Sentiment Analysis from Speech in the Albanian Language: A Deep Learning Approach for Low-Resource Scenarios

Ilma Lili [1], Mirvjen Ulqinaku [1], Florijan Qosja [2], Alex Thomo [3], Ana Ktona [1], Arber Ceni [1], Anxhela Kosta [1]

1.University of Tirana, Faculty of Natural Sciences, Department of Informatics, Bulevardi " Zogu I ",Tiranë, Albania

2. Senior Software Engineer at Level Financial Technology, London, Great Britain

3.University of Victoria, Faculty of Engineering and Computer Science, Department of Computer Science, PO Box 3055, STN CSC, Victoria, BC, Canada

***Abstract* –**Understanding emotions and sentiment in speech is becoming increasingly important in fields such as education, mental health, customer support, and online safety. Although there have been significant advancements in speech analysis for the world's most widely spoken languages, the Albanian language has received very little attention in this regard. In this paper, we present a novel approach to recognizing sentiment directly from spoken Albanian. We created a small dataset of audio recordings, each labeled as positive, neutral, or negative, and used this data to train and test deep learning models. Our experiments explored both traditional and modern methods: starting with feature extraction using Mel Frequency Cepstral Coefficients (MFCCs), and extending to self-supervised models such as HuBERT, wav2vec 2.0 and it's improved version WavLM, These features were then used to train lightweight neural classifiers, including a two-layer feed-forward network and a bidirectional LSTM with attention, for sentiment classification. Despite the limited linguistic resources for the Albanian language, the results are encouraging, indicating that these models can effectively detect sentiment in speech. By making our dataset and model code publicly available, we aim to support and encourage future research and help expand sentiment and emotion analysis tools to include under-resourced languages such as Albanian

Keywords: *Speech sentiment analysis, Albanian language, Low resource speech, MFCC, Self Supervised speech representations.*

I. INTRODUCTION

Speech carries powerful para-linguistic cues intonation, pitch, energy, and timbre, that reveal emotion beyond words. These acoustic signals often convey sentiment more effectively than text transcripts, especially when transcription is imperfect or unavailable (Zhao et al., 2022). Recognizing emotion from speech enables empathetic and natural interactions with machines. This is essential in applications like virtual assistants, educational platforms, healthcare, and customer service, where responding to emotional nuance can significantly improve user experience (Atmaja and Sasou, 2022). In scenarios where speech is the only modality (e.g., phone calls), sentiment detection directly from audio empowers real-time feedback systems, accessibility tools, and automated support systems to respond appropriately, without needing text transcription (Atmaja and Sasou, 2022). Fusing speech emotion data with text sentiment leads to more nuanced understanding and higher accuracy. Multimodal systems combining acoustic features with NLP models show significantly better performance in recognizing emotional states (Resende Faria et al., 2024). Self-supervised learning techniques—such as HuBERT, wav2vec, or UniSpeech-SAT—demonstrate strong capability in capturing emotion and sentiment cues directly from speech, even in the absence of text (Atmaja and Sasou, 2022). Speech emotion recognition (SER) has emerged as a crucial technology with significant applications across multiple domains. In healthcare, SER enables early detection of mental health issues and monitoring of psychological well-being, particularly important given increased mental health problems

during the COVID-19 pandemic (Elsayed et al., 2022). The technology enhances human-computer interaction by creating more empathetic AI systems and improving Intelligent Virtual Personal Assistant services (Mangalmurti et al., 2024) In educational contexts, emotion recognition from speech plays a vital role in understanding student behavior and engagement. Educators can leverage sentiment analysis to assess students' emotional states during classroom activities, helping them tailor approaches to maximize learning performance and improve educational outcomes (Anjali Verma and Bappaditya Jnna, 2025) (Kerkeni et al., 2017). Since emotions significantly influence cognitive processes and learning, SER systems enable teachers to better support student well-being and academic success (Kerkeni et al., 2017). These applications demonstrate SER's importance in creating more inclusive, emotionally intelligent digital environments across healthcare, education, and technology sectors.

II. Albanian as a Low-resource Language

The Albanian language has very little annotated text for tasks such as Sentiment Analysis, NER (named-entity recognition), classification, etc. The rate of sentiment works is relatively low compared to major languages (Kastrati and Biba, 2022) (Kadriu et al., 2022). For identifying a low resource language as Albania by using Artificial Neural Networks (ANN) and Convolutional Neural Networks (CNN) both models have shown good capabilities to learn Albania Language patters (Binjaku et al., 2022). Cases through different scientific articles presents better the situation as follows:

- There are very few NLP tools such as spaCY, Stanza, HuggingFace models that are mainly for other languages and as far as the Albanian language is concerned, researchers have to build tokenizers, lemmatizers, and pipelines manually (Shehu et al., 2025).
- Compared to English benchmark Albanian has very few labeled resources NER(Named Entity Recognition) dataset (AlbNER) with ~900 annotated sentences (Çano, 2023a)
- Sentiment dataset (AlbMoRe) with 800 movie reviews (Çano, 2023b)
- Topic classification dataset (AlbNews) with 600 labeled titles (Çano and Lamaj, 2024)
- Albanian has two major dialects, Gheg and Tosk and the lack of balanced representation makes it difficult to build a robust model

III. Categories of Acoustic and Learned Features in Speech Emotion Recognition

Speech Emotion Recognition (SER) relies on the extraction of informative features that capture both the physical properties of speech signals and the high-level semantic representations learning by deep models. Broadly, these features can be categorized into:

1. Prosodic features like Pitch(F_0), intensity/loudness, speech rate, pause, tone contour are important elements for identifying emotions, mainly pitch and intensity, along with LPC and MFCC (Mel-Frequency Cepstral Coefficients) (Madanian et al., 2023).
2. Spectral Cepstral Features here we mention MFCCs, Spectrograms, Chroma, Spectral contrast, Tonnetz and represent timbre, frequency structure and emotion-laden acoustic patterns and in one study these are listed among key features extracted for speech emotion recognition (Rezapour Mashhadi and Osei-Bonsu, 2023). While in another study the combination of MFCC, Chroma, Mel-spectrogram, Tonnetz, and spectral contrast enhances the accuracy of emotion classification by 93% on RAVDESS. (Gondohanindijo et al., 2023).
3. Perceptual (Psychoacoustic) features related to what is perceived by humans. It is based on the ITU PEAQ (international telecommunication unit perceptual evaluation of Audio Quality) standards where features such as partial loudness, emotional difference-to-mask ratio, envelope variation, harmonic content, and temporal occurrence of emotional blocks serve to identify this perception by humans (Sezgin et al., 2012).
4. Multi-Feature Fusion Approaches which involves combining several types of features to achieve the best possible performance. Studies specify that the use of features of different types improves emotion classification by around 95% using Bayesian Regularized ANN(Gondohanindijo et al., 2023). Another study using the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech

and Song) dataset operates on preprocessed audio, where framing, silence removal etc. may have been done to create an emotion recognition dataset (Colunga-Rodriguez et al., 2025).

