

Article

Emotion Recognition Using Acoustic Features and Deep Learning: A Speaker-Independent Study

Marcin Kołodziej ^{1,*} , Andrzej Majkowski ¹  and Tomasz Rywik ²

¹ Institute of Theory of Electrical Engineering, Measurement and Information Systems, Faculty of Electrical Engineering, Warsaw University of Technology, ul. Koszykowa 75, 00-662 Warsaw, Poland; andrzej.majkowski@pw.edu.pl

² Department of Heart Failure and Transplantology, National Institute of Cardiology, ul. Alpejska 42, 04-628 Warsaw, Poland

* Correspondence: marcin.kolodziej@pw.edu.pl

Abstract

This study compares the effectiveness of two approaches to speech emotion recognition for three affective states in Polish: sad, neutral, and happy. Both a set of acoustic features—capturing prosodic, phonatory, temporal, spectral, and cepstral properties—and representations learned by self-supervised models (wav2vec 2.0 and WavLM) were analyzed. Experiments were conducted on the nEMO corpus, comprising 2327 recordings from nine speakers, using a rigorous leave-one-subject-out protocol to evaluate cross-speaker generalization. In the feature-based approach, 107 acoustic features were used, and classification was performed with logistic regression and, additionally, SVM variants. In the deep learning approach, the wav2vec2-base and WavLM-base models were fine-tuned for the three-class task. The best results were achieved by the self-supervised models: WavLM reached a global balanced accuracy of 0.727 and a macro-F1 score of 0.710, while wav2vec 2.0 achieved 0.722 and 0.695, respectively. Both outperformed the feature-based approach (BAcc = 0.627, macro-F1 = 0.584). Confusion matrix analysis showed that the greatest difficulty lies in distinguishing the neutral class from the sad and happy classes, whereas sad and happy classes are more clearly separable. Feature utility analysis (SFS under the LOSO protocol) indicated the significant role of cepstral features (MFCCs and their derivatives), complemented by selected prosodic and temporal features. An additional comparison of SVM classifiers suggested that the main limitation of this approach lies in the signal representation itself rather than solely in the choice of classifier. Explainability analyses of the deep models, using layer-wise probing and integrated gradients, showed that affective information is best represented in intermediate layers, and that model decisions rely on locally salient segments of the signal. Furthermore, a speaker adaptation experiment demonstrated that personalization significantly improves classification performance, highlighting the potential of such methods for long-term monitoring of affective expression changes in the same individual.



Academic Editor: Jozef Juhár

Received: 26 May 2026

Revised: 7 July 2026

Accepted: 10 July 2026

Published: 14 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: emotion recognition; wav2vec 2.0; WavLM; speaker-independent; LOSO (leave-one-subject-out); acoustic features; mood monitoring

1. Introduction

Speech conveys not only linguistic information but also a range of acoustic features—prosodic, phonatory, and spectral—that jointly shape the perception of emotional expression [1,2]. Changes in emotional state may manifest simultaneously in fundamental

frequency, intonation dynamics, phonation stability, temporal organization of speech, and the distribution of energy across the acoustic spectrum [1]. In this study, emotion recognition is defined as the assignment of a speech signal to one of three labels: sad, neutral, or happy. A key challenge in this type of research is separating emotional information from speaker-specific characteristics. Models evaluated using simple random splits often achieve inflated performance, as they learn patterns specific to individual speakers rather than generalizable properties related to emotional state [3,4]. For this reason, speaker-independent evaluation is essential, as it allows one to determine whether a model truly recognizes emotions rather than speaker identity or idiosyncratic speaking style [5].

The importance of such research extends beyond technological applications like voice-based interfaces. The affective content of speech may also serve as an easily accessible, low-invasive indicator of changes in mental state over time [6]. This is particularly relevant in depressive disorders, where deterioration may be accompanied by reduced vocal energy, narrower prosodic dynamics, increased monotony, and changes in voice quality. Therefore, systems analyzing speech affect should primarily be viewed as tools supporting long-term monitoring of mood and expressive changes, rather than as autonomous diagnostic systems [7,8].

The literature includes numerous studies based on classical acoustic features, and in recent years an increasing number of works have employed deep learning models such as wav2vec 2.0 and WavLM [9,10]. However, studies combining three elements simultaneously—Polish emotional speech data, rigorous speaker-independent evaluation, and interpretable analysis of classical acoustic descriptors—remain relatively scarce [11]. This gap forms the direct motivation for the present work.

This study makes three main contributions. First, it presents a comparison of a classical model, wav2vec 2.0, and WavLM on the Polish nEMO corpus using the LOSO protocol. Second, it identifies acoustic features and feature groups that most strongly differentiate the analyzed classes. Third, it examines how the distribution of acoustic features differs between the three affective states, using an auxiliary LDA analysis to visualize whether the classes form separable regions in the feature space. In other words, the study addresses the problem of comparing classical and self-supervised approaches to recognizing three affective states in Polish speech, with evaluation conducted under rigorous cross-speaker generalization conditions.

2. Related Works

Contemporary research on speech emotion recognition (SER) has evolved along several parallel directions: from classical approaches based on hand-crafted acoustic features, through deep learning architectures operating on time–frequency representations, to self-supervised learning (SSL) models that learn representations without requiring fully labeled data [12–16]. The traditional line of work relied on descriptors such as fundamental frequency (F0), jitter, shimmer, harmonic-to-noise ratio (HNR), signal energy, and mel-frequency cepstral coefficients (MFCCs). Their standardization has been reinforced by feature sets such as GeMAPS and tools like openSMILE [12].

A key advantage of classical representations lies in their physical interpretability. Individual features can be directly linked to properties of the speech signal, such as intonation, phonation stability, spectral structure and the energy pattern of the utterance [12]. However, this approach is inherently limited, as it provides only a partial description of the signal, constrained by a predefined set of descriptors. For this reason, classical features remain an important baseline, but are increasingly complemented or replaced by automatically learned representations [13,14].

A major breakthrough in the field has come with self-supervised models. The wav2vec 2.0 model demonstrated that speech representations can be effectively learned by masking segments of the signal and predicting their quantized latent counterparts in a contrastive framework [13]. Building on this idea, WavLM extends self-supervised learning with additional mechanisms, including denoising, and provides representations useful across a wide range of speech processing tasks, including paralinguistic applications [14].

In SER applications, SSL-based representations have been shown to often outperform classical acoustic features, particularly when training data are limited or when the task requires stronger generalization [15,16]. Pepino et al. demonstrated that features derived from wav2vec 2.0 provide a strong basis for emotion classification in speech [15]. Atmaja and Sasou, in turn, conducted a comprehensive comparison of multiple self-supervised representations and showed that models such as WavLM, HuBERT, and UniSpeech-SAT can achieve highly competitive performance relative to classical descriptors [16]. At the same time, more recent studies indicate that combining learned representations with low-level acoustic features can further improve classification performance, suggesting complementarity rather than strict competition between these approaches [10].

Despite these advances, the limited availability of labeled emotional speech remains one of the major challenges in speech emotion recognition. Existing emotional speech corpora are typically small, often imbalanced, and vary substantially in terms of language, recording conditions, and speaker characteristics. These factors make the development of robust deep learning models particularly challenging [17,18].

Recent review papers indicate that overcoming data scarcity requires not only data augmentation, but also the use of transfer learning, self-supervised learning, contrastive learning, domain adaptation, multimodal learning, and synthetic data generation. These approaches have been shown to substantially improve model robustness and cross-speaker generalization, and they have become the dominant strategies for training deep learning models on limited emotional speech datasets [18–20].

In parallel, deep learning architectures based on spectrograms, attention mechanisms, and multi-scale modeling have continued to evolve. In recent years, effective solutions have been reported using CNN–LSTM–attention hybrids, transformer-based networks operating on mel-spectrogram representations, multi-task learning models, and graph neural networks [21–25]. Although these studies often report high performance on benchmark datasets, their results should be interpreted with caution, as outcomes strongly depend on the experimental protocol, dataset characteristics, and the degree of speaker separation between training and test sets [16,26].

There is a growing emphasis in the literature on rigorous evaluation protocols for SER models. It has been shown that data partitioning strategies can significantly influence reported performance, and that speaker-independent scenarios are substantially more challenging than speaker-dependent ones [26]. Consequently, protocols such as Leave-One-Subject-Out (LOSO) are considered more reliable for assessing a model's ability to generalize to unseen speakers. Additionally, cross-corpus studies demonstrate that models achieving high performance on a single dataset often degrade when transferred to different datasets, languages, or recording conditions [9].

Another important research direction concerns model interpretability and the analysis of what types of acoustic information are encoded in self-supervised representations. Studies show that SSL representations capture meaningful phonetic and acoustic information, but that its distribution depends on the model layer and the task [27,28]. This suggests that layer-wise analysis and attempts to relate learned representations to classical acoustic features may lead to a deeper understanding of how emotional information is encoded in speech signals.

In the context of clinical applications, speech analysis is increasingly considered a potential source of biomarkers for affective states, including depression. At the same time, systematic reviews highlight substantial heterogeneity in methods, datasets, and validation procedures, as well as the need to clearly distinguish emotion classification from diagnostic inference [29]. In practice, this means that SER results may inform clinical applications but should not be interpreted as a direct substitute for diagnosis.

For the Polish language, resources for emotional speech remain limited. The nEMO dataset constitutes one of the few available sources of Polish emotional speech and thus provides an important foundation for SER research in this language [30,31]. However, there is still a lack of systematic studies that, under speaker-independent conditions, compare classical acoustic features with SSL representations and evaluate their utility in a methodologically rigorous manner. This gap provides a strong motivation for further research in this area.

3. Aim of the Article

The aim of this paper is to systematically compare two different approaches to speech emotion recognition in Polish for three affective states: sadness, happiness, and neutrality. The first approach is based on hand-crafted acoustic features, while the second uses modern deep learning models operating on raw audio representations, namely wav2vec 2.0 and WavLM. The study does not aim to provide an exhaustive comparison of all neural network architectures proposed for speech emotion recognition. Instead, it focuses on comparing a representative classical feature-based approach with contemporary self-supervised learning models under a rigorous speaker-independent evaluation protocol. Additional CNN models trained from scratch are included later in the study as complementary reference models. Their purpose is to assess the contribution of learned speech representations relative to network architecture alone, rather than to serve as the primary methods of comparison.

Figure 1 presents the overall experimental pipeline. The starting point is the nEMO dataset, which contains Polish speech recordings annotated with three emotional classes. In the first stage, standard signal preprocessing was applied (mono conversion, resampling, and normalization) to ensure consistency of input data across all analyzed methods. Next, the data were partitioned according to the LOSO protocol to ensure speaker-independent evaluation [5,32,33]. Within the classical approach (left branch of the pipeline), a set of acoustic features was extracted, including pitch, energy, jitter, shimmer, harmonic-to-noise ratio (HNR), spectral features, and MFCCs [34]. These features were then statistically aggregated (mean and standard deviation), yielding a fixed-dimensional feature vector used as input to a logistic regression classifier [35,36].

In parallel (the middle and right branches in Figure 1), deep models wav2vec 2.0 and WavLM were analyzed [37,38]. In both cases, the input is the raw speech waveform, and the models rely on pretrained representations. For the three-class emotion classification task, fine-tuning was applied by modifying the classification layer to produce predictions for one of the three target classes.

In the final stage, all approaches were evaluated using the same performance metrics: balanced accuracy (BAcc), overall accuracy (Accuracy), and macro-F1 score [39]. This set of metrics enables reliable evaluation under potential class imbalance and allows for direct comparison between classical and deep models within an identical experimental protocol.

From a conceptual perspective, the study addresses three main questions. First, under the conditions of the nEMO corpus for Polish speech and the LOSO protocol, do deep learning models truly outperform the classical approach? Second, which groups of classical acoustic features contribute most to differentiating the analyzed affective classes? Third, does incorporating a small number of recordings associated with known affective states for

the same user significantly improve recognition performance, which would be relevant for long-term monitoring of an individual?

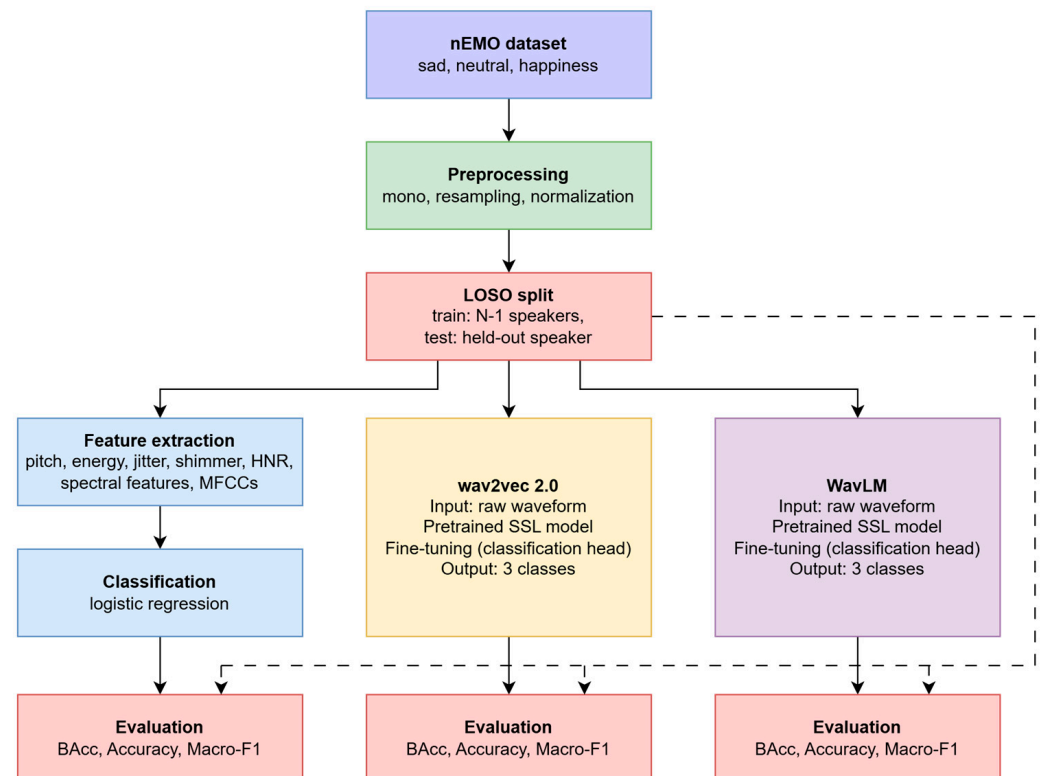


Figure 1. Experimental setup.

