A Comparative Study of ML and DL Algorithms for Speech Emotion Recognition

[1] Nam Thai NGUYEN, [1] Hoang Khang LE, [1] Tran Viet Thu LE,
[1] Thien Gia Bao TU

[1] University of Economics Ho Chi Minh City (UEH), Ho Chi Minh City, Vietnam

Abstract. This paper will investigate both traditional machine learning (SVM, Random Forest) and deep learning (CNN, MLP) for speech emotion recognition, using the Ravdess and TESS datasets. We achieved up to 97% prediction accuracy with the SVM model, demonstrating that it can be effective when used to study audio and emotion-related datasets.

Keywords: Emotion recognition, deep learning, machine learning, classification, mel-frequency cepstral coefficients, MLP, SVM, RF, RAVDESS, TESS.

1 Introduction

People use spoken language all the time to connect and share info. It’s not just about daily chats — it’s also key in practical places like call centers or Business Process Outsourcing (BPO) [1]. In these situations, catching the speaker’s emotions can really help. For example, emotion detection can show if customers are happy or upset, improve how computers talk back, clear up any confusion in speech, and let systems adjust based on the person’s mood.

In this study, we try to build models that pick out the main emotion in a speech clip. While many others use things like computer vision or text processing, we mainly focus on raw audio signals. To do this, we use Mel-Frequency Cepstral Coefficients (MFCCs) [6], which turn speech into numbers that computers can analyze easily.

However, there are still a few limitations, such as the issue of noise in the dataset. As mentioned in [4], it is necessary to select static features based on the spectral images of the dataset for those sets that contain excessive noise. In this study, we will use a noise-filtered and normalized dataset to shorten the execution time.

The paper is split into several parts. Section 2 covers past research about speech emotion recognition. Section 3 explains the data and how we put the system together. Section 4 talks about the methods and algorithms we used. Section 5 shares the results and some discussion. Finally, Section 6 wraps up with conclusions.

2 Related Work

In the context of increasingly developing human-computer interaction, mental health care, and smart education applications, speech emotion recognition (SER) has become a key research area, for example, in [1]. To evaluate the effectiveness of SER methods, the RAVDESS dataset is used as a benchmark by many authors. RAVDESS consists of voice recordings and facial expressions of 24 actors (12 male, 12 female) expressing 8 different emotional states, with high recording quality and gender diversity. Regarding the method, Iqbal et al. [2] deployed three traditional algorithms, including Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Gradient Boosting on male voice subsets, female voice subsets, and mixed data. SVM [7] demonstrates stability and efficiency by achieving absolute accuracy (100%) for the emotions “angry” and “neutral” in male voices, maintaining high performance for the emotion “angry” in female voices, and remaining stable on mixed data. Or, PCA and KCA on the other hand [3]. KNN is good for unambiguous emotions but drops sharply in performance for emotions like “happy” and “sad” in female voices due to frequency variation. Gradient boosting excels in recognizing complex emotions like “excited” and “sad” but has poor stability on scattered data. In addition, Prabhakar [8] developed a multi-input deep learning model, training three separate neural networks for facial still images, sound waves, and a model combining both. The results show that the accuracy with audio data alone is 66.41%, while combining both signals improves it up to 90%, demonstrating the effective complementarity of multimodal data. Another approach is the domain-specific classification strategy, which uses a Dag-SVM architecture combined with multi-task learning to automatically partition data into feature domains such as gender or accent and then applies a domain-specific classifier. This method improves recognition performance on highly distributed and heterogeneous datasets, which is a major challenge in practice. In summary, SVM is the most stable algorithm for diverse datasets, gradient boosting is suitable for strong sentiments but needs to be careful with distributed data, and KNN should be used when sentiments are clearly expressed. The multi-input deep learning model and domain-specific classification strategy open up a more flexible and accurate way to apply SER in complex real-world environments, so, the conclusion is similar to Sharan's. [9]. The combination of traditional algorithms and deep learning models, together with domain-specific classification strategies, promises to improve the accuracy and applicability of speech emotion recognition systems in the fields of human–computer interaction, healthcare, and smart education. We have taken parameters from the latest research by Wang, N. and D. Yang [10] to improve our model, resulting in promising outcomes.

3 Implementation detail

3.1 Dataset

For the purpose of training and evaluating our emotion recognition model, we utilized two widely recognized datasets: RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and TESS (Toronto Emotional Speech Set). These datasets offer high-quality audio recordings representing various emotional expressions, which are crucial for training a model capable of identifying emotions in speech.