5. Deep Audio Embeddings, which are modern approaches that leverage what is learned from neural networks. Models like L3-Net and VGGish, can capture high-level emotional semantics in music without handcrafted features (Koh and Dubnov, 2021).
6. Studime te tjera perdorin ASR (Automatic Speech Recognition) model embeddings for emotion recognition and sometimes outperform classical feature sets such as eGeMAPS (extended Geneva Minimalistic Acoustic Parameter Set, which is a standard set of acoustic features for performing speech and emotion analysis) (Tits et al., 2018).

IV. FEATURE EXTRACTION METHODOLODY FOR EMOTION RECOGNITION FROM AUDIO

To analyze an audio, there are several important elements such as amplitude or energy that identifies how high the sound is, zero-crossing rate (ZCR) that means how often the signal crosses zero to identify voiced vs unvoiced sounds, and Short-time Energy that measures the energy in short fragments and helps in detecting speech activity. These features are extracted directly from the waveform. Normally, the focus is on features related to emotion detection. In fact, features are numerical representations of an audio signal. They serve as an input for the Machine Learning or Deep Learning model. Most well known accoustic feature are:

- Tone:A medium or high tone reflects emotions such as anger, excitement, etc. while a lower tone suggests sadness, boredom, etc.(Kaloga and Kodrasi, 2025).
- Energy: identifies the intensity of emotions such as sadness, joy that have high energy levels
- Spectral feature: Mel-spectograms, MFCCs, spectral centroid, flux provide timbral and frequency content cues (Venkataramanan and Rajamohan, 2019).
- Prosodic feature: Duration, speech rate, pauses, intonation patterns help distinguish emotional states. The importance lies in the context "how it was said and not what was said"
- Voice quality metrics: mention is made for the vibration of the voice, breathing or even other elements that add emotional meaning to the voice with different nuances.

For feature extraction there are several traditional and modern methods. The methods chosen in this article are MFFC (Mel-frequency Cepstral Coefficients) as a traditional method and HuBERT (Hudden-Bnit ERT), wav2vec 2.0 wavLM as modern methods. HuBERT is a Self-Supervised Speech Model (SSM) that extracts embeddings from audio. In this case, the audio segment is given to the model and a vector with fixed dimensions (768-1024) is extracted. With this method, complex features are extracted without the need for labeling. Meanwhile, the traditional MFCC method extracts 13-40 coefficients per frame, the average and standard deviation are calculated for each segment.

V. DATASET PREPARATION

A. DATA COLLECTION

The recorded voice is of a female person. The audio is based on the reading of several neutral sentences or even with positive and negative emotional annotations. Three basic classes have been created where these audios are also grouped according to emotions like positive, negative and neutral. For each class, ~300 audio samples were created with time intervals from 3 to 9 sec. To calculate audio duration, the ratio between number of frames and sample rate is used as figure 1.

$$\text{Duration (sec)} = \frac{\text{Number of frames}}{\text{Sample rate}}$$

Fig 1 Audio duration formula

The calculation of descriptive statistics includes the minimum as a linear algorithm that iterates through all values and keeps the smallest one, the maximum that finds the largest value, the median that performs the

ranking of the list (usually O(nlogn)) and selects the middle value dhe total duration per class as a division of the sum of the time of all audios over 3600 to get the total in hours.

	Positive	*Negative*	*Neutral*
Max Duration	10.14 s	9.25 s	11.10 s
Min Duration	2.82 s	3.09 s	2.55 s
Median Duration	4.99 s	5.59 s	6.59 s
Total Duration	0.44 h	0.46 h	0.54 h
Nr of files	295	296	295

Tab 1-Descriptive Statistics for recorded audio

The tool used for recording the audio is a laptop with Microphone Array AMD Audio device. The texts that were registred are content found online randomly. Input setting of the microphone are format 2 Channels, 16 bit, 48000 Hz (DVD Quality). The importance does not lie only in recording the audio but also in creating an excel dataset with csv format (comma separated values). The fields of this format are **path**, as the location of the audios, and **labels**, as a tag for each audio. The preparation of the dataset also includes the process of removing noise, normalization or even the creation of segments and frames. Before extracting features, it is important that the audios are 16 kHz mono as a suitable format to be used in Automatic Speech Recognition and Speech Emotion Recognition datasets .

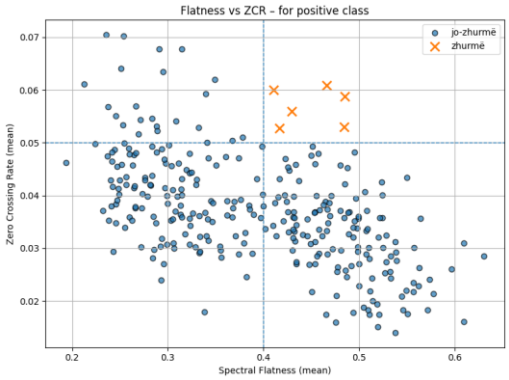

Fig 2 Mean values for S.Flatness and ZCR (Negative)

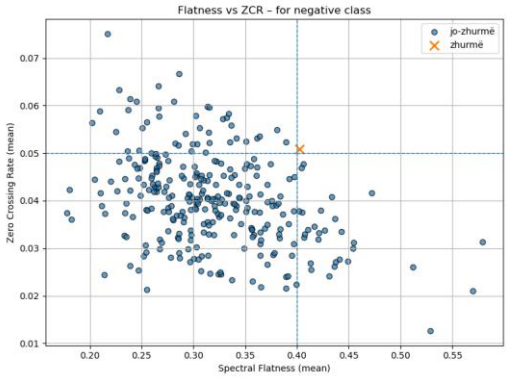

Fig 3 Mean values for S.Flatness and ZCR (Positive)

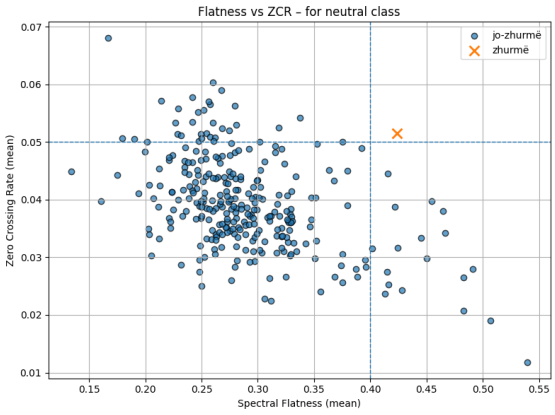

Fig 4 Mean values for S.Flatness and ZCR (Neutral)

For sentiment classification from audio, many acoustic features must be considered. The graphs below present the average of two features such as Spectral Flatness and Zero Crossing Rate based on the data of the created dataset. The first to distinguish tonality from noise while the second distinguishes quiet voice from busy one.