The main contributions of the paper can be summarized as follows:

- A comparison of a classical approach based on acoustic features and logistic regression with two approaches based on wav2vec 2.0 and WavLM on the Polish nEMO corpus using the leave-one-subject-out protocol, enabling evaluation of cross-speaker generalization.
- Computation and analysis of a 107-dimensional representation of classical acoustic features, linked to classification performance and interpretation.
- Analysis of feature usefulness and stability using Sequence Forward Selection, identifying the most relevant descriptors.
- Inclusion of LDA analysis and feature space visualizations to interpret data structure and the difficulty of the neutral class.
- Extension of the analysis to include a comparison of different classifiers, disentangling the effects of representation and decision model.
- Complementary explainability analysis of deep models using layer-wise probing and integrated gradients.
- A speaker adaptation experiment demonstrating the potential of model personalization.
- A comparison with CNN models trained from scratch, analyzing the impact of input data representation and the advantages of self-supervised models.

4. Materials

The experiments were conducted on the Polish emotional speech corpus nEMO. From the dataset, samples corresponding to three affective classes—sad, neutral, and happy—were selected. As a result, the dataset comprised 2327 recordings from nine speakers labeled EB0, IC0, KD0, KS0, MK0, MP0, SM0, TP0, and WR0. The nEMO corpus

contains acted emotional speech in Polish, where speakers produce short, semantically neutral sentences (lacking explicit emotional content), modulating only their affective expression. The material includes utterances realized in different emotional states—sad, neutral, and happy—while preserving identical linguistic content across emotions. The class distribution is approximately balanced, reducing the risk of dominance by any single category during training and justifying the use of balanced accuracy as the primary evaluation metric. The average recording duration is 2.532 s (SD = 0.662 s), with a range from 0.912 s to 5.459 s.

In the analyzed subset, most speakers exhibit full balance across emotional classes. Speakers EB0, IC0, KD0, KS0, MK0, and TP0 each contribute 90 utterances per class (sad, neutral, happy), totaling 270 samples per speaker. Minor deviations occur in three cases. For speaker SM0, the counts are 90 (sad), 90 (neutral), and 89 (happy), yielding 269 utterances. Similarly, speaker WR0 has 90 (sad), 89 (neutral), and 90 (happy), also totaling 269 samples. The largest imbalance is observed for speaker MP0, with 49 (sad), 90 (neutral), and 30 (happy) utterances, for a total of 169 samples. The mean utterance duration per speaker ranges from approximately 2.39 s (TP0) to 2.67 s (MP0), with standard deviations between roughly 0.62 and 0.72 s.

At the data level, two limitations regarding generalization should be noted: the relatively small number of speakers and the reliance on a single corpus. Table 1 summarizes the characteristics of the analyzed dataset, including class counts, their percentage distribution, and recording duration statistics (mean $\pm$ standard deviation and range), indicating a balanced distribution across the sad, neutral, and happy classes, as well as comparable sample durations.

Table 1. Characteristics of the analyzed dataset.

Class	Number of Recordings	Percentage	Average Duration $\pm$ SD [s]	Range [s]
sad	769	33.047	2.691 $\pm$ 0.698	1.026–5.459
neutral	809	34.767	2.557 $\pm$ 0.641	0.912–4.642
happy	749	32.186	2.342 $\pm$ 0.597	0.917–4.542
total	2327	100.000	2.532 $\pm$ 0.662	0.912–5.459

5. Methods

The study was conducted using three parallel approaches, while maintaining an identical evaluation procedure. In the first approach, a set of acoustic features was extracted, followed by classification using multiclass logistic regression. In the second approach, the wav2vec 2.0 model was fine-tuned for a three-class affective state classification task, while in the third approach, an analogous procedure was applied to the WavLM model.

For each LOSO split, the same sequence of steps was performed: one speaker was held out as the test set, the remaining speakers formed the training pool, the model was trained and then evaluated on the test set. Finally, results were averaged across all LOSO splits [33]. This design ensured experimental consistency and eliminated the risk of information leakage between training and test data. The overall workflow of the study is presented in this section, organizing the successive stages from data selection to the aggregation of final results.

5.1. Emotion Recognition Using Acoustic Speech Features

In the classical approach, a two-stage representation of the speech signal was adopted. In the first stage, the signal was analyzed in short time frames, enabling the capture of its non-stationary properties [40]. In the second stage, statistics of these features were

computed over the entire utterance, yielding a fixed-dimensional feature vector [21]. For most descriptors, the mean and standard deviation were used, while for pitch-related features the percentile-based range (P90–P10) was additionally included. This formulation preserves both the typical level of a feature and its temporal variability. As a result, 107 acoustic features were extracted, describing prosodic, phonatory, temporal, spectral, and cepstral properties [41].

The speech signal was converted to a monophonic form and analyzed at a sampling rate of 24,000 Hz. Amplitude normalization to RMS = 0.100 was applied to reduce the impact of loudness differences across recordings [42].

Pitch analysis was conducted using the parameters `time_step` = 0.010 s (analysis step), `pitch_floor` = 75 Hz (lower bound of fundamental frequency estimation), and `pitch_ceiling` = 400 Hz (upper bound). Based on this, the mean fundamental frequency, its variability, and range were computed [43]. Voice quality measures such as jitter and shimmer quantified irregularities in the period and amplitude of vocal fold vibrations, respectively [44]. Harmonicity measures (the ratio of periodic to noise energy) were computed using the To Harmonicity (cc) method (an autocorrelation-based algorithm available in Praat) with parameters `time_step` = 0.010 s and minimum pitch = 75 Hz [44]. Additionally, long-term average spectrum (LTAS) features were analyzed to capture the global energy distribution across frequencies [45].

Spectral and cepstral features were computed using frames of 2048 samples with a hop size of 512, based on the Fourier transform [46]. A set of 20 MFCC coefficients describing the spectral envelope was extracted, along with their temporal derivatives (delta MFCC), representing spectral dynamics [47]. Additional features included RMS energy, zero-crossing rate, spectral centroid (center of mass of the spectrum), spectral bandwidth (energy dispersion around the centroid), spectral roll-off (frequency below which a specified portion of spectral energy is contained), and spectral flatness (a measure of tonality versus noisiness) [34].

Temporal features related to silence and utterance structure were computed using a detection threshold of `top_db` = 30 (a relative threshold defining silence as segments at least 30 dB below the maximum signal level). Based on this, the pause ratio, mean and maximum pause duration, and the number of silence segments were calculated. Additionally, the rate of acoustic changes was analyzed using onset-related parameters (moments of sudden energy increase), such as the number of events per second and statistics of inter-onset intervals [48].

Feature selection was guided by complementarity and the ability to capture different aspects of speech production. Spectral and cepstral descriptors constituted the largest portion of the representation, while prosodic, phonatory, and temporal features played a complementary role, enabling a more comprehensive, multidimensional description of the signal. A detailed description of the acoustic features and their interpretation is provided in Table 2. The acoustic feature dataset supporting this study has been deposited in Zenodo and is available at: [<https://doi.org/10.5281/zenodo.20065078>]. The repository contains 107 acoustic features extracted from 2327 Polish speech utterances, together with metadata, feature descriptions, speaker-level files, and utterance-level JSON records.

For the classical machine-learning classifiers, all acoustic features were standardized prior to model training. In each LOSO fold, the `StandardScaler` was fitted exclusively on the training speakers and subsequently applied to the held-out test speaker.

The resulting standardized feature matrix was then used to train the logistic regression and SVM classifiers. Importantly, this feature-level standardization was independent of the RMS normalization applied earlier to the audio signal before acoustic feature extraction.

Table 2. Classification of the 107 classical features into groups and their interpretative significance.

Group of Features	Number of Features	Descriptors	Description
Duration	1	duration_s	total duration of the speech
Pitch	3	f0_mean_hz, f0_std_hz, f0_range_p90_p10_hz	intonation level and variability
Voice quality	4	jitter_local, shimmer_local, hnr_mean_db, hnr_std_db	phonation stability and harmonic component contribution
Pauses	4	pause_ratio, pause_mean_s, pause_max_s, n_pauses	the proportion of silence and the fragmentation of speech
Rate	3	onset_rate_per_s, inter_onset_mean_s, inter_onset_std_s	the rate of acoustic changes and the rhythm of speech
Energy	2	rms_mean, rms_std	the level and variability of energy
Voicing	2	zcr_mean, zcr_std	approximate noise level and signal tonality
Spectrum	8	spectral_centroid_mean_hz, spectral_centroid_std_hz, spectral_bandwidth_mean_hz, spectral_bandwidth_std_hz, spectral_rolloff_mean_hz, spectral_rolloff_std_hz, spectral_flatness_mean, spectral_flatness_std	energy distribution and spectral characteristics
Spectral contrast	14	spectral_contrast_band1_mean, spectral_contrast_band2_mean, spectral_contrast_band3_mean, spectral_contrast_band4_mean, spectral_contrast_band5_mean, spectral_contrast_band6_mean, spectral_contrast_band7_mean, spectral_contrast_band1_std, spectral_contrast_band2_std, spectral_contrast_band3_std, spectral_contrast_band4_std, spectral_contrast_band5_std, spectral_contrast_band6_std, spectral_contrast_band7_std	variation in energy between spectral bands
LTAS	1	ltas_spectral_slope	global slope of the long-term spectrum
MFCC	40	range ‘mfcc1_mean-mfcc20_mean’ and ‘mfcc1_std-mfcc20_std’	the spectral envelope and its variability
Delta MFCC	25	range ‘dmfcc1_mean-dmfcc20_mean’ and ‘dmfcc1_std-dmfcc5_std’	the temporal dynamics of cepstral changes
Total	107	full vector of classical features used in LOSO experiments	the input representation of the classical model

In the classical approach, multiclass logistic regression [49] was used as the baseline classifier. This choice was deliberate. First, a linear model provides a simple and well-interpretable reference framework in which the influence of individual acoustic features on

the classification outcome can be analyzed without interference from the complex structure of more advanced models. Second, given the relatively small number of speakers and the use of a rigorous speaker-independent evaluation scheme, linear models are less prone to overfitting to specific individuals. Third, logistic regression facilitates the interpretation of feature importance, as the contribution of individual descriptors can be directly linked to the classifier's decision. For these reasons, the linear model was treated as the most appropriate baseline for analyzing the usefulness of classical features. Figures 2–4 present the same utterance produced by a single speaker across three affective classes, illustrating differences in pitch trajectories and the time–frequency structure of the signal as observed in the spectrograms.

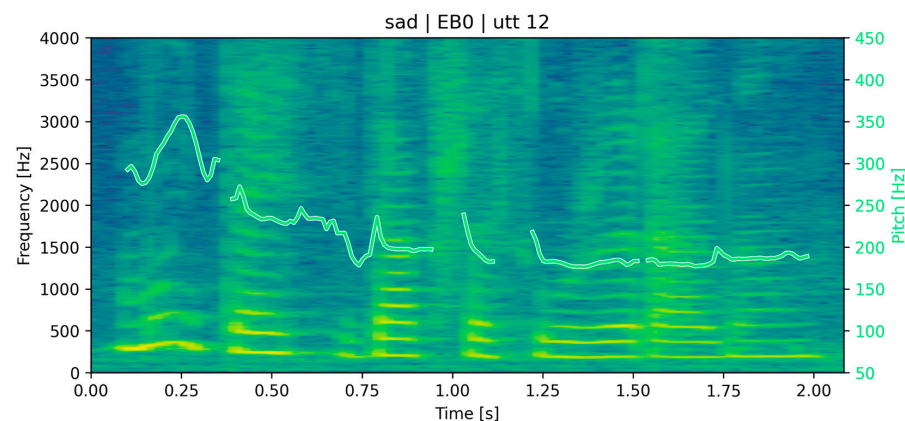


Figure 2. Spectrogram of a sad utterance with pitch contour.

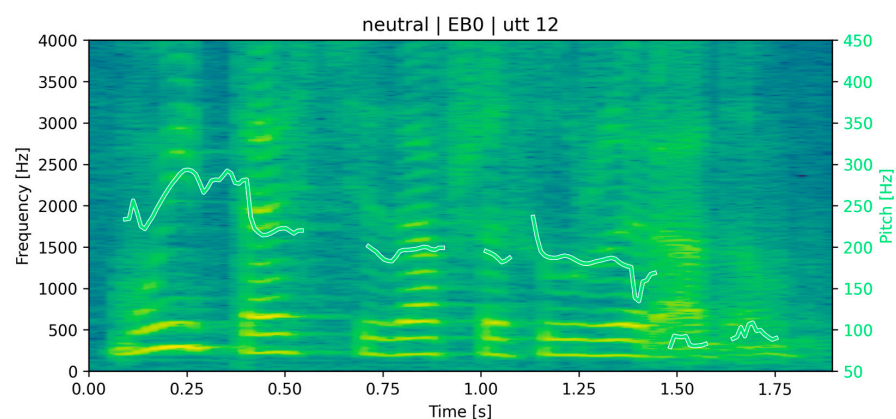


Figure 3. Spectrogram of a neutral utterance with pitch contour.

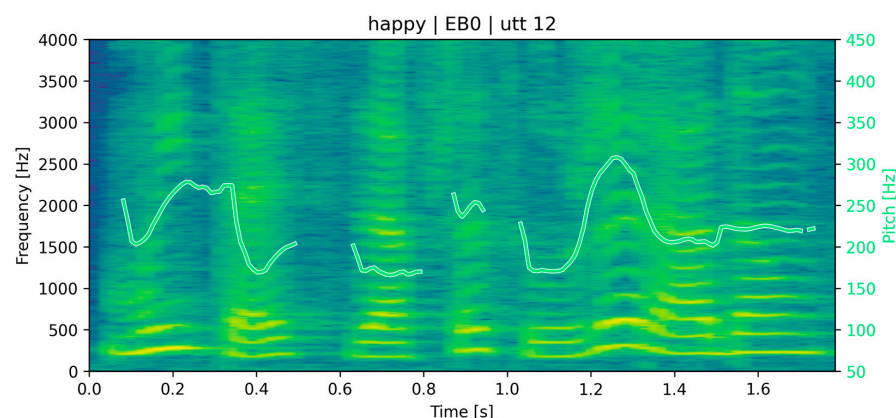


Figure 4. Spectrogram of a happy utterance with pitch contour.