RAVDESS: This dataset consists of audio and video recordings of actors expressing eight emotions: happy, sad, angry, fearful, disgusted, surprised, calm, and neutral. The RAVDESS dataset includes both male and female speakers and provides both speech and song recordings.

TESS: The TESS dataset includes emotional speech data from different actors, recorded in a controlled environment. It consists of speech samples for emotions such as happy, sad, angry, and fearful, providing an essential resource for training speech emotion recognition models.

We will be making predictions for the following emotions: 0 - Neutral, 1 - Calm, 2 - Happy, 3 - Sad, 4 - Angry, 5 - Fearful, 6 - Disgust, 7 - Surprised.

Filename identifiers

Statement (01 : "Kids are talking by the door", 02 : "Dogs are sitting by the door").

Repetition (01 - 1st repetition, 02 - 2nd repetition)

Actor (01 to 24. Odd-numbered actors are male, even-numbered actors are female)

3.2 Data Preprocessing

In this study, we use MFCCs (Mel-Freq Cepstral Coefficients) as the main input feature for emotion classification models. This is also easily decided because MFCC is one of the most popular features in such studies, MFCC simulates very well how humans perceive sounds on the mel frequency scale..

MFCC [6] is another representation of the cepstrum in the Mel frequency scale, and has been proven to be one of the most effective methods in representing audio signals for automatic speech recognition tasks. MFCC coefficients have the outstanding advantage of compressing the spectral information of the sound wave into a compact vector, which is easily processed by machine learning algorithms.

MFCC, each audio file is divided into short signal frames (frames) with a fixed window size to ensure statistical stability. The amplitude spectrum is then transformed into a Mel frequency [5] scale to highlight the frequency bands that are more important to human auditory perception.

Specifically, from each audio file, 40 MFCC features are extracted. The original audio data is converted into a real-time series, which then forms a series MFCC [6] coefficients. This feature matrix is transposed and then horizontally averaged, resulting in an input vector representing the generalization of each input audio sample.

Specifically, each audio set is normalized and brought to 16kHz frequency, cut to a fixed length of 2 seconds. Then, we will extract (in this case 40) MFCC coefficient features for each frame, combined with delta to capture the feature variation over time to form a feature vector.

3.3 Algorithms

CNN: is a deep learning algorithm, capable of automatically extracting spatial and temporal features through convolutional and pooling layers.

MLP: is the most basic deep learning model of artificial neural network with many hidden layers, using activation function and all layers are fully connected, simple structure and easy to apply.

Random Forest: is a machine learning algorithm, a type of ensemble model consisting of multiple decision trees that work together to increase accuracy, with each tree trained on a different subset of data, then performing a "vote" action to decide the output. This model is highly effective with small data sets and does not require much normalization.

SVM: is a robust classification model that performs well when the number of features is larger than the number of samples. This model is based on the principle of optimally separating hyperplanes between classes, which we use specifically as a non-linear classifier using Kernel trick (RBF).