In the negative class according to spectral flatness most of the points are between 0.20-0.50 with moderate structure while ZCR has values between 0.02-0.07 which are typical for normal speech. As can be seen only one point has crossed the threshold and is considered as noise. In conclusion the negative class is relatively clean and there are very few segments that look like noise. In the Positive class spectral flatness varies from 0.25 to over 0.5 which means there is more noise like. Meanwhile ZCR has a wider distribution where some points exceed 0.06. This class has more perceived noise compared to the others or more precisely it has more acoustically blurred elements, so it includes non-clean segments. In the neutral class, most of the spectral flatness points have values around 0.25-0.35 and this is similar to the negative class, while the ZCR has less variation. This class has cleaner audio compared to the negative class. As a result, the class that has more acoustic noise with more segments where flatness is>0.45 and ZCR higher than 0.06 is the positive class. The calculation of ZCR (Zero crossing rate) is performed using a linear O(n) algorithm and its value is higher when there is noise in the audio, while the calculation of spectral flatness is based on the O(nlogn) algorithm. The closer to the value 1, the more "flat" (noise-like) the spectrum; the closer to the value 0, the more "tonal/structured" (speech, music).

The graph in Figure 5 shows the average Zero Crossing Rate (ZCR) for three audio classes: Positive, Negative, and Neutral. X-axis (0–1): normalized time (from the beginning to the end of the audio, regardless of the absolute length). Y-axis (0–0.08): average ZCR, which shows how often the audio signal crosses zero in each unit of time. The three lines (blue, orange, green): average ZCR for the Positive, Negative, and Neutral classes. To interpret the graph, we divide the Y-axis (which is ZCR) into three segments.

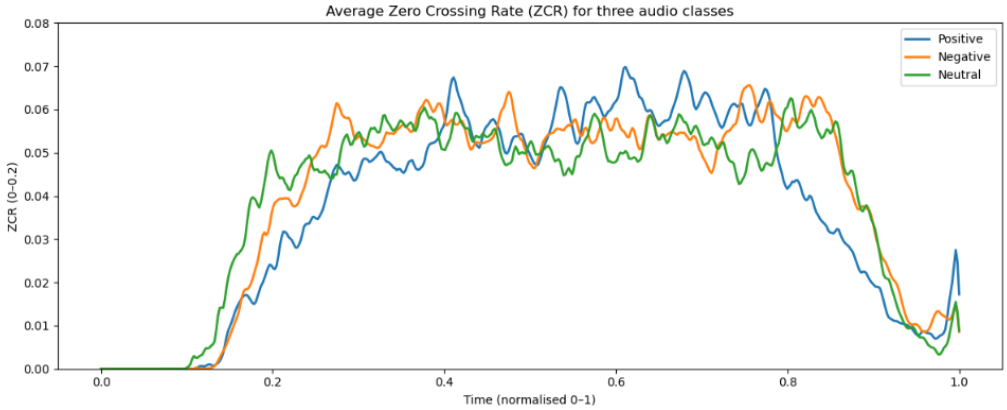

Fig 5 Average ZCR for three audio classes

First segment (0–0.2) Where a rapid increase in ZCR is observed for all classes. This phenomenon indicates that at the beginning of the audio there are many fast signal transitions (word onset, breathing, sharp sounds). The Neutral class (green) tends to have a slightly higher ZCR, which may suggest more diffuse or less stable sounds at the input.

Central segment (0.2–0.8) The lines are more stable and approach values of 0.05–0.07 on average. This stabilization suggests more stable vocal parts, with small differences between classes. The differences between classes are most noticeable around 0.3–0.5: Positive and Negative reach higher points than Neutral, which may be related to greater intensity or more energetic sounds in these emotions.

Final segment (0.8–1.0) It is found that ZCR falls significantly for all classes, indicating a decrease in signal activity (word ending or pause). Positive falls earlier and more strongly, which may suggest smoother word closure compared to Neutral and Negative.

All three classes follow a similar profile (initial rise → central plateau → decline at the end), which is characteristic of the general structure of speech. Neutral has slightly higher ZCR at the beginning, which

may reflect scattered sounds or shorter word beginnings. Positive and Negative ZCRs show more pronounced fluctuations during the central segment, which is associated with more energetic vocal production and perhaps changes in frequency due to emotional load. Positive's ZCR shows a more pronounced decline at the end, suggesting greater clarity at the end of speech.

B. LABELLING STRATEGIES IN SPEECH EMOTION AND SENTIMENT ANALYSIS

In sentiment and emotion analysis from audio, the quality and manner of data labeling is as important as the choice of model or acoustic features. An important debate is: should labeling be done at the original audio level (utterance-level) or after preprocessing/segmentation (segment- or frame-level)? Most classical works do labeling at the utterance level (e.g. a full recording is labeled “positive”, “negative” or “neutral”), and then preprocessing is used to segment the signal into smaller units for training the model. Segments often inherit the label of the utterance, although they may not contain all the emotional load. This process is simple and efficient, but creates a risk of label noise when the segments are heterogeneous. Therefore, for audio segments, a distance metric algorithm, such as Euclidean distance, was used to assess the heterogeneity of the segmentations.

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

Fig 3 Euclidean Distance formula

This algorithm was applied to each class by grouping them according to the file name. Based on the values where in cases mean_distance >=280 cases are identified where the audio segments are heterogeneous, while when this distance is >=100 but <280 segments that have similar parts within the group are categorized and finally if mean_distance is <100, the audio segments are evaluated as having high similarity. The table below presents the values of mean_distance according to classes.

	Positive	*Negativ*	*Neutral*
Heterogen	0	0	0
Some Similarities	85	100	95
Similar	207	196	199
Total	**293**	**296**	**294**

Table 2 Mean distance between segmentation

One article proposes using a method for frame-level pseudo-emotion labels (clustering) and then using the labels to refine the model. This method directly addresses the fact that “not all frequencies/frames” within an utterance represent the full emotionality (Li et al., 2024). So the process involves tagging the utterance before the audio is preprocessed.