Classical features were classified using multiclass logistic regression [50]. The model employed L2 regularization with a parameter $C = 1.0$, which controls the penalty imposed on large weights and thus reduces the risk of overfitting. Optimization was performed using the L-BFGS algorithm (a quasi-Newton method that approximates the Hessian matrix), under the multinomial setting (i.e., direct modeling of the multi-class problem within a single model, without decomposition into binary tasks).

Training was conducted with the maximum number of iterations set to 4000, ensuring convergence of the solution. To guarantee reproducibility, the `random_state` parameter was fixed at 1234, corresponding to the initialization of the random number generator.

5.2. Emotion Recognition Using the wav2vec 2.0 Model

To evaluate the performance of a deep learning approach, we employed an architecture based on the wav2vec 2.0 base model, using the pretrained facebook/wav2vec2-base checkpoint. The model was originally pretrained in a self-supervised manner on approximately 960 h of the LibriSpeech corpus and was fine-tuned in this study only for the emotion classification task. The wav2vec 2.0 base model is one of the standard approaches to self-supervised speech representation learning [50]. Because it is pretrained on large-scale speech data, it can extract informative acoustic representations without manual feature engineering, making it a suitable reference model for speech emotion classification. The model was evaluated using the LOSO protocol.

A fixed random seed (`random_state = 1234`) was used to make the internal splits reproducible. This is important because different random selections of validation recordings may slightly change training dynamics and the epoch selected as best. Keeping the same seed, together with unchanged hyperparameters across all folds, made the experimental procedure consistent and repeatable.

The recordings were converted to a single-channel (mono) format, resampled to 16,000 Hz. The model architecture consisted of a pretrained feature extraction module, followed by temporal pooling to obtain a single vector representation for each recording. This was followed by a linear layer reducing the dimensionality to 256, a ReLU activation function, a dropout mechanism with a rate of 0.200, and a final linear layer performing classification into three classes.

During fine-tuning, a two-stage training strategy was applied. In the first stage, all parameters of the pretrained component (both the feature extractor and the encoder) were frozen, and only the parameters of the added classification layer (weights and bias) were trained. In the second stage, the last three layers of the encoder, along with their corresponding normalization layers, were unfrozen, allowing their parameters to be updated during training, while the remaining layers remained frozen. This approach enabled gradual adaptation of higher-level representations to the specific task, while preserving the general properties of the pretrained model. The parameters of the applied wav2vec 2.0 model are summarized in Table 3. The objective of this study was not to compare different pretrained wav2vec variants but rather to evaluate a representative self-supervised speech model against classical acoustic features under identical experimental conditions. Other pre-trained variants, such as wav2vec2-large, XLS-R, or multilingual models trained on larger and more diverse speech corpora, may provide improved cross-speaker or cross-lingual representations and constitute an interesting direction for future work.

Table 3. Parameters of the wav2vec 2.0 model used in this study.

Element	Parameter	Value
Base model	architecture	wav2vec2-base
Audio preparation	sampling rate	16,000 Hz
Audio preparation	maximum recording length	6.000 s
Classification head	hidden layer dimension	256
Regularization	dropout	0.200
Tuning	unfrozen encoder layer	3
Learning	learning rate of the head	1×10^{-4}
Learning	core learning rate	1×10^{-5}
Learning	weight_decay	1×10^{-4}
Learning	maximum number of epochs	12
Learning	patience	3
Learning stabilization	max_norm	1.000

5.3. Emotion Recognition Using the WavLM Model

The second self-supervised variant was based on the WavLM-base model [51]. Signal preprocessing followed the same pipeline as for wav2vec 2.0: conversion to mono, resampling to 16,000 Hz, truncation to a maximum length of 6.0 s, and preparation of inputs with sequence-length masking. The classification head (responsible for assigning each recording to a class) remained unchanged.

During fine-tuning of the WavLM model, the last three encoder layers, the normalization layers, and the feature projection layer were unfrozen to allow further training on the analyzed data. The remaining optimization parameters were kept consistent with those used for wav2vec 2.0 to enable a fair comparison between the two models under similar conditions. Table 4 presents the parameters of the WavLM model used in this study.

Table 4. Parameters of the WavLM model used in this study.

Element	Parameter	Value
Base model	architecture	WavLM
Audio preparation	sampling rate	16,000 Hz
Audio preparation	maximum recording length	6.000 s
Classification head	hidden layer dimension	256
Regularization	dropout	0.200
Tuning	unfrozen encoder layer	3
Tuning	additionally unfrozen layers	normalization and feature projection
Learning	learning rate of the head	1×10^{-4}
Learning	core learning rate	1×10^{-5}
Learning	weight_decay	1×10^{-4}
Learning	maximum number of epochs	12
Learning	patience	3
Learning stabilization	max_norm	1.000

5.4. Evaluation Metrics and Computational Environment

Model performance was evaluated using overall accuracy, balanced accuracy (BAcc), macro-F1 score, and detailed per-class metrics, including sensitivity, specificity, and precision. This enabled a comprehensive assessment of classification effectiveness, particularly under class imbalance conditions. Balanced accuracy was defined as the mean of per-class sensitivities, reducing the impact of class size disparities and better reflecting the difficulty of recognizing intermediate classes. Sensitivity was interpreted as the proportion

of correctly identified instances of a given class, specificity as the proportion of correctly rejected instances not belonging to that class, and precision as the proportion of correct predictions among all instances assigned to a given class. Per-class metrics were computed in a one-vs-rest manner based on confusion matrices aggregated across all LOSO runs.

Global metrics were calculated from all predictions aggregated over LOSO runs, while fold-wise metrics were reported as the mean, standard deviation, median, and range of BAcc across individual runs, in order to capture variability in generalization across speakers. Experiments were conducted in an Anaconda environment (env_glos) using Python 3.10.19, NumPy 1.26.4, pandas 2.3.3, scikit-learn 1.7.2, PyTorch 2.2.2, torchaudio 2.2.2, librosa 0.11.0, transformers 4.57.3, and parselmouth 0.4.7. Training of self-supervised models was performed on an NVIDIA GeForce RTX 4070 SUPER GPU. The source code implementing the speaker-independent speech emotion recognition experiments, including the pipeline based on 107 classical acoustic descriptors with logistic regression and linear SVM, as well as fine-tuning procedures for Wav2Vec2 and WavLM models under the leave-one-speaker-out (LOSO) protocol, has been made publicly available on: [<https://github.com/kolodzima/EmotionRecognitionUsingAcousticFeaturesandDeepLearning>] (accessed on 1 July 2026).

6. Results and Discussion

This chapter presents the results obtained for the analyzed methods of affective state recognition and interprets them in terms of classification performance, the usefulness of acoustic features, the structure of the feature space, the explainability of self-supervised models, and their potential for adaptation to individual speakers.

6.1. Comparison of Emotion Recognition Methods

In the first step, the performance of three main approaches to affective state recognition was compared, taking into account both global metrics and the stability of results across individual LOSO runs. Table 5 presents a comparison of classification performance for the considered models using selected quality metrics.

Table 5. Comparison of classification performance across models using selected evaluation metrics.

Model	Global Accuracy	Balanced Global Accuracy	Global Macro-F1	BAcc: Average Over LOSO Runs	Standard Deviation	Median BAcc
Classic model	0.623	0.627	0.584	0.629	0.098	0.607
wav2vec 2.0 model	0.719	0.722	0.695	0.725	0.122	0.778
WavLM model	0.723	0.727	0.710	0.730	0.098	0.752

The classical model achieved a global accuracy of 0.623, a global balanced accuracy of 0.627, and a global macro-F1 score of 0.584. The wav2vec 2.0 model reached a global accuracy of 0.719, a BAcc of 0.722, and a macro-F1 of 0.695, while the WavLM model achieved slightly higher results of 0.723, 0.727, and 0.710, respectively.

The results indicate a clear advantage of deep learning models over the classical acoustic representation. The improvement in balanced accuracy relative to the classical model is not marginal, suggesting that representations learned directly from the signal better preserve information relevant for cross-speaker generalization. Figure 5 illustrates a comparison of the global values of the analyzed metrics for the individual models.

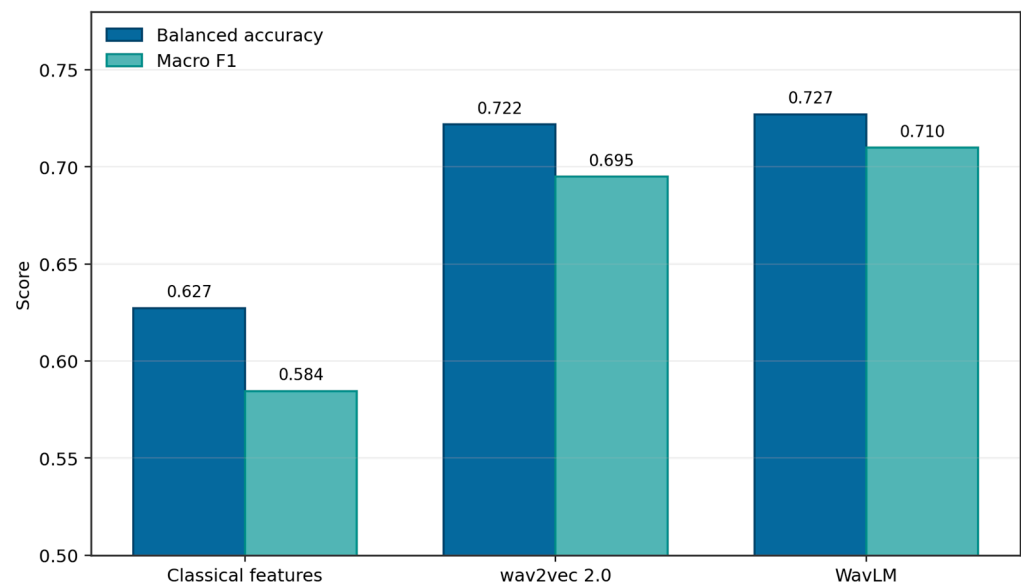


Figure 5. Comparison of models in terms of balanced accuracy and macro-F1 score.

To assess whether the models differ significantly in their performance, the Friedman test was applied ($\chi^2 = 10.467$; $p = 0.015$). This result indicates that, overall, the models are not equivalent, i.e., at least one of them achieves significantly different results. Since the Friedman test does not identify which specific models differ, additional pairwise comparisons were conducted using the Wilcoxon signed-rank test with Holm correction for multiple comparisons. This analysis did not reveal statistically significant differences between specific pairs of models [52,53]. In particular, the difference between WavLM and wav2vec 2.0 was not statistically significant ($p = 0.820$). This means that WavLM should be interpreted as the variant with the highest average performance, but not as a model that is unequivocally superior to wav2vec 2.0 in a statistical sense. It should also be noted that, due to the small number of folds ($n = 9$), the statistical power of these tests remains limited.

At the level of individual speakers, considerable variability in classification accuracy was observed, confirming that the challenge of inter-speaker generalization persists even when using more powerful representations.

Table 6 presents a comparison of classification performance across individual speakers using the balanced accuracy (BAcc) metric for three approaches: the classical model and two self-supervised models (wav2vec 2.0 and WavLM). Overall, deep learning models achieve higher BAcc values than the classical model for most speakers, indicating the advantage of representations learned from large-scale data. The highest score in the entire comparison was obtained for speaker EB0 using wav2vec 2.0 (0.867), while very high performance was also observed for KS0 with WavLM (0.856). High BAcc values are likewise evident for TP0 and MP0, particularly in the self-supervised variants. The classical model achieves relatively competitive results only for selected speakers, notably KS0 (0.789) and TP0 (0.774), where the differences compared to deep models are small. In other cases, its performance is clearly lower, with the weakest results observed for WR0 (0.509) and SM0 (0.519).

Among the self-supervised models, wav2vec 2.0 more frequently attains the highest BAcc values, although this is not consistent across all speakers. WavLM outperforms wav2vec 2.0 for certain speakers, including KD0 (0.707 vs. 0.652) and KS0 (0.856 vs. 0.785), suggesting that its architecture may generalize better in specific cases.

Table 6. Comparison of classification performance across models using the BAcc metric.

Speaker	Classic Model	wav2vec 2.0	WavLM
EB0	0.607	0.867	0.789
IC0	0.607	0.463	0.511
KD0	0.593	0.652	0.707
KS0	0.789	0.785	0.856
MK0	0.596	0.644	0.689
MP0	0.662	0.801	0.798
SM0	0.519	0.729	0.706
TP0	0.774	0.807	0.763
WR0	0.509	0.778	0.752

The lowest performance among the self-supervised models is observed for speaker IC0, where wav2vec 2.0 achieves only 0.463 and WavLM 0.511—both lower than the classical model (0.607). This highlights substantial inter-speaker variability and suggests that deep models may be sensitive to speaker-specific data characteristics. No detailed phonetic or voice-quality analysis was performed for individual speakers. Therefore, the lower performance observed for speaker IC0 cannot be attributed to any specific vocal characteristic and may instead reflect natural inter-speaker variability or differences in the acoustic realization of emotional expressions.

In summary, the results demonstrate a clear advantage of self-supervised models in most cases, while also revealing significant variability across speakers and no consistent dominance of a single deep model under all conditions. Figures 6–8 present the confusion matrices, facilitating direct analysis of misclassification patterns between classes.

The confusion matrices confirm that the primary challenge lies in distinguishing the neutral class from the sad or happy classes. In the classical model, neutral was confused with both sad and happy in a nearly symmetrical manner, reflecting its intermediate position in the feature space.

True class	Sadness	563 73.2%	164 21.3%	42 5.5%
	Neutral	209 25.8%	327 40.4%	273 33.7%
	Happiness	42 5.6%	148 19.8%	559 74.6%
		Sadness	Neutral	Happiness
		Predicted class		

Figure 6. Confusion matrix for the classical approach.

True class	Sadness	613 79.7%	151 19.6%	5 0.7%
	Neutral	193 23.9%	506 62.5%	110 13.6%
	Happiness	22 2.9%	173 23.1%	554 74.0%
		Sadness	Neutral	Happiness
		Predicted class		

Figure 7. Confusion matrix for the wav2vec 2.0 model.

True class	Sadness	623 81.0%	142 18.5%	4 0.5%
	Neutral	139 17.2%	553 68.4%	117 14.5%
	Happiness	6 0.8%	237 31.6%	506 67.6%
		Sadness	Neutral	Happiness
		Predicted class		

Figure 8. Confusion matrix for the WavLM model.