4 Methodology

4.1 Overview of the system design flowchart

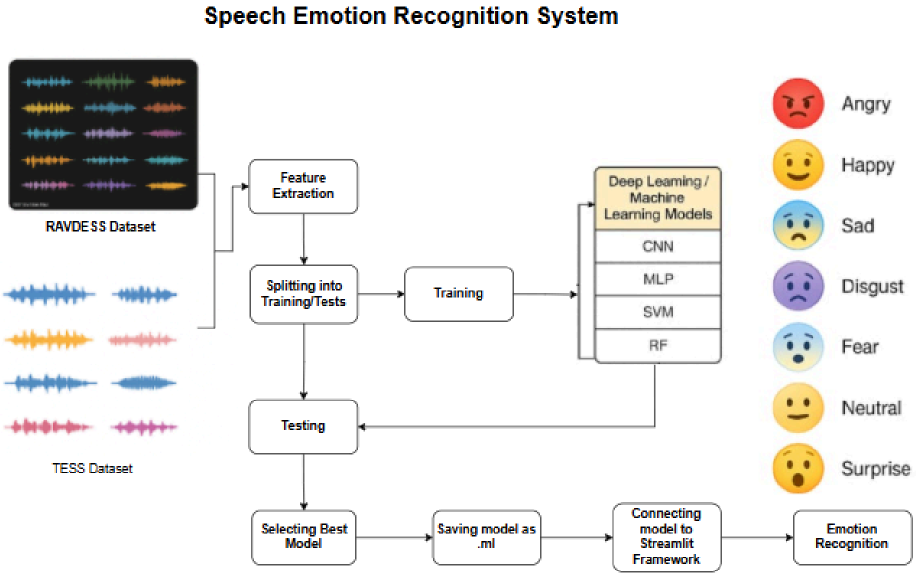


Fig 1. System design flowchart

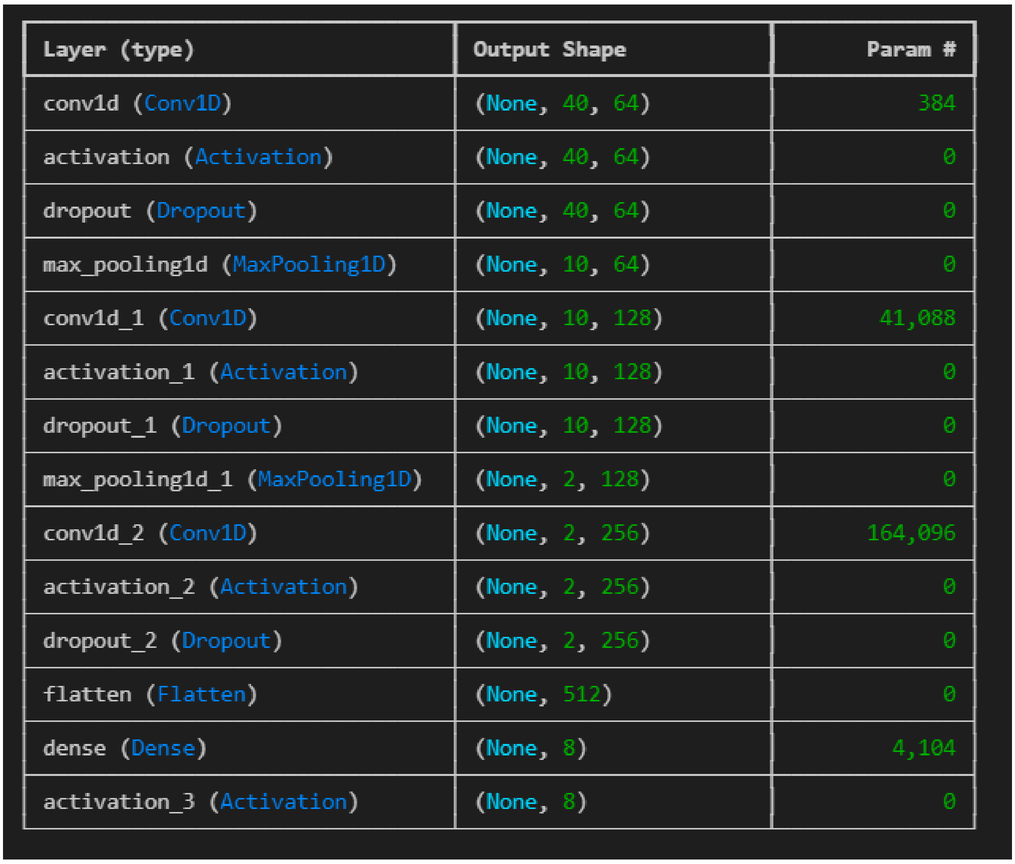
4.2 Methodology detail

CNN

The architecture of the Convolutional Neural Network (CNN) used in this study for emotion classification is illustrated in Figure 1. The input data is 40 MFCC features extracted from 2-second audio clips, with each audio being represented as a vertical feature vector (40x1).

The convolutional layer is the first layer, which is used to scan through the input feature vectors and learn important feature patterns, using the ReLu function, defined as follows: g(z) = max(0,z). The purpose of this layer is to create non-linearity to help the network learn more complex relationships.

The next layer is a Max Pooling layer, which reduces the size of the output dataset of the first layer, to avoid overfitting, followed by a flattening layer and a fully connected layer to perform classification based on 8 emotion classes: Neutral - 0, Calm - 1, Happy - 2, Sad - 3, Angry - 4, Fear - 5, Disgust - 6 and Surprise - 7.