C. PREPROCESSING

Audio processing goes through several stages. First is the normalization of the amplitude (0-1) or [-1,1]. By normalization of the amplitude it is meant that the signal becomes of a uniform scale and is safe for processing, e.g. unified in volume. By applying linear transformation f(x)=2x+1it results that if x∈[−1,1] then f(x)∈[0,1]. Then is the creation of 1-2 second segments. 1-2 second segments are used because they provide enough emotional context (intonation, rhythm, energy) without mixing too many different states in the same example and without overloading the memory during training. Since emotion changes slowly in speech, fragments of several seconds are necessary to capture prosodic patterns and this is emphasized in works that require long time dependencies in SER (Tang et al., 2021). Longer segments increase "label nose" where different emotions can exist within the same utterance, and fragmenting them into shorter parts reduces the problem of this mix-up (Madanian et al., 2023). Removing unnecessary noise and silence (denoising or silence removal) because real environments are noisy and this leads to a significant decrease in performance (Leem et al., 2024) . Noise distorts features (changes energy in bands, peaks, etc.) so

classifiers get confused (Bhattacharjee et al., 2016). Finally, there is the sampling rate, which in the case of HuBERT is 16 kHz, while for MFCC it is 16-44 kHz. Samples rate means how many samples we take per second from the sound wave. It is important because it specifies how much detail we capture in the sound wave within a second. As result by normalization is created a numerical stability and comparable amplitudes, by Silece removal &denoise can achieve higher effective SNR and less label noise in segments. By resampling is provided a consistent time-frequency support for MFCC/SSL feature and by segmentation for batching will be fixed shapes which enables segment or frame label labelling.

VI. EXTRACTING FEATURE WITH MFCC AND HUBERT.

Comparing the MFCC and HuBert methods for audio feature extraction, it was found that HuBert typically outperforms MFCC with 1.33-10.46% higher accuracy across applications, although HuBert requires more computational resources and there are cases where its advantages may decrease depending on the domain. In a study on dysarthria detection and severity classification (Javanmardi et al., 2024) reported that HuBERT features yielded 1.33–2.86% higher detection accuracy and 6.54–10.46% greater severity classification accuracy than MFCC and other baselines. During Automatic Speech Recognition in the article by Kumar it was observed that HuBert's performance varies depending on the domain, accent and language, while Mel-filterbanks sometimes performs better (Kumar et al., 2022). Other studies evaluate HuBert in speech recognition (Hsu et al., 2021) and identifying emotions (Chakhtouna et al., 2024) reported favorable metrics such as a word error rate of up to 4.6-7.6% and a recognition rate of 82.6%, although no direct comparison of MFCC was provided. In the within-language scenario, SSL models (e.g. HuBERT) achieved up to 34.4% improvement in articulatory feature (AF) testing over MFCC, while in the cross-language scenario the improvement is 26.7% over MFCC (Ji et al., 2022).

VII. LIGHTWEIGHT NEURAL CLASSIFIER FOR AUDIO SENTIMENT ANALYSIS

A lightweight neural classifier is a simple neural network that is used to classify features extracted with MFCC, HuBERT, etc. and offers a balance between performance and cost. Such models are preferred when datasets are small or when real-time inference is required.

A. TWO LAYER FEED FORWARD NETWORK (MLP)

MLP (Multi-Layer Perception) is a type of Artificial Neural Network composed of multiple layers of neurons (perceptors) where each layer is fully connected to the other (fully connected/dense layer). Initially, there is a layer that receives features (e.g. vectors from MFCC or embedding from HuBERT) which is considered the input layer, then there are one or more hidden layers that store non-linear information of the data and the output layer where there is a layer that produces classes (positive, negative, neutral). Two Layer Feed Forward Netwrok is a simple case of MLP that contains 2 hidden layers. Two Layer Feed Forward MLP is used for audio feature classification (Tzirakis et al., 2017). The MLP classifier is imported from scikit-learn, and the hidden layer size is set to 128 which is the number of neurons. In audio sentiment analysis, the ReLU (Rectified Linear Unit) activation function is used to ensure stability in the hidden layers. The form of the ReLU activation function is as follows figure 4:

:

$$f(x) = \max(0, x)$$

Fig 4.: Activation function ReLU

The ReLU activation function passes positive values and zero to negative ones. It is very fast and helps in avoiding the problem of vanishing gradients. So in the hidden layer we have Dense Layer which mathematically is represented by the following formula in figure 5 along with the ReLU activation function.

$$h = f(Wx + b)$$

Fig 5. Dense layer function

B. Classification Model Two Layer Feed Forward for MFCC and HuBERT

Initially, each csv file was loaded separately, for MFCC and HuBERT in order to compare the performance depending on what is obtained from these feature extractors. Meanwhile, the focus is on the stratification for splitting the dataset (e.g. train/validation/test split or cross-validation). Stratification is the process that ensures that the percentage of classes is kept the same in each split as in the original dataset. If stratification is not applied, then the split will be random and non-proportional distributions can be produced. It guarantees that each class is represented in all splits, ensuring reliability in the accuracy, F1, Confusion Matrix metrics and avoiding the risk of "biased splits" where a model may appear better or worse simply because a class has not been tested. The stratification used is based on the Random Stratified Sampling Algorithm. The next step is the application of the Two-Layer Feed-Forward classifier and performance evaluation that includes accuracy, Macro-F1, Classification Report and Confusion Matrix.

	Precision			*F1 -score*		
Classification Models	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
MFCC-MLP(128)	0.681	0.717	0.681	0.671	0.74	0.657
HuBERT-MLP(128)	0.652	0.754	0.677	0.668	0.754	0.655