Self-supervised models significantly reduced these misclassifications, although they did not eliminate them entirely. In particular, the WavLM variant exhibits a clear asymmetry: the happy class is still relatively often confused with neutral, whereas sadness is almost never confused with happiness.

Table 7 presents a detailed analysis of classification performance across emotional classes, considering three metrics: sensitivity, specificity, and precision. For the classical model, the highest performance is observed for the sad and happy classes, with sensitivities of 0.732 and 0.746, respectively. At the same time, the neutral class is clearly more difficult to recognize, as reflected by its low sensitivity (0.404) and moderate precision (0.512). This indicates a tendency of the model to frequently miss or misclassify neutral states.

Table 7. Per-class sensitivity, specificity, and precision for the evaluated recognition methods.

Model	Class	Sensitivity	Specificity	Precision
Classic model	sad	0.732	0.839	0.692
Classic model	neutral	0.404	0.795	0.512
Classic model	happy	0.746	0.800	0.640
wav2vec 2.0	sad	0.797	0.862	0.740
wav2vec 2.0	neutral	0.626	0.787	0.610
wav2vec 2.0	happy	0.740	0.927	0.828
WavLM	sad	0.810	0.907	0.811
WavLM	neutral	0.684	0.750	0.593
WavLM	happy	0.676	0.923	0.807

Self-supervised models show an overall improvement over the classical approach, particularly for the neutral class. For wav2vec 2.0, its sensitivity increases to 0.626, and for WavLM to 0.684, suggesting a better ability to capture subtle characteristics of neutral speech. At the same time, both models maintain high performance for the sad class, reaching sensitivities of 0.797 (wav2vec 2.0) and 0.810 (WavLM). The highest precision for the happy class is achieved by wav2vec 2.0 (0.828), along with very high specificity (0.927), indicating strong separation of this class from the others. WavLM achieves comparable specificity (0.923) and high precision (0.807), but with lower sensitivity (0.676), meaning that more instances of this class are missed. WavLM stands out with the highest specificity for the sad class (0.907) and the highest precision (0.811), indicating stable identification of this emotion. In contrast, wav2vec 2.0 provides more balanced performance across classes, particularly for happiness.

In summary, self-supervised models outperform the classical model across most evaluated metrics, with the most notable improvement observed for the neutral class. At the same time, a trade-off between sensitivity and precision is evident depending on the model and class, reflecting differences in the types of errors made by each approach.

6.2. Usefulness of Classical Acoustic Features in Speaker-Independent Conditions

To determine which classical acoustic descriptors carry stable affective information under cross-speaker generalization conditions, a feature selection analysis was conducted using a LOSO scheme. A multiclass logistic regression model with prior standardization was used as the evaluation model, allowing the usefulness of the features themselves to be assessed with minimal influence from classifier complexity. Feature selection was performed using Sequential Forward Selection (SFS).

Applying feature selection led to a slight improvement in performance (BAcc from 0.627 to 0.633), while reducing the number of features to an average of 15.3 (median: 16). This result indicates redundancy in the full representation and the existence of a more informative subset of descriptors, although the selection process itself does not fundamentally alter the nature of the classical model.

A key outcome of the analysis was the stability of feature selection. While no identical set of descriptors appeared across all LOSO runs, a consistent core of features emerged, selected independently of the specific training data composition. The most stable feature was *spectral_flatness_mean* (9/9 folds), while high selection frequencies were also observed for *mfcc9_mean* (7/9), *f0_mean_hz* (6/9), and *inter_onset_std_s*, *mfcc17_std*, and *mfcc19_mean* (each 5/9). A summary of the most frequently selected features is provided in Table 8. These findings indicate that affective information is primarily encoded in the spectral–cepstral structure of the signal, represented by MFCCs and their variability, while prosodic and temporal features play a complementary role. In particular, *spectral_flatness_mean* proved especially important, suggesting the relevance of the balance between tonal and noise components of the signal [54].

Table 8. Most frequently selected features in the nested SFS procedure.

Feature	Number of Runs	Frequency
<i>spectral_flatness_mean</i>	9	1.000
<i>mfcc9_mean</i>	7	0.778
<i>f0_mean_hz</i>	6	0.667
<i>inter_onset_std_s</i>	5	0.556
<i>mfcc17_std</i>	5	0.556
<i>mfcc19_mean</i>	5	0.556
<i>mfcc2_std</i>	4	0.444
<i>pause_ratio</i>	4	0.444
<i>spectral_bandwidth_std_hz</i>	4	0.444

In summary, classical acoustic features do contain useful affective information; however, it is concentrated within a limited and relatively stable subset of descriptors. Feature selection improves the interpretability of the representation but does not lead to a substantial change in model performance.

To provide a more detailed illustration of the properties of the selected features, the *spectral_flatness_mean* descriptor was analyzed. This feature, which characterizes the noisiness of the signal spectrum, demonstrated high stability and consistency across emotional states in the conducted experiments. Lower values of this feature correspond to signals with a more structured, harmonic spectral composition, while higher values indicate a flatter, more noise-like spectrum. As such, it serves as a useful indicator of subtle changes in the spectral organization of speech signals.

The analysis was performed on a dataset of 735 complete triplets of recordings, in which the same speaker produced the same utterance in three emotional states: sad, neutral, and happy. The results revealed a clear, monotonic decreasing trend in *spectral_flatness_mean* values from sadness to happiness (Figure 9). The mean values were 0.2723 for sad, 0.2601

for neutral, and 0.2415 for happy. This indicates that as emotional valence becomes more positive, the signal spectrum becomes, on average, less noise-like and more tonal, which can be interpreted as an increase in the regularity of its harmonic structure.

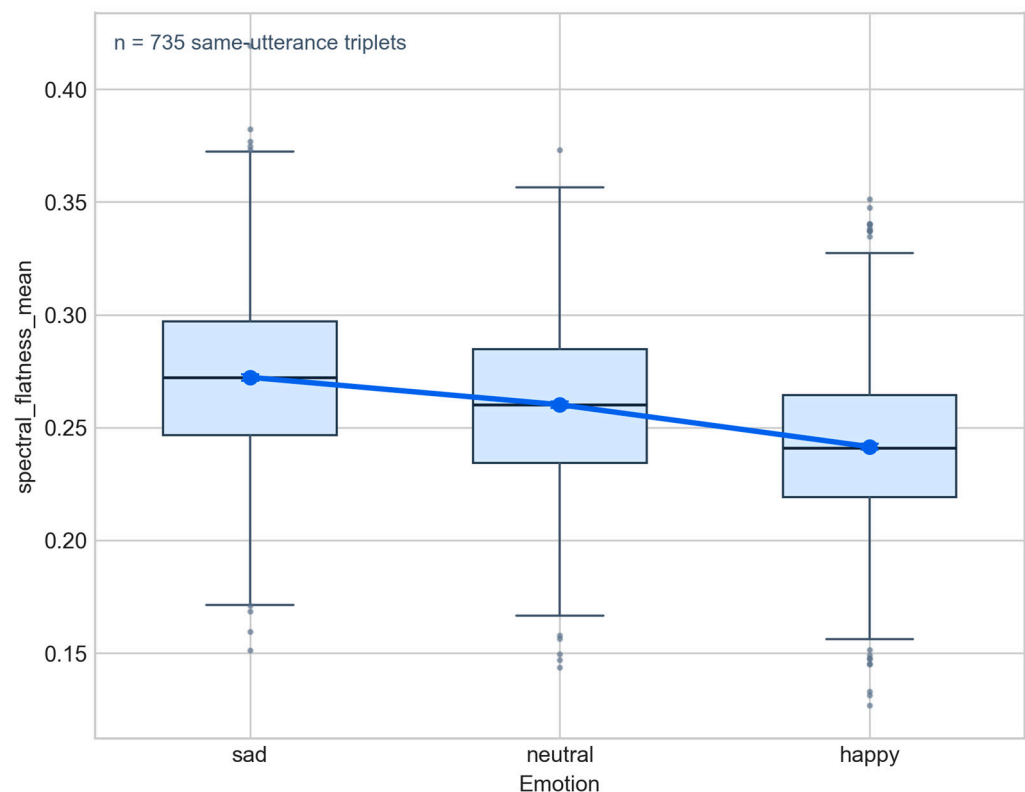


Figure 9. Trend of *spectral_flatness_mean* across the three emotional states.

The observed relationship was consistent and not attributable to random variability across recordings. In 640 out of 735 cases (87.1%), the value of *spectral_flatness_mean* for happiness was lower than for sadness. Additionally, in 339 cases, a fully monotonic pattern (sad > neutral > happy) was observed, indicating that, in a substantial portion of the data, this transition occurs in an orderly manner even when the intermediate state is considered. Importantly, this effect was present for all analyzed speakers, with the proportion of consistent cases at the individual level ranging from 65.6% to 98.9% (mean: 87.7%), confirming its cross-speaker stability.

Differences in *spectral_flatness_mean* across the three emotional states (sad, neutral, happy) were also confirmed statistically. The Friedman test for dependent samples showed a significant global effect ($p \ll 0.001$), while post hoc pairwise comparisons (Wilcoxon signed-rank test and paired *t*-test) revealed significant differences for all emotion pairs. This indicates that the observed decreasing trend is not random but represents a systematic property of the data.

6.3. LDA Analysis and Feature Space Visualization

Linear Discriminant Analysis (LDA) was used as an interpretative tool rather than as a primary classification method [55]. Its purpose was to examine whether reducing the 107-dimensional feature space to two discriminant axes would preserve a clear class structure and whether such a visualization could help explain the misclassification patterns observed in the classification models. A simpler visualization based on two strongly discriminative classical features is shown in Figure 10, while the LDA-based projection is presented in Figure 11.

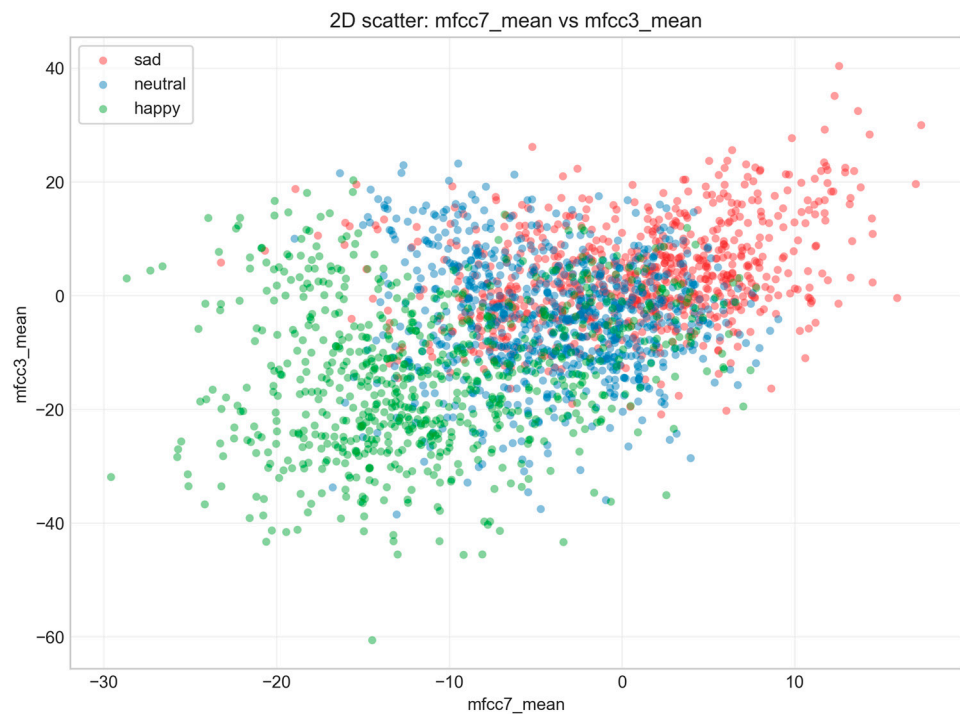


Figure 10. Two-dimensional visualization based on two features: *mfcc7_mean* and *mfcc3_mean*.

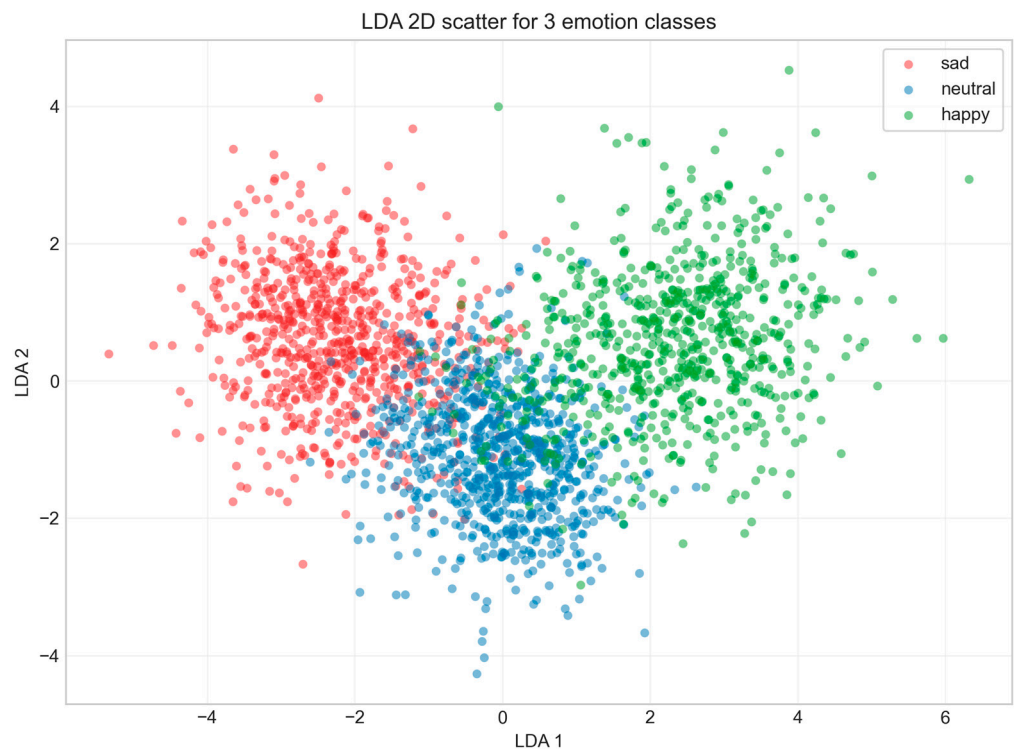


Figure 11. LDA projection for the three emotional classes.

