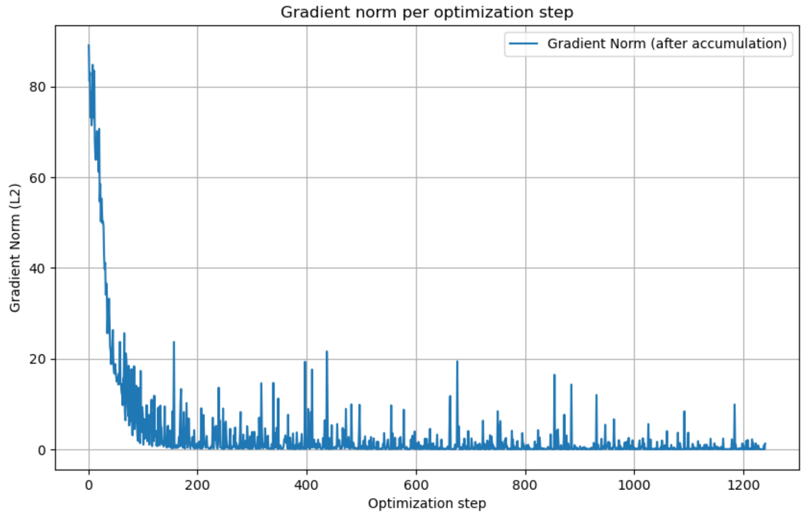
	WER	**CER**
Before	**17,12%**	**4,88%**
After	**9,29%**	**2,13%**

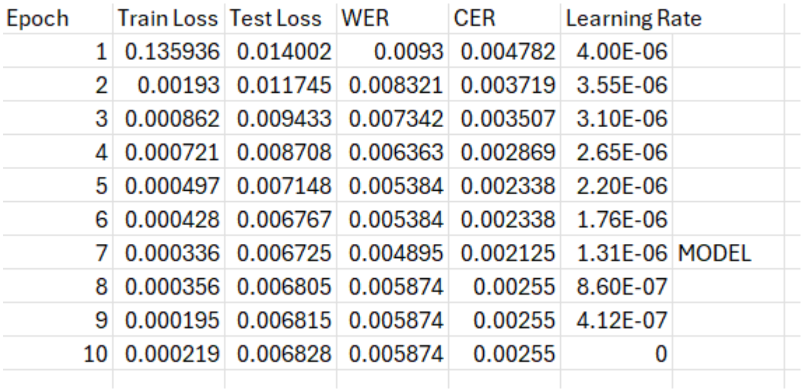
7.4. Whisper-large-v3-turbo_eng fine-tuning











Epoch	Train Loss	Test Loss	WER	CER	Learning Rate	
1	0.135936	0.014002	0.0093	0.004782	4.00E-06	
2	0.00193	0.011745	0.008321	0.003719	3.55E-06	
3	0.000862	0.009433	0.007342	0.003507	3.10E-06	
4	0.000721	0.008708	0.006363	0.002869	2.65E-06	
5	0.000497	0.007148	0.005384	0.002338	2.20E-06	
6	0.000428	0.006767	0.005384	0.002338	1.76E-06	
7	0.000336	0.006725	0.004895	0.002125	1.31E-06	MODEL
8	0.000356	0.006805	0.005874	0.00255	8.60E-07	
9	0.000195	0.006815	0.005874	0.00255	4.12E-07	
10	0.000219	0.006828	0.005874	0.00255	0	

	WER	CER
Before	**10.32%**	**3.34%**
After	**0.48%**	**0.21%**

8. Conclusion

The complete implementation and integration of the multimodal system for Speech Emotion Recognition (SER) and automated Speech-to-Text transcription have demonstrated outstanding results. The proposed model was evaluated and optimized, achieving competitive performance:

1. **Speech Emotion Recognition (SER):**
   - The system achieved good results in bot classification and regression, with stable convergence between training and validation losses.
   - The proposed architecture, which integrates WavLM for audio, Roberta for text, and CNN for spectrograms, proved highly effective in capturing diverse modalities of data.
2. **Speech-to-Text Transcription:**
   - The Whisper-large-v3-turbo model, fine-tuned for the transcription task, achieved remarkable metrics.
   - Fine-tuning resulted in a significant improvement over the initial values demonstrating the effectiveness of adapting the model to the datasets.
3. **Multimodal Integration:**
   - The integration between SER and transcription makes the system robust and scalable for real-world applications.
   - The system uses Whisper-generated transcriptions during inference, highlighting its capability to operate autonomously in practical scenarios.

In conclusion, the proposed solution represents a significant step toward developing a fully functional speech analysis system, showcasing state-of-the-art performance for both emotion recognition and speech transcription. This system can be applied to various domains, such as virtual assistants, conversational analysis, or emotional monitoring, serving as a successful example of combining cutting-edge models for practical solutions.