Table 3 Classification report

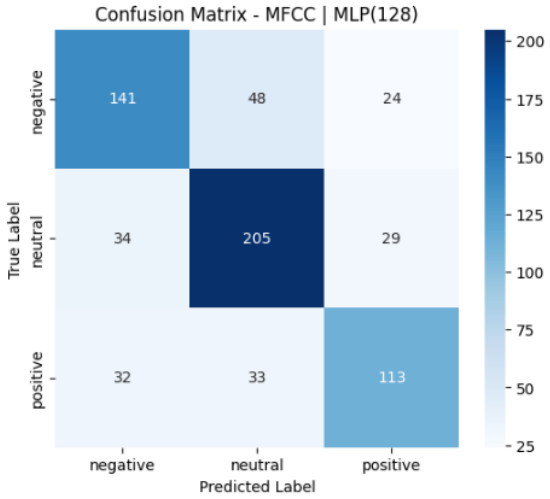

Fig 6-Confusion Matrix for MFCC for MLP classifier

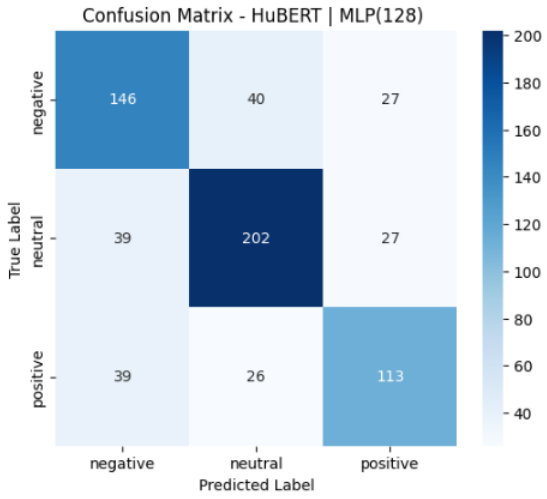

Fig 7-Confusion Matrix for HuBERT for MLP Classifier

Confusion matrix is a standard tool for evaluating the performance of classifiers, providing more detailed information than a single metric such as accuracy or F1-score. It shows where classes are confused, which is not visible from accuracy. It is presented as a matrix where the rows represent the real labels (ground truth), while the columns represent the labels predicted by the model. Diagonal elements (e.g. "negative → negative") indicate the number of cases correctly classified, while off-diagonal elements indicate specific model errors (e.g. "positive" classified as "neutral"). This tool is useful for understanding the distribution of errors and identifying classes that the model often confuses. In the case of our results, both configurations (MFCC + MLP, HuBERT + MLP) show that the neutral class has the highest accuracy, while the negative and positive classes are often confused with each other. This is to be expected in sentiment analysis from audio, where the line between negative and positive emotions is often thin and depends on nuances of intonation (El Ayadi et al., 2011). Furthermore, the comparison of the confusion matrix between the models clearly shows that using MLP on HuBERT embeddings gives a more balanced distribution of correct

classifications, reducing cross-class errors. This highlights the advantage of using self-supervised models compared to traditional MFCC features (Mohamed et al., 2022).

		Accurancy		*Macro F1*	
		Test	**Valid**	**Test**	**Valid**
HuBERT	**MLP(128)**	0.7	0.728	0.692	0.721
MFCC	**MLP(128)**	0.697	0.729	0.689	0.727

Table 4 Accurancy and Macro-F1

The experimental results, Table 4, showed that the HuBERT + MLP(128) model achieved a macro-F1 of about 0.69 on the test set, while models based on MFCC alone or linear classifiers (LogReg) showed somewhat lower performance (0.64–0.68). This level of performance is consistent with the existing literature on speech sentiment/emotion recognition in resource-limited languages, where typical models based on MFCC alone report F1 in the range of 0.60–0.70 (Zhang et al., 2018) (El Ayadi et al., 2011). Meanwhile, the use of embeddings derived from self-supervised models such as HuBERT or wav2vec 2.0 has been shown to significantly improve performance compared to traditional MFCC, but without reaching the levels of languages with very large datasets (Mohamed et al., 2022). The small difference between valid and test performance (around 2–3%) suggests that overfitting is not a problem, and that the model is able to generalize consistently. These results show that, although F1=0.69 is not very high compared to "high-resource" scenarios, it constitutes a reliable and expected performance for sentiment analysis from audio in contexts where the dataset is limited and the language is low-resource, as in the case of Albanian.

C. BIDIRECTIONAL LSTM WITH ATTENTION LAYER

Bi-directional LSTM (Bi-LSTM) is one of the most widely used architectures today in speech emotion recognition (SER) and sentiment analysis (Senthilkumar et al., 2022). Bi-LSTM is a variant of RNN (Recurrent Neural Networks) that captures long-term dependencies in temporal sequences. These sequences are processed in both directions (start◊end and end◊start). So it is of particular importance if information is obtained from both the past and the future within the segment to explain the intonation and rhythm of the audio (Graves and Schmidhuber, 2005). The problem with LSTMs is that not all MFCC frames (or HuBERT embeddings) have the same weight for the classification task, so an Attention Layer is considered which allows the model to "focus" attention on frames that have more information, such as where strong changes in energy or pitch occur that indicate emotion. Mathematically, attention calculates a distribution of weights over all time steps and combines the hidden states into a weighted representation (Bahdanau et al., 2014). The BiLSTM+Attention combination has become standard in many speech sentiment/emotion recognition applications. BiLSTM captures the two-way context while Attention selects the most informative frames, thereby increasing interpretability and distinguishing which parts of the audio most influence the model's decision (Mirsamadi et al., 2017). BiLSTM with Attention is a hybrid architecture that significantly improves performance compared to simple BiLSTM, especially in tasks where temporal dynamics and acoustic strengths play a key role.