In the analyzed task, the first LDA axis explained 0.836 of the discriminative variance, and the second 0.164. This indicates that most of the class separation can be described by a single dominant direction, although a complete representation requires at least two dimensions. As shown in Figure 11, the sad and happy samples are positioned more toward opposite extremes, while neutral occupies an intermediate region and partially overlaps with both classes. This observation is consistent with the confusion matrix results

and confirms that the main challenge of the task lies not in distinguishing the emotional extremes, but in reliably isolating the intermediate region.

The two-dimensional visualization based solely on two features, shown in Figure 10, leads to a very similar conclusion, although in a less geometrically structured manner. The pair *mfcc7_mean* and *mfcc3_mean* already allows for relatively good separation between sad and happy classes, but the clusters corresponding to the neutral class still overlap with both extremes. This indicates that even simple two-dimensional projections capture the core difficulty of the task: the extreme classes differ clearly in acoustic terms, whereas neutral represents a transitional region that cannot be easily separated by a single decision boundary. The comparison of Figures 10 and 11 is methodologically important. Figure 10 shows that part of the class structure is already present in individual, well-chosen cepstral descriptors, while Figure 11 reveals that a more complete picture emerges only when the full multidimensional feature space is considered. In practice, this means that classical features are interpretable and help explain the sources of separation, but their usefulness increases when they are analyzed jointly rather than in isolation.

However, the interpretative usefulness of LDA did not translate into improved classification performance. In the LDA + logistic regression variant, a global balanced accuracy of 0.612 and a macro-F1 score of 0.564 were obtained—both lower than the results achieved using the full classical feature space without dimensionality reduction. This indicates that while LDA effectively organizes the data for visualization purposes, reducing the representation to two axes leads to a loss of information relevant for classification. From the perspective of this study, this is an important methodological finding: LDA is a valuable tool for interpreting and presenting data structure, but it does not replace the full feature representation in the task of affective state recognition.

To determine whether the lower performance of the classical approach stems primarily from limitations of the feature representation itself or rather from the linearity of the classifier, an additional analysis was conducted under the same LOSO protocol using several variants of support vector machines (SVMs). The evaluated models included a linear SVM, an SVM with a radial basis function (RBF) kernel, as well as second- and third-degree polynomial variants [56]. For each LOSO run, hyperparameters were tuned exclusively on the training speakers using internal GroupKFold validation, and standardization was fitted only on the training data. This ensured that the comparison followed the same rigorous speaker-independent protocol as the main part of the study.

The results, summarized in Table 9, show that none of the tested SVM variants improved upon the classification performance achieved by logistic regression. The best-performing SVM was the linear variant, yet its global balanced accuracy remained lower than that of the logistic model. Nonlinear variants not only failed to close the gap to the self-supervised models, but in most cases led to further decreases in BAcc and macro-F1, as well as greater variability across LOSO runs. This suggests that, within the analyzed 107-dimensional classical feature space, more complex decision boundaries do not compensate for the limitations of the representation itself.

This finding has important implications for the overall interpretation of the study. The additional analysis indicates that the advantage of wav2vec 2.0 and WavLM over the classical approach is not merely due to the use of an overly simple linear classifier. If even more flexible decision boundaries do not improve performance, the most plausible limitation lies in the feature representation itself. In other words, self-supervised models gain their advantage not only from classification, but primarily from a richer and more generalizable representation of the signal.

Table 9. Additional comparison of classifiers for the 107 classical features.

Classifier	Global Accuracy	Balanced Global Accuracy	Global Macro-F1	BAcc: Average over LOSO Runs	Standard Deviation: BAcc
Logistic regression	0.623	0.627	0.584	0.629	0.098
Linear SVM	0.609	0.615	0.575	0.616	0.103
Quadratic SVM	0.568	0.578	0.518	0.583	0.139
Cubic SVM	0.577	0.584	0.523	0.589	0.169

At the same time, this analysis does not imply that all possible classification variants for classical features have been exhausted. While a representative set of linear and nonlinear classifiers was evaluated, ensemble methods and more advanced tabular deep learning architectures were not considered. Nevertheless, these additional results reinforce the main conclusion of the study: the difference between the classical approach and deep models is primarily representational rather than purely due to the choice of classifier.

6.4. Deep Learning Models and the Issue of Explainability

The results obtained with wav2vec 2.0 and WavLM indicate that the advantage of self-supervised learning becomes most apparent in areas where the classical approach reaches the limits of interpretable, hand-crafted descriptors. Both models improved the recognition of the neutral class, i.e., the region of the emotional space that was most ambiguous for the classical model. This suggests that self-supervised models leverage information distributed simultaneously across time, spectral structure, and articulatory–prosodic dynamics—information that cannot be easily captured by a small set of global statistical features.

At the same time, higher performance does not automatically imply better explainability. In the classical approach, it is possible to directly identify which features drive class separation and relate them to phonetic intuition, such as higher or more variable F0, changes in the spectral envelope, phonation stability, or the proportion of harmonic components. In wav2vec 2.0 and WavLM, such interpretation is no longer straightforward, as the input representation is formed within a deep architecture and does not have a simple correspondence to individual acoustic descriptors.

For this reason, the explainability of self-supervised models was analyzed using two indirect methods. The first involved so-called layer-wise probing, where representations from each layer of the model were extracted and evaluated in terms of how well a simple linear classifier could decode affective information from them [57]. The results of this analysis are shown in Figure 12.

In both models, the earliest layers performed the worst, while the highest balanced accuracy and macro-F1 values were observed in the middle layers, around layer 6. For wav2vec 2.0, the best-performing layer achieved a mean BAcc of 0.779, and for WavLM, 0.807. This indicates that the information most relevant for affective state recognition is not concentrated at the input level or solely in the final layer, but emerges most clearly in intermediate representations.

Importantly, not only is the peak of the curve in Figure 12 relevant, but also its overall shape. In both models, the improvement in performance from the early to the middle layers is clear and systematic, suggesting a gradual transition from representations closely tied to local acoustic structure toward representations that are more informative for affective content. Beyond this point, the curves flatten or slightly decline, which can be interpreted as an indication that higher layers become increasingly task-specific and less suitable for simple linear decoding of emotions. WavLM exhibits a slightly higher and more stable

profile in this regard than wav2vec 2.0; however, this difference is quantitative rather than qualitative.

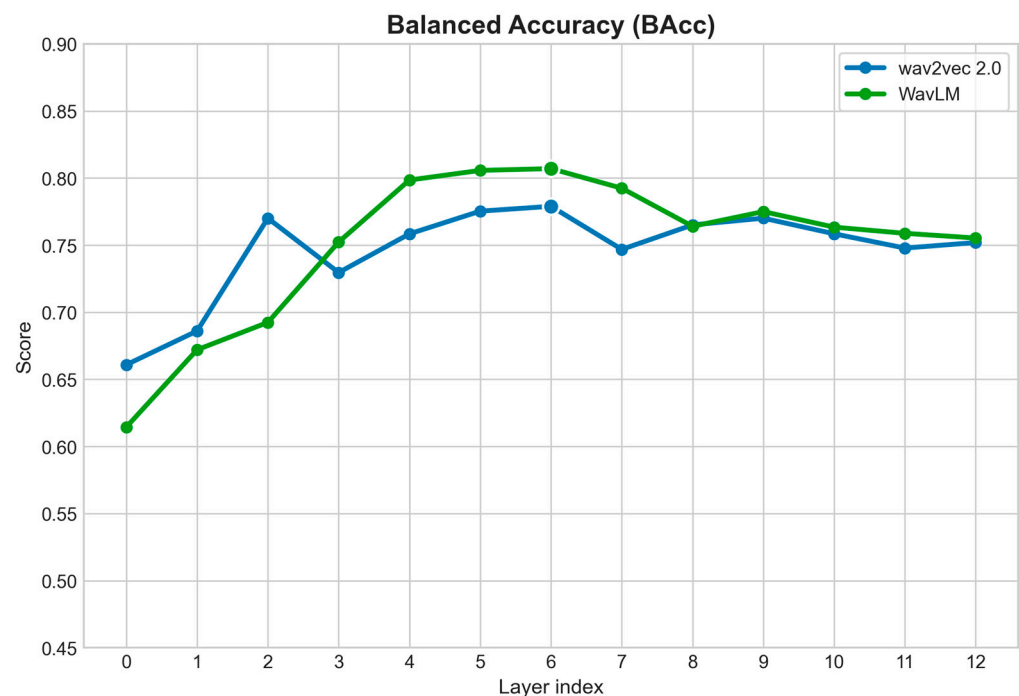


Figure 12. Layer-wise probing results: balanced accuracy (BAcc) as a function of layer index for wav2vec 2.0 and WavLM.

The second analysis was local in nature and relied on the Integrated Gradients method, which makes it possible to estimate which parts of the signal most strongly influenced the model's decision for a given sample. This method is based on analyzing the gradient of the model output with respect to the input, indicating how changes in the signal at a given point affect the classification result [58]. Instead of computing gradients for a single input, Integrated Gradients averages them along a continuous path from a reference signal (e.g., silence) to the actual recording, providing a more stable and reliable estimate of the contribution of individual signal segments. For qualitative analysis, Integrated Gradients was used to identify the segments of the speech signal with the greatest influence on classification outcomes. The method assigns each segment a contribution score by integrating gradients along the path from a baseline signal to the input recording. Figures 13–15 present three representative examples corresponding to the sad, neutral, and happy emotions. In each case, the same input signal was analyzed using two self-supervised models: WavLM and wav2vec 2.0. The upper panel corresponds to WavLM, and the lower panel to wav2vec 2.0. The gray curve represents the normalized acoustic waveform, while the colored overlay shows the distribution of Integrated Gradients values, interpreted as the relative contribution of different temporal segments to the model's decision. Regions labeled T1, T2, and T3 indicate the three most influential segments for a given sample and model.

The visualizations show that both models do not rely uniformly on the entire utterance but instead focus on selected segments with high diagnostic value. For the sad emotion, both models highlight similar key regions, particularly in the early and middle parts of the recording, suggesting agreement on the most relevant acoustic cues. For neutral, the alignment of the most important regions is also high, indicating stable and consistent recognition by both models. For happiness, a similar concentration of attribution is observed in the central part of the signal, although slight differences appear in the location of the third

most important region, which may reflect differences in how later parts of the utterance are weighted.

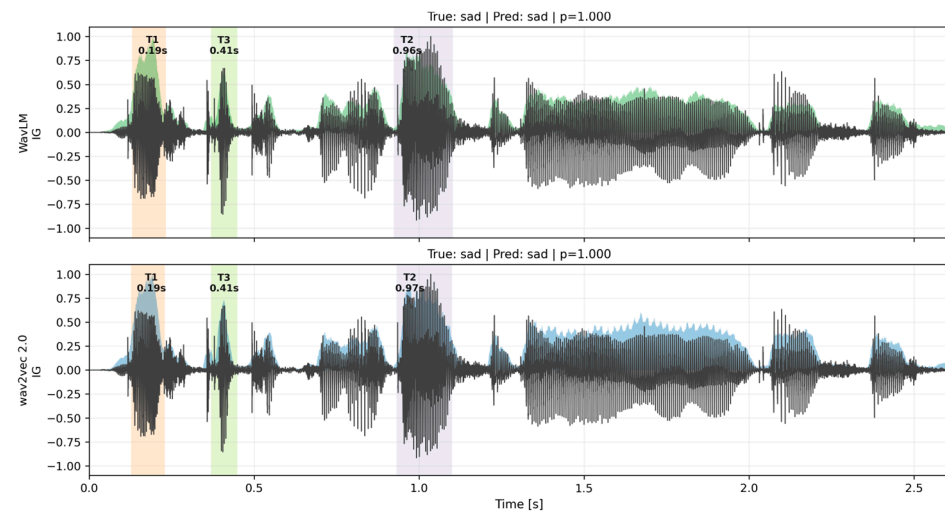


Figure 13. Integrated Gradients for the sad emotion (WavLM: top panel; wav2vec 2.0: bottom panel).

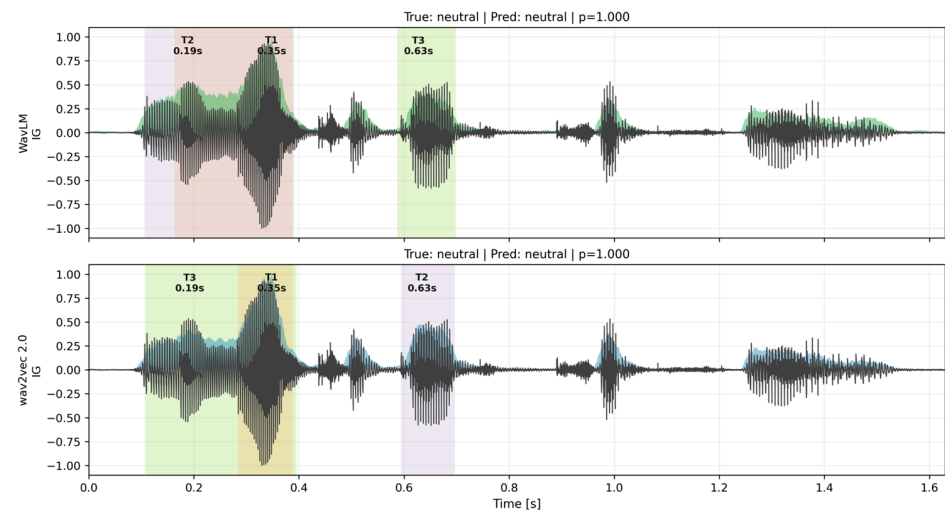


Figure 14. Integrated Gradients for the neutral emotion (WavLM: top panel; wav2vec 2.0: bottom panel).