Layer (type)	Output Shape	Param #
conv1d (Conv1D)	(None, 40, 64)	384
activation (Activation)	(None, 40, 64)	0
dropout (Dropout)	(None, 40, 64)	0
max_pooling1d (MaxPooling1D)	(None, 10, 64)	0
conv1d_1 (Conv1D)	(None, 10, 128)	41,088
activation_1 (Activation)	(None, 10, 128)	0
dropout_1 (Dropout)	(None, 10, 128)	0
max_pooling1d_1 (MaxPooling1D)	(None, 2, 128)	0
conv1d_2 (Conv1D)	(None, 2, 256)	164,096
activation_2 (Activation)	(None, 2, 256)	0
dropout_2 (Dropout)	(None, 2, 256)	0
flatten (Flatten)	(None, 512)	0
dense (Dense)	(None, 8)	4,104
activation_3 (Activation)	(None, 8)	0

Fig 2. Detail description of CNN

MLP

For the MLP model, we extract MFCC, Chroma, Mel-Spectrogram audio features and then combine the three features into a single feature vector. This helps to convert the raw audio data into a suitable input for machine learning and deep learning models (each audio segment is divided into many small frames, then averaged over time to form a single feature vector). Next, assign emotion labels based on file names according to the data set's identification rules. After that, we split the training and testing sets at a ratio of 75/25.

RF

The Random Forest model for classifying emotions from speech requires a thorough data processing, namely the representation of audio features combining three common types of features into a single vector, similar to the way we preprocessed the data for other models. After feature extraction, the data is divided into training and testing sets. The Random Forest model is trained on the input feature vectors, using optimized parameters to achieve high accuracy, however, this model is quite sensitive to noise so the accuracy is not very high.

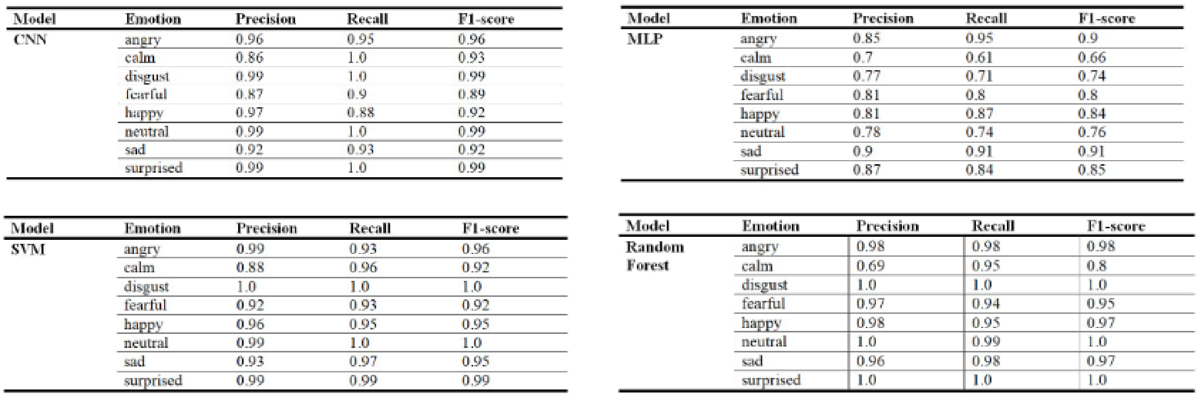
SVM

Due to the complex audio data, we decided to choose the SVM model used as Kernel (Radial Basis Function) because it is very strong in linear classification, with the parameter c = 10 increasing the weight for errors, making it difficult for the model to miss difficult-to-classify samples, improving the sensitivity of the model.

5 Experimental Results

5.1. Discussion

The results show that all four models perform well, but there are still clear differences between the models in terms of efficiency. SVM is the model with the highest performance with an accuracy of 97.4% and an average F1-score of up to 0.97 in Fig 3. This shows that SVM is very suitable for the emotion recognition problem on the current Ravdess and Tess datasets (The special environment has little noise and the data is well represented by extracted features such as MFCC). For Random Forest, unlike SVM, RF achieves very good results with an accuracy of 96%, but in the "Angry" and "Surprised" classes, instead of in the stable classes such as "Neutral" and "Sad". This also proves that this model is stable and less affected by noise, due to the aggregation feature based on many decision trees. The CNN model performs quite well with temporal data like audio, but using only one convolutional layer does not outperform more classical models, even though MFCC reduces the complexity of the input. The final model used for comparison is the MLP model, which has the lowest accuracy (82%). The results are quite poor in some layers like the Calm layer, indicating that the model does not generalize well across the entire dataset. Although this model is a simple deep learning model similar to CNN, it does not have a spatial feature learning mechanism, so it may not be suitable for time-series audio data.