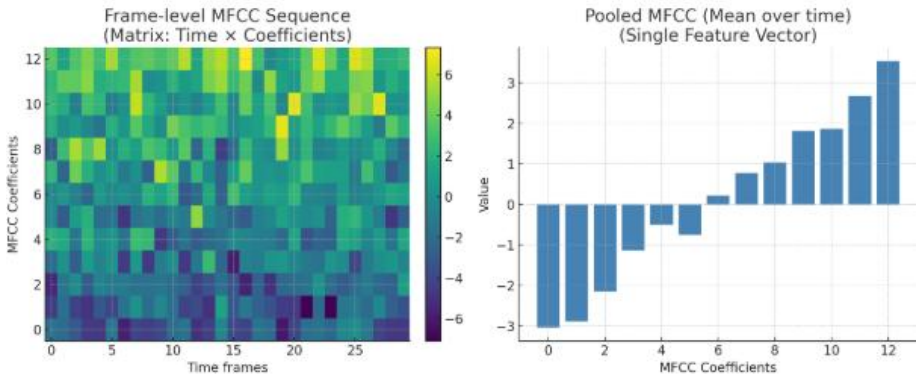

Fig 8 Visual perception of matrix with time frame and vector

We have many audio clips and each one is converted into a sequence of frames. Each frame has 5 coefficients (MFCC_0..MFCC_4) → vector with dimension D=5 but the clips have different lengths (T_i ≠ T_j). Neural models usually expect tensors with fixed length. Temporal ordering is performed according to path and frame to preserve the temporal order of the frames (time "mixing" is avoided). For each file p, a matrix (T,5)—T frame, 5 features each is obtained. Label is obtained once for the entire clip (assumption: the entire clip has the same sentiment). Label encoding is performed by converting them from textual to indices (e.g. neutral=0, positive=1, negative=2). This is necessary for classification loss (e.g. CrossEntropy). The correct dtypes for PyTorch are provided (float32 for inputs, long for labels). Therefore, throughout this paper, speech has been treated as a multidimensional time series; each frame is a vector of acoustic features. When it is necessary to perform sentiment classification (negative, neutral, positive) a fixed vector is required for decision making, but the sequences have different lengths and therefore the BiLSTM+pooling model is used. This model aims to read the entire sequence of features (MFCC) by taking context from the past and the future. The sequence of length T is condensed into a single fixed vector (pooling in time) and this vector is projected into the class space through a linear+softmax layer for probabilities. The architecture includes Input X ∈ ℝ^{B×T×D} where B is the batch size, T the number of frames (which varies in clips) and D=feat_dim which in our case 5 features were used for each frame. After the input there is the LSTM that processes the sequence in two directions: left→right and right→left. The output for each time is: H_t ∈ ℝ^{2H} (the union of the two directions) while the total output is: H_all ∈ ℝ^{B×T×2H}. Pooling in time (average) is A ∈ ℝ^{B×2H} averaging over the time axis making the representation invariant to length:

$$\overline{h} = \frac{1}{T}\sum_{t=1}^{T} H_t$$

Fig 9 Arithmetic mean of a sequence of vectors

The last in this architecture is the linear+softmax classifier Z=WH+b. Softmax gives p(y|x)=softmax(Z). While the loss is treated with Cross-Entropy : L=CrossEntropy(Z,y)

Fig 10 Train vs validation loss curbe-BiLSTM+Attention+MFCC

To understand the behavior of the model during training, the visual representation with the graph above is used. The graph presents two lines: Blue (train loss) and Orange (Val loss). Train loss is observed to have a continuous decrease from ≈1.08 → ≈0.92 and this shows that the model is learning well on the training data. Meanwhile, Val Loss starts around 1.06, drops slightly to ≈1.04, and then stays almost flat which means that the model does not improve much on the validation data. From the graph it is concluded that there are no strong signs of overfitting because according to classical overfitting the train loss line would continue to fall while val loss would start to increase clearly. In our case, the train loss drops, while the val loss remains stable, assuming that the model may have reached its maximum capacity with these parameters. Given that the difference between train and val loss is small (~0.1), we say that the improvement has stalled, so the model continues to learn in training but does not benefit in validation. This suggests that the architecture (BiLSTM 128, average pooling) or features no longer provide enough information to further reduce the val loss. Since the val loss does not improve after several epochs, stopping is justified because it saves time and avoids over-training. For optimization, it is necessary to increase the number of features, e.g. 13-20 MFCC + Δ + ΔΔ, instead of average pooling, attention-pooling is tried or the last hidden state is taken. The capacity of the model can be increased with more units, e.g. 256 or two BiLSTM layers. In conclusion, based on the results, we say that the model is learning because the train loss is continuously decreasing and the model is adapting to the training data. The performance in validation is problematic where the model is not improving its generalization ability on new data. It is estimated that there are no signs of overfitting because the wave loss does not increase much but we have partial underfitting because the model does not manage to learn enough structures to significantly reduce the wave loss. Model training stops after epoch 3-4 when the improvement in validation stops. To save computational resources and to prevent overtraining, early stopping techniques are used

.

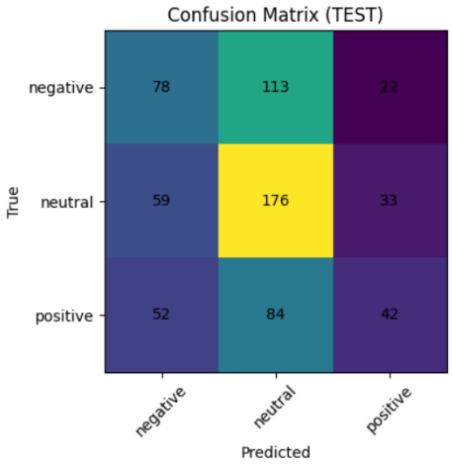

Fig 11. Confusion matrix for BiLSTM with Attention before class weight loss

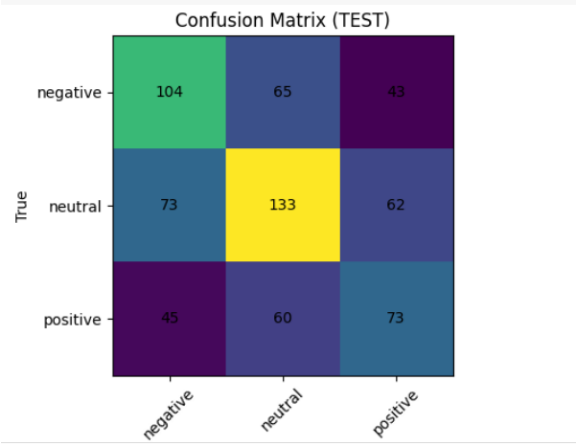

Fig 12. Confusion matrix for BiLSTM with Attention after class weight loss

On the held-out test set the BiLSTM-attention model achieves 45% accuracy (macro-F1 0.41). The confusion matrix reveals a systematic bias toward the *neutral* class (373 predicted), with frequent false-neutral assignments from both *negative* and *positive*. Class-wise performance is uneven (F1: neutral 0.55, negative 0.39, positive 0.31), indicating limited discriminability of the current MFCC representation and/or class imbalance. Improvements should target feature quality and class-imbalance mitigation. But by setting class weight we see an improvement in the confusion matrix because it favors the minority class. With standard cross-entropy the loss for a sample of class c is:

$$\ell = -\log p_\theta(y{=}c \mid x).$$

Fig 13 Cross entropy formula

Meanwhile with class weights w_c the formula is:

$$\ell = -w_c \log p_\theta(y{=}c \mid x).$$

Fig 14 Formula with class weight loss

According to the classification report, it most likely shows underfitting and a slight imbalance/ambiguity of labels, not necessarily overfitting because the overall performance is average.