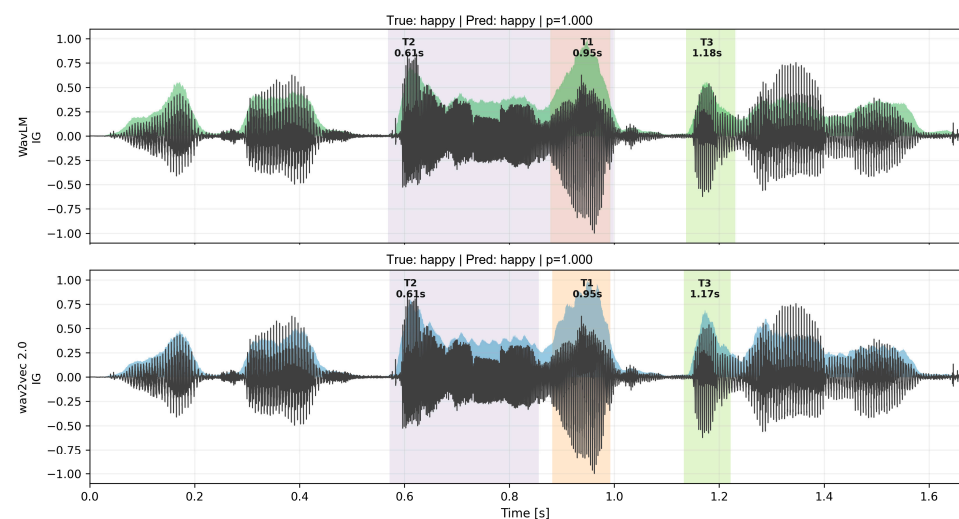


Figure 15. Integrated Gradients for the happy emotion (WavLM: top panel; wav2vec 2.0: bottom panel).

Overall, the comparison suggests that WavLM and wav2vec 2.0 rely on partially overlapping temporal segments as carriers of emotional information, confirming that certain prosodic and energy-related features are consistently identified as diagnostic. At the same time, the attribution profiles are not identical, indicating differences in how the models distribute attention across the signal. In particular, WavLM tends to assign greater importance to temporally extended regions, whereas wav2vec 2.0 sometimes emphasizes more localized peaks. These findings suggest that both models achieve correct classification based on partially similar, but not identical, affective representations, and that interpretability analysis provides additional insight into how emotional information is encoded in self-supervised models.

This form of explainability is functional rather than semantic. It does not allow one to conclude that a specific hidden unit corresponds, for example, to pitch or HNR, but it does make it possible to identify at which stage of the network affective information becomes most linearly accessible, as well as which segments of the signal contribute to the classification decision. When this observation is combined with classical feature analysis, the overall picture remains consistent: self-supervised models do not contradict the acoustic interpretation derived from MFCCs, delta MFCCs, pitch, and voice quality features, but rather extend it by capturing richer time–frequency patterns.

From the perspective of practical applications, the most important finding is that self-supervised models distinguish the neutral class from the extreme classes more effectively than the classical approach. In depression monitoring, the key value does not lie in assigning a single label to an individual recording, but in tracking changes in affective expression over time. If, for a given individual, the proportion of utterances resembling sad increases, intonational dynamics decrease, or the range of affective expression narrows, this may serve as an additional signal supporting the assessment of mood deterioration. This highlights the practical relevance of research on speech affect: it provides a foundation for developing tools that support longitudinal monitoring, rather than merely performing one-off emotion classification.

6.5. Speaker-Specific Adaptation and Its Impact on Classification Performance

An additional experiment on speaker-specific adaptation was conducted to assess whether, after obtaining a limited number of labeled recordings from an end user, it is possible to significantly improve affective state recognition for that individual [59]. This scenario does not replace rigorous speaker-independent evaluation but reflects an important practical application. In long-term monitoring, a system may begin with a short calibration session and subsequently compare incoming recordings against an individualized vocal profile.

The experiment was performed for all speakers in the corpus. For each speaker, approximately half of the recordings from each class were assigned to the adaptation set, while the remaining samples were used as the final test set, resulting in an approximate 50/50 split within each class. In practice, the number of samples varied across speakers, so the split was adjusted to maintain as balanced a division as possible between adaptation and test data.

In the classical approach, the adaptation data were simply added to the training set. For wav2vec 2.0 and WavLM, the checkpoint obtained from the corresponding LOSO run was used, followed by brief additional fine-tuning exclusively on the adaptation recordings of the given speaker. This setup should be interpreted as a LOSO-based protocol extended with a speaker adaptation stage, rather than as an alternative to a purely speaker-independent evaluation.

In the table and figures presented in this section, the original LOSO results are used as the reference point, rather than the results obtained on the test subset prior to adaptation. This approach allows for a direct assessment of the performance gain resulting from personalization relative to a fully speaker-independent setting. It should be noted, however, that this comparison involves two different evaluation scenarios: a global LOSO baseline and results obtained after adaptation, where the model is additionally fine-tuned for each speaker using their adaptation data and then evaluated on the remaining recordings of that same individual. Therefore, this comparison should be interpreted as an illustration of the potential benefits of personalization, rather than a strict comparison of two identical experimental procedures.

Table 10 and Figure 16 present the balanced accuracy in two settings: the original LOSO scheme and a model adaptation setting using a portion of data from the target speaker. In the latter case, approximately 50% of the recordings from each class were used for fine-tuning, while the remaining samples served as an independent test set. This visualization facilitates interpretation by showing the absolute performance levels in both scenarios. In this experiment, improvements were observed across all model families, although their magnitude varied.

Table 10. Comparison of the original LOSO results with those obtained after bidirectional speaker adaptation.

Model	BAcc in Original LOSO	BAcc After Adaptation	Delta BAcc	Macro-F1 in Original LOSO	Macro-F1 After Adaptation
Logistic regression	0.627	0.831	0.204	0.584	0.827
LDA + Logistic regression	0.612	0.815	0.203	0.564	0.810
Linear SVM	0.615	0.824	0.209	0.575	0.821
Quadratic SVM	0.578	0.895	0.317	0.518	0.894
Cubic SVM	0.584	0.902	0.318	0.523	0.898
SVM RBF	0.569	0.903	0.334	0.505	0.900
wav2vec 2.0	0.722	0.899	0.177	0.695	0.901
WavLM	0.727	0.855	0.128	0.710	0.859

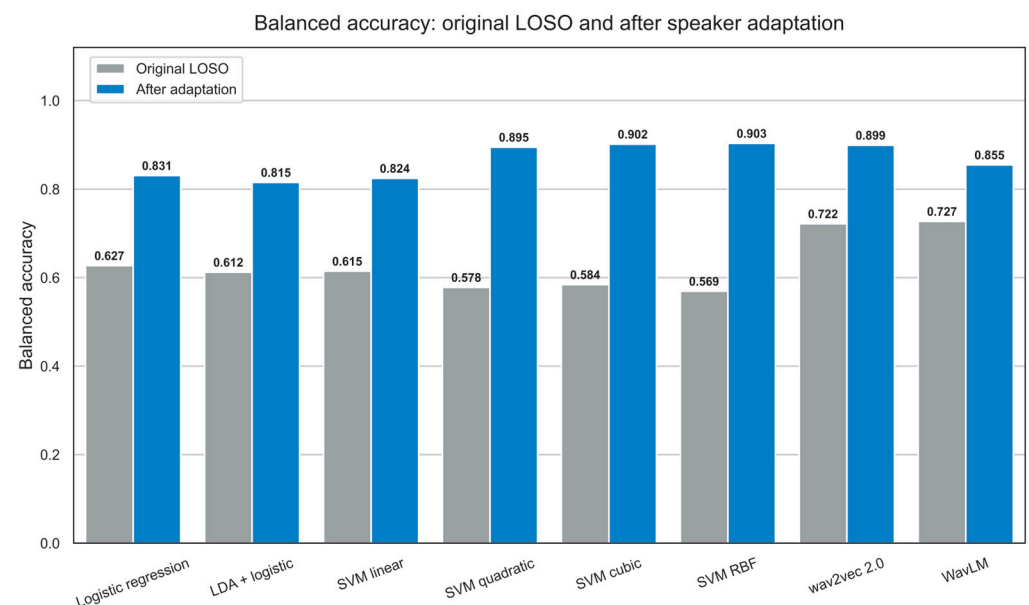


Figure 16. Comparison of BAcc in the original LOSO setting and after bidirectional speaker adaptation.

The largest average gains in BAcc were achieved by nonlinear SVM variants (from +0.317 to +0.334). This demonstrates that classical acoustic descriptors become substantially more effective when the system can tailor decision boundaries to an individual speaker's vocal profile. Among self-supervised models, the average improvement was +0.177 for wav2vec 2.0 and +0.128 for WavLM. Thus, deep models maintain an advantage already at the baseline LOSO level, but they also benefit from personalization. In other words, in the speaker-adapted setting, the gain is not driven solely by the quality of the representation, but also by its alignment with speaker-specific characteristics. This finding is particularly relevant for applications involving longitudinal monitoring of an individual's emotional state. In practice, a more realistic scenario than one-off, fully speaker-independent classification is one in which a reference database is available for a given user, collected at different time points—for example, during earlier assessments or periods of relative stability. The system can then not only classify the current affective state but also detect deviations from the individual's vocal profile. The results suggest that such a personalized approach may increase sensitivity to subtle temporal changes and thereby support long-term mood monitoring.

At the same time, caution is required in interpreting these findings. Speaker adaptation does not replace speaker-independent evaluation and does not imply that affect recognition has become an easy problem in general. Rather, it shows that once some user-specific data are available, the decision space becomes significantly more stable. Therefore, this experiment is best viewed as a translational result: it strengthens the argument that speech-based affect analysis may be particularly useful for monitoring the same individual over time, rather than solely for one-off, population-level emotion recognition.

6.6. Additional Comparison Using CNN Models

To extend the main comparative analysis, a group of neural models trained from scratch was included, serving as an intermediate point between classical acoustic features and large self-supervised models. This experimental design had important methodological implications, as it enabled several key questions to be addressed simultaneously. First, whether a relatively lightweight end-to-end model can serve as a reliable baseline for speech emotion recognition. Second, whether the choice of input representation significantly affects the performance of models trained from scratch. Third, how large the performance gap remains between such models and self-supervised learning approaches when stronger neural baselines are considered.

Three deep learning variants were evaluated: a CNN operating on log-Mel spectrograms, a CNN using STFT spectrograms, and a temporal model (Time CNN) operating directly on the waveform [60]. All CNN models were trained and evaluated using the same LOSO protocol. Additionally, 12% of the training data was set aside as a validation set using class-stratified sampling. Model selection was based on validation performance rather than test results.

All recordings underwent identical preprocessing. Audio signals were converted to mono, resampled to 16 kHz. Longer recordings were truncated, while shorter ones were zero-padded to a fixed length of 64,000 samples. Each signal was normalized with respect to its maximum absolute amplitude. Using a fixed input length simplified training and batching, while ensuring a consistent input format.

Balanced accuracy was used as the primary evaluation metric due to class imbalance. Standard accuracy can be dominated by more frequent classes, whereas balanced accuracy treats all classes equally by averaging per-class sensitivities. Accuracy and macro-F1 were also reported, but balanced accuracy served as the main comparison criterion.

The log-Mel CNN model used logarithmic Mel spectrograms as input, representing the time–frequency energy distribution while incorporating perceptual properties of human hearing. A Fourier transform of length 1024 was applied, determining frequency resolution. The analysis window length was 960 samples, providing a balance between time and frequency resolution, with a hop size of 240 samples controlling temporal sampling density. The spectrum was projected onto 64 Mel bands within the 20–8000 Hz range. Logarithmic scaling provided dynamic range compression, stabilizing training. The representation was further converted to decibel scale ($top_db = 80$) and normalized to the $[0, 1]$ range. The architecture consisted of four 2D convolutional blocks with increasing channel sizes: 24, 48, 96, and 128. Each block included 3×3 convolutions, BatchNorm2D, and ReLU activation, with MaxPool2D and Dropout2D applied in the first three blocks.

At the output stage, the feature maps produced by the final convolutional block were transformed into a compact recording-level representation using two parallel adaptive global pooling operations: AdaptiveAvgPool2D (1,1) and AdaptiveMaxPool2D (1,1). Both operations reduced each feature map to a single scalar value, independently of its original time-frequency dimensions. Adaptive average pooling computed the mean activation across the entire feature map, capturing the overall presence of a learned acoustic pattern, whereas adaptive max pooling preserved the strongest activation corresponding to the most salient occurrence of that pattern. Consequently, the model retained both global and peak activation information from each learned feature map. The outputs of the average and max pooling operations were concatenated, resulting in a 256-dimensional feature vector. This representation was subsequently passed to a fully connected classifier with the structure $256 \rightarrow 128 \rightarrow 3$. The first linear layer reduced the feature dimensionality from 256 to 128 units and was followed by ReLU activation and dropout with a probability of 0.35 to reduce overfitting. The final linear layer mapped the 128-dimensional representation to the three target emotion classes.

The model was trained for 14 epochs using the AdamW optimizer with a learning rate of 8×10^{-4} and weight decay of 10^{-4} . Weighted CrossEntropyLoss was applied to compensate for class imbalance, while label smoothing with a factor of 0.05 was used to reduce overconfident predictions. During training, SpecAugment was additionally employed by masking selected temporal and frequency regions of the spectrograms, encouraging the network to learn more robust and generalized acoustic representations.

The STFT CNN differed only in its input representation. It used STFT spectrograms computed with a 512-point Fourier transform, a window length of 400 samples, and a hop size of 160 samples [61]. The spectrum was represented in terms of power (squared magnitude), followed by logarithmic transformation and normalization. This resulted in 257 linearly spaced frequency bins. The architecture was kept as similar as possible to the log-Mel CNN to isolate the effect of input representation. The model was trained for 12 epochs using AdamW and a weighted CrossEntropyLoss to account for class imbalance. Label smoothing of 0.05 was applied to reduce overconfident predictions, while a ReduceLROnPlateau scheduler automatically lowered the learning rate when validation performance stopped improving. Time-domain augmentation was used during training to improve robustness to variations in the raw waveform signal.

The Time CNN model operated directly on the waveform [62]. The first layer was defined as Conv1D (1, 24, $kernel_size = 21$, $stride = 5$, $padding = 10$), followed by BatchNorm1D and GELU activation. Four convolutional blocks with residual connections were used, with channel sizes of 48, 96, 128, and 192. The main convolutions used kernel sizes of 9, 7, 5, and 5 with stride 4, while shortcut connections were implemented via 1×1 projections. At the output stage, the final one-dimensional feature maps were summarized using two parallel adaptive pooling operations: AdaptiveAvgPool1D and AdaptiveMaxPool1D. In this case,

adaptive average pooling computed the mean activation across the temporal dimension, capturing the overall presence of each learned waveform pattern, whereas adaptive max pooling retained the strongest activation, preserving the most salient occurrence of that pattern in the signal. The outputs of both pooling branches were concatenated, producing a 384-dimensional representation of each recording.