Model	Emotion	Precision	Recall	F1-score
CNN	angry	0.96	0.95	0.96
	calm	0.86	1.0	0.93
	disgust	0.99	1.0	0.99
	fearful	0.87	0.9	0.89
	happy	0.97	0.88	0.92
	neutral	0.99	1.0	0.99
	sad	0.92	0.93	0.92
	surprised	0.99	1.0	0.99

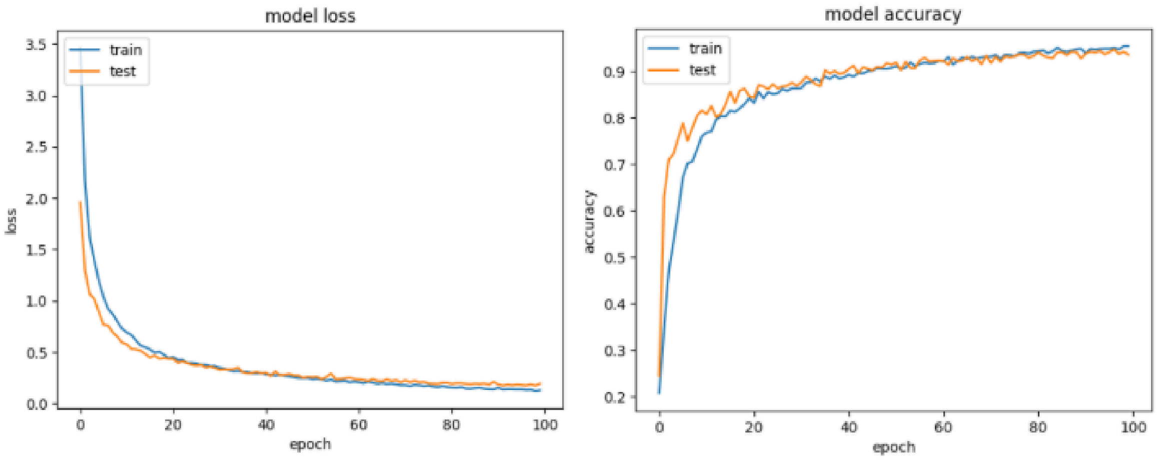
Model	Emotion	Precision	Recall	F1-score
MLP	angry	0.85	0.95	0.9
	calm	0.7	0.61	0.66
	disgust	0.77	0.71	0.74
	fearful	0.81	0.8	0.8
	happy	0.81	0.87	0.84
	neutral	0.78	0.74	0.76
	sad	0.9	0.91	0.91
	surprised	0.87	0.84	0.85

Model	Emotion	Precision	Recall	F1-score
SVM	angry	0.99	0.93	0.96
	calm	0.88	0.96	0.92
	disgust	1.0	1.0	1.0
	fearful	0.92	0.93	0.92
	happy	0.96	0.95	0.95
	neutral	0.99	1.0	1.0
	sad	0.93	0.97	0.95
	surprised	0.99	0.99	0.99

Model	Emotion	Precision	Recall	F1-score
Random Forest	angry	0.98	0.98	0.98
	calm	0.69	0.95	0.8
	disgust	1.0	1.0	1.0
	fearful	0.97	0.94	0.95
	happy	0.98	0.95	0.97
	neutral	1.0	0.99	1.0
	sad	0.96	0.98	0.97
	surprised	1.0	1.0	1.0

Fig 3. SVM, RF, CNN and MLP classification evaluation metrics table

For the CNN model, the Accuracy and Loss plots show the effectiveness of the model on the input data. In the first 20 epochs, the accuracy increases sharply, especially on the test set, showing that the model learns quickly and has good generalization ability from an early age. From epoch 20 onwards, the training and test set curves show asymptote, proving that the model is not overfitting and achieves stable performance on both sets. Coming to the second plot - the loss plot, both the train and test curves fluctuate at a very low level, not much different from each other, especially from epoch 30 onwards. Both models show that the model is really effective and even on both data sets, CNN still does not produce results that are superior to SVM for MFCC-normalized data.