	Precision			*F1 -score*		
Classification Models	**Negative**	**Neutral**	**Positive**	**Negative**	**Neutral**	**Positive**
MFCC-BiLSTM+Attention)	0.41	0.47	0.43	0.39	0.55	0.31
Weight class loss instead standard entropy	0.41	0.52	0.46	0.42	0.50	0.37
HuBERT- BiLSTM+Attention)	0.55	0.70	0.59	0.57	0.64	0.61

Table 5. Classification Report

Table 5 compares the performances of three different approaches to classifying emotion recognition models based on voice, using standard metrics such as Precision and F1-score for the three classes (positive, negative and neutral). The MFCC-BiLSTM+Attention model has a modest result for precision (0.41-0.47)

and a low F1-score mainly for the Positive class (0.31). This means that although MFCC provides a solid base of acoustic features, there are limitations in distinguishing more complex emotions. The other case was considered to improve the results with MFCC. Instead of standard entropy, weight class loss was used and the results showed slight improvements compared to standard entropy, mainly for the neutral and positive classes (precision 0.46, F1 score 0.37). Weight class loss helped in balancing the datasets although the effect remains limited. The third case is HuBERT-BiLSTM+Attention and this is the combination that presents the best performance with precision 0.70 for the Neutral class and F1-score up to 0.64 for Neutral and 0.61 for Positive. It is best reflected that self-service models such as HuBERT capture rich semantic and prosodic information compared to traditional models. Improvement was observed especially for the Positive class which was often more challenging to identify.

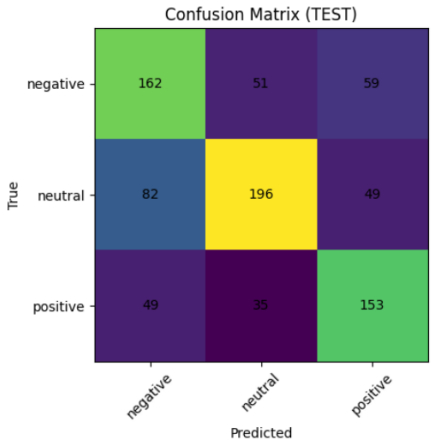

Fig 15.Confusion matrix for HuBERT+BiLSTM with Attention

VIII. RESULTS

Results show that HuBERT-based models combined with BiLSTM+Attention outperform traditional MFCC-based models by increasing precision, F1-score, and accuracy. This confirms the fact that the use of features extracted from self-supervised models significantly improves the performance in classifying speech emotions in a resource-limited language such as Albanian. Although the dataset used is balanced in the number of samples, the performance for the "Positive" class remains lower compared to "Neutral" and "Negative". This is not related to the lack of balance but to the nature of the acoustic cues themselves: positive emotions very often have similarities with neutral tones. As a result, the model has difficulty identifying a clear acoustic profile that belongs to this category, which suggests the need to use additional techniques such as acoustic augmentation or the inclusion of prosodic features.

IX. RECOMMENDATION

The findings of this study lead to several recommendations for researcher and practitioners working on speech-based sentiment analysis, especially in low-resource languages such as Albanian. One recommendation is to adopt self-supervised embeddings rather then relying solely on traditional MFCC feature because they have shown superior performance in prosodic nuances. Improve and robust the model by applying data augmentation techniques to address the limited size of available dataset. Consider prosodic features to distinguish emotions that are accoustically similar (in our case positive and neutral).

X. FUTURE WORK

In the next steps, it is intended to precisely define the acoustic elements that should be taken into account when building the dataset for classification into three emotional categories (positive, negative and neutral). An important research direction is to evaluate the possibility of labeling at the level of shortened audio segments after the preprocessing process, instead of labeling only at the utterance level. This would allow a more detailed overview of prosodic and emotional variations within the same utterance, as well as provide

an opportunity to evaluate the impact of this approach on the final performance of the model. It is also suggested to integrate semantic analysis of the audio, taking into account not only acoustic cues but also the linguistic context of the utterances. This combination of prosodic features with semantic information is expected to contribute in increasing the accuracy of the models and their ability to better distinguish similar emotional overtones. A multimodal approach that combines acoustic and textual sentiment cues is expected to yield higher performance.

REFERENCES

Anjali Verma, Bappaditya Jnna, 2025. Sentiment Analysis for Education and Behavior Analysis Based on Their Sentiment State. Int J Sci Res Sci & Technol 12, 585–587. https://doi.org/10.32628/IJSRST2411480

Atmaja, B.T., Sasou, A., 2022. Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations. Sensors 22, 6369. https://doi.org/10.3390/s22176369

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate. https://doi.org/10.48550/ARXIV.1409.0473

Bhattacharjee, U., Gogoi, S., Sharma, R., 2016. A statistical analysis on the impact of noise on MFCC features for speech recognition, in: 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE). Presented at the 2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE), IEEE, Jaipur, India, pp. 1–5. https://doi.org/10.1109/ICRAIE.2016.7939548

Binjaku, K., Janku, J., Mece, E.K., 2022. Identifying Low-Resource Languages in Speech Recordings through Deep Learning, in: 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM). Presented at the 2022 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), IEEE, Split, Croatia, pp. 1–6. https://doi.org/10.23919/SoftCOM55329.2022.9911376

Çano, E., 2023a. AlbNER: A Corpus for Named Entity Recognition in Albanian. https://doi.org/10.48550/ARXIV.2309.08741

Çano, E., 2023b. AlbMoRe: A Corpus of Movie Reviews for Sentiment Analysis in Albanian. https://doi.org/10.48550/ARXIV.2306.08526

Çano, E., Lamaj, D., 2024. AlbNews: A Corpus of Headlines for Topic Modeling in Albanian. https://doi.org/10.48550/ARXIV.2402.04028

Chakhtouna, A., Sekkate, S., Adib, A., 2024. Unveiling embedded features in Wav2vec2 and HuBERT msodels for Speech Emotion Recognition. Procedia Computer Science 232, 2560–2569. https://doi.org/10.1016/j.procs.2024.02.074

Colunga-Rodriguez, A.A., Martínez-Rebollar, A., Estrada-Esquivel, H., Clemente, E., Pliego-Martínez, O.A., 2025. Developing a Dataset of Audio Features to Classify Emotions in Speech. Computation 13, 39. https://doi.org/10.3390/computation13020039

El Ayadi, M., Kamel, M.S., Karray, F., 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44, 572–587. https://doi.org/10.1016/j.patcog.2010.09.020

Elsayed, N., ElSayed, Z., Asadizanjani, N., Ozer, M., Abdelgawad, A., Bayoumi, M., 2022. Speech Emotion Recognition using Supervised Deep Recurrent System for Mental Health Monitoring, in: 2022 IEEE 8th World Forum on Internet of Things (WF-IoT). Presented at the 2022 IEEE 8th World Forum on Internet of Things (WF-IoT), IEEE, Yokohama, Japan, pp. 1–6. https://doi.org/10.1109/WF-IoT54382.2022.10152117