This representation was passed to a fully connected classifier with the structure $384 \rightarrow 160 \rightarrow 3$. The first linear layer reduced the pooled feature vector from 384 to 160 units, followed by GELU activation and dropout with a probability of 0.35 to reduce overfitting. The final linear layer mapped the 160-dimensional representation to the three target emotion classes.

The model was trained for 12 epochs using the AdamW optimizer. A weighted CrossEntropyLoss was applied to account for class imbalance, and label smoothing with a value of 0.05 was used to reduce overconfident predictions. Time-domain augmentation was additionally used during training to improve robustness to small variations in the raw waveform signal.

Figure 17 presents a comparison of all methods in terms of balanced accuracy. The best results were achieved by self-supervised models: wav2vec 2.0 (BAcc = 0.725) and WavLM (BAcc = 0.730). The classical model based on hand-crafted features achieved BAcc = 0.629. Among models trained from scratch, Time CNN performed best (BAcc = 0.623), while spectrogram-based models achieved lower results: log-Mel CNN (0.522) and STFT CNN (0.539).

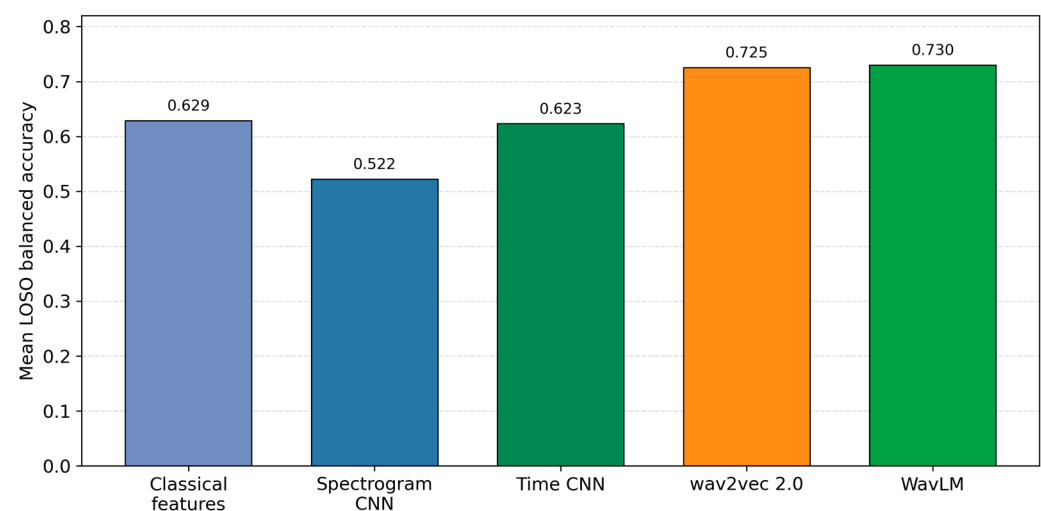


Figure 17. Comparison of mean LOSO balanced accuracy across selected evaluated approaches. Pretrained SSL models achieved the highest performance, with WavLM slightly outperforming wav2vec 2.0. Among models trained from scratch, Time CNN obtained the best result.

The log-Mel CNN achieved the lowest performance, while STFT CNN performed slightly better, suggesting that linear-frequency representations retain additional useful information. However, both spectrogram-based models remained clearly inferior to Time CNN. This result indicates that, for compact models trained from scratch, direct waveform modeling may be more effective than relying on simple time–frequency representations.

The sequential analysis shows that classical acoustic features remain a strong baseline. Simple CNN models operating on spectrograms do not reach comparable performance. Only the Time CNN significantly narrows this gap, while SSL models provide the largest improvement due to rich representations learned during pretraining.

Including additional CNN models reinforces the main conclusions of the study. The differences between methods are driven not only by classifier architecture, but primarily by

the quality of the input representation and the availability of pretraining. The Time CNN represents an important intermediate point, approaching classical methods but clearly underperforming compared to SSL models, confirming their advantage stemming from self-supervised learning.

6.7. Computational Complexity Analysis

In addition to classification performance, the computational cost of the analyzed methods is an important aspect of their evaluation. In practical applications of speech emotion recognition systems, both the time required to classify a single recording and the time needed to train the model are relevant. Therefore, the analysis considered two scenarios: the classification time for a single audio file and the training time for a single model.

The classification time for a single file included the complete preparation of the input data required by a given method, followed by prediction using a previously loaded model. For the classical methods, the measurement included the automatic extraction of 107 acoustic descriptors and classification using either logistic regression or a linear SVM. For the CNN models, the measured time included preparation of the corresponding signal representation, namely the raw waveform, STFT spectrogram, or log-Mel spectrogram, followed by a single forward pass through the neural network. For the self-supervised wav2vec 2.0 and WavLM models, the measurement included preparation of the input signal and a single forward pass through the transformer-based model. Classification time was reported as the median of multiple measurements to reduce the influence of one-time GPU initialization overhead.

The training time for a single model referred to one training run in a speaker-independent setting, in which the data from one speaker served as the test set, while recordings from the remaining speakers were used to train the model. The total time required for the full LOSO procedure was not reported, because it represents repeated executions of the same training process for successive held-out speakers and does not reflect the actual cost of preparing a model in a typical practical application. Table 11 presents a comparison of the classification time for a single recording and the training time for a single model across all analyzed methods.

Table 11. Comparison of the classification time for a single audio file and the training time of individual methods.

Method	Classification Time	Training Time
Classical 107 features + logistic regression	29.3 ms	66.3 s
Classical 107 features + linear SVM	29.3 ms	79.0 s
Time-domain CNN	1.9 ms	29.9 s
Log-Mel CNN	3.2 ms	23.5 s
STFT CNN	2.9 ms	105.4 s
wav2vec 2.0	10.6 ms	188.9 s
WavLM	11.9 ms	218.6 s

The obtained results show that the analyzed methods differ substantially in their computational profiles. In the classical approach, the extraction of 107 acoustic descriptors accounts for the largest part of the classification time for a single recording. The prediction performed by the linear classifier itself takes less than 1 ms; however, the preceding computation of parameters describing, among others, fundamental frequency, jitter, shimmer, HNR, energy, pauses, speech rate, spectral features, and MFCC coefficients increases the total classification time for a single recording to approximately 29 ms.

The shortest inference times were obtained for the CNN models, for which the classification of a single recording required between 1.9 and 3.2 ms. For these models, the

computational cost is mainly associated with preparing the input representation and performing a single forward pass through a relatively small neural network. The fastest solution was the Time CNN model, which operates directly on the time-domain waveform.

The self-supervised wav2vec 2.0 and WavLM models required longer inference times, namely 10.6 ms and 11.9 ms, respectively. This is due to the greater complexity of the transformer-based architecture and the substantially larger number of parameters. Nevertheless, the classification time for a single recording still remains within several milliseconds, which makes these models suitable for near-real-time applications, provided that the model has already been loaded into memory.

Even larger differences were observed during model training. The shortest training time was obtained for the Log-Mel CNN model (23.5 s), followed by the Time CNN model (29.9 s). Classical logistic regression required approximately 66 s, with acoustic feature extraction being the dominant component of this time. The variant based on a linear SVM was slightly more computationally demanding (79 s) because of the hyperparameter selection procedure. The highest computational cost was observed for the self-supervised models, whose fine-tuning required approximately 3.1 min for wav2vec 2.0 and 3.6 min for WavLM.

Overall, the analysis shows that the choice of model involves a trade-off between computational cost, interpretability, and classification performance. Classical methods remain easy to interpret and relatively inexpensive to train; however, their classification time is dominated by the extraction of hand-crafted acoustic features. CNN models provide the shortest inference times, but their classification performance remains lower than that of the self-supervised models. In contrast, wav2vec 2.0 and WavLM require the greatest computational effort during fine-tuning, but they provide the best classification performance while maintaining short classification times for individual recordings after training has been completed.

6.8. Limitations of the Study

The interpretation of the results should be approached with caution. First, the study is based on a small number of speakers and a single corpus, which limits the ability to draw broad conclusions about generalization beyond the nEMO conditions. The findings have not been validated on an external dataset, restricting their cross-corpus generalizability. Second, the data are emotional rather than clinical in nature; therefore, any potential clinical applications remain speculative and would require separate validation on spontaneous, longitudinal, and clinically annotated data. Third, the feature importance analysis was exploratory, due to the lack of full independence between observations. Fourth, although the additional comparison included several SVM variants, this study was not intended to provide an exhaustive comparison of all machine learning architectures for speech emotion recognition. Alternative approaches, including other tabular classifiers, radial basis function networks, probabilistic neural networks, extreme learning machines, transformer-based models beyond those investigated here, graph neural networks, and hybrid architectures, were beyond the scope of this work.

Finally, internal validation for the self-supervised models was conducted at the sample level rather than at the speaker level.

7. Conclusions

The conducted study demonstrates that, in the task of recognizing three affective states in Polish speech, self-supervised learning models clearly outperform approaches based on classical acoustic features. The wav2vec 2.0 and WavLM models achieved higher balanced

accuracy and macro-F1 scores under the LOSO protocol, indicating better cross-speaker generalization. The differences between the two models were not statistically significant.

The neutral class proved to be the most challenging, as it occupies an intermediate position between the sad and happy states, making it difficult to separate reliably. Self-supervised models improved performance in this region, which is particularly important from a practical application perspective.

Analysis of classical features showed that MFCCs and their derivatives provide the greatest discriminative power, highlighting the dominant role of spectral–cepstral information. Pitch-related features and voice quality measures played a complementary role. Increasing classifier complexity did not lead to performance improvements, suggesting that the main limitation lies in the representation rather than in the decision model.

Although this work compares representative classical and deep learning approaches, it does not exhaust the wide range of machine learning and neural network architectures proposed for speech emotion recognition. Future studies may extend this comparison by including alternative models, such as radial basis function networks, probabilistic neural networks, extreme learning machines, and other specialized architectures evaluated under the same speaker-independent protocol.

The speaker adaptation experiment demonstrated a clear performance gain across all approaches, emphasizing the importance of personalization in monitoring systems. However, the results should be interpreted with caution due to the limited number of speakers, although they point to the potential of speech as a non-invasive indicator of affective changes.

Author Contributions: Conceptualization, M.K., A.M. and T.R.; methodology, M.K., A.M. and T.R.; software, M.K.; validation, M.K., A.M. and T.R.; formal analysis, M.K., A.M. and T.R.; investigation, M.K. and A.M.; resources, M.K. and A.M.; data curation, M.K., A.M. and T.R.; writing—original draft preparation, M.K. and A.M.; writing—review and editing, M.K., A.M. and T.R.; visualization, M.K. and A.M.; supervision, M.K., A.M. and T.R.; project administration, M.K.; funding acquisition, M.K. and A.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The acoustic feature dataset supporting this study has been deposited in Zenodo and is available at: [<https://doi.org/10.5281/zenodo.20065078>]. The repository contains 107 acoustic features extracted from 2327 Polish speech utterances, together with metadata, feature descriptions, speaker-level files, and utterance-level JSON records. The source code implementing the speaker-independent speech emotion recognition experiments, including the pipeline based on 107 classical acoustic descriptors with logistic regression and linear SVM, as well as fine-tuning procedures for Wav2Vec2 and WavLM models under the leave-one-speaker-out (LOSO) protocol, has been made publicly available on: [<https://github.com/kolodzima/EmotionRecognitionUsingAcousticFeaturesandDeepLearning>] (accessed on 1 July 2026).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACC	Accuracy
AdamW	Adaptive Moment Estimation with weight decay
BACC	Balanced accuracy
CNN	Convolutional Neural Network
F0	Fundamental frequency

F1	F1-score
FDR	False Discovery Rate
FFT	Fast Fourier Transform
GELU	Gaussian Error Linear Unit
GroupKFold	Grouped k-fold cross-validation
HNR	Harmonics-to-noise ratio
IG	Integrated Gradients
jitter	Cycle-to-cycle variation in fundamental frequency
LDA	Linear Discriminant Analysis
LOSO	Leave-one-subject-out
LTAS	Long-Term Average Spectrum
mean	Arithmetic mean
Mel	Mel-scale spectrogram
MFCC	Mel-frequency cepstral coefficients
one-vs-rest	One-versus-rest classification strategy
pitch	Fundamental frequency contour
RBF	Radial Basis Function
ReLU	Rectified Linear Unit
RMS	Root Mean Square
rolloff	Spectral roll-off
SD	Standard deviation
shimmer	Cycle-to-cycle variation in amplitude
SpecAugment	Spectrogram augmentation method
SSL	Self-Supervised Learning
STFT	Short-Time Fourier Transform
SVM	Support Vector Machine
Time CNN	Temporal Convolutional Neural Network
top_db	Threshold in decibels for silence detection
ZCR	Zero-Crossing Rate
Δ MFCC	Delta Mel-frequency cepstral coefficients

References

1. Ververidis, D.; Kotropoulos, C. Emotional Speech Recognition: Resources, Features, and Methods. *Speech Commun.* **2006**, *48*, 1162–1181. [\[CrossRef\]](#)
2. El Ayadi, M.; Kamel, M.S.; Karray, F. Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases. *Pattern Recognit.* **2011**, *44*, 572–587. [\[CrossRef\]](#)
3. Rybka, J.; Janicki, A. Comparison of Speaker Dependent and Speaker Independent Emotion Recognition. *Int. J. Appl. Math. Comput. Sci.* **2013**, *23*, 797–808. [\[CrossRef\]](#)
4. Yuan, Z.; Philip Chen, C.L.; Li, S.; Zhang, T. Disentanglement Network: Disentangle the Emotional Features from Acoustic Features for Speech Emotion Recognition. In *Proceedings of the ICASSP 2024—2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024*; IEEE: Piscataway, NJ, USA, 2024; pp. 11686–11690.
5. Portal, F.; De Lope, J.; Graña, M. A Performance Benchmarking Review of Transformers for Speaker-Independent Speech Emotion Recognition. *Int. J. Neural Syst.* **2025**, *35*, 2530001. [\[CrossRef\]](#) [\[PubMed\]](#)
6. Low, D.M.; Bentley, K.H.; Ghosh, S.S. Automated Assessment of Psychiatric Disorders Using Speech: A Systematic Review. *Laryngoscope Investig. Otolaryngol.* **2020**, *5*, 96–116. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Kappen, M.; Vanderhasselt, M.-A.; Slavich, G.M. Speech as a Promising Biosignal in Precision Psychiatry. *Neurosci. Biobehav. Rev.* **2023**, *148*, 105121. [\[CrossRef\]](#) [\[PubMed\]](#)
8. Gumus, M.; DeSouza, D.D.; Xu, M.; Fidalgo, C.; Simpson, W.; Robin, J. Evaluating the Utility of Daily Speech Assessments for Monitoring Depression Symptoms. *Digit. Health* **2023**, *9*, 20552076231180523. [\[CrossRef\]](#) [\[PubMed\]](#)
9. Pastor, M.A.; Ribas, D.; Ortega, A.; Miguel, A.; Lleida, E. Cross-Corpus Training Strategy for Speech Emotion Recognition Using Self-Supervised Representations. *Appl. Sci.* **2023**, *13*, 9062. [\[CrossRef\]](#)
10. Smietanka, L.; Maka, T. Enhancing Embedded Space with Low-Level Features for Speech Emotion Recognition. *Appl. Sci.* **2025**, *15*, 2598. [\[CrossRef\]](#)