a.Trend of the cost function of CNN over 100 epochs

b.Trend of the accuracy of CNN over 100 epochs

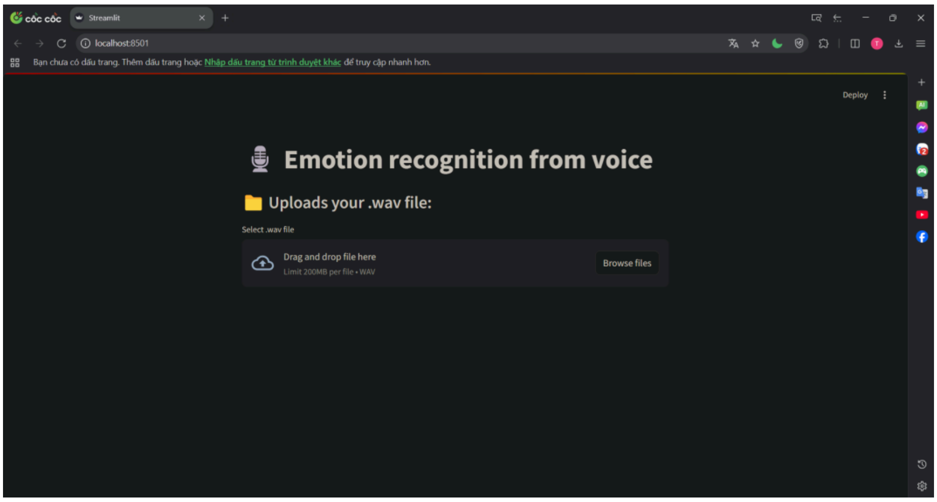
Fig 4. Model loss and accuracy of the CNN model

In terms of emotion class discrimination, classes such as "Disgust", "Neutral", "Surprised" are always classified with high accuracy on both SVM, RF and CNN, showing that these are emotion classes with distinct, easily recognizable audio features. In contrast, classes "Calm" and "Sad" are the two most difficult to predict, with lower precision on both MLP and CNN models, showing that these models easily confuse emotions with mild or similar audio expressions.

5.2 Application development

Development environment: Python language development tool, Streamlit framework.

Apply SVM model - the model with the best results, to build a voice emotion recognition model. Propose integrating the model into applications that support the diagnosis of depression, or study the satisfaction of users of specific applications.



Bạn chưa có dấu trang. Thêm dấu trang hoặc Nhập dấu trang từ trình duyệt khác để truy cập nhanh hơn.

Emotion recognition from voice

Uploads your .wav file:

Select .wav file

Drag and drop file here

Limit 200MB per file • WAV

Fig 5. Emotion recognition from voice interfaces

6 Conclusion

With the current dataset, the machine learning model (SVM) is superior to the deep learning model, thanks to its ability to separate the edges in the extracted feature space (MFCC). Meanwhile, deep learning models such as CNN only really work effectively when the dataset is larger and the architecture is more complex. MLP, because it does not take advantage of the spatial or temporal structure of audio data, has significantly lower performance, and is more suitable as a baseline. Choosing the right model depends on the data and the application goal. If high accuracy and short training time are required, SVM and RF are the optimal choices. If the data is rich and there is a need for expansion, CNN is a potential model.

References

1. International, O. T. (2024). "Retracted: A Classroom Emotion Recognition Model Based on a Convolutional Neural Network Speech Emotion Algorithm." Occup Ther Int **2024**: 9825450.

2. Jarvers, I., et al. (2024). "Impact of alexithymia, speech problems and parental emotion recognition on internalizing and externalizing problems in preschoolers." PLoS One **19**(9): e0310244.

3. Kingeski, R., et al. (2024). "Fusion of PCA and ICA in Statistical Subset Analysis for Speech Emotion Recognition." Sensors (Basel) **24**(17).

4. Leem, S. G., et al. (2024). "Selective Acoustic Feature Enhancement for Speech Emotion Recognition With Noisy Speech." IEEE/ACM Trans Audio Speech Lang Process **32**: 917-929.

5. Li, H., et al. (2024). "MelTrans: Mel-Spectrogram Relationship-Learning for Speech Emotion Recognition via Transformers." Sensors (Basel) **24**(17).

6. Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

7. Parlak, C. (2025). "Cochleogram-Based Speech Emotion Recognition with the Cascade of Asymmetric Resonators with Fast-Acting Compression Using Time-Distributed Convolutional Long Short-Term Memory and Support Vector Machines." Biomimetics (Basel) **10**(3).

8. Prabhakar, S. K. and D. O. Won (2024). "A Methodical Framework Utilizing Transforms and Biomimetic Intelligence-Based Optimization with Machine Learning for Speech Emotion Recognition." Biomimetics (Basel) **9**(9).

9. Sharan, R. V., et al. (2024). "Emotion Recognition from Speech Signals by Mel-Spectrogram and a CNN-RNN." Annu Int Conf IEEE Eng Med Biol Soc **2024**: 1-

10. Wang, N. and D. Yang (2025). "Speech emotion recognition using fine-tuned Wav2vec2.0 and neural controlled differential equations classifier." PLoS One **20**(2): e0318297.