Gondohanindijo, J., -, M., Noersasongko, E., -, P., Setiadi, D.R.M., 2023. Multi-Features Audio Extraction for Speech Emotion Recognition Based on Deep Learning. IJACSA 14. https://doi.org/10.14569/IJACSA.2023.0140623

Graves, A., Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks 18, 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

Hsu, W.-N., Bolte, B., Tsai, Y.-H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A., 2021. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. IEEE/ACM Trans. Audio Speech Lang. Process. 29, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Javanmardi, F., Kadiri, S.R., Alku, P., 2024. Pre-trained models for detection and severity level classification of dysarthria from speech. Speech Communication 158, 103047. https://doi.org/10.1016/j.specom.2024.103047

Ji, H., Patel, T., Scharenborg, O., 2022. Predicting within and across language phoneme recognition performance of self-supervised learning speech pre-trained models. https://doi.org/10.48550/ARXIV.2206.12489

Kadriu, F., Murtezaj, D., Gashi, F., Ahmedi, L., Kurti, A., Kastrati, Z., 2022. Human-annotated dataset for social media sentiment analysis for Albanian language. Data in Brief 43, 108436. https://doi.org/10.1016/j.dib.2022.108436

Kaloga, Y., Kodrasi, I., 2025. Towards interpretable emotion recognition: Identifying key features with machine learning. https://doi.org/10.48550/ARXIV.2508.04230

Kastrati, M., Biba, M., 2022. Natural language processing for Albanian: a state-of-the-art survey. IJECE 12, 6432. https://doi.org/10.11591/ijece.v12i6.pp6432-6439

Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M.A., 2017. A review on speech emotion recognition: Case of pedagogical interaction in classroom, in: 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP). Presented at the 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), IEEE, Fez, Morocco, pp. 1–7. https://doi.org/10.1109/ATSIP.2017.8075575

Koh, E., Dubnov, S., 2021. Comparison and Analysis of Deep Audio Embeddings for Music Emotion Recognition. https://doi.org/10.48550/ARXIV.2104.06517

Kumar, P., Sukhadia, V.N., Umesh, S., 2022. Investigation of Robustness of Hubert Features from Different Layers to Domain, Accent and Language Variations, in: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Singapore, Singapore, pp. 6887–6891. https://doi.org/10.1109/ICASSP43922.2022.9746250

Leem, S.-G., Fulford, D., Onnela, J.-P., Gard, D., Busso, C., 2024. Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 917–929. https://doi.org/10.1109/TASLP.2023.3340603

Li, Q., Gao, Y., Wang, C., Deng, Y., Xue, J., Han, Y., Li, Y., 2024. Frame-Level Emotional State Alignment Method for Speech Emotion Recognition, in: ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Seoul, Korea, Republic of, pp. 11486–11490. https://doi.org/10.1109/ICASSP48485.2024.10446812

Madanian, S., Chen, T., Adeleye, O., Templeton, J.M., Poellabauer, C., Parry, D., Schneider, S.L., 2023. Speech emotion recognition using machine learning — A systematic review. Intelligent Systems with Applications 20, 200266. https://doi.org/10.1016/j.iswa.2023.200266

Mangalmurti, S., Saxena, O., Singh, T., 2024. Speech Emotion Recognition using CNN-TRANSFORMER Architecture, in: 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI). Presented at the 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), IEEE, Gwalior, India, pp. 1–6. https://doi.org/10.1109/IATMSI60426.2024.10503276

Mirsamadi, S., Barsoum, E., Zhang, C., 2017. Automatic speech emotion recognition using recurrent neural networks with local attention, in: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the 2017 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP), IEEE, New Orleans, LA, pp. 2227–2231. https://doi.org/10.1109/ICASSP.2017.7952552
Mohamed, A., Lee, H., Borgholt, L., Havtorn, J.D., Edin, J., Igel, C., Kirchhoff, K., Li, S.-W., Livescu, K., Maaløe, L., Sainath, T.N., Watanabe, S., 2022. Self-Supervised Speech Representation Learning: A Review. IEEE J. Sel. Top. Signal Process. 16, 1179–1210. https://doi.org/10.1109/JSTSP.2022.3207050
Resende Faria, D., Weinberg, A.I., Ayrosa, P.P., 2024. Multimodal Affective Communication Analysis: Fusing Speech Emotion and Text Sentiment Using Machine Learning. Applied Sciences 14, 6631. https://doi.org/10.3390/app14156631
Rezapour Mashhadi, M.M., Osei-Bonsu, K., 2023. Speech emotion recognition using machine learning techniques: Feature extraction and comparison of convolutional neural network and random forest. PLoS ONE 18, e0291500. https://doi.org/10.1371/journal.pone.0291500
Senthilkumar, N., Karpakam, S., Gayathri Devi, M., Balakumaresan, R., Dhilipkumar, P., 2022. Speech emotion recognition based on Bi-directional LSTM architecture and deep belief networks. Materials Today: Proceedings 57, 2180–2184. https://doi.org/10.1016/j.matpr.2021.12.246
Sezgin, M.C., Gunsel, B., Kurt, G.K., 2012. Perceptual audio features for emotion detection. J AUDIO SPEECH MUSIC PROC. 2012, 16. https://doi.org/10.1186/1687-4722-2012-16
Shehu, D., Luarasi, T., Kyratsis, P., 2025. Designing a natural language processing-driven communication system for urban planning: A case study. JCAU 0, 8417. https://doi.org/10.36922/jcau.8417
Tang, D., Kuppens, P., Geurts, L., Van Waterschoot, T., 2021. End-to-end speech emotion recognition using a novel context-stacking dilated convolution neural network. J AUDIO SPEECH MUSIC PROC. 2021, 18. https://doi.org/10.1186/s13636-021-00208-5
Tits, N., Haddad, K.E., Dutoit, T., 2018. ASR-based Features for Emotion Recognition: A Transfer Learning Approach. https://doi.org/10.48550/ARXIV.1805.09197
Venkataramanan, K., Rajamohan, H.R., 2019. Emotion Recognition from Speech. https://doi.org/10.48550/ARXIV.1912.10458
Zhang, Shiqing, Zhang, Shiliang, Huang, T., Gao, W., 2018. Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. IEEE Trans. Multimedia 20, 1576–1590. https://doi.org/10.1109/TMM.2017.2766843
Zhao, P., Liu, F., Zhuang, X., 2022. Speech Sentiment Analysis Using Hierarchical Conformer Networks. Applied Sciences 12, 8076. https://doi.org/10.3390/app12168076