11. Majkowski, A.; Kołodziej, M. Emotion Recognition from Speech in a Subject-Independent Approach. *Appl. Sci.* **2025**, *15*, 6958. [CrossRef]
12. Eyben, F.; Scherer, K.R.; Schuller, B.W.; Sundberg, J.; André, E.; Busso, C.; Devillers, L.Y.; Epps, J.; Laukka, P.; Narayanan, S.S.; et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. *IEEE Trans. Affect. Comput.* **2016**, *7*, 190–202. [CrossRef]
13. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2020; Volume 33, pp. 12449–12460.
14. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
15. Pepino, L.; Riera, P.; Ferrer, L. *Emotion Recognition from Speech Using Wav2vec 2.0 Embeddings*; International Speech Communication Association (ISCA): Grenoble, France, 2021; pp. 3400–3404.
16. Atmaja, B.T.; Sasou, A. Evaluating Self-Supervised Speech Representations for Speech Emotion Recognition. *IEEE Access* **2022**, *10*, 124396–124407. [CrossRef]
17. Abbaschian, B.J.; Sierra-Sosa, D.; Elmaghraby, A. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors* **2021**, *21*, 1249. [CrossRef] [PubMed]
18. Kakuba, S.; Han, D.S. Addressing Data Scarcity in Speech Emotion Recognition: A Comprehensive Review. *ICT Express* **2025**, *11*, 110–123. [CrossRef]
19. Latif, S.; Rana, R.; Khalifa, S.; Jurdak, R.; Qadir, J.; Schuller, B. Survey of Deep Representation Learning for Speech Emotion Recognition. *IEEE Trans. Affect. Comput.* **2023**, *14*, 1634–1654. [CrossRef]
20. Akinpelu, S.; Viriri, S. Deep Learning Framework for Speech Emotion Classification: A Survey of the State-of-the-Art. *IEEE Access* **2024**, *12*, 152152–152182. [CrossRef]
21. Wang, Y.; Huang, J.; Zhao, Z.; Lan, H.; Zhang, X. Speech Emotion Recognition Using Multi-Scale Global–Local Representation Learning with Feature Pyramid Network. *Appl. Sci.* **2024**, *14*, 11494. [CrossRef]
22. Makhmudov, F.; Kutlimuratov, A.; Cho, Y.-I. Hybrid LSTM–Attention and CNN Model for Enhanced Speech Emotion Recognition. *Appl. Sci.* **2024**, *14*, 11342. [CrossRef]
23. Li, H.; Li, J.; Liu, H.; Liu, T.; Chen, Q.; You, X. MelTrans: Mel-Spectrogram Relationship-Learning for Speech Emotion Recognition via Transformers. *Sensors* **2024**, *24*, 5506. [CrossRef] [PubMed]
24. Wang, C.; Shen, X. Feature-Enhanced Multi-Task Learning for Speech Emotion Recognition Using Decision Trees and LSTM. *Electronics* **2024**, *13*, 2689. [CrossRef]
25. Yan, J.; Li, H.; Xu, F.; Zhou, X.; Liu, Y.; Yang, Y. Speech Emotion Recognition Based on Temporal-Spatial Learnable Graph Convolutional Neural Network. *Electronics* **2024**, *13*, 2010. [CrossRef]
26. Atmaja, B.T.; Sasou, A. Effect of Different Splitting Criteria on the Performance of Speech Emotion Recognition. In *Proceedings of the TENCON 2021—2021 IEEE Region 10 Conference (TENCON)*, Auckland, New Zealand, 7–10 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 760–764.
27. Martin, K.; Gauthier, J.; Breiss, C.; Levy, R. *Probing Self-Supervised Speech Models for Phonetic and Phonemic Information: A Case Study in Aspiration*; International Speech Communication Association (ISCA): Grenoble, France, 2023; pp. 251–255.
28. Choi, K.; Pasad, A.; Nakamura, T.; Fukayama, S.; Livescu, K.; Watanabe, S. *Self-Supervised Speech Representations Are More Phonetic than Semantic*; International Speech Communication Association (ISCA): Grenoble, France, 2024; pp. 4578–4582.
29. JMIR Mental Health. Performance of Automatic Speech Analysis in Detecting Depression: Systematic Review and Meta-Analysis. Available online: <https://mental.jmir.org/2025/1/e67802> (accessed on 23 April 2026).
30. Christop, I. nEMO: Dataset of Emotional Speech in Polish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*; European Language Resources Association (ELRA): Paris, France, 2024.
31. Center for Artificial Intelligence, Adam Mickiewicz University. nEMO: Dataset of Emotional Speech in Polish. Available online: <https://csi.amu.edu.pl/en/datasets/nemo-dataset-of-emotional-speech-in-polish> (accessed on 23 April 2026).
32. Lin, N.; Zheng, M.; Li, L.; Hu, P.; Gao, W.; Sun, H.; Xu, C.; Yuan, G.; Liang, Z.; Dong, Y.; et al. An EEG Dataset for Interictal Epileptiform Discharge with Spatial Distribution Information. *Sci. Data* **2025**, *12*, 229. [CrossRef] [PubMed]
33. Pentari, A.; Kafentzis, G.; Tsiknakis, M. Speech Emotion Recognition via Graph-Based Representations. *Sci. Rep.* **2024**, *14*, 4484. [CrossRef] [PubMed]
34. Bhangale, K.; Kothandaraman, M. Speech Emotion Recognition Based on Multiple Acoustic Features and Deep Convolutional Neural Network. *Electronics* **2023**, *12*, 839. [CrossRef]

35. Eyben, F.; Wöllmer, M.; Schuller, B. Opensmile: The Munich Versatile and Fast Open-Source Audio Feature Extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*; Association for Computing Machinery: New York, NY, USA, 2010; pp. 1459–1462.
36. Rumagit, R.Y.; Alexander, G.; Saputra, I.F. Model Comparison in Speech Emotion Recognition for Indonesian Language. *Procedia Comput. Sci.* **2021**, *179*, 789–797. [[CrossRef](#)]
37. Wang, N.; Yang, D. Speech Emotion Recognition Using Fine-Tuned Wav2vec2.0 and Neural Controlled Differential Equations Classifier. *PLoS ONE* **2025**, *20*, e0318297. [[CrossRef](#)] [[PubMed](#)]
38. Lu, C.; Zhang, K.; Lian, H.; Li, S.; Qi, T.; Zong, Y.; Zheng, W. Speaker-Independent Speech Emotion Recognition Using Group Sparse-Based Adversarial Local Fisher Discriminant Analysis. *Pattern Recognit.* **2026**, *178*, 113419. [[CrossRef](#)]
39. Farhadpour, S.; Warner, T.A.; Maxwell, A.E. Selecting and Interpreting Multiclass Loss and Accuracy Assessment Metrics for Classifications with Class Imbalance: Guidance and Best Practices. *Remote Sens.* **2024**, *16*, 533. [[CrossRef](#)]
40. Byun, S.; Lee, S.-P. Frame and Utterance Emotional Alignment for Speech Emotion Recognition. *Future Internet* **2025**, *17*, 509. [[CrossRef](#)]
41. Nugroho, K.; Al Amin, I.H.; Noviasari, N.A.; Setiadi, D.R.I.M. Prosodic Spatio-Temporal Feature Fusion with Attention Mechanisms for Speech Emotion Recognition. *Computers* **2025**, *14*, 361. [[CrossRef](#)]
42. Hashem, A.; Arif, M.; Alghamdi, M. Speech Emotion Recognition Approaches: A Systematic Review. *Speech Commun.* **2023**, *154*, 102974. [[CrossRef](#)]
43. Lobmaier, J.S.; Klatt, W.K.; Schweinberger, S.R. Voice of a Woman: Influence of Interaction Partner Characteristics on Cycle Dependent Vocal Changes in Women. *Front. Psychol.* **2024**, *15*, 1401158. [[CrossRef](#)] [[PubMed](#)]
44. Marsano-Cornejo, M.-J.; Roco-Videla, Á. Variation of the Acoustic Parameters: F0, Jitter, Shimmer and Alpha Ratio in Relation with Different Background Noise Levels. *Acta Otorrinolaringol. (Engl. Ed.)* **2023**, *74*, 219–225. [[CrossRef](#)] [[PubMed](#)]
45. Supreetha, L.; Jayakumar, T. A Multiparametric Acoustic and Self-Perceptual Measure of Laryngeal Hydrations in Speech-Language Pathologists: A Preliminary Longitudinal Investigation. *J. Voice* **2024**, *in press*. [[CrossRef](#)] [[PubMed](#)]
46. Liu, T.; Yuan, X. Paralinguistic and Spectral Feature Extraction for Speech Emotion Classification Using Machine Learning Techniques. *J. Audio Speech Music Proc.* **2023**, *2023*, 23. [[CrossRef](#)]
47. Wang, F.; Shen, X. Research on Speech Emotion Recognition Based on Teager Energy Operator Coefficients and Inverted MFCC Feature Fusion. *Electronics* **2023**, *12*, 3599. [[CrossRef](#)]
48. Liu, S.; Nakajima, Y.; Chen, L.; Arndt, S.; Kakizoe, M.; Elliott, M.A.; Remijn, G.B. How Pause Duration Influences Impressions of English Speech: Comparison Between Native and Non-Native Speakers. *Front. Psychol.* **2022**, *13*, 778018. [[CrossRef](#)] [[PubMed](#)]
49. Zhu-Zhou, F.; Gil-Pita, R.; García-Gómez, J.; Rosa-Zurera, M. Robust Multi-Scenario Speech-Based Emotion Recognition System. *Sensors* **2022**, *22*, 2343. [[CrossRef](#)] [[PubMed](#)]
50. Jakubec, M.; Lieskovska, E.; Jarina, R.; Spisiak, M.; Kasak, P. Speech Emotion Recognition Using Transfer Learning: Integration of Advanced Speaker Embeddings and Image Recognition Models. *Appl. Sci.* **2024**, *14*, 9981. [[CrossRef](#)]
51. Yang, J.; Liu, J.; Huang, K.; Xia, J.; Zhu, Z.; Zhang, H. Single- and Cross-Lingual Speech Emotion Recognition Based on WavLM Domain Emotion Embedding. *Electronics* **2024**, *13*, 1380. [[CrossRef](#)]
52. Liu, J.; Xu, Y. T-Friedman Test: A New Statistical Test for Multiple Comparison with an Adjustable Conservativeness Measure. *Int. J. Comput. Intell. Syst.* **2022**, *15*, 29. [[CrossRef](#)]
53. Eisinga, R.; Heskies, T.; Pelzer, B.; Te Grotenhuis, M. Exact P-Values for Pairwise Comparison of Friedman Rank Sums, with Application to Comparing Classifiers. *BMC Bioinform.* **2017**, *18*, 68. [[CrossRef](#)] [[PubMed](#)]
54. Gudivaka, S.R.; Polipogu, P.; Kodali, R.; Dhulipalla, V.R.; Tatavarty, V.S.K.; Penumudi, J.; Gogulamudi, P.R. Speech Emotion Recognition in Adults and Children: A Comprehensive Review of Traditional Features and Raw Waveform Models. *Int. J. Speech Technol.* **2026**, *29*, 21. [[CrossRef](#)]
55. Zhao, S.; Zhang, B.; Yang, J.; Zhou, J.; Xu, Y. Linear Discriminant Analysis. *Nat. Rev. Methods Prim.* **2024**, *4*, 70. [[CrossRef](#)]
56. Ravi; Taran, S. A Nonlinear Feature Extraction Approach for Speech Emotion Recognition Using VMD and TKEO. *Appl. Acoust.* **2023**, *214*, 109667. [[CrossRef](#)]
57. Aldabergen, A.; Kynabay, B.; Kadyrov, S. Layer-Wise Probing of Paralinguistic Attributes in Fine-Tuned Whisper for Kazakh Speech. *Eng. Technol. Appl. Sci. Res.* **2026**, *16*, 33399–33404. [[CrossRef](#)]
58. Enguehard, J. Sequential Integrated Gradients: A Simple but Effective Method for Explaining Language Models. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023*; Rogers, A., Boyd-Graber, J., Okazaki, N., Eds.; Association for Computational Linguistics: Kerrville, TX, USA, 2023; pp. 7555–7565.
59. Fang, S.-X.; Shen, L.-Y.; Lin, Y.-C.; Chou, H.-C.; Lee, H. Meta-PerSER: Few-Shot Listener Personalized Speech Emotion Recognition via Meta-Learning; International Speech Communication Association (ISCA): Grenoble, France, 2025; pp. 136–140.
60. Kim, B.; Kwon, Y. Searching for Effective Preprocessing Method and CNN Based Architecture with Efficient Channel Attention on Speech Emotion Recognition. *Sci. Rep.* **2025**, *15*, 32689. [[CrossRef](#)] [[PubMed](#)]

61. Begazo, R.; Aguilera, A.; Dongo, I.; Cardinale, Y. A Combined CNN Architecture for Speech Emotion Recognition. *Sensors* **2024**, *24*, 5797. [[CrossRef](#)] [[PubMed](#)]
62. Kilimci, Z.H.; Bayraktar, Ü.; Küçükmanisa, A. Evaluating Raw Waveforms with Deep Learning Frameworks for Speech Emotion Recognition. *Multimed. Tools Appl.* **2025**, *84*, 45119–45149